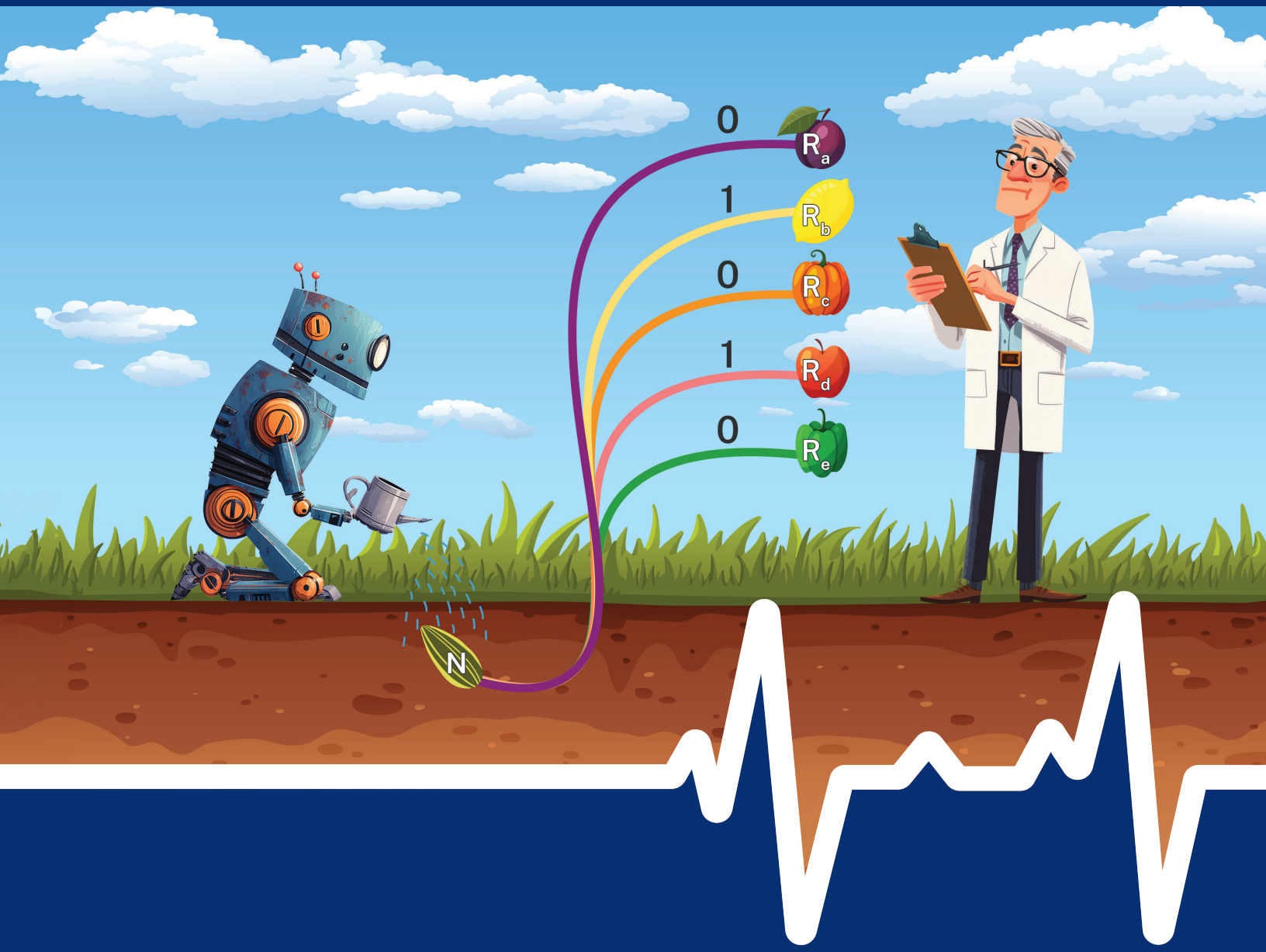


Targeted Postmarket Surveillance:

The Way Toward Responsible AI Innovation in Health Care

Kev Coleman Michael J. Pencina

PARAGON HEALTH INSTITUTE

SEPTEMBER 2025

The authors are grateful to the Paragon team for their exceptional comments and work in review of the paper. The authors also benefited from many spirited conversations with colleagues as well as feedback from reviewers who donated their time generously. We'd like to particularly thank David Page, Charles M. Silver, Naomi Lopez, Nicoleta Economou, Monica Agrawal, Ricardo Henao, Frank Rockhold, and Lee Fleisher.

Copyright © 2025 Paragon Health Institute

All parts of this publication are protected by copyright and/or related rights. You are free to reproduce, store in a retrieval system, or transmit in any form or by any means — electronic, mechanical, photocopy, recording, or any other — as well as modify the work as long as proper attribution is given to the authors and Paragon Health Institute as the original publication source, and a link to the Paragon Health Institute website (www.paragoninstitute.org) is provided.

(703) 527-2734
paragoninstitute.org

ABOUT THE AUTHORS

Kev Coleman is a Research Fellow whose policy foci include artificial intelligence, association health plans, and health insurance. Kev is a recognized health care leader, having been named one of "The 20 most Creative People in Insurance." He was also the conceptual architect of the internet's first private Medicare insurance marketplace in 2006, years before the launch of the government's Healthcare.gov marketplace. His health care research has been cited in top newspapers and media across the country and referenced in congressional health reform discussions.

A veteran of multiple technology companies, Mr. Coleman consults within the health care market on issues ranging from start-up product evaluation to business models. His health care research has spanned artificial intelligence applications in health care, Medicare plan designs and formularies, the Affordable Care Act, employer-based health plans, dental insurance, and telemedicine. His expertise in association health plans is highlighted in his monograph "Association Health Plans & The Future of American Health Insurance."

Michael J. Pencina, PhD, is Duke Health's chief data scientist and serves as vice dean for data science, director of Duke AI Health, and professor of biostatistics and bioinformatics at the Duke University School of Medicine. Dr. Pencina is an internationally recognized authority in the evaluation of AI algorithms. Thomson Reuters/Clarivate Analytics acknowledges Dr. Pencina as one of the world's "highly cited researchers" in clinical medicine and social sciences, with over 400 publications cited 135,000 times.

EXECUTIVE SUMMARY

What This Paper Covers

Artificial intelligence (AI) unpredictability has prompted calls for extensive AI regulation within health care. AI unpredictability, in this context, refers to the variability of outputs some health AI medical devices produce in response to *identical* inputs (e.g. a mammogram or EKG recording), whether those inputs occur in immediate succession or after prolonged intervals. Although not ubiquitous in AI, unpredictability is not confined to a single programming architecture or training process.

The risk unpredictable AI poses for patient safety is a major concern for both regulators and health care providers. However, unpredictability is a double-edged sword. It can produce genuine patient hazards, but, unfortunately, hazards can also arise from a suboptimal regulatory response. Suboptimal regulatory responses would include:

- delaying or restricting market access for new lifesaving medical devices;
- failing to prevent mass patient injuries because of rules that ignore the root causes of unpredictability;
- increasing compliance costs in instances where AI output variability neither endangers patients nor impairs clinical value;
- failing to identify which AI devices are susceptible to unpredictabilities that do not manifest during premarket reviews by the Food and Drug Administration (FDA); and
- failing to adequately address AI systems that, after deployment[1] in the marketplace, continually update themselves based on observed outcomes and other new data to improve their own performance.

The FDA's current review system was built for an earlier era — physical devices and software whose outputs are predictable and consistent. Its premarket[2] validation, while still necessary, is not sufficient for AI systems whose unpredictability may take time to manifest due to incremental data changes or irregular occurrence. The FDA has contemplated issues related

1 Deployment refers to the implementation of an AI device at a health system for real-world use with patients.

2 Premarket refers to the period before a medical device passes FDA review and may be commercially distributed.

to unpredictability in the form of adaptive algorithms that continuously learn,[3] but its guidance has been critiqued for being incomplete.[4] A more effective review process would augment a pre-deployment evaluation with an inspection method that scrutinizes post-deployment performance of those AI medical devices for which there are objective concerns regarding the device's output consistency.

"Former FDA Commissioner Robert Califf has commented on the FDA's need for private sector collaboration on the postmarket oversight of AI, saying it is not something the FDA can do on its own."

The FDA does have an existing pathway for postmarket monitoring. Under the authority of Section 522 of the Federal Food, Drug, and Cosmetic Act, the FDA may oblige a manufacturer to collect and analyze data on a marketed medical device. However, the agency has pointed out that its authority to conduct such surveillance is limited. The FDA has acknowledged that it has inadequate resources to significantly expand such surveillance. Recognizing this labor constraint, former FDA Commissioner Robert Califf has commented on the FDA's need for private sector collaboration on the postmarket oversight of AI, saying it is not something the FDA can do on its own.

For a public-private postmarket surveillance to succeed, private parties (device manufacturers, health care providers, etc.) must have sufficient incentives to voluntarily contribute to the FDA's mission. Additionally, costs should be minimized, as AI device development and compliance is already an expensive process, and additional costs could negatively affect AI adoption. If postmarket surveillance is too costly or burdensome for staff, then postmarket surveillance will not be widely implemented.

The first step in successfully implementing postmarket surveillance is to properly scope the surveillance effort. Not all AI is unpredictable, and for those devices that are, some do not present a significant risk for patient harm depending on the tasks they perform. This paper proposes a framework that concentrates surveillance on those AI devices where output unpredictability intersects with prospects for meaningful patient harm. Thus, regulatory efficiency begins with reliance on existing, general safety processes for medical devices where additional AI-specific surveillance would produce little benefit. The remaining AI

3 FDA, "Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions," December 4, 2024, https://www.fda.gov/media/166704/download.

4 See King & Spaulding, "FDA Publishes Final Predetermined Change Control Plan Guidance for AI-Enabled Device Software Functions," December 13, 2024, https://www.kslaw.com/news-and-insights/fda-publishes-final-predetermined-change-control-plan-guidance-for-ai-enabled-device-software-functions.

devices that do manifest unpredictability and have a medium to high risk for patient harm[5] would be evaluated for either periodic device revalidations or performance monitoring.

Periodic Device Revalidations

For AI devices whose programming architectures[6] do not contribute to unpredictability but may adapt their outputs based on open-ended data analysis,[7] we recommend **periodic device revalidations.** Manufacturer test data that was originally used for a device's premarket review could be employed in periodic reiterations of its testing to confirm that the device's latest outputs have remained within acceptable parameters (but still allowing for further testing using health-system-supplied data). The reuse of existing test data and its labels reduces surveillance costs for health systems and eliminates the need for advanced data science consulting (and the related consulting costs) in test data assembly. The time intervals at which these periodic device revalidations occur would progressively increase so that those devices whose outputs are highly unstable could be identified early while moving more stable devices toward a less-frequent maintenance testing schedule. To preserve the confidentiality of patient data as well as the manufacturer's intellectual property and testing acceptance criteria, the periodic revalidation neither requires health systems to see the device programming code (and introduce intellectual property concerns) nor manufacturers to access patient data (and introduce privacy concerns).

Performance Monitoring

We recommend a second type of postmarket surveillance, **performance monitoring**, for AI devices with output unpredictability that is intrinsic to the devices' programming (model, parameterization, routing, etc.). Performance monitoring, unlike periodic revalidation, uses clinical output information gathered post-deployment. Performance monitoring complements the FDA's capture of serious adverse outcomes by leveraging health system infrastructure. As a part of such infrastructure, an EHR can monitor and collect malfunctions that, while not resulting in patient harm, provide early signals pertaining to a device's output reliability and its impact on care delivery. Specifically, performance monitoring collects data on erroneous outputs, safety events, indications of model degradation, and undesirable outcomes.

Because variability in local population health and procedures for AI use can affect the performance of an AI medical device, the aggregated (not individual patient) performance data produced from both revalidations and monitoring should be compared across health

5 Regarding FDA risk classes, see FDA, "How to Study and Market Your Device," October 12, 2023, https://www.fda.gov/medical-devices/device-advice-comprehensive-regulatory-assistance/how-study-and-market-your-device

6 A programming architecture is a comprehensive structure of a software application from its specific algorithms and operations to the relationships among them.

7 These issues are discussed in detail within Section III of the paper.

systems. These comparisons can spot trends or outliers that may indicate when AI has a problem that is not necessarily technical. Accordingly, we recommend that these postmarket surveillance processes be performed within an "aggregated outcome data registry" shared among health systems that have deployed the same AI device. Such a network would — in a process compliant with privacy guidelines in the Health Insurance Portability and Accountability Act (HIPAA) — facilitate analysis and identify negative events or trends with alerts transmitted, as appropriate, to the FDA as well as the manufacturer and health systems within the network.[8]

An aggregated outcome data registry is similar to a federated health data network — a secure way for providers in different places to share data and resources without exposing private systems to the connected parties. Within an aggregated outcome data registry, manufacturers and health systems can collaboratively monitor the performance of AI medical devices at risk for unpredictability. Ideally, this monitoring could be facilitated by an EHR system provider or other data aggregation specialist with relevant data voluntarily provided by multiple health systems employing the same AI medical device.

The FDA could set high-level goals, while day-to-day operations are managed by coalitions of AI adopters (health systems), manufacturers (AI developers), and technology providers (EHR vendors, cloud platforms, data managers). This effort would build on and expand the FDA's Sentinel Initiative. Within the aggregated outcome data registry, a software routine (such as an AI agent) accessing individual health records related to an AI medical device could identify relevant data and transform this information into anonymized summary data that can be shared among the providers, manufacturer, and FDA without patient privacy violations. Together with this structured data, the routine could further combine unstructured data such as notes pertaining to deployment challenges or errors.

AI manufacturers have strong motivation to join the kind of voluntary surveillance outlined in this paper. First, there is the desire to avoid device failures that can result in legal and financial liabilities as well as reputational damage. These liabilities are further magnified by the absence of the Learned Intermediary Rule for many complex AI medical devices. Under the Learned Intermediary Rule, a medical device manufacturer may be shielded from some legal accountability for a patient injury given that a health care provider made the decision that the AI device was appropriate for the patient's needs in light of the manufacturer's disclosure of risks and benefits. However, for complex AI systems with low explainability, this doctrine would not apply, as the provider may not be able to assess the complete scope of risk represented by the AI medical device.

8 The model proposed in this paper can be utilized for AI devices not yet commercialized as well as those that have been approved by the FDA.

Health care providers, alongside their interest in patient welfare, share manufacturers' liability concerns. AI devices do not have an extended history of medical use, and without this history, some providers worry about their long-term safety. Although these devices pass FDA's premarket review, the agency's past lack of transparency — especially through the Alternative Summary Reporting program, which hid from the public many adverse events until 2019 — has left providers uneasy. Postmarket surveillance provides a potential means by which negative device trends can be detected before they reach the point of patient injury. Additionally, the postmarket surveillance advocated here offers health care providers an alternative to market safety programs carrying high costs and high organizational disruption.

For the FDA, this postmarket surveillance proposal avoids several limitations of competing oversight schemes. First, it avoids duplicating the FDA's premarket validation. Consequently, there is no need for a massive financial investment to create new and independent test regimes for all the types of AI medical devices under the FDA's purview, thus making this model of surveillance much more affordable and scalable through a concentration of scope. Second, this concentration intentionally avoids unnecessary costs that will amplify the competitive advantage that well-funded health systems have over smaller and less-resourced health systems with respect to AI technology purchases. The total cost of ownership for health care AI, instead, is treated as an important factor affecting AI adoption and minimized as much as possible in the effort to achieve the goals of appropriate postmarket surveillance. Finally, this proposal's validation and monitoring do not impose heavy labor costs or require providers to acquire specialized AI development expertise.

To move our vision forward, the following next steps are needed:

1. Secure the FDA's public support for the proposed postmarket surveillance framework
2. Identify health systems (or other health care organizations)[9] with AI adoption and an interest in developing an efficient postmarket surveillance system
3. Identify AI manufacturers willing to join postmarket surveillance pilots
4. Identify technology, data management, and AI monitoring partners ready to work with AI adopters to syndicate the postmarket surveillance system
5. Define the technical, security, and data standards that will underpin the system

9 See Kev Coleman, "Could the VA Be the Key to Lowering the Cost of American Health Care?" Paragon Health Institute, July 16, 2025, https://paragoninstitute.org/public-health/could-the-va-be-the-key-to-lowering-the-cost-of-american-health-care/

6. Develop financial models and incentive structures to sustain the effort, including funding for methods that improve AI unpredictability assessment
7. Conduct well-scoped pilots to optimize surveillance implementation and acquire practical experience and lessons

INTRODUCTION

The unpredictability of artificial intelligence (AI) has provoked calls for its extensive regulation within health care.[10] ***Unpredictability*, in this context, refers to the output variability some health AI medical devices may produce in response to identical inputs, whether those inputs occur in immediate succession or after prolonged intervals.** The adjective "some" is noteworthy because it is a reminder that unpredictability is not present for all types of AI. With regard to "inputs," they can vary by device. For example, an AI medical device determining the risk of lung cancer may use a chest scan as an input, but a different device calculating the probability of sepsis, on the other hand, may use results of a blood culture along with other vital signs. The outputs AI devices produce based upon such inputs include important critical functions such as illness predictions, diagnoses, and treatment recommendations. Unpredictability presents the concern all future outputs cannot be reliably extrapolated at the time of premarket review of an AI medical device by the Food and Drug Administration (FDA). In other words, outputs observed before market approval may not be consistent with the outputs observed after the product is used in the market by health systems.

"AI output variability for identical inputs, in some cases, can be a byproduct of desired functionality. In fact, it can be highly beneficial. Some AI devices, for example, can adapt over time and improve their accuracy."

If an AI medical device's outputs are inconsistent, there is the unavoidable question "Is this unreliability a threat to public health?" The answer is not simple as it may appear, as this behavior is not a programming defect similar to the "bugs" of traditional software. AI output variability for identical inputs, in some cases, can be a byproduct of desired functionality. In fact, it can be highly beneficial. Some AI devices, for example, can adapt over time and improve their accuracy.

Unpredictability is not intrinsic to all AI, but neither is it localized to a single programming architecture or a single dataset used to train AI. Moreover, its occurrence in health care can pose genuine patient hazards, but misregulation presents equal dangers. They include, at the very least, problems such as:

10 See Kev Coleman and Michael Pencina, "The Regulation of Uncertainty," Paragon Health Institute, February 5, 2025, https://paragoninstitute.org/private-health/the-regulation-of-uncertainty/.

- delaying or restricting market access for new lifesaving medical devices;
- failing to prevent mass patient injuries because of rules that ignore the root causes of unpredictability;
- increasing compliance costs in instances where AI output variability neither endangers patients nor impairs clinical value;
- failing to identify which AI devices are susceptible to unpredictabilities that do not manifest during premarket reviews by the Food and Drug Administration (FDA); and
- failing to adequately address AI systems that, after deployment in the marketplace, continually update themselves based on observed outcomes and other new data to improve their own performance.

Given the possibility of a long delay before the expression of AI unpredictability, a mechanism of postmarket surveillance is preferable to a review process that attempts a safety attestation that is duplicative of premarket FDA review but more extensive.[11] However, a postmarket surveillance must not only avoid the issues that misregulation can produce but also navigate concerns over regulatory capture and cost that present barriers to AI adoption in health care. We propose in this paper a postmarket surveillance framework that can successfully operate within these constraints. The foundation of this framework begins with an understanding of the factors causing the technology's unpredictability. Failing to grasp these factors risks disastrous AI policies that carry even more disastrous health care consequences.

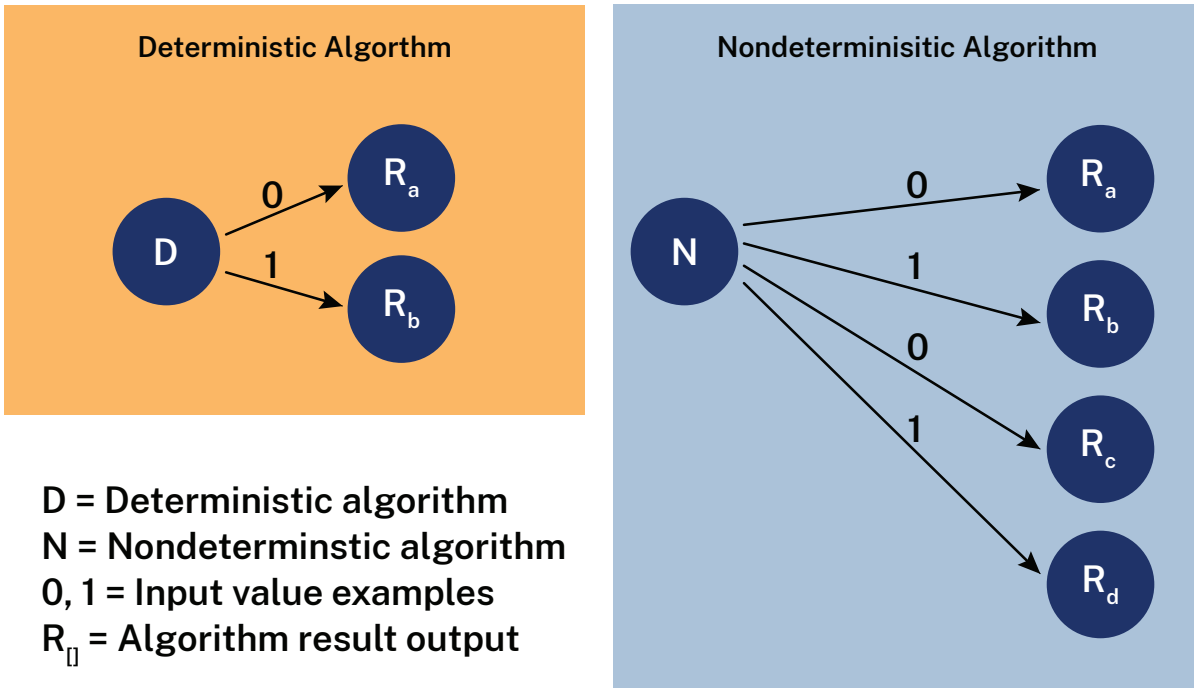
SECTION I: THE AI UNPREDICTABILITY PROBLEM

The foundation of AI functionality is the underlying machine learning algorithms that process system inputs. Those inputs can be formal data points or items that are transformed into data, such as medical images and the words spoken by patients. Each machine learning algorithm is its own set of procedures for converting an input value into an output value through statistics, probability, logic, calculus, or some combination thereof. The processing itself performs activities such as a classification, prediction, or inference. Some machine learning algorithms — such as decision trees, linear regression, and support vector machines — are deterministic functions where identical inputs consistently produce the same predicted output. In other words, there is no randomness in the equations or the outputs they produce.

When the algorithms within an AI system are trained on data, the result is an AI model. This AI model encompasses the parameters, weights, and data relationships that facilitate

11 *Premarket review* refers to the FDA's processes that establish the safety and effectiveness of medical devices under its regulatory oversight

Figure 1: Deterministic Versus Nondeterministic Algorithms



SOURCE: Author's original illustration.

successful processing of new inputs. A model may be static or adaptive, the latter of which will be discussed regarding training data that continually adjusts the model and changes outputs over time.

1. Unpredictability Concerns

Alongside deterministic algorithms are those that are stochastic — that is to say, involving a degree of randomness or unpredictability in outputs. In a nondeterministic algorithm (see Figure 1 above), an input can produce multiple output possibilities (or different states if the algorithm is embedded within a larger algorithm or model). Multiple output possibilities means that the output cannot be definitively known before the output is generated. In contrast, the eventual output of a deterministic system is known based on knowledge of the input.

Within health care AI, two programming architectures — large language models (LLMs) and generative AI — have attracted the greatest concern regarding unpredictability.[12] These architectures, while occurring independent of one another, have also been combined in systems described as foundation models.[13] The broad utility of LLMs, generative AI, and foundation models have made these technologies very attractive to software manufacturers because of their ability to reduce time and expense when developing new health care solutions. Most importantly, the language interpretation abilities and learning built into the foundation model can be transferred to new (but analogous) contexts and problems. However, very public errors[14] produced by LLMs and generative AI have given policymakers serious reservations regarding their safety in health care settings. Further, while not discrediting the technology as a whole, there have been instances of egregious health care AI errors such as IBM's Watson for Oncology where the system produced "multiple examples of unsafe and incorrect treatment recommendations" to doctors.[15] Perhaps more infamous was the Epic Sepsis Model's performance in a 2021 study where the AI's predictive accuracy of sepsis for hospitalized patients was "substantially worse than the performance reported by its developer."[16]

AI errors have introduced a new meaning to an old word: *hallucination*. Misleading both in its inferences of consciousness as well as uniformity of error, an *AI hallucination*[17] is actually a generic term describing several different AI anomalies:

- Unintelligible language outputs. A ChatGPT prompt that requested a family biography on Michael Jackson outputted "Schwittendly, the sparkle

12 Regarding LLMs, see the "Non-Reproducibility" section of the meta study "Current applications and challenges in large language models for patient care" that discusses inconsistent outputs across multiple iterations of the same input. Felix Busch et al, "Current Applications and Challenges in Large Language Models for Patient Care: A Systematic Review," *Communications Medicine*, January 21, 2025, https://www.nature.com/articles/s43856-024-00717-2. Regarding generative AI's capacity for indeterminacy, see Quanhan Xi and Benjamin Bloem-Reddy, "Indeterminacy in Generative Models: Characterization and Strong Identifiability," Proceedings of the 26th International Conference on Artificial Intelligence and Statistics, Valencia, Spain, 2023, https://proceedings.mlr.press/v206/xi23a/xi23a.pdf. See also a study specific to GPT-4 (the most used instance of generative AI technology) by Samuel J. Aronson et al., "GPT-4 Performance, Nondeterminism, and Drift in Genetic Literature Review," *New England Journal of Medicine AI* 1, no. 9 (August 8, 2024), https://ai.nejm.org/doi/full/10.1056/AIcs2400245.

13 The Stanford Institute for Human-Centered AI's influential definition of *foundation model* is more vague, removing reference to specific AI technologies and, instead, focusing on training data characteristics and extensibility. It describes an AI foundation model as "any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks." Rishi Bommasani et al., "On the Opportunities and Risks of Foundation Models," Center for Research on Foundation Models, 2021, https://crfm.stanford.edu/assets/report.pdf.

14 A well-publicized failure involved an AI-enabled voice recognition ordering system at more than 100 McDonald's fast food restaurants that resulted in "misinterpreted orders ranging from bacon-topped ice cream to hundreds of dollars' worth of chicken nuggets." Tom Gerken, "Bacon Ice Cream and Nugget Overload Sees Misfiring McDonald's AI Withdrawn," *BBC*, June 18, 2024, https://www.bbc.com/news/articles/c722gne7qngo.

15 Casey Ross and Ike Swetlitz, "IBM's Watson Supercomputer Recommended 'Unsafe and Incorrect' Cancer Treatments, Internal Documents Show," *STAT News*, July 25, 2018, https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/.

16 Andrew Wong et al., "External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients," *JAMA Internal Medicine* 181, no. 8 (June 21, 2021), https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2781307.

17 Cf. IBM, "What Are AI Hallucinations?," https://www.ibm.com/topics/ai-hallucinations. For a more detailed treatment see Yujie Sun et al., "AI Hallucination: Towards a Comprehensive Classification of Distorted Information in Artificial Intelligence-Generated Content," *Humanities and Social Sciences Communications*, September 27, 2024, https://www.nature.com/articles/s41599-024-03811-x.

of tourmar on the crest has as much to do with the golver of the 'moon paths' as it shifts from follow."[18]

- Plausible, but factually inaccurate, claims.
- Answers that are accurate but are misaligned with the intent of the end user's questions. An LLM is given a prompt that requests the optimal resting heart rate for a healthy adult male, and the LLM answers, "The optimal blood pressure for a healthy adult male is below 120 over 80."
- Citations of resources that do not exist. A 2023 study in *The American Economist* found that 20 percent of ChatGPT citations were false for prompts based on *Journal of Economic Literature* categories.[19]

Hallucinations qualify as unpredictability only if their occurrence is irregular despite identical inputs (otherwise known as prompts). In the absence of irregularity, hallucinations are important types of software errors outside the bounds of this discussion.

Although various strategies[20] have been developed to reduce or eliminate hallucinations, one of their most frequent causes illuminates an important contributor to AI unpredictability: output creativity. **Creative outputs are an intentional feature in LLMs and generative AI** and thus the potential for unpredictability within creative outputs has a structural cause rather than a programming accident. Without creativity, AI systems would be unable to generate novel ideas and solutions that are not just extrapolations of existing knowledge and patterns.

2. Structural Unpredictability: Irregularity Arising from Software Design

Structural unpredictability, in and of itself, is not a programming error in the traditional sense. It is a purposeful aspect of software design that fulfills a desired system operation (e.g. originality or creativity) but also has the capacity to produce erroneous outputs. Given the seeming irreconcilability of a design feature being both intentional and erroneous, it is worthwhile to review several examples of structural unpredictability. (**Non-technical readers are invited to skip to the next section "Data issues and unpredictability."**) However, before reviewing these examples it is important to discuss a more fundamental issue lying behind structural unpredictability and, in fact, AI itself: statistical uncertainty.

Statistics, which is a basis for much of AI, uses a finite number of observations (i.e., data) to construct rules that produce predictions, categorizations, or inferences based on information

18 Aman Sharma, "LLM vs Generative AI Insights for a Robust AI Tech Stack," Lamatic.ai, December 12, 2024, https://blog.lamatic.ai/guides/llm-vs-generative-ai/.

19 Joy Buchanan et al., "ChatGPT Hallucinates Non-Existent Citations: Evidence from Economics," *The American Economist* 69, no. 2 (November 17, 2023), https://www.researchgate.net/publication/376855338_ChatGPT_Hallucinates_Non-existent_Citations_Evidence_from_Economics.

20 Examples include Retrieval-Augmented Generation, highly specialized data training, and reducing prompt ambiguity.

related to a new situation. For example, a manufacturer trained AI on a combination of vital signs (pulse, respiratory rate, temperature, blood pressure, etc.) from past patients who did and did not develop sepsis to predict for new patients the condition's risk within the next 48 hours.[21] Because the data used to construct the rules are based on a limited data selection and the dataset itself may include some randomness, there is a degree of uncertainty intrinsic to the rules. This shortcoming has occasioned various uncertainty-sensitive mitigations such as Bayesian artificial neural networks using Markov Chain Monte Carlo (MCMC) algorithms. The basic Bayesian framework updates its rule parameters based on new data and calculates outcomes in terms of probabilities (which quantifies the persistence of uncertainty in its results). An MCMC algorithm seeks to improve the accuracy of a Bayesian network through a data sampling method whose product is reflective of the likelihood of each datum's occurrence — that is to say, the target distribution. Thus, just as the outcomes of a basic Bayesian framework express uncertainty, the MCMC algorithm incorporates uncertainty (in the form of probability-distributed data) at the level of the framework's rule parameterization while simultaneously operating to improve the accuracy of the Bayesian network.

AI unpredictability is on the continuum of statistical uncertainty given that its expression may be due to probabilistic calculations, limitations in the representativeness of training data, or both. For example, generative AI systems and LLMs have a parameter known as *temperature*. Temperature affects how random an output will be produced during the inference process whereby the AI system responds to a prompt. In an LLM, the temperature setting modifies (increasing or decreasing) the differentiation between tokens. A *token*, in the context of an LLM, may be a word, a phrase, a segment of a word, punctuation, or even a character[22] used in language processing alongside other operations such as grammatical parsings. If an LLM's temperature is set low, then the LLM will construct an output that orders tokens based on the highest statistical likelihood[23] for their places within a language sequence. A low temperature setting increases the determinism in token selection but does not eliminate unpredictability completely. For example, while uncommon, two tokens could tie for having the highest probability for being correct within a given utilization, and the system could select one of the tied tokens arbitrarily. This specific token choice might not be repeated if the same scenario recurs. Likewise, two prompts making the same request but using different wordings can

21 Christopher Barton et al, "Evaluation of a machine learning algorithm for up to 48-hour advance prediction of sepsis using six vital signs," Comput Biol Med., April 24, 2019, https://pmc.ncbi.nlm.nih.gov/articles/PMC6556419/

22 A token, in the context of an LLM, may be a word, a phrase, a segment of a word, or even a character. Michael Humor, "Understanding 'Tokens' and Tokenization in Large Language Models," *Medium*, September 10, 2023, https://blog.devgenius.io/understanding-tokens-and-tokenization-in-large-language-models-1058cd24b944.

23 This determination is affected by the number of tokens used by the LLM in the "context window" for text generation.

produce differing outputs,[24] in part due to the attention mechanism within an LLM.[25] The attention mechanism calculates relevance weights for tokens based on their interrelationships among one another and provides context sensitivity, which, in turn, determines the correct semantic relationships among words within an output. The weighting is multi-directional: It is bidirectional for tokens within a sentence and trans-directional across sentences. Given that, different prompt constructions can produce different weightings for the same tokens and, as a consequence, output variability.

Output variability can be purposely encouraged by setting an LLM temperature to high. A high temperature LLM reduces the differences among numeric values assigned to individual tokens, resulting in more novel (and less deterministic) outputs. The temperature principle applies also to generative AI except that the output being generated may be an image, sound, video, music, or other item as opposed to just language.

Temperature can contribute to unpredictable outputs, but it is not the only factor. Sampling, where an LLM creates an output[26] where tokens are selected because of a probability distribution instead of the highest score among competitors, interferes with deterministic predictions of LLM outputs. Other structural factors within an AI algorithm can also insert indeterminacy into an LLM. A *mixture of experts* (MoE) architecture, by way of illustration, portions its artificial neural network into multiple subnetworks, each serving as an "expert" specializing in a subset of input data but collaborating with other experts on the completion of a task.[27] An MoE attempts to route a prompt-related activity to an appropriate expert subnetwork that is neither under-trained nor overfitted.[28] To accomplish this comparative homogeneity, the MoE must prevent a subset of experts from receiving a disproportionate amount of tasks during training and becoming much better suited to these tasks than competing experts. If not, the MoE will exacerbate this advantage, because future task assignments will be biased toward the best trained experts and reinforce this advantage. One

24 A particularly worrisome instance of this within health care concerned brand versus generic drug names. A 2024 study observed "a surprising drop in the performance of LLMs on common medical benchmarks when the drug names are swapped from generic to brand names: 4% drop in accuracy on average." Jack Gallifant et al., "Language Models Are Surprisingly Fragile to Drug Names in Biomedical Benchmarks," *Findings of the Association for Computational Linguistics: EMNLP 2024*, November 2024, https://aclanthology.org/2024.findings-emnlp.726.pdf.

25 Ashish Vaswani et al., "Attention Is All You Need," Association for Computing Machinery, December 4, 2017, https://dl.acm.org/doi/10.5555/3295222.3295349.

26 This output creation is formally known as a decoding strategy.

27 Dave Bergmann, "What Is Mixture of Experts?," IBM, April 5, 2024, https://www.ibm.com/think/topics/mixture-of-experts. An individual expert may combine a specialization in a less common task alongside general expertise in tasks that are common, with this expertise replicated among other experts.

28 Bergmann, "What Is Mixture of Experts?" Under-training, otherwise known as underfitting, describes an AI algorithm that fails to fully detect the patterns and relationships among training data and, as a consequence, produces poor quality outputs (e.g., predictions, classifications, decisions). Overfitting describes an AI algorithm that is too closely tied to its training data so that its generalization to new data is limited.

means to prevent the privilege of a subset of experts is to inject some Gaussian noise[29] into the calculations affecting individual expert selection for a task and, thus, incorporate some randomness into the system.[30] The manifestation of this randomness in outputs for identical prompts may be inconsequential or material.

3. Data Issues Affecting Unpredictability

The causes of LLM and generative AI unpredictability extend beyond programming architecture to the data. In AI medical devices, algorithms transform training data into software rules that the device then uses to evaluate real-world inputs and produce outputs. The characteristics of training data, as well as subsequent inputs, can potentially cause output unpredictability.

Assuming an absence of errors in deployment of inputs, major data issues related to AI unpredictability include the following:

- *Inadequate training data "scrubbing" and normalization.* When the data resources used to train an AI model originate from multiple sources, there is the possibility for quantitative and qualitative inconsistencies such as divergent terminology, information labels, measurements, time scales, etc. Data scrubbing is the process by which inaccurate, noisy,[31] and redundant data are remedied, and data normalization is the efficient organization of data according to standardized metrics. Inadequate scrubbing and normalization in training data may result in AI output unpredictability.[32]
- *Open dataset training data.* Adaptive AI systems continue to learn through ongoing use, because the system continues to train *after* implementation by a health system (i.e., deployment). An adaptive design means that the training dataset for adaptive AI is open (*open* in this context conveying no predetermined boundary on the amount of data points used). Through mechanisms such as reinforcement learning — as well as evolutionary algorithms that can modify weightings and biases — adaptive AI systems attempt to become better with experience (i.e., more accurate predictions, classifications, etc.). By virtue of being adaptive, such systems have the

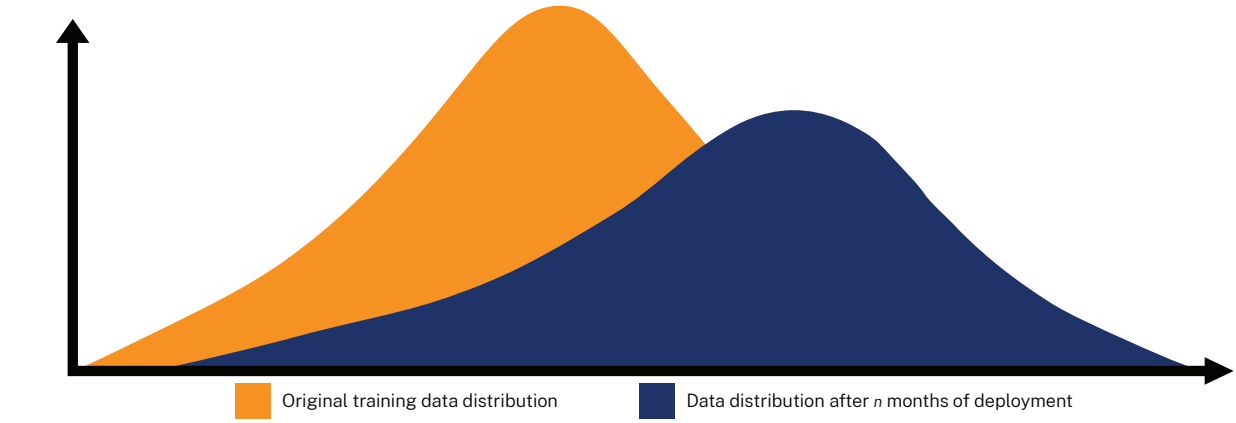
29 Gaussian noise injection is the augmentation of operational data with additional random values that follow a normal, or Gaussian, probability distribution.

30 Bergmann, "What Is Mixture of Experts?" Other data conditions can produce expert assignments, most notably individual expert capacity and the routing of tasks to secondary choice experts. See also Yanqi Zhou, "Mixture-of-Experts with Expert Choice Routing," Google Research, November 16, 2022, https://research.google/blog/mixture-of-experts-with-expert-choice-routing/.

31 *Noisy* refers to data that lacks a discernible pattern and, thus, interferes with algorithm training. Data collection errors and statistical outliers are variables that can contribute to noise.

32 Cf. Gallifant et al., "Language Models Are Surprisingly Fragile."

Figure 2: Data Drift



SOURCE: Author's original illustration.
NOTE: This image illustrates a change in distribution over time. Being a generic example, this diagram does not include axis labels.

possibility of outputs changing in unpredicted ways over time for the identical inputs. While the assumption is that such output adaptations will be an improvement, there is the risk of "data drift" (otherwise known as covariate shift). *Data drift* refers to a change in training data distribution that can affect system outputs and, in some cases, impair their accuracy.

- *Incomplete training data.* When training data is inadequate to train an AI system on all the input possibilities for which it will produce outputs, the system may make unreliable generalizations due to the system's epistemic deficits.[33] This may be the case even when manufacturers augment insufficient real-world training data with "synthetic data."[34] With respect to radiology, for example, incomplete training data may result in inconsistent outputs for medical images that do not align closely with training data. As a result, the AI system may not dependably identify the same pattern(s) as most relevant for disease classification. In addition, remedying incomplete training data with synthetic data may reduce system accuracy as compared to relying on an adequate supply of real-world data.[35] Another dimension of unpredictability resulting from incomplete training data is unanticipated performance problems for populations underrepresented in training data. Another way of expressing

33 In statistics, an epistemic deficit is a lack of information that produces uncertainty and the possibility of an incorrect conclusion (e.g. a categorization or prediction).

34 *Synthetic data* refers to data derived from AI generation as opposed to real-world data collection.

35 Debbie Rankin et al., "Reliability of Supervised Machine Learning Using Synthetic Data in Health Care: Model to Preserve Privacy for Data Sharing," *JMIR Medical Informatics* 8, no. 7 (July 2020), https://medinform.jmir.org/2020/7/e18910/.

this is incomplete training data may create a situation where the outputs appropriate for some populations may not be satisfactory for others. This situation may argue for certain AI devices to be selectively deployed in order to avoid patient injury.[36] Selective deployment means advocating an AI device for certain populations but not others so that the latter is not harmed. An example of this would be a breast cancer detection device whose training data was largely female and, thus, not advocated for men despite their own potential of breast cancer.[37] With respect to foundation models, incomplete training data may be better categorized as a deficit of specialization within the training data.

- *Synthetic data.* AI-generated data, known as synthetic data, reflects the patterns AI observes in real-world data but lacks the diversity that real-world data exhibits. When used to train an AI model, synthetic data can contribute to an early-or late-model collapse where the performance of the AI system degrades.[38]
- *Input data ambiguity or complexity.* Both academics and the public have noticed a relationship between higher degrees of input semantic uncertainty (i.e., ambiguity) and an LLM's increased likelihood to produce an arbitrary or otherwise incorrect output.[39] Ambiguous inputs may fail to provide the necessary context or intent for an AI system to produce consistent outputs for its use. Input data ambiguity can also intersect with incomplete training data in instances where the input utilizes a word or words outside the vocabulary of the training data. Complex inputs, like input ambiguity, present similar challenges in interpretation and "may result in different degrees of processing difficulty and thus also lead to variation in the interpretation process."[40]
- *Input data uncertainty.* AI manufacturers cannot necessarily predict all real-world inputs, especially with respect to questions (prompts) asked of LLMs, generative AI, and foundation models. Consequently, these systems may have incomplete or underspecialized data whose deficits manifest in unpredictable outputs.

36 Robert Vandersluis and Julian Savulescu, "The Selective Deployment of AI in Healthcare," *Bioethics* 38, no. 5 (February 16, 2024), https://onlinelibrary.wiley.com/doi/10.1111/bioe.13281.

37 Ibid.

38 Ilia Shumailov et al., "AI Models Collapse When Trained on Recursively Generated Data," *Nature* 631 (July 24, 2024), https://www.nature.com/articles/s41586-024-07566-y.

39 See Lance Eliot, "The Best Prompt Engineering Techniques for Getting the Most out of Generative AI," *Forbes*, May 9, 2024, https://www.forbes.com/sites/lanceeliot/2024/05/09/the-best-prompt-engineering-techniques-for-getting-the-most-out-of-generative-ai/.

40 Joris Baan et al., "Uncertainty in Natural Language Generation: From Theory to Applications," arXiv, July 28, 2023, https://arxiv.org/pdf/2307.15703.

Although there are strategies for both quantifying and remediating[41] unpredictabilities resulting from training and input data conditions, they are still emergent and unperfected. Accordingly, unpredictability remains an issue for medical applications of AI and a risk for patient safety.

4. Unpredictability and Regulation

The FDA considers AI to be a medical device (technically "software as a medical device") when it is used to diagnose disease and health conditions; affects either the structure or function of the body; or addresses disease through prevention, mitigation, treatment, or cure.[42] The FDA's safety review process for medical devices was shaped by physical appliances (e.g. x-ray machines, MRIs, artificial joints, etc.) and deterministic software. Consequently, the review process is biased toward premarket validation. In contrast, AI's unpredictability — whether intentional (in the case of adaptive systems) or unintentional (in the case of systems with nondeterministic factors) — argues for reviews that extend into the market after AI is approved for commercial use. The FDA already has a pathway for postmarket monitoring. Under Section 522 of the Federal Food, Drug, and Cosmetic Act, the FDA may oblige a manufacturer to perform a study that collects and analyzes data on a marketed medical device.[43] A Section 522 study could be justified under one of the following conditions[44] if the AI device:

- is likely to have serious adverse health consequences,
- includes significant use in pediatric populations,
- is implanted in a patient for more than one year (such as a pacemaker), or
- is a life-sustaining or life-supporting device used outside a device user facility.

While the need for AI postmarket monitoring has been widely discussed in health care, "there is little consensus on how to design effective monitoring systems for the post-deployment setting."[45]

41 An example of a remediation strategy that attempts to address unpredictability from data inconsistencies can be seen in Jihye Choi et al., "MALADE: Orchestration of LLM-Powered Agents with Retrieval Augmented Generation for Pharmacovigilance," arXiv, August 3, 2024, https://arxiv.org/pdf/2408.01869.

42 FDA, "How to Determine If Your Product Is a Medical Device," September 29, 2022, https://www.fda.gov/medical-devices/classify-your-medical-device/how-determine-if-your-product-medical-device.

43 FDA, "Postmarket Surveillance Under Section 522 of the Federal Food, Drug, and Cosmetic Act," October 2022, https://www.fda.gov/regulatory-information/search-fda-guidance-documents/postmarket-surveillance-under-section-522-federal-food-drug-and-cosmetic-act.

44 FDA, "522 Postmarket Surveillance Studies Program," October 6, 2022, https://www.fda.gov/medical-devices/postmarket-requirements-devices/522-postmarket-surveillance-studies-program.

45 Jean Feng et al., "Not All Clinical AI Monitoring Systems Are Created Equal: Review and Recommendations," *New England Journal of Medicine AI* 2, no. 2 (January 23, 2025), https://ai.nejm.org/doi/full/10.1056/AIra2400657.

SECTION II: EXISTING EFFORTS AND PROPOSALS: STRENGTHS AND GAPS

The FDA's draft guidance on Artificial Intelligence-Enabled Device Software Functions, issued in January 2025, stresses the need for a total product life cycle[46] (TPLC) approach and clearly lays out the expectations for submission materials expected of AI manufacturers. The TPLC operates within a larger matrix of FDA market pathways for medical devices, whether AI or not. These pathways provide formal review tracks as well as limited exemptions.

Given the uncertainty associated with outputs generated by some AI technologies, several approaches have been proposed to address the potential safety concerns. These range from centralized approaches (described as "assurance laboratories") to decentralized proposals that adapt the Clinical Laboratory Improvement Amendments (CLIA) concept to AI. While these approaches contain important proposals and expectations, they are not without limitations.

1. Current Regulatory Process

The January 2025 FDA draft guidance outlines the process to receive approval to market AI-enabled medical devices in the United States. This process broadly parallels the FDA's approach to regulating medical devices in general. It starts with early FDA engagement, leading to context of use definition, which needs to identify clearly the application, target population, and types of data that will be analyzed. The FDA expects the manufacturers to perform risk-based credibility assessment — involving evaluation of the model's performance, reliability, and relevance to the context of use — and to provide comprehensive documentation of the AI model's development process, including data sources, algorithms, validation methods, and performance metrics. These need to be informed by rigorous premarket validation studies in real-world scenarios that generate evidence supporting the AI model's safety, effectiveness, and quality based on predefined performance criteria. The new guidance emphasizes the importance of TPLC management, which includes plans for monitoring the device after it is deployed in the market and updating its AI model.

After the above steps are completed, the manufacturer submits a detailed regulatory dossier for FDA's review. This process may involve multiple rounds of review and feedback, with marketing approval as the desired conclusion.

46 Lifecycle refers to the entire period from initial planning, development, and testing to a medical device's use in the marketplace after FDA review (including iterative improvements or decommissioning).

Table 1: Current FDA Medical Device Review Pathways

FDA Pathway	Use Case	Risk Level(s)	Examples
Exempt from FDA Submission	Premarket review formally exempted for a low-risk medical device with the FDA	Low	Surgical apparel
Enforcement Discretion	Premarket review not enforced for a low-risk medical device with the FDA reserving the right to enforce regulation in the future	Low	Mobile apps enabling a patient to send an alert or general emergency notification to first responders
Premarket Notification (510(k))	Premarket review for a device that is substantially equivalent to an FDA product that was not subject to premarket approval	Intermediate	Radiological computer-assisted detection/ diagnosis software for fracture (AI medical device)
Premarket Authorization	The most rigorous premarket review pathway reserved for high-risk medical devices	High	Imagio Breast Imaging System (AI medical device)
De Novo	Premarket review of devices of low to moderate risk and not substantially equivalent to any FDA-approved medical device	Low to intermediate	Sepsis ImmunoScore (AI medical device)
Humanitarian Device Exemption	An exemption of FDA effectiveness requirements for a medical device (with no competitive alternatives) that benefits patients suffering from rare diseases	Low to high	PulseRider Aneurysm Neck Reconstruction Device
Breakthrough Devices Program	Expedited review in existing FDA pathways for devices offering more effective treatment or diagnosis of life-threatening conditions and irreversibly debilitating diseases	Low to high	EVOQUE Tricuspid Valve Replacement System

SOURCE: Author's original table.

1a. Total Product Life Cycle Management

While the primary focus on the January 2025 FDA draft guidance was to provide more clarity on marketing submissions for AI-enabled device software functions, it emphasized the importance of TPLC view for these technologies. The goal of the TPLC approach is to ensure that AI-enabled devices are safe, effective, and reliable throughout their entire life cycle, from

development to decommissioning. Based on another paper by FDA authors,[47] the life cycle of AI development consists of seven phases:

1. Planning and design
2. Data collection and management
3. Model building and tuning
4. Verification and validation
5. Model deployment
6. Operation and monitoring
7. Real-world performance evaluations

With its primary focus on marketing submissions, the draft guidance spends the most time on the premarket portions of TPLC and provides important details on the FDA's expectations regarding the development and testing information necessary for marketing authorization. Still, the guidance recognizes that devices deployed in real-world settings "may change or degrade over time, presenting a risk to patients." It acknowledges that these changes may be caused by many factors, including changes in patient populations, disease patterns, or data drift. It does not explicitly list the inherent uncertainty associated with some generative AI technologies. Importantly, the FDA acknowledges, "Because the performance of AI-enabled devices can change as aspects of the environments in which they are approved or cleared for use in may change over time, it may not be possible to completely control risks with development and testing activities performed premarket (prior to device authorization and deployment)."

To address these concerns, the draft guidance states, "As part of their ongoing management of AI-enabled devices manufacturers should proactively monitor, identify, and address device performance changes, as well as changes to device inputs and the context in which the device is used that could lead to changes in device performance. In addition, sponsors must develop and implement plans for comprehensive risk analysis programs and documentation consistent with the Quality System Regulation (21 CFR Part 820) to manage risks related to undesirable changes in device performance for AI-enabled devices." It also requires that manufacturers report to the FDA "information about deaths, serious injuries, and malfunctions in accordance with 21 CFR Parts 803 and 806."

Furthermore, the draft guidance describes the concept of performance monitoring and highlights several components that such plans should include:

47 Manesh R. Patel, Suresh Balu, Michael J. Pencina, "Translating AI for the Clinician," JAMA, October 15, 2024, https://jamanetwork.com/journals/jama/article-abstract/2825145

- Description of the data collection and analysis methods for identifying, characterizing, and assessing changes in model performance and monitoring potential causes of undesirable changes in performance;
- Description of robust software life cycle processes that include mechanisms for monitoring in the deployment environment;
- Plans for deploying updates, mitigations, and corrective actions; and
- Description of the procedures for communicating the results of performance monitoring and any mitigations to device users.

Acknowledging that the AI-enabled device manufacturers do not control the environments in which their products are deployed, the FDA encourages (but does not mandate) the inclusion of performance monitoring plans with marketing submissions. Still, it leaves the door open for *requiring* performance monitoring plans in some circumstances, including premarket authorization and de novo submissions.

2. Limitations on FDA Oversight of AI-Enabled Devices

The FDA draft guidance on Artificial Intelligence-Enabled Device Software Functions is an important step in improving and clarifying the regulatory environment for health AI. Its promotion of the TPLC approach is critically important, and the thoughtful section on postmarket monitoring raises numerous important points. However, there exist objective limitations, which may complicate progress in the health AI ecosystem.

First, as acknowledged by some FDA authors,[48] the agency lacks sufficient workforce, both in numbers and expertise, to evaluate all health AI device technologies that are in scope for its authority. With the increased momentum for national debt reduction, it is unlikely that the agency will grow in size. Thus, even if developers submit strong and clear proposals according to the recommendations contained in the guidance, the FDA will either have to limit the scope of what it will review or create bottlenecks that slow down approvals. Either scenario is likely to hamper progress, the former by increasing consumer risk and potentially leading to a negative backlash against health AI technologies and the latter by slowing down adoption of health technologies that save lives or make the national health complex more efficient.[49]

Second, despite FDA's efforts to define the scope of its jurisdiction,[50] there still exists a large grey area of AI-enabled software with unclear responsibility for verification of its

48 Haider J. Warraich et al., "FDA Perspective on the Regulation of Artificial Intelligence in Health Care and Biomedicine," JAMA 333, no. 3 (October 15, 2024), https://jamanetwork.com/journals/jama/article-abstract/2825146.

49 See Kev Coleman, "Healthcare AI Regulation: Guidelines for Maintaining Public Safety and Innovation," Paragon Health Institute, December 2024, https://paragoninstitute.org/private-health/healthcare-ai-regulation/.

50 FDA, "Clinical Decision Support Software: Guidance for Industry and Food and Drug Administration Staff," September 28, 2022, https://www.fda.gov/media/109618/download.

performance. This includes higher-risk technologies developed by a local health system for exclusive use within that health system (for example, sepsis risk prediction tools) but not yet commercialized for general market use, algorithms offered to customers by EHR vendors (for example, Epic's suite of risk prediction tools), and technologies at the intersection of clinical and operational applications (for example, ambient voice recognition tools or automated pre-authorization technologies). Currently, the FDA does not review these technologies, and it is unclear who is responsible for assuring their quality and monitoring their performance.

Third, while the FDA's draft guidance contains several important recommendations for postmarket monitoring, the agency stops short — and for objectively good reasons — of making it a requirement that it will enforce. As explained in the draft guidance and suggested elsewhere,[51] the performance of health AI depends on the local context and requires local data for meaningful monitoring. We argue that there is almost a gradation, from least to greatest, between drugs, medical devices, and health AI-enabled technologies in terms of their dependence on the local context. Thus, it is impossible for a central governmental agency to unilaterally create a system that enables high-quality postmarket monitoring. This limits the FDA's primary focus to the necessary but not sufficient space of premarket testing, which does not address the local context (e.g. population health trends in the area), local data (most importantly patient data), and unpredictability issues associated with health AI technologies.

3. Postmarket Surveillance for Drugs, Devices, and Biologics: The Sentinel Initiative

The problem of postmarket surveillance is neither new nor limited to AI medical devices. The FDA Amendment Act of 2007 stipulated that the agency establishes an active postmarket risk identification and analysis system for products under its jurisdiction. In response, in 2008 the FDA launched the Sentinel Initiative to detect early signals of adverse safety events in pharmaceuticals, devices, and biologics that have received marketing approval. The Sentinel Initiative uses a distributed data approach, where data remains with local data owners (i.e., insurance companies, health care providers etc.), helping protect patient privacy and ensure data security.[52] Sentinel incorporates and standardizes data from multiple sources (insurance claims, EHRs, patient registries) using a common data model. This standardization was intended to increase efficiency of the safety assessments, which are performed using multiple analytical techniques, including routine querying, statistical analyses (propensity score analyses, case series, sequential testing, distributed regression) and machine learning

51 Alexey Youssef, Michael Pencina, Anshul Thakur, Tingting Zhu, David Clifton and Nigam H. Shah, "External validation of AI models in health should be replaced with recurring local validation," Nature Medicine, October 18, 2023, https://www.nature.com/articles/s41591-023-02540-z

52 FDA, "About the Food and Drug Administration (FDA) Sentinel Initiative," https://www.sentinelinitiative.org/about.

approaches (classification algorithms, natural language processing methods for unstructured data).

Sentinel is a collaborative network, involving academic institutions, health care organizations, industry partners, and regulatory bodies, with the Harvard Pilgrim Health Care Institute serving as the Sentinel Operations Center. Some successes of the Sentinel program include detection of safety signals related to anticoagulants (bleeding risks), some diabetes medications (increased risk of heart failure), and opioids, though the program has also had failures.[53]

> "... performance unpredictability is unique to the type of product itself. This means that, at least for some of the AI-enabled devices, it will be necessary to have a surveillance system that is more specific to a particular device, which is not how the current Sentinel network is designed."

The Sentinel Initiative began launched as a mini-Sentinel pilot in 2014 and transitioned to the full Sentinel system in 2016,[54] both dates predating AI's accelerating health care use in the 2020s. Understandably, there does not exist an equivalent of the Sentinel Initiative for AI-enabled medical devices. Although Sentinel could be pointed to capture safety signals associated with AI-enabled devices, it is not clear that the entire framework can address the requirements specific for this category of devices. Unlike pharmaceuticals, biologics, or even "traditional" medical devices, the performance of AI-enabled devices is heavily influenced by the local application context (i.e., a health system's protocols around AI use, staff training, and the local population health trends of the patients it serves). Moreover, as outlined above, performance unpredictability is unique to the type of product itself. This means that, at least for some of the AI-enabled devices, it will be necessary to have a surveillance system that is more specific to a particular device, which is not how the current Sentinel network is designed. However, the general Sentinel model — with its distributed, standardized data model and collaborative network of diverse stakeholders — offers valuable blueprints for what might be needed for AI-enabled devices.

53 Sheila Kaplan, "Failure to Warn: An Early Warning System for Drug Risks Falls Flat," *STAT News*, June 6, 2017, https://www.statnews.com/2017/06/06/sentinel-fda-drug-risks/.

54 FDA, "FDA's Sentinel Initiative," March 8, 2024, https://www.fda.gov/safety/fdas-sentinel-initiative.

4. Assurance Labs

Recognizing that no governmental agency will be able to evaluate all AI technologies being developed[55] and inspired by industry self-regulation exemplified by entities like the Underwriters Laboratories, the Coalition for Health AI (CHAI) proposed the concept of AI quality assurance labs.[56] The proposal postulates a diverse set of new nonprofit and for-profit organizations that would aggregate sufficiently representative data from multiple health systems, execute and evaluate health AI algorithms on these data, and create an algorithm performance report according to pre-defined criteria. Potential adopters could then use this report to decide if a given AI technology meets the standards for local implementation.

The clear advantages of this proposal include taking some of the burden off the FDA by promoting an instance of industry self-regulation, which can help expedite the premarket review process as well as complement safety practices effectively imposed through the insurance conditions of product liability coverage. In a world where health systems and other potential users of health AI technologies struggle with differentiating solutions that offer true value from those that are of poor quality or unclear benefits, having an objective source of truth is a desirable offering. CHAI itself does not intend to be a quality assurance lab, but it plans to be a central entity that would certify individual labs for adherence to responsible AI principles.

5. Criticisms and Limitations of Assurance Labs

Although the concept of quality assurance laboratories did envision them operating in the implementation and monitoring phases of the health AI life cycle, most of the focus — and criticism — was concentrated on the premarket phase. Some authors affiliated with the Health AI Partnership have argued that assurance labs have an "equity problem" arising from the centralized nature of the process, which may be prone to domination by larger and more affluent health systems.[57] Indeed, if assurance labs failed to gather diverse and representative sets of data tailored to AI type and medical application context, the results would lack the necessary generalizability. In addition, the internal subject matter expertise needed to certify data collections and algorithmic performance reporting for all the different clinical applications of AI is a considerable operational challenge, even for a federal agency such as the FDA. Furthermore, if the focus were solely on retrospective rather than prospective or postmarket data, the assurance service would lack the real-world local

55 Pencina, M.J., Goldstein, B.A., D'Agostino, R.B.: Prediction models-development, evaluation, and clinical application. N. Engl. J. Med. 382(17), 1583–1586 (2020)

56 Nigam H. Shah et al., "A Nationwide Network of Health AI Assurance Laboratories," *JAMA* 331, no. 3 (2024), https://jamanetwork.com/journals/jama/fullarticle/2813425.

57 Mark P. Sendak et al., "AI Assurance Labs Intended to Test Health Care Technology Have an Equity Problem," *STAT News*, February 7, 2024, https://www.statnews.com/2024/02/07/ai-assurance-laboratories-onc-fda-equity/.

application context and be prone to the same limitations described in the previous section. Instead, it would operate as a type of premarket certification parallel to that of the FDA.

An author affiliated with venture capital as well as some congressional representatives have raised concerns related to potential regulatory capture resulting from large technology companies playing a significant role within an AI review process.[58] As an alternative some members of the AI community have proposed an alternative process based on localized quality assurance that "would provide resources to allow every provider to operate its own review process, rather than consolidating these reviews with a handful of big tech companies and academic medical centers."[59] Related to these issues are conflicts of interest and intellectual property concerns. It would not be desirable for a given entity to provide a platform or funding for an assurance lab while simultaneously developing similar, competing products to those products being evaluated by the lab.

Another untested feature of the assurance lab proposal is the financial model. In the drug and device arena, developers must pay for clinical studies that demonstrate the safety and efficacy of their products. This precedent does not extend to health AI technologies, except those regulated by the FDA. The expectation that the users (health systems, etc.) would pay for this service is unrealistic given the financial strains under which they operate and the generally poorly articulated value proposition of the emerging AI technologies. For health systems, the expense of certifying AI through a third party would be factored into the technology's total cost of ownership and, consequently, negatively affect AI adoption. The market needs to verify developers' readiness to absorb the costs: They may be willing to take on some of them, with the hope of passing them on to users once their products have been demonstrated to be of high value and applicability.

6. Decentralized CLIA-Type AI Ops Units

A contrasting approach to the Assurance Labs proposal has been inspired by the Clinical Laboratory Improvement Amendments (CLIA) model put together to ensure the reliability of laboratory testing.[60] Unlike the centralized assurance labs model, the CLIA-like approach to health AI governance is based on a decentralized model where local AI operations units serve as the accountable parties that could be accredited by existing health care accreditation agencies.[61] These local units would oversee validation and verification as well as calibration and quality control. Validation and verification would be based on local data already in the

58 Julie Yoo, "Oversight of Health AI Must Be Democratic, Not Done by the Big Tech Companies," *STAT News*, June 17, 2024, https://www.statnews.com/2024/06/17/health-ai-oversight-democratic-process-not-controlled-by-big-tech-companies/.

59 Yoo, "Oversight of Health AI Must Be Democratic."

60 Brian R. Jackson et al., "Regulation of Artificial Intelligence in Healthcare: Clinical Laboratory Improvement Amendments (CLIA) as a Model," *Journal of the American Medical Informatics Association* 32, no. 2 (February 1, 2025), https://pubmed.ncbi.nlm.nih.gov/39657218/.

61 Ibid.

possession of the health system using the AI device, greatly reducing the concerns about appropriate real-world context and data acquisition costs. However, staffing of the local AI units would require medical directors with appropriate clinical and informatics background.

This approach has several appealing features: It aligns with the principle of subsidiarity, which in this case means AI governance will be performed as close as possible to the local unit that implements the technology. Additionally, it assures the real-world applicability, meaning the desired recurrent local validation comes as a standard. Assuming appropriate staffing and resources, bottlenecks could be reduced with more decisions pushed to the local level.

7. Limitations of the Fully Decentralized Model

Despite its many attractive features, the local AI ops model is not without limitations. First, local AI governance, which we fully support, would benefit from *but does not require* exclusive reliance on local validation. In fact, local validation of an AI system's ability to generalize performance beyond its original training data (and, thus, establish that the AI model is not statistically overfit) will likely be assisted by the AI manufacturer in many cases. If an AI device fails to generalize — that is to say, successfully apply its functionality to a broader population than what was represented in its training data — the device's manufacturer will risk contract non-renewals as well as lost new sale opportunities due to word-of-mouth testimonies of device performance problems. In the case of AI startup companies, these conditions would represent an existential threat. This reality may bias the CLIA model's applicability to AI devices provided by noncommercial academic institutions and large health care systems.

Second, we worry about its feasibility. While shifting the validation and verification work to the local units unburdens the central system, including the regulators, it adds burden to local users, which in many cases are resource-strapped health systems. Specialists are needed to perform the test data labeling needed to confirm the AI device's accuracy, but the vast majority of health systems do not have access to the right personnel to staff the local AI ops units with testing expertise tailored to AI type. Existing health system staff with domain expertise are ill-equipped to evaluate AI technology whose computational operations may be able to detect early signs of illness years before an experienced clinician can.[62] Thus, the health systems must pay for considerable consulting expense from external third parties or staffing expense if the expertise is to be brought in house. In both cases, the expenses increase as the number of different AI devices are utilized, because the subject matter

62 See Zoe Kleinman, "NHS AI Test Spots Tiny Cancers Missed by Doctors," *BBC*, March 20, 2024, https://www.bbc.com/news/technology-68607059; Berkeley Lovelace Jr. et al., "Promising New AI Can Detect Early Signs of Lung Cancer That Doctors Can't See," *NBC News*, April 11, 2023, https://www.nbcnews.com/health/health-news/promising-new-ai-can-detect-early-signs-lung-cancer-doctors-cant-see-rcna75982.

expertise needed in relation to sepsis prediction, for example, is different than the expertise needed in prostate cancer diagnosis. Thus, we arrive at what amounts to be a different form of assurance lab functions: Rather than providing centralized premarket evaluation, they would offer help with local validations. The right financial model is still needed because inflated AI consulting expenses would reduce AI adoption at health systems.

Third, while the decentralized local AI ops model is sensible for local algorithms developed internally by a health system for local use, we are concerned about extending it to externally-developed solutions due to difficulties with market-level aggregation and learning. It is unclear how signals of algorithmic malfunction or drift observed at one local AI ops unit could be shared with other institutions using the same algorithmic solution. Moreover, it will be challenging for local AI ops units to adhere to the same standards, potentially leading to a fractured environment, reminiscent of what happened with the adoption of the EHR. We are also concerned about the efficiency of this approach. **If health system users perform their own validations of a given AI solution, there will be many concurrent validation studies running at the same time. This is not sustainable unless only a small percentage of health AI solutions are being validated.** It may also mean that the aggregate, national cost of validations will greatly exceed that of the development of the solution, something that our strained health care ecosystem cannot afford and under which health AI manufacturers cannot flourish. Finally, this single-point-in-time validation model is not suited to unpredictability that may manifest over an extended period.

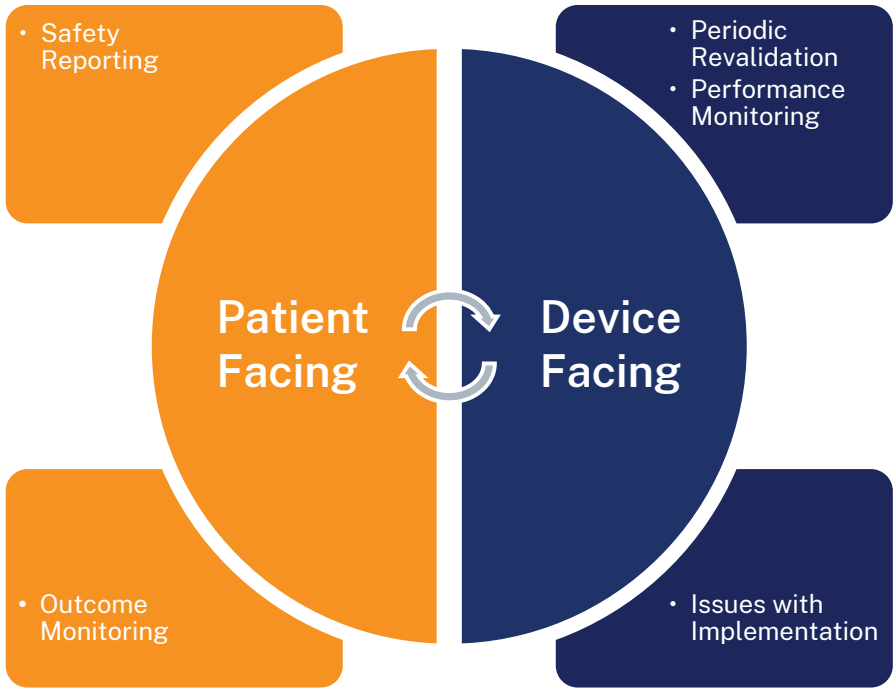
SECTION III: TARGETED POSTMARKET SURVEILLANCE

The discussions related to the Sentinel Initiative, Assurance Labs, and CLIA models have been valuable, both in exposing the safety complications associated with AI and unintentionally illuminating areas of strong industry disagreement. The ensuing national debates and challenges related to the optimal funding models suggest that any successful effort for responsible implementation of health AI must balance numerous competing concerns among manufacturers, health care providers, and regulators. To create an efficient responsible AI ecosystem and meet the needs and address concerns of the various stakeholders, we propose a model built upon postmarket surveillance performed through a public-private partnership.

Foremost in priority among the stakeholders to be appeased is the FDA. The FDA has expressed a desire to collect performance data after an AI medical device is approved and has entered the market.[63] However, the agency has pointed out it only has "authority to conduct

63 U.S. Government Accountability Office, *Federal Regulation: Selected Emerging Technologies Highlight the Need for Legislative Analysis and Enhanced Coordination*, January 2024, https://www.gao.gov/assets/d24106122.pdf.

Figure 3: Postmarket Surveillance Paradigm



SOURCE: Author's original illustration.

this postmarket surveillance in specific circumstances, such as in the case of an adverse event or if the device is recalled."[64] Even if this limitation is removed by law or regulation, there is still another impediment for the FDA: inadequate resources. Recognizing this labor constraint, former FDA Commissioner Robert Califf has commented on the FDA's need for collaboration on the postmarket oversight of AI, saying that "this is not something the FDA can do on its own. We're going to need clinical health systems and professional societies to get very involved in self-regulation just like we do on the farms, where if you're a farmer growing vegetables and there's a farm upstream that has cows contaminating the water, it's your responsibility to take that into account, and it's no different here in this postmarket phase."[65]

A public-private AI surveillance effort could help the FDA quickly identify AI device problems.

While the FDA's openness to a public-private AI surveillance is positive, incentives are still required for the other stakeholders to participate voluntarily. Health care providers, as users

64 Government Accountability Office, *Federal Regulation*. See also FDA, "522 Postmarket Surveillance Studies Program."

65 Roy Perlis and Jennifer Abbasi, "FDA Commissioner Robert Califf on Setting Guardrails for AI in Health Care," *JAMA* 332, no. 23 (November 22, 2024), https://jamanetwork.com/journals/jama/fullarticle/2827144.

Table 2: Model-Focused Postmarket Surveillance Recommendations for Medium- to High-Risk AI Medical Devices

	Postmarket Surveillance Category				
	Existing Safety Practices	Periodic Revalidation		Performance Monitoring	
AI Model/ Algorithm Type	Adapt standard safety reporting and other existing systems to capture AI-related adverse events, patient outcomes, and malfunctions	Deterministic OR Probabilistic Nonadaptive	Probabilistic Adaptive	Deterministic OR Probabilistic Nonadaptive	Probabilistic Adaptive
Training Dataset Characterization		Fixed	Open	Fixed	Open
Synthetic Data in Training Data?		True	False	False	True
Training Data Representative of Patient Cohorts?		True	False	True	False
Possible LLM Input Complexity or Semantic Ambiguity?		True	False	False	True
Structural Output Unpredictability?		True	False	False	True

Legend

- Condition strongly justifies category-specific surveillance
- Condition justifies category-specific surveillance
- Condition does not necessitate surveillance beyond existing safety practices
- Conditions covered by existing regulatory and health system protocols

SOURCE: Author's original table.

of AI medical devices, are potentially exposed to reputational, financial, and other legal liabilities in the event AI unpredictability leads to a patient injury or other adverse event.[66] Participation in postmarket surveillance of an AI device with the capacity for unpredictability could help reduce a provider's liability, as the surveillance is a good faith effort to avoid patient harm. AI device manufacturers would benefit for the same reason, which is especially important given that they do not enjoy protection from the Learned Intermediary Rule often accompanying medical device use. Under the Learned Intermediary Rule, a device manufacturer may be shielded from some legal accountability in a patient injury given that a health care provider made the decision that the device was appropriate for the patient's needs considering its risks and benefits. However, given the low explainability of many complex AI

66 The FDA defines *adverse event* as any undesirable experience related to the use of a medical product in a patient. Adverse events include death, permanent damage, and hospitalization. FDA, "What Is a Serious Adverse Event?," May 18, 2023, https://www.fda.gov/safety/reporting-serious-problems-fda/what-serious-adverse-event.

systems, this doctrine would not apply, because the provider may not be able to assess the complete scope of risk represented by the AI medical device.

1. Components of a Postmarket Surveillance System

For approved devices already deployed in the market, the FDA already mandates that undesirable experiences (e.g., permanent injury, hospitalization, death) be reported by device manufacturers, device user facilities, and device importers.[67] The agency further "encourages health care professionals, patients, caregivers and consumers to submit voluntary reports about serious adverse events that may be associated with a medical device, and use errors, product quality issues, and therapeutic failures."[68] A new postmarket surveillance framework enhancing the effectiveness of the FDA's existing medical device reporting efforts should be comprised of the following four components, two of which (outcome monitoring and adverse event reporting) are already part of hospital safety surveillance protocols (Figure 3):

1. Documenting adverse events (e.g., patient was prescribed the wrong medicine by an AI device)
2. Monitoring outcome s(e.g., the rate of re-hospitalizations increased after AI technology was introduced)
3. Identifying AI implementation issues (e.g., erroneous AI outputs occurring after an update to the device or the IT systems in which it operates)
4. Detecting troublesome performance issues (e.g., model discrimination degradation) through periodic revalidations and/or performance monitoring of AI devices at risk for unpredictability

Instead of needlessly increasing industry costs by recommending all AI devices be subject to the same level of postmarket surveillance, a new framework should increase its prospects for adoption by concentrating new surveillance interventions only on those clinical use cases that portend the highest risk to patients and health care delivery organizations and those AI devices whose structural design and/or training data characteristics present a reasonable prospect for output unpredictability given that unpredictability may not be observed during premarket review. A risk-based process for resource allocation is critical here to avoid AI technologies that present low risk (or no risk) for patient harm competing for the same limited resources with those devices whose failures have far worse repercussions. Although, in theory, the postmarket surveillance discussed in this paper may be used for AI devices beyond its proposed scope, such use would operate outside the original priorities shaping the process.

67 FDA, "What Is a Serious Adverse Event?"

68 FDA, "What Is a Serious Adverse Event?"

The above table differentiates three distinct forms of AI postmarket surveillance: existing safety practices, periodic revalidation, and performance monitoring. Their respective recommendations are decided according to six AI medical device attributes.

1. **AI model/algorithm type.** This field in the above matrix has three possibilities: deterministic, probabilistic adaptive, or probabilistic nonadaptive. An algorithm is a procedure by which a function (e.g., a categorization or prediction) is accomplished. An AI model is a set of algorithms after training data has parameterized them (e.g., determined their weights and biases). A deterministic model always delivers the same output for a specific input. Its underlying computations may be rules-based or employ another algorithm type (such as linear regression) where there are fixed relations between inputs and outputs. A probabilistic adaptive model, in contrast, may potentially produce different outputs for the same input because, as an adaptive system, its model alters over time. Such alterations, and their associated effect on outputs, is absent for a probabilistic nonadaptive model where the probabilities generated are not subject to any input randomization or structural stochasticity (e.g., randomized data sampling).[69]
2. **Training dataset characterization.** This field has two possibilities: closed and open. A closed dataset indicates that the training data has a finite number of elements and has parameterized the AI system prior to its deployment in the market. An open dataset, in contrast, is a training data collection that not only expands after deployment but can also change AI system performance after its original training.
3. **Synthetic data in training data.** This field has two possibilities: yes or no. Synthetic data is derived from AI generation as opposed to real-world data collection. When used in training data, synthetic data increases certain risks, including the possibility of model collapse (i.e., discontinuation of desired functionality).
4. **Training data representative of patient cohorts.** This field has three possibilities: yes, no, or not applicable. Training data, when derived from information produced from human beings (e.g., medical images, test results, etc.), can be representative or unrepresentative. "Yes" indicates the training data proportionally resembles the principal demographic characteristics of the patient populations served by the AI system.

69 Were a probabilistic algorithm to employ randomness in data input or in its processing of data, it would be classified as "probabilistic adaptive" with respect to the postmarket surveillance framework.

5. **LLM input complexity or semantic ambiguity.** This field has two possibilities: yes or no. "Yes" indicates that the AI system is or contains an LLM that can receive a verbal or textual prompt that is either complicated or semantically vague. A textual or verbal prompt to an AI system that is not an LLM, and where the prompt must match a predetermined value in order to initiate a function, would not have the potential for complexity or semantic ambiguity.
6. **Structural output unpredictability.** This field has two possibilities: yes or no. "Yes" indicates that the AI system has a programming architecture (e.g., generative AI or LLM) whose structure may produce inconsistent outputs for the same inputs.

AI devices whose attributes correspond to one or more cells colored orange or red within Table 2 are the ones where the justification for postmarket surveillance is most compelling.

2. Adapting Existing Safety Practices

The way in which AI technologies are deployed, as well as the local population health context, can have a significant bearing on AI performance. AI technologies, when deployed by a health system, are governed by protocols guiding their use. These protocols may extend beyond direct technology interaction with a patient and include staff training as well as oversight and audits. Likewise, the protocols may operate alongside multiple competing protocols pertaining to the physician and other medical devices and, thus, be integrated within a larger workflow. Having access to, and the prospect of modifying, protocols is a prerequisite for realizing opportunities for AI-facilitated health care spending reductions. This also recognizes that any negative outcomes related to AI technologies may be the result of the AI technology itself or they may have been affected by the way the AI was implemented.

Health care organizations have general mechanisms for safety reporting and patient outcome monitoring. Given the anticipated ubiquity of AI technologies (they might soon be a part of most technology systems deployed by health care providers), the most sensible and efficient approach is to adapt these existing systems to capture safety events, adverse patient outcomes, and other malfunctions related to the deployment of AI into the workflow. Such adaptation, in the context of reliance on existing systems, would treat those AI technologies that are not expected to be of higher risk on par with other potential causes of adverse patient and health system experiences.

3. Periodic Revalidation

Periodic revalidation (contemplated in the FDA draft guidance as periodic re-evaluation) is the simpler of the two modes of proposed postmarket surveillance and is envisioned for adaptive AI with an open dataset. Being probabilistic, adaptive AI models make determinations based on likelihoods, and in the case of open training datasets, these likelihoods change (ideally improving) over time through use of real-world data. These changes manifest in the market without formal FDA review as opposed to traditional software, where a programming update that modifies a medical device's effectiveness is typically obligated to file a new 510(k) submission[70] to the FDA. According to the agency:

> If a manufacturer modifies their device with the intent to significantly affect the safety or effectiveness of the device (for example, to significantly improve clinical outcomes, to mitigate a known risk, in response to adverse events, etc.), submission of a new 510(k) is likely required. A change intended to significantly affect the safety or effectiveness of the device is considered to be a change that "could significantly affect the safety or effectiveness of the device" and thus requires submission of a new 510(k) regardless of the considerations outlined below.[71]

After adaptive AI has been deployed, periodic revalidation would repeat the testing submitted to the FDA in connection with the premarket review process. Because the device manufacturer would have already supplied the test data and the acceptance criteria for outputs corresponding to the test data, this low-effort revalidation does not require additional data collection expense nor the consulting labor and informatics expertise to determine what the proper outputs should be for new test data. If, however, the health system supplements (or replaces) the test data with its own, then this would not be the case.

Periodic revalidation would be performed at scheduled intervals by the manufacturer working in collaboration with the provider (a health system, academic medical center, etc.). If possible (given workplace constraints as well as the nature of the AI device), the first test could be conducted a month after deployment, followed by progressively longer intervals — the third month, the sixth month, the twelfth month, and annually thereafter. This schedule would identify data drift problems early in the case of very unstable adaptive models while safely moving toward lower frequency surveillance for models that demonstrate ongoing accuracy with respect to the testing. As such, this surveillance activity can inexpensively reduce the incidence of adverse outcomes due to data drift and the liabilities that attend such events.

70 See FDA, "Premarket Notification 510(k)," August 22, 2024, https://www.fda.gov/medical-devices/premarket-submissions-selecting-and-preparing-correct-submission/premarket-notification-510k.

71 FDA, "Deciding When to Submit a 510(k) for a Software Change to an Existing Device," October 25, 2017, https://www.fda.gov/media/99785/download.

The manufacturer, **with the health system's approval**, would re-execute the testing on the deployed AI with the health system having full access to the results of the testing. The capture of data related to these revalidations is described in a later section, where it can be related to both modes of postmarket surveillance outlined in Table 2. If, for some reason, periodic revalidation requires additional health system data, then the periodic revalidation could employ software privacy measures (e.g. access permissions assigned at the user level) so that health systems do not see the AI code and the manufacturers do not see the patient data. This would preserve the confidentiality of patient data as well as the manufacturer's intellectual property and acceptance criteria.

Given the various conditions that can spawn irregular outputs, there is need for a second mode of postmarket surveillance tailored for unpredictable AI systems that would not be adequately safeguarded by periodic revalidations. This second mode, **performance monitoring**, distinguishes itself from periodic revalidations by continuous monitoring of outputs generated from real-world inputs (as opposed to test data). Unlike a CLIA-like certification process based on the results obtained from a single point in the AI device's history, performance monitoring focuses on unpredictability throughout an AI system's product life cycle (as encouraged by the FDA).[72] This life-cycle bias, along with the use of real-world data from at-risk AI systems, makes performance monitoring more practical to implement. Specifically, this performance monitoring approach avoids the need for:

- new test regimes for every type of AI device in health care,
- monitoring systems whose algorithms and datasets would **not** produce unpredictability, and
- external informatics specialists to consult on test data as well as results analysis.

Performance monitoring mitigates the risk for AI deployment delays due to a lack of availability of certification specialists that would emerge if all health care AI devices were subject to certification.

At a very basic level, performance monitoring would extend AI surveillance beyond the Sentinel Initiative and the Safe Medical Devices Act's existing requirement on manufacturers and device user facilities for reporting adverse events. As every possible valuable data point cannot be conceived of (let alone preemptively stipulated, given AI's numerous clinical settings), performance monitoring would, at a minimum, track trends for two subsets of anonymized clinical outcomes: false positives and false negatives. A false positive (for most

72 See Alex Youssef et al., "All Models Are Local: Time to Replace External Validation with Recurring Local Validation," arXiv, May 2023, https://arxiv.org/pdf/2305.03219.

AI systems) would be an incorrect positive diagnosis, prediction, or classification of a condition or disease. Given the issues surrounding AI output unpredictability and the diversity of AI applications, the definition of *false positive* should be expanded to also include errors in prediction, decision-making, and recommendations. For an LLM, however, a false positive would coincide with the previously discussed categories of hallucinations:

- Unintelligible language outputs
- Plausible, but factually inaccurate, claims
- Answers that are accurate but are misaligned with the intent of the end user's questions
- Citations of resources that do not exist

As evidenced above, there is not a direct correlate of false negatives for LLMs, while for many other types of AI the phrase would retain its canonical definition: an incorrect determination that a condition or disease is absent. Though, in the case of an LLM, the definition of *false negative* would still include inaccurate prescriptions or diagnoses.

4. Aggregated Outcome Data Registry

The full value of the proposed process will not be realized unless the outcomes collected at the local level (i.e., the hospital system deploying the AI) can be aggregated and fed back to health system users and device manufacturers. Moreover, creating a standardized data architecture that is common to all (or many) AI users, while desirable, would be labor-intensive. Instead, we propose to utilize AI agents that would sit on top of outcome data collected by local users, extract the relevant information in aggregated data form, and feed it into an aggregated outcome data registry (see Figure 4). As a fundamental first step, the agents would start with extracting and aggregating data from the existing safety reports. Then, they would be trained to extract and aggregate data from periodic revalidations and performance monitoring. Although relevant data could be manually transferred from a health system user to the registry, a more automated process (whether by Application Programming Interface or AI agent) is a preferable alternative.

> "The first review by providers and manufacturers can, in theory, eliminate the FDA reporting of negative trends driven by population health issues and deployment failings, thus avoiding alert fatigue on the part of the FDA for problems that are not directly attributable to a device deficiency."

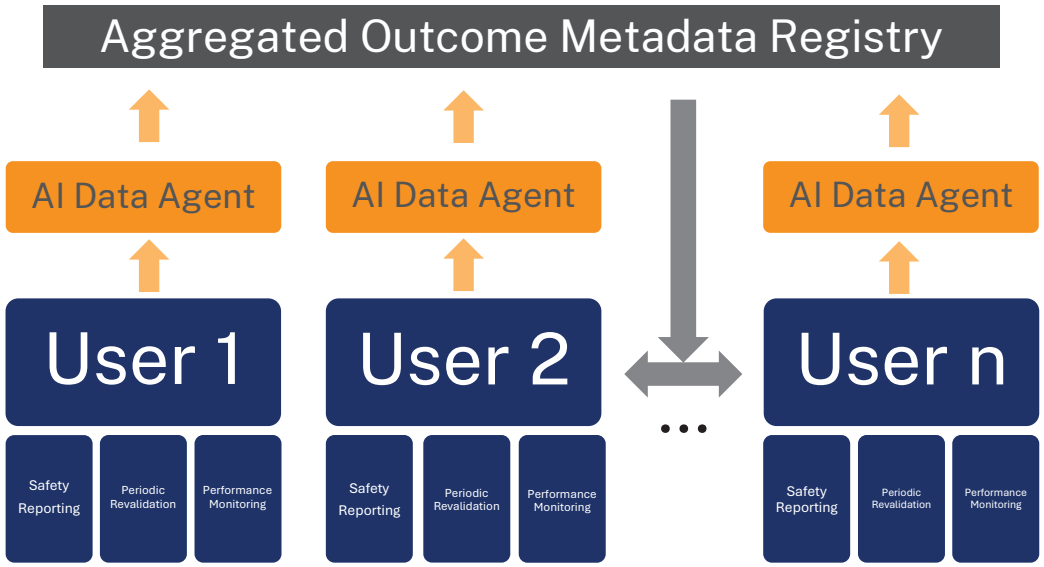
The aggregated data in the registry would alert the associated network of providers and the manufacturer of a given AI device to emerging negative trends (below the threshold of adverse events) that require investigation — given that they may be related either to the AI technology, the specifics of deployment, or the health characteristics of the local population. For those negative trends that are system-related, they can be reported to the FDA and evaluated internally at the provider. The first review by providers and manufacturers can, in theory, eliminate the FDA reporting of negative trends driven by population health issues and deployment failings, thus avoiding alert fatigue on the part of the FDA for problems that are not directly attributable to a device deficiency. Because the negative trend is drawn from a larger registry representing multiple health systems, the FDA can request additional comparative outcome information at other providers to assist in their review.

The FDA could provide high-level oversight and guidance for such repositories. Day-to-day management activities could be delegated to coalitions of AI adopters (i.e., health systems) partnered with manufacturers (i.e., industry developers of AI solutions) and platform or technology providers (EHR vendors, cloud providers, data management facilitators). This effort could leverage the experiences with the Sentinel Initiative designed for postmarket surveillance of existing FDA-regulated products. The participation of multiple providers within the same registry allows providers to compare the results of their revalidations.

This aggregated data sharing has multiple benefits. First, it allows the manufacturer to determine if poor test performance is either an outlier or a trend for the AI system. If poor performance is a trend, the manufacturer needs to take concrete remedial actions. If, on the other hand, the issue is an outlier, the provider and manufacturer can collaboratively determine if the performance is programming-related or specific to the characteristics of the local population being served by the provider. Second, the registry allows providers to compare the performance of their deployment against others without HIPAA violations of protected health information. There is the added benefit that participating health systems can inquire after clinical protocols that may have contributed to better performance among one or more of the other health systems in the group.

The registry framework could be further developed into a federated aggregated outcome data network for a given AI technology, which would operate on health systems' individual-level data (and not just aggregated data) and further automate periodic revalidation and performance monitoring. The development of such a network could be greatly enhanced through the assistance of EHR vendors. A lack of vendor participation would necessitate software development expenditures to address the needs of data extraction and reporting. In either scenario, possible sources of financial support in this effort are medical liability

Figure 4: Aggregated Outcome Data Registry



SOURCE: Author's original illustration.

insurance companies and possibly health plans.[73] Both groups not only have a financial interest in preventing patient injuries, but, in the case of malpractice insurance, the conditions of their coverage influence medical practice.[74] They also have a history of backing patient safety groups and other patient safety initiatives through financial contributions.[75]

SECTION IV: ADVANTAGES OF THE PROPOSED NEW FRAMEWORK VS PREVIOUS PROPOSALS

We believe that the voluntary, risk-based postmarket surveillance model proposed here addresses several limitations of the existing alternatives and has the potential to establish an efficient and functional ecosystem promoting innovation while safeguarding quality and prioritizing patient safety. It possesses several attractive features worth enumerating.

1. Enhanced Patient Safety and Expedited Capture of AI-Related Adverse Events

Given the newness of AI technologies, the national health care ecosystem has few safeguards related to identification, capture, and remediation of AI-related adverse events. This increases risk not only to patients and health care delivery organizations but also AI manufacturers,

73 The authors would like to thank Professor Charles M. Silver, from the University of Texas at Austin, for this insight.

74 Tom Baker and Charles Silver, "How Liability Insurers Protect Patients and Improve Safety," *DePaul Law Review* 68, no. 209 (2019), https://scholarship.law.upenn.edu/cgi/viewcontent.cgi?article=2996&context=faculty_scholarship.

75 Baker and Silver, "How Liability Insurers Protect Patients."

which have to rely on anecdotal evidence or their own efforts to learn about adverse events as well as any untoward experiences related to their products. This void can explain why many manufacturers have focused their attention on low-risk, operational AI tools rather than higher-risk, patient-facing technologies. Broad adoption of our framework could rapidly improve this situation.

Adapting existing safety reporting and patient documentation systems to detect negative outcomes associated with AI technologies provides an economical means to create a safer ecosystem that can attract wide support. The leveraging of existing systems also reduces the disruption the surveillance process represents for health care providers. Periodic revalidation or performance monitoring for higher-risk AI technologies additionally provides access to information on negative device trends earlier than would be the case with formal agency announcements, resulting in higher levels of assurance for patients and health systems. It also provides manufacturers with access to real-world postmarket data to quickly address emerging problems before more serious events occur. Given the nature of the postmarket surveillance, not only can device defects be captured but also issues that can indicate problems with deployment or the representativeness of training data.

2. Availability of Testing Expertise Tailored to AI Type and Medical Context

As discussed earlier, the FDA, any other national or local agency, or individual AI implementers (health systems, payers, etc.) lack sufficient expertise (technical, clinical, etc.) or scale (workforce size, training, etc.) to thoroughly monitor all health care AI technologies. Our proposed model, manufacturer participation provides technical knowledge, and health care system (as implementer) participation augments clinical subject matter proficiency, leading to a robust ecosystem that operates as a partnership.

Our framework creates a “golden middle” between the more implementer-focused CLIA lab-like proposals (which would likely lack sufficient expertise from AI developers and increase cost burdens on implementer organizations) and the more developer-focused assurance lab concepts (which might struggle with access to real-world data and deployment experiences of AI technologies used by health systems or other entities).

3. Scalability

Our proposed framework is also flexible, intending to direct resources toward health AI technologies that pose the greatest risk. It reduces postmarket surveillance obligations to reliance on already existing systems for technologies where the risks are minor and the benefits of monitoring are unlikely to outweigh the costs. The proposed collaboration between developers and implementers lessens the burden on already stretched FDA

resources and does not require more internal expertise than the agency can hire, addressing the concerns of limitations in size and scope.

With its three-layer surveillance scheme (existing safety practices, periodic revalidation, and postmarket monitoring), our new framework emphasizes efficiency while minimizing the operational disruption that a surveillance effort can bring. For many lower-risk AI technologies, postmarket surveillance would be based on adapting existing safety reporting systems to accommodate collecting information about adverse AI events and patient outcomes and other malfunctions. This approach avoids excessive labor and unnecessary data collection. Periodic revalidation encourages collaboration between manufacturers and implementers by using already created databases and structural frameworks, reducing the burden on implementers unlikely to possess sufficient expertise. The most stringent form of surveillance, which we call performance monitoring, builds efficiency by relying on an aggregated outcome data registry, thus reducing the burden on individual health system users or manufactures of health AI technologies.

Additional efficiencies are introduced by inviting EHR vendors to help develop the proposed aggregated outcome data registry. Given their proximity to the data that would power the network, they are ideally positioned to support this process and thus offer additional value to their health system customers and AI developing partners. Importantly, as many EHR vendors also develop or enable implementation of health AI technologies, to avoid conflicts of interest, it is critical that they facilitate rather than run the process. The same is true for cloud providers, which could also be significant contributors here by providing data environments and structures that would facilitate the proposed network.

4. Reducing Concerns Related to Conflicts of Interest, Intellectual Property, and Market Capture

Adapting existing safety practices to enable the first layer of safety surveillance eliminates concerns related to conflicts of interest, leakage of intellectual property, or market capture by dominant cloud or EHR vendors.

With periodic revalidation and performance monitoring there arises the need for exchange of data or information about the health AI technology that should be safeguarded. New tools, including confidential compute or “clean room” frameworks, can be employed alongside traditional cybersecurity measures to protect manufacturers from intellectual property disclosures and implementers from data losses, privacy and security concerns, or HIPAA violations.

Furthermore, as a voluntary, affordable, and independent process, with active industry participation and support from the FDA, our framework prevents one or more large AI companies from overtly influencing the design and execution of the postmarket surveillance process. Instead, the design would follow consensus standards developed by public-private partnerships with diverse representation. The framework avoids any hint of regulatory capture that creates rules with which only the most resourced AI companies comply. Instead, our framework encourages the participation of small startups by not making postmarket surveillance an excessively labor intensive or expensive process that favors large companies with considerable financial resources. We do acknowledge, however, that the process would benefit from — and in some ways depends on — a wide-scale adoption related to the quality and volume of monitoring data that can be aggregated to identify meaningful signals.

5. Financial Sustainability

The efficiency and scalability of the proposed framework should translate into cost savings for key stakeholders. Indeed, without an efficient framework in place, the burden of health AI technology evaluation is already increasing. For example, before adopting ambient voice scribes, numerous institutions performed more or less detailed assessments without coordinating or sharing their results. Although emerging, independent premarket-focused assurance laboratories might help reduce some of the burden, they face the risk of unproven business models that rely on funding from developers or implementers. On the other end of the spectrum, CLIA-lab-like entities require appropriately trained workforce or third-party organizations to make the system work, increasing the cost of AI ownership and shifting the burden of evidence generation to the implementers.

Our framework, focused on postmarket surveillance, might partially solve these problems by creating a funding ecosystem with all parties — including larger cloud and EHR vendors — contributing in kind. Our proposal would be further enhanced through participation of payers as well as health insurance and malpractice insurance providers, which stand to gain meaningfully from decreased patient injuries and knowledge of the safest health AI technologies. Initial funding from federal or state government agencies could be helpful to establish a proof of concept or seed initial development. At the same time, commercial entities, including venture capital firms investing in AI, might be interested in supporting this endeavor given additional downstream benefits related to real-world data networks that can be used only for postmarket surveillance but also — with the right business and privacy models — for development, validation, and enhancement of health AI solutions.

To achieve this vision, the right incentives need to be in place for all stakeholders to participate. For example, provisions can be created to reduce legal and financial liability for

vendors that opt to participate in the surveillance system. For example, the FDA could reclassify premarket risk assessment and the ensuing approval pathway based on a manufacturer commitment to postmarket surveillance. Similarly, liability for adverse outcomes may be lessened for health systems by virtue of the postmarket surveillance process insofar as it represents a risk mitigation effort for providers who have used AI in good faith without prior indication of adverse trends. Generally, access and potential to monetize real-world data created in the process would be attractive to both developers and implementers. Cloud and EHR vendors would see their participation as a further enablement of their customers, creating a new value stream to incentivize their customers to continue working with them. Furthermore, as mentioned above, payers and health insurance providers would benefit from better and more efficient patient care enabled by high-quality AI technologies.

6. Enhanced Local Governance with Common Standards

As a "golden middle" between CLIA lab-like approaches and national-assurance models, our framework would bring the best aspects of both: It would enhance local governance, while at the same time promote common standards. Operationalizing the proposed framework would require every user organization to keep an inventory of implemented health AI solutions and create a linkage mechanism to existing safety reporting systems. Furthermore, performance monitoring would be streamlined by an outcomes registry within each health system, enhancing local governance. At the same time, operationalization of the aggregated outcome data registry would require common standards adopted across user organizations, thus establishing an ecosystem where these standards are shared. This in turn allows aggregation of information on a national level and sharing with relevant stakeholders: other users, manufacturers, and regulators, increasing overall transparency. Additionally, there is the prospect of sharing insights with small hospitals and rural health care facilities that do not have the financial resources for AI assessments and optimization. This would help combat a digital divide from developing between the use of AI in large health systems and the use of AI in smaller health systems with limited financial resources.

SECTION V: CONCLUSIONS AND NEXT STEPS

The current regulatory paradigm, with its heavy emphasis on premarket assessment, *is necessary but not sufficient* for effective and efficient regulation of AI-enabled devices or answering the safety concerns raised at the beginning of this paper. Given the uncertainty associated with some AI technologies and their context-dependent performance, we propose an alternative framework that augments premarket evaluation with efficient, risk-based postmarket surveillance organized in a centralized manner that consolidates surveillance

among multiple health systems that employ the same AI medical device. We believe that, compared with existing alternatives, the proposed surveillance framework offers the most effective way to enhance the quality of developed tools and safeguard patient safety. We acknowledge that the national framework proposed here would not be implemented all at once. Smaller-scale, voluntary pilots can provide valuable information to tweak and enhance the concepts. These could occur as part of already-contemplated large-scale AI initiatives (e.g., project Stargate[76]), be connected to transformation efforts occurring at federal agencies (e.g., the Department of Veterans Affairs)[77] or state-level efforts where regulators wish to promote innovation while maintaining appropriate safety.

> "We believe that the proposed surveillance framework offers, compared with existing alternatives, the most effective way to enhance the quality of developed tools and safeguard patient safety" to "We believe that, compared with existing alternatives, the proposed surveillance framework offers the most effective way to enhance the quality of developed tools and safeguard patient safety."

To move our vision forward, we believe that the following next steps are necessary:

1. Secure the FDA's public support for the proposed postmarket surveillance framework
2. Identify health systems (or other health care organizations) with AI adoption and an interest in developing an efficient postmarket surveillance system
3. Identify AI manufacturers willing to join postmarket surveillance pilots
4. Identify technology, data management, and AI monitoring partners ready to work with AI adopters to syndicate the postmarket surveillance system
5. Define the technical, security, and data standards that will underpin the system
6. Develop financial models and incentive structures to sustain the effort, including funding for methods that improve AI unpredictability assessment
7. Conduct well-scoped pilots to optimize surveillance implementation and acquire practical experience and lessons

76 Joe Edwards, "Trump Backs $500B Stargate Project, Transforming Abilene into AI Epicenter," *Dallas Express*, May 29, 2025, https://dallasexpress.com/state/trump-backs-500b-stargate-project-transforming-abilene-into-ai-epicenter/.

77 See Coleman, "Could the VA Be the Key to Lowering the Cost of American Health Care?"

Health AI is moving quickly from concept to implementation. To maintain or accelerate the pace of innovation and at the same time safeguard patient benefits and safety, an agile and flexible regulatory framework is required. We argue that redefining the current process with a more balanced view of pre-and postmarket assessments can benefit both manufacturers and adopters of AI technologies, increasing the quality, reliability, and speed of delivered technologies for the benefit of all patients.