THE COLUMBIA
SCIENCE & TECHNOLOGY
LAW REVIEW

VOLUME 27 STLR.ORG NUMBER 2

ARTICLE

MONITORING, OVERSIGHT, AND LEARNING IN MEDICAL AI

W. Nicholson Price II*

When medical AI errs, it often goes unnoticed. If there's a specific patient injury, and the link to AI is obvious, that problem might be reported to the Food and Drug Administration (FDA), but not always. And many other types of problems, like worse performance on specific groups or ineffective integration into health system workflows, simply don't fall within the contours of regularized reporting. Even if they are noticed by the health system—far from a given—there's no obvious way to share that information more broadly. Against this backdrop, there are justified calls for better oversight and reporting. But there's the opportunity to do more. If now is the time to build more robust surveillance systems and standards for sharing that information, it should also be the time to build systems to share information about positive performance and learning, so that AI can help enable the vision of a learning health system that not only fixes mistakes but also constantly improves.

* Professor of Law, Michigan Law School. For helpful comments and conversations, I thank Ana Bračič, Doni Bloomfield, Yong Lim, Rebecca Wexler, Christopher Yoo, and participants in the Columbia University Science and Technology Law Review 2025 Symposium. For excellent research assistance, I thank Ryan DoyLoo and Eduard Toderescu. This work was supported by the National Research Foundation of Korea (2022R1A5A7083908) and by the Novo Nordisk Foundation (NNF23SA0087056). All errors are my own.

I. INTRODUCTION

AI in medicine has tremendous potential that is being increasingly realized, but it also has serious challenges. Some of those are typical to medical devices or products generally—it has to be built right, and tested right, and it has to work, and sometimes it doesn't. Most medical technologies we treat as stable and stably effective once they've been reviewed, approved, and implemented. Sure, they might be used wrong, but they are what they are, and they'll (roughly) stay that way. Not so with AI. How well AI performs, what it gets wrong and right, can vary substantially from place to place and over time to an extent that seems quite new, at least relative to how we treat other medical technologies.[1] And AI can change much more quickly than most other medical technologies.[2]

Unfortunately, existing mechanisms to oversee AI for safety problems are deeply inadequate. Initial review is not particularly searching, and oversight of errors that arise from the use of AI in real-world settings is quite minimal.[3] In recognition of this problem, many proposals have arisen for how to improve safety oversight, including continuous monitoring and broadening the kind of performance evaluation that takes place.[4] These proposals encourage sharing information, especially to FDA but also more broadly to allow this sort of oversight. This task will require enabling a new type of sharing, with standards required to ease the task of conveying information that is genuinely useful for oversight.

---

[1] *See infra* Part II.A.1. To be sure, at least some of this is because we treat *non*-AI technologies as more stable than they really are with respect to variation in place, time, and changing patient populations. But that's an argument for another day.

[2] *See infra* Part II.A.2.

[3] *See infra* Part II.B.

[4] *See infra* Part III.

But there's an opportunity in sharing more information that we shouldn't miss. Just oversight for safety and error leaves on the table the possibility of using AI to enable a rapidly-learning health system that improves based on the information gleaned in real world practice. Sharing positive information is necessary for this type of learning, and that task, too, will require standards for sharing information.

Before jumping in, I'll pause to forestall a quibble. I'll be drawing a distinction between safety monitoring, exemplified by the reduction in errors or harmful adverse events, and learning for improvement. These two things can also, of course, be viewed as different framings of the same thing. Decreasing wrong answers often means increasing right answers, in some basic sense.[5] And yet. There is a meaningful difference between trying to reduce errors—mistakes against a baseline—and trying to improve and move the baseline itself up. Not botching surgeries is different from trying to improve surgical techniques. Not missing diagnoses in an AI system is different from trying to enable that system to diagnose more things. There are many instances where this distinction fuzzes, but it's useful for describing and evaluating how we regulate and monitor and what we're trying to do to make medical AI better.

This piece proceeds in four Parts. Part II describes the sources of failures for medical AI once it's implemented in the real world—and how existing efforts at monitoring miss many problems. Part III describes a range of proposals to improve monitoring of medical AI systems, and how those proposals tend to focus on catching problems and fixing mistakes. Part IV presents an extended vision, imagining how data sharing and data collection about the performance of medical AI systems could be used not only to monitor for problems, but proactively to help realize the vision of a learning health system where AI enables forward-looking improvement. Part V addresses some complexities: funding, ownership, and international harmonization. A few brief thoughts conclude.

II. VARIABLE FAILURES FOR MEDICAL AI

Given the fact of imperfect AI in differently imperfect systems, errors are going to happen. Tumors will be missed, diagnoses will be botched, recommendations will be wrong, and patients will be hurt. This is the reality of medicine. Some of these errors will be negligent; others not.[6] In the context of AI, errors arise from

---

[5] True, it's often much more complicated. False positives, true positives, false negatives, and true negatives are mathematically related but different concepts, and there's a good reason that AI classifiers are typically evaluated by how well they manage tradeoffs between different success and error types, rather than a simple "how often is it right" rate. *See* Seong Ho Park & Kyunghwa Han, *Methodologic Guide for Evaluating Clinical Performance and Effect of Artificial Intelligence Technology for Medical Diagnosis and Prediction*, 286 RADIOLOGY 800, 802-03 (2018) (describing the receiver operator characteristic and the area under the curve as a way of including both sensitivity and specificity in a single performance metric). Nevertheless, there remains some truth to "less wrong = more right."

[6] *See* W. Nicholson Price II & I. Glenn Cohen, *Locating Liability for Medical AI*, 73 DEPAUL L. REV. 339, 343-48; *see also* Charlotte Tschider, *Medical Device Artificial Intelligence: The New Tort Frontier*, 46 BYU L. REV. 1551, 1573-81 (2020) (describing regulatory preemption of some

many sources, including poor design or mistakes in the initial process, transposition of systems across different health-care environments, and changes in the AI system itself that introduce new errors.

In a system that prioritizes effective care and minimizing patient harm (as well as other things, like being efficient and enabling access), errors need to be observed and monitored so that problems can be fixed and patient harm can be reduced going forward. Monitoring is essential; it is also quite difficult to implement effectively. This Part discusses the sources of AI errors, focusing on problems that arise post-approval, then turns to the challenges and deficiencies of existing monitoring efforts.

This Part, like the rest of the piece, focuses primarily on evaluation and oversight by the U.S. Food and Drug Administration (FDA), as the most prominent regulator of medical technology, including AI. FDA regulates "medical devices," a broad term[7] that includes many forms of medical AI.[8] The precise contours of which AI fit into the definition of medical devices are complex, contested, and changing.[9] However, FDA has authorized over 1,350 AI-enabled medical devices for marketing as of January 13, 2026.[10] To be sure, there are other sources of

---

forms of tort liability).

[7] A medical device is:

> an instrument, apparatus, implement, machine, contrivance, implant, in vitro reagent, or other similar or related article, including any component, part, or accessory, which is—
>
> (A) recognized in the official National Formulary, or the United States Pharmacopeia, or any supplement to them,
>
> (B) intended for use in the diagnosis of disease or other conditions, or in the cure, mitigation, treatment, or prevention of disease, in man or other animals, or
>
> (C) intended to affect the structure or any function of the body of man or other animals, and
>
> which does not achieve its primary intended purposes through chemical action within or on the body of man or other animals and which is not dependent upon being metabolized for the achievement of its primary intended purposes. The term "device" does not include software functions excluded pursuant to section 360j(o) of this title.

Federal Food, Drug, and Cosmetic Act, 21 U.S.C. § 321(h).

[8] *See* U.S. FOOD & DRUG ADMIN., ARTIFICIAL INTELLIGENCE IN SOFTWARE AS A MEDICAL DEVICE, https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-software-medical-device [https://perma.cc/34UH-WVME] (last visited Feb. 23, 2026).

[9] *See* W. Nicholson Price II, Rachel E. Sachs & Rebecca S. Eisenberg, *New Innovation Models in Medical AI,* 99 WASH U. L. REV. 1121, 1141-51 (2022) (discussing the scope of FDA regulation of medical AI); W. Nicholson Price II, *An Incidental Standard for Medical AI*, 8 PENN. J.L. INNOVATION __, Part I (forthcoming 2026) (describing the evolution of FDA regulation of AI clinical decision support software in a 2019 draft guidance and a significantly changed 2022 final guidance) [hereinafter Price, *Incidental Standard*]; *see also* U.S. FOOD & DRUG ADMIN., CLINICAL DECISION SUPPORT SOFTWARE: GUIDANCE FOR INDUSTRY AND FOOD AND DRUG ADMINISTRATION STAFF (2026), https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software [https://perma.cc/Q7TT-37WK] (last visited Mar. 11, 2026) (revising again the agency's thinking on which clinical decision support software will be regulated as a medical device).

[10] U.S. FOOD & DRUG ADMIN. ARTIFICIAL INTELLIGENCE-ENABLED MEDICAL DEVICES, https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices [https://perma.cc/2W4J-B2KM] (last visited Feb. 23, 2026).

governance, including by other federal agencies (such as the Centers for Medicare and Medicaid Services under the Clinical Laboratories Improvements, or the Federal Trade Commission for false advertising), nongovernmental entities (such as the Coalition for Health AI, or the Health AI Partnership), local health systems, or even individual providers.[11] I focus here on FDA because it controls market access for many medical AI systems and sets influential guidelines.[12]

A. AI Systems, Varying Contexts, and Drift Over Time

AI systems err for different reasons. Most simply, there are a host of potential errors in the initial development process including poor data, unrepresentative data, sloppy programming, problematic choices of outcomes or data labels, or mischaracterizing the problem being addressed.[13] These challenges can lead to various forms of poor performance or bias.[14] Ideally, initial review by FDA (for those medical AI systems that are medical devices and undergo FDA review[15]) addresses these challenges. It doesn't always—there are substantial problems with FDA review of medical AI devices, as with other medical devices, based on the depth of scrutiny and the available resources of expertise. These challenges have been the subject of substantial critique,[16] and so I won't canvas them in depth here. Instead, I'll turn to problems that typically show up once the AI system has been authorized for marketing by FDA. First, AI runs into problems when it is used in different health care environments with different patients, resources, and providers.

---

[11] *See* W. Nicholson Price II, *Clinicians in the Loop of Medical AI*, 74 EMORY L.J. 1265, 1272-78 (2021) [hereinafter Price, *Clinicians in the Loop*] (describing human clinicians as an imperfect source of governance in individual instances); *see also* W. Nicholson Price II, Mark P. Sendak, Suresh Balu & Karandeep Singh, *Enabling Collaborative Governance of Medical AI*, 5 NATURE MACH. INTEL. 821, 821 (2023) (describing collaborative governance between different entities).

[12] Price, *Incidental Standard*, *supra* note 9.

[13] *See* Richard J. Chen et al., *Algorithmic Fairness in Artificial Intelligence for Medicine and Healthcare*, 7 NATURE BIOMEDICAL ENG'G 719, 723-28 (2023) (discussing AI drawbacks resulting from data bias and incomplete data); *see also* Harriet Evans & David Snead, *Why Do Errors Arise in Artificial Intelligence Diagnostic Tools in Histopathology and How Can We Minimize Them?*, 84 HISTOPATHOLOGY 279, 280-83 (2023) (describing errors in AI diagnostic tools resulting from data labelling errors, lack of generalizability, incorrect programming by humans, etc.).

[14] *See, e.g.*, Sharona Hoffman & Andy Podgurski, *Artificial Intelligence and Discrimination in Health Care*, 19 YALE J. HEALTH POL'Y, L. & ETHICS 1, 17-22 (2019) (describing how algorithmic bias can function in unanticipated ways that lead to discrimination against particular groups); *see also* Ana Bracic, Shawneequa L. Callier & W. Nicholson Price II, *Exclusion Cycles: Reinforcing Disparities in Medicine*, 377 SCI. 1158, 1158-60 (2022) (demonstrating that biased and unrepresentative training data can cause AI systems to perform poorly and discriminatorily); *see also* Khiara M. Bridges, *Race in the Machine: Racial Disparities in Health and Medical AI*, 110 VA. L. REV. 243, 274-94 (2024) (describing bias in medical AI).

[15] *See supra* notes 4-7 and accompanying text.

[16] *See* Sara Gerke, *Health AI for Good Rather Than Evil? The Need for a New Regulatory Framework for AI-Based Medical Devices*, 20 YALE J. HEALTH POL'Y, L. & ETH. 433, 498-510 (2021); W. Nicholson Price II, *Medical AI and Contextual Bias*, 33 HARV. J.L. & TECH 65, 90-97 (2019) [hereinafter Price, *Contextual Bias*]; Vijaytha Muralidharan et al., *A scoping review of reporting gaps in FDA-approved AI medical devices*, 7 NPJ DIGIT. MED. 273, 273 (2024).

Second, AI systems can themselves change over time, and those changes can bring problems of their own.

1. Contextual Changes, Holding AI Constant

For any given individual AI system, even one that doesn't change at all from the way it was initially developed and brought onto the market, performance can vary substantially across place, patient population, and time,[17] a problem that has been referred to as "contextual bias"[18] or "dataset shift."[19]

When an AI system is brought into practice and implemented in different environments, that system will typically perform differently for multiple reasons. First, patients. AI systems are trained on datasets that are typically underrepresentative of the population generally; it's easier to collect data for AI training in, say, high-resource academic medical centers than in small rural health centers, some groups are more likely to seek care in situations where their data are collected, and some groups are less likely to consent to data collection when given a choice.[20] And AI systems typically perform worse on populations different from those on which they were trained,[21] whether that's about skin color and dermatology apps[22] or patient deterioration predictors in different hospitals.[23] There is some evidence that this problem has been improving in time, but it's equivocal[24]—and remarkably enough, it's tough to even measure that change, because it remains the case that most AI systems cleared by FDA report distressingly little information about the data on which they've been trained. According to a study of all 903 devices authorized by FDA in the U.S. as of August 2024, fewer than one in three reported sex-specific performance, under a quarter

---

[17] *See* Price, *Contextual Bias, supra* note 16, at 90-97.

[18] *Id.* at 98-100.

[19] *See* Samuel G. Finlayson et al., *The Clinician and Dataset Shift in Artificial Intelligence*, 385 NEW ENGL. J. MED. 283, 283 (2021) ("Dataset shift occurs when a machine-learning system underperforms because of a mismatch between the data set with which it was developed and the data on which it is deployed.")

[20] *See* Kayte Spector-Bagdady et al., *Respecting Autonomy and Enabling Diversity: The Effect of Eligibility and Enrollment on Research Data Demographics*, 40 HEALTH AFF. 1892, 1893-95 (2021).

[21] *See* Fereshteh Hasanzadeh, Colin B. Josephson, Gabriella Waters, Demilade Adedinsewo, Zahra Azizi & James A. White, *Bias Recognition and Mitigation Strategies in Artificial Intelligence Healthcare Applications*, 8 NPJ DIGIT. MED. 154, 154 (2025).

[22] *See* Roxana Daneshjou et al., *Disparities in Dermatology AI Performance on a Diverse, Curated Clinical Image Set*, SCI. ADVANCES, Aug. 12, 2022, at 1.

[23] *See* Patrick Rockenschaub et al., *External Validation of AI-Based Scoring Systems in the ICU: A Systematic Review and Meta-Analysis*, BCM MED. INFORMATICS AND DECISION MAKING, Jan. 2025, at 1.

[24] *E.g.*, Shivam Sharma et al., *Genetic Ancestry and Population Structure in the All of Us Research Program Cohort*, 16 NATURE COMM., May 2025, at 5 (describing the All of Us research study's efforts to increase research data from a genetically diverse population).

age-specific performance, and vanishingly few on racial- or ethnic-group-specific performance.[25]

Second, workplace. Different clinical care environments have substantially different workflows and resources, which change how AI performs *in situ*. An academic medical center may have substantial resources to follow up on AI recommendations, which resources may simply be absent in the context of a lower-resource, small care setting.[26] Patterns of health care practice vary substantially, such that recommendations that work well in one place simply don't in another. Workflow integration also differs[27]—what may be a useful alert in one place may simply be an obnoxious alarm leading to alert fatigue in another.[28] Notably, while most reports of problems with AI systems made to FDA's reporting system were of technical issues, those issues which were about human use or workflow problems were four times as likely to result in harm to patients.[29] A substantial fraction of FDA-authorized AI device recalls occur within a year of clearance, and frequently those recalls are prompted by malfunctioning software, problems with integration into the clinical workflow, and real-world clinical performance worse than that demonstrated in development.[30]

Over time, the performance of a static AI system will change too—typically and unfortunately, performing worse on a technical level as time passes.[31] To take a well-known example, when the coronavirus pandemic struck in 2020, the performance of AI systems predicting sepsis in hospital patients grew dramatically worse, because patient populations looked very different than they had months before.[32] But the changing nature of care in time can come in many forms, as patient populations change demographically, related data acquisition systems change in a health setting, treatment patterns or the standard of care change,

---

[25] Daniel Windecker et al., *Generalizability of FDA-Approved AI-Enabled Medical Devices for Clinical Use*, 8 JAMA NETW. OPEN, Apr. 2025, at 1; *see also* Vijaytha Muralidharan et al., *A Scoping Review of Reporting Gaps in FDA-approved AI Medical Devices*, 7 NPJ DIGIT. MED. 273, 273 (2024) (finding frequent omission of data about demographic subgroup performance in 692 AI devices authorized by FDA between 1995 and 2023).

[26] *See* Price, *Contextual Bias*, *supra* note 16, at 91-94.

[27] *See* Mark Sendak et al., *Editorial: Surfacing Best Practices for AI Software Development and Integration in Healthcare*, FRONTIERS IN DIGIT. HEALTH, Feb. 2023, at 1.

[28] *See* Amanda L. Joseph et al., *Alert Fatigue and Errors Caused by Technology: A Scoping Review and Introduction to the Flow of Cognitive Processing Model*, 13 KNOWLEDGE MGMT. & E-LEARNING 500, 500-17 (2001) (describing the phenomenon of alert fatigue with respect to health technology)

[29] David Lyell, Ying Wang, Enrico Coiera & Farah Magrabi, *More than Algorithms: An Analysis of Safety Events Involving ML-Enabled Medical Devices Reported to the FDA*, 30 J. AM. MED. INFORM. ASS'N 1227, 1227 (2023).

[30] *See id.;* Branden Lee et al., *Early Recalls and Clinical Validation Gaps in Artificial Intelligence–Enabled Medical Devices*, 6 JAMA HEALTH F., no. 8: e253172, 2025, at 1.

[31] Daniel Vela et al., *Temporal Quality Degradation in AI Models*, 12 SCI. REP. 11654 (2022).

[32] Finlayson et al., *supra* note 19, at 283.

clinician or patient behavior changes (perhaps in response to the AI system), or any other number of possibilities.[33]

All these performance differences need to be monitored, overseen, and understood. AI is simply not a set-it-and-forget-it technology. And even that view imagines that the AI itself is a constant in a shifting environment. But it's not; the AI systems change too.

2. Changing AI

A second major form of variation is when the AI system itself changes. AI systems are software, and they're therefore much easier to update than, say, a physical device or a drug. Accordingly, developers—like with other software—can implement changes to AI systems to account for new data or new needs. But also like with other software, AI system changes can be both a solution to and a source of problems. New updates can break existing workflow or introduce new errors, requiring their own oversight and monitoring.

Unlike most software, updates for those medical AI systems that are regulated as medical devices need to proceed through FDA scrutiny. Through a regulatory lens, changes to a marketed AI device can occur in two principal ways. First, and more traditionally, the developers of an AI system can submit proposed changes to the system in a new regulatory filing to FDA, typically under the 510(k) pathway. Under this pathway, the new product—here, the updated AI system—must be substantially equivalent to an already marketed product—here, the existing version of the system. Substantial equivalence operates sequentially and is therefore not transitive; the facts that product B is equivalent to earlier product A and that product C is equivalent to earlier product B do not together imply that product C is equivalent to product A, a phenomenon known as "predicate creep."[34]

Second, and still in development, FDA is implementing what it describes as "Predetermined Change Control Plans" (PCCPs), which create a pathway for changes to an AI system that are not individually reviewed by the agency but would otherwise require a new marketing submission.[35] In a PCCP, a developer submits to FDA as part of its initial marketing submission a plan for proposed

---

[33] *Id.* at 284-85 (providing a helpful table of possible changes with examples, including potential recognition and mitigation strategies).

[34] Urs J. Muehlematter et al., *FDA-Cleared Artificial Intelligence and Machine Learning-Based Medical Devices and Their 510(k) Predicate Networks*, 5 LANCET DIGIT. HEALTH e618, e618 (2023). Predicate creep can be very substantial for devices cleared on the basis of predicates that were quite different; for instance, an AI system to identify breast tissue abnormalities based on MRI data served as a predicate for an AI system to identify brain tissue abnormalities based on CT data. *Id*. at e623. For our purposes here, I am considering only changes to an individual AI system produced by a single developer.

[35] U.S. FOOD & DRUG ADMIN., MARKETING SUBMISSION RECOMMENDATIONS FOR A PREDETERMINED CHANGE CONTROL PLAN FOR ARTIFICIAL INTELLIGENCE-ENABLED DEVICE SOFTWARE FUNCTIONS: GUIDANCE FOR INDUSTRY AND FOOD AND DRUG ADMINISTRATION STAFF (2025), https://www.fda.gov/media/166704/download [https://perma.cc/RL7K-WAKJ] [*hereinafter* FDA, AI PCCP GUIDANCE].

modifications, including what those modifications will be, how the modifications will occur, and how the impact of those modifications will be evaluated.[36] If a later modification to the product falls within those guardrails, the manufacturer can implement it without submitting a new marketing document to FDA.[37] Modifications can't change the device's intended use or make it non-equivalent to a predicate device used for initial authorization.[38] For example, in an AI system to characterize skin lesions to aid in diagnosis when used by a clinician, a PCCP could allow the system to be used with additional smartphone cameras meeting the PCCP's specifications, after analytical validation—but the manufacturer could not market a version to be used directly by patients without a new marketing submission, since that was not part of the initial intended use.[39] FDA has issued a set of guiding principles for PCCPs in the AI context, recommending that they be focused and bounded, risk-based, evidence-based, transparent, and created using a "total product lifecycle" (TPLC) approach.[40]

FDA has not yet authorized any adaptive AI systems—that is, a system that can change its own parameters automatically.[41] FDA acknowledges that its regulatory system "was not designed for adaptive artificial intelligence and machine learning technologies."[42] That issue is being rethought over time; PCCPs are a path forward that can theoretically include adaptive updates (though not yet), for instance.[43] And there have been calls for more adaptive regulatory oversight that can allow truly continuous software updates.[44]

Questions of AI systems changing become more complex and harder in the context of LLMs. In one sense, LLMs are more fixed than most software, because they are trained on immense corpora of data, and that training is incredibly

---

[36] *See* U.S. FOOD & DRUG ADMIN., PREDETERMINED CHANGE CONTROL PLANS FOR MEDICAL DEVICES: DRAFT GUIDANCE FOR INDUSTRY AND FDA STAFF 9 (2024), https://www.fda.gov/media/180978/download [https://perma.cc/6ZXW-MDRD].

[37] *Id.* at 14.

[38] *Id.* at 19.

[39] FDA, AI PCCP GUIDANCE, *supra* note 35, at 40-41.

[40] U.S. FOOD & DRUG ADMIN., PREDETERMINED CHANGE CONTROL PLANS FOR MACHINE LEARNING-ENABLED MEDICAL DEVICES: GUIDING PRINCIPLES (2025), https://www.fda.gov/medical-devices/software-medical-device-samd/predetermined-change-control-plans-machine-learning-enabled-medical-devices-guiding-principles [https://perma.cc/ZM58-2J8C].

[41] U.S. FOOD & DRUG ADMIN., ARTIFICIAL INTELLIGENCE IN SOFTWARE AS A MEDICAL DEVICE (Mar. 3, 2025), https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-software-medical-device [https://perma.cc/WMN4-RBD2] [*hereinafter* FDA, AI SAMD GUIDANCE] ("Many changes to artificial intelligence and machine learning-driven devices may need a premarket review."); *See* Boris Babic et al., *Algorithms on Regulatory Lockdown in Medicine*, 366 SCIENCE 1202, 1202 (2019) (describing FDA's disinclination to authorize adaptive AI systems and noting that this approach limits the potential of medical AI).

[42] FDA, AI SAMD GUIDANCE, *supra* note 41.

[43] *See* Snigdha Santra et al., *Navigating Regulatory and Policy Challenges for AI Enabled Combination Devices*, 6 FRONT. MED. TECHNOL., no. 1473350, 2024, at 6.

[44] *Id.*

expensive and time-consuming. Training a new flagship model takes an estimated tens to hundreds of millions of dollars and many months.[45] But in another sense, LLMs can change exceedingly rapidly, including in the course of a single interaction with a user. The output of an LLM can change based on the information it's given in a prompt—not just the wording of a particular query, but also information embedded in the prompt or in prior prompts. A user could upload multiple labeled images and state that those images were the gold standard for image labeling, for instance, and the LLM would take that information into account in providing future answers. In addition, retrieval-augmented generation (commonly known as RAG) is a technique whereby documents can be quickly provided for an LLM to reference and use in its answers.[46] Taken together, these technologies can let an LLM change substantially quite quickly. FDA has acknowledged this difficulty and is still grappling with how to regulate LLMs as medical devices.[47]

B. Inadequate Monitoring Under Current Systems

The current systems for monitoring AI performance have substantial gaps.[48] The most important monitoring is federal: FDA maintains a centralized system for reporting adverse events associated with medical devices, the Manufacturer and User Facility Device Experience (MAUDE). Under the Medical Device Reporting Regulation,[49] manufacturers, importers, and user facilities (like hospitals or nursing homes)[50] must report specific adverse events involving medical devices, including death, serious injury, or malfunction (if the malfunction would be likely to cause

---

[45] Dorothy Neufeld, *Charted: The Surging Cost of Training AI Models*, VISUAL CAPITALIST (Apr. 24, 2025), https://www.visualcapitalist.com/the-surging-cost-of-training-ai-models/ [https://perma.cc/4BZ3-HXDW].

[46] *See* Michael Klesel & H. Felix Wittmann, *Retrieval-Augmented Generation* (RAG), 67 BUS. & INFO. SYS. ENG'G 551, 551-52 (2025).

[47] *See* FDA Digital Health Advisory Committee Meeting (Nov. 20-21, 2024), https://www.fda.gov/advisory-committees/advisory-committee-calendar/november-20-21-2024-digital-health-advisory-committee-meeting-announcement-11202024 [https://perma.cc/9EZP-HCH2]. *See also* Daria Onitiu et al., *Walking Backward to Ensure Risk Management of Large Language Models in Medicine*, 53 J.L. MED. ETH. 454, 454-61 (2025) (arguing that a markedly different approach is needed to regulate LLMs in the EU); *but see* Hannah Louise Smith & W. Nicholson Price II, *Do Specialized Medical LLMs Demand a Radically New Approach Under the EU's Medical Device Regulation?*, 53 J.L. MED. ETH. 465, 465-66 (2025) (noting LLM similarities to existing medical device issues and suggesting a different approach). Other regulatory issues arise as well, such as the fact that many LLMs are general-purpose tools with substantial medical-device-like uses and outputs. *See*, *e.g.,* Gary E. Weissman et al., *Unregulated Large Language Models Produce Medical Device-Like Output*, 8 NPJ DIGIT. MED. Art., no. 148, 2025, at 1.

[48] *See* Boris Babic et al., *A General Framework for Governing Marketed AI/ML Medical Devices*, 8 NPJ DIGIT. MED. 328, 328 (2025).

[49] 21 C.F.R. § 803 (2025).

[50] A device user facility includes "a hospital, ambulatory surgical facility, nursing home, outpatient diagnostic facility, or outpatient treatment . . . which is not a physician's office." 21 C.F.R. § 803.3(d) (2025).

serious injury death if it were to recur).[51] Neither clinicians nor patients are under any direct legal obligations to report problems, though FDA does encourage voluntary reporting by those individuals.[52] Almost 97% of reports are submitted by manufacturers.[53] Reports of adverse events are collected in MAUDE.

MAUDE is the most comprehensive source for evaluating the performance of medical AI (and all other medical devices)—but note immediately what it focuses on: adverse events, or recognized individual failures of a system that cause some sort of problem or injury, rather than broader metrics about performance as a whole.[54] That is, even if MAUDE were working perfectly, it would include only a fraction of the information needed for effective monitoring of how well AI systems are actually performing in the real world.

But MAUDE is not working perfectly. There are substantial gaps in monitoring data. In a 2025 study of all AI/ML-device-associated adverse events reported to MAUDE between 2010 and 2023, Boris Babic and colleagues found that reporting was deeply inadequate. Ninety-eight percent of reports were from just five devices.[55] In describing the error, the majority of reports just noted, "malfunction."[56] And there was a large amount of missing data.[57] Jessica L. Handley and colleagues similarly found that MAUDE reporting lacked detailed data that were specific to algorithms.[58] Among reports on MAUDE generally, almost a third are submitted late, with most of the late submissions over half a year behind schedule[59]—a substantial lag, especially in the fast-moving field of medical AI, and given the reality that many AI device recalls nonetheless occur within the first year of marketing.[60]

---

[51] 21 C.F.R. § 803.10(a)-(c) (2025). The reporting requirements are slightly different between the three entities and for the three categories.

[52] *See* U.S. FOOD & DRUG ADMIN., MEDICAL DEVICE REPORTING (MDR): HOW TO REPORT MEDICAL DEVICE PROBLEMS (2025), https://www.fda.gov/medical-devices/medical-device-safety/medical-device-reporting-mdr-how-report-medical-device-problems [https://perma.cc/9VAP-66R6].

[53] Meital Mishali et al., *Evaluation of Reporting Trends in the MAUDE Database: 1991 to 2022*, 11 DIGIT. HEALTH 1, 2 (2025).

[54] *Id.* at 1.

[55] Boris Babic et al., *A General Framework for Governing Marketed AI/ML Medical Devices*, 8 NPJ DIGIT. MED. No. 328, 2005, at 3.

[56] *Id.*

[57] *Id.* (highlighting a significant concern with the extent of missing data in MAUDE database fields, namely Event Location, Health Professional Reporter, Event Date, and Report Occupation fields).

[58] Jessica L. Handley et al., *Artificial Intelligence Related Safety Issues Associated with FDA Medical Device Reports*, 7 NPJ DIGIT. MED. no. 351, 2024, at 1.

[59] Alexander O. Everhart et al., *Late Adverse Event Reporting from Medical Device Manufacturers to the US Food and Drug Administration: Cross Sectional Study*, 388 BMJ no. e081518, 2025, at 1.

[60] Lee et al., *supra* note 30, at 1 (finding 43.4% of all AI device recalls occurred within the first 12 months of device clearance).

Monitoring takes resources, especially to do more systematically rather than just reporting particular adverse events. When some form of monitoring is required, as by FDA in the context of MAUDE, an affirmative obligation exists, and manufacturers, importers, and device user facilities must report.[61] But hospitals and health systems don't receive any particular benefit in sharing that information, and such reports are voluntary—and frequently absent.[62] With respect to broader monitoring of performance, both within a health system and ideally with information sharing more broadly, there is real misalignment of incentives and resources. As I have argued in previous work with Mark Sendak and Suresh Balu, insurance reimbursement is rarely available for developing or using the infrastructure for continuous performance monitoring.[63] Accordingly, health systems with lower resources may not have the wherewithal to do such monitoring, and health systems with the resources may simply lack the incentives.[64]

III. BETTER MONITORING FOR OVERSIGHT

There is growing recognition of problems with oversight of medical AI by FDA, both in the context of initial approval and ongoing oversight. Accordingly, calls for improvement have been suggested by academics, nongovernmental organizations, and government agencies alike (including FDA itself). These calls typically focus on safety monitoring and ensuring that performance does not get worse over time, though some view the possibility of reporting more broadly.

A. Academic Proposals

A substantial and growing literature on the regulation and governance of medical AI has recognized these problems with changing AI performance in different contexts and over time and has generally responded with calls for substantial improvement in postmarket surveillance.

Some scholars note that better, more standardized reporting of characteristics at the time of clearance of approval enables better approaches to oversight of variable performance.[65] For instance, knowing the demographic breakdown of training data makes it easier to tell when a system is being used in a patient population that differs substantially from that initial dataset.

---

[61] *See* Everhart et al, *supra* note 59, at 1; 21 C.F.R. §803.10.

[62] *See* Meital Mishali et al., *Evaluation of Reporting Trends in the MAUDE Database: 1991 to 2022*, 11 DIG. HEALTH 39850626, 7 (2025) (stating that "[u]ser facilities are not required to report malfunctions to the FDA," and that "voluntary users . . . contribute only 0.7%" of adverse event reports).

[63] Mark P. Sendak, W. Nicholson Price II & Suresh Balu, *A market failure is preventing efficient diffusion of health care AI software*, STAT (2022), https://www.statnews.com/2022/05/24/market-failure-preventing-efficient-diffusion-health-care-ai-software/ [https://perma.cc/SM55-FQJH]

[64] *Id.*

[65] *See* Muralidharan et al., *supra* note 16, at 7; Windecker et al., *supra* note 25, at 1.

Others call specifically for more robust reporting of performance in the real world.[66] Vijaytha Muralidharan and colleagues suggest formal harmonized standards for performance reporting, including a breakdown by demographics and explicit documentation of any postmarket findings.[67] Rawan Abdulibdeh and colleagues call for mandatory postmarket clinical validation, reduction in bias over time, transparency across the lifecycle of AI devices, and greater stakeholder involvement in the review process generally.[68] I have called in prior work for greater sharing of performance data specifically to enable collaborative governance not only by regulators but also by other actors like academics or professional organizations.[69] Branden Lee and coauthors suggest that more responsive surveillance is needed to detect errors earlier and decrease patient harm.[70]

Still other authors look to the role of surveillance within larger systems. Sara Gerke and colleagues suggest that AI systems should be treated and governed as whole systems, not merely devices, and that surveillance is needed to oversee and monitor that whole system.[71] The Algorithm-Based Clinical Decision Support (ABCDS) Oversight framework at Duke Medicine similarly recommends and includes ongoing monitoring as part of the lifecycle deployment system.[72]

B. NGO and Government Proposals

In parallel with academic proposals various governmental bodies and nongovernmental organizations have issued guidelines for the development and oversight of AI that includes calls for more robust postmarket monitoring and surveillance.

The World Health Organization (WHO), in its *Regulatory Considerations on Artificial Intelligence for Health*,[73] emphasizes the need for regulatory oversight across the entire lifecycle of medical AI. It calls for risk assessment in post-market

---

[66] *E.g.,* Vidhi Singh et al., *United States Food and Drug Administration Regulation of Clinical Software in the Era of Artificial Intelligence and Machine Learning*, 3 MAYO CLIN. PROC. DIGIT. HEALTH 100231, 5 (2025); Handley et al., *supra* note 58, at 1.

[67] *See* Muralidharan et al., *supra* note 16, at 273.

[68] Rawan Abulibdeh et al., *The Illusion of Safety: A Report to the FDA on AI Healthcare Product Approvals*, 4 PLOS DIGIT. HEALTH no. 6: e0000866, 1 (2025).

[69] Price et al., *Enabling Collaborative Governance*, *supra* note 11, at 821.

[70] Lee et al., *supra* note 30, at 3.

[71] Sara Gerke et al., *The Need for a System View to Regulate Artificial Intelligence/Machine Learning-Based Software as Medical Device*, 3 NPJ DIGIT. MED. Art. No. 53, 1 (2020); Lyell et al., *supra* note 29, at 1.

[72] Armando D. Bedoya et al., *A Framework for the Oversight and Local Deployment of Safe and High-Quality Prediction Models*, 29 J. AM. MED. INFORM. ASS'N 1631, 1631 (2022); *Algorithm-Based Clinical Decision Support (ABCDS) Oversight*, DUKE HEALTH, https://healthaigovernance.duke.edu/abcds-oversight [https://perma.cc/DJ8B-U4HX] (last visited Feb. 16, 2026).

[73] WORLD HEALTH ORG., REGULATORY CONSIDERATIONS ON ARTIFICIAL INTELLIGENCE FOR HEALTH 9 (2023), https://www.who.int/publications/i/item/9789240078871 [https://perma.cc/D5S4-GLSX], 9.

surveillance as well as in earlier stages of development.[74] WHO notes that adverse event reporting is largely inadequate, and instead advocates for more proactive monitoring: "Regulators must be notified of reportable incidents (adverse events), and findings from more continuous monitoring using real-world data may help developers and regulators better understand and assure the safety and performance of these devices in real-world use."[75] WHO acknowledges this won't be cheap: "For prospective monitoring of real-world data, significant investment will be required in prospectively curating and labelling validation data."[76]

FDA similarly recognizes the need for more robust postmarket monitoring as well and is actively working on the issue. In autumn 2025, FDA's Digital Health Center of Excellence issued a request for public comment on "Measuring and Evaluating AI-Enabled Medical Device Performance in the Real World," seeking input from stakeholders on how best to measure real-world performance, including "strategies for identifying and managing performance drift, such as detecting changes in input and output."[77] The agency seeks input on performance changes based on "clinical usage patterns and user interactions" over time,[78] performance metrics,[79] the combination of human and automated monitoring,[80] the data used for postmarket evaluation of ongoing performance,[81] and how to incorporate "clinical outcomes and user feedback into model updates."[82] Notably, while FDA's standards only apply directly to medical devices that fall within its regulatory ambit—a contestable set that is notably less than all of the AI systems being used in health care today—the standards FDA sets can nevertheless influence practices even for devices it doesn't directly oversee.[83]

The National Institute of Standards and Technology has issued its *Artificial Intelligence Risk Management Framework*, which also addresses the entire lifecycle of AI, though it focuses on AI generally rather than specifically in the health context.[84]

---

[74] *Id.* at 24-25.

[75] *Id.* at 25.

[76] *Id.*

[77] *Request for Public Comment: Measuring and Evaluating Artificial Intelligence-enabled Medical Device Performance in the Real-World,* U.S. FOOD & DRUG ADMIN, (Sept. 30, 2025), https://www.fda.gov/medical-devices/digital-health-center-excellence/request-public-comment-measuring-and-evaluating-artificial-intelligence-enabled-medical-device [https://perma.cc/XV4R-4XF5].

[78] *Id.* at 3.

[79] *Id.* at 2.

[80] *Id.*

[81] *Id.* at 3.

[82] *Id.*

[83] *See* W. Nicholson Price II, *An Incidental Standard for Medical AI*, 8 J.L. INNOV. (forthcoming 2026); Ariel Dora Stern & W. Nicholson Price II, *Regulatory Oversight, Causal Inference, and Safe and Effective Health Care Machine Learning*, 21 BIOSTATISTICS 363, 365 (2019).

[84] NATL. INST. STANDARDS & TECH., ARTIFICIAL INTELLIGENCE RISK MANAGEMENT

And there are others. The National Academy of Medicine, in its *AI Code of Conduct for Health and Medicine*, calls for continuous assessment of medical AI systems.[85] The Coalition for Health AI, in its *Blueprint for Trustworthy AI*, calls for continual monitoring.[86] It recommends implementing this in part through a nationwide network of assurance laboratories that can not only initially test AI models, but can also monitor them and validate them across the lifecycle and in different contexts.[87] And the Health AI Partnership recommends monitoring AI performance (in addition to the work environment[88]) to make sure the system remains effective and can potentially be improved; it has a step-by-step guide including "routine checks on data quality, algorithm performance, user satisfaction, and any potential biases or ethical concerns," identifying "relevant performance metrics to track and assess the tool's effectiveness . . . over time," assigning a multidisciplinary team, taking into account regulatory best practices, and sharing the outcome of monitoring to end users and developers (potentially in an automated process).[89]

IV. MONITORING FOR LEARNING

Monitoring for oversight and safety is crucial, but it shouldn't be where policymakers stop. AI has the potential (for good and ill) to adapt and change much more readily and much more quickly than many other medical technologies.[90] Above, that change was described as a challenge, and it surely is; it means that monitoring is necessary to make sure that changes work and don't raise new problems. It's also an opportunity—the ability for AI to change flexibly and

---

FRAMEWORK AI RMF (AI RMF1.0), NIST AI 100-1 (2023), https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf [https://perma.cc/ZY77-PXSV].

[85] NATIONAL ACADEMY OF MEDICINE, AN ARTIFICIAL INTELLIGENCE CODE OF CONDUCT FOR HEALTH AND MEDICINE: ESSENTIAL GUIDANCE FOR ALIGNED ACTION: ESSENTIAL GUIDANCE FOR ALIGNED ACTION (2025) [hereinafter NATIONAL ACADEMY OF MEDICINE].

[86] COALITION FOR HEALTH AI, BLUEPRINT FOR TRUSTWORTHY AI IMPLEMENTATION GUIDANCE AND ASSURANCE FOR HEALTHCARE 11 (2023).

[87] Nigam H. Shah, John D. Halamka & Suchi Saria, *A Nationwide Network of Health AI Assurance Laboratories*, 331 JAMA 245, 246 (2024).

[88] Jenna Burrell, *Guide: Monitor Work Environment*, HEALTH AI PARTNERSHIP (Mar. 7, 2026, 11:29 AM), https://healthaipartnership.org/guiding-question/monitor-work-environment [https://perma.cc/K8T7-6BD2].

[89] David Vidal & Mark Lifson, *Guide: Monitor AI performance*, HEALTH AI PARTNERSHIP (Mar. 7, 2026, 11:32 AM), https://healthaipartnership.org/guiding-question/monitor-ai-performance [https://perma.cc/A43S-REE5].

[90] It can certainly change much more rapidly than drugs can, where each new formulation requires additional clinical trials and a substantial regulatory review process. Physical devices can be modified more quickly, whether through the 510(k) clearance pathway or through the PMA supplement pathway. *See* George Horvath, *Medical Device Dangers: Choosing Ignorance in the Courts and at the FDA*, 67 WM. & MARY L. REV. 669, 684-88 (forthcoming). But AI can change much more rapidly still, as evidenced by FDA's Predetermined Change Control Plan, acknowledging the possibility of AI changes that do not require regulatory pre-evaluation, so long as they fit within predetermined parameters. *See supra* notes 35-40 and accompanying text.

adaptively means that AI is nearly ideal for incorporation into the idea of a learning health system, and that such learning can happen both within and across systems.[91]

Let me back up for a moment and explain learning health systems. The idea has been around for a few decades, and was expressed perhaps most clearly in a landmark report by the Institute of Medicine in 2007, which characterized a learning health system as: "one in which knowledge generation is so embedded into the core of the practice of medicine that it is a natural outgrowth and product of the healthcare delivery process and leads to continual improvement in care."[92] Several other Institute of Medicine reports elaborated the basic idea.[93] The idea, to add a bit more detail, is that in a learning health system data are continuously captured about patient care and how it goes (rather than letting that information slip by),[94] and that data are analyzed to learn what works well and where problems are,[95] and crucially, those analyses are actually used to change care going forward.[96] As Charles Friedman and colleagues characterize learning health systems, they involve the stages of Performance to Data (capturing what happens in the health system), Data to Knowledge (assembling and analyzing those data and interpreting those results in light of external evidence), and Knowledge to Performance (designing an intervention and taking access).[97]

A key difficulty with establishing true learning health systems is the gap in the final step of the cycle, back to implementation. Here, AI has the chance to shine, because it offers the opportunity to rapidly embed new learning into a system that is already incorporated into practice. It can change how it behaves without the need to, for instance, train clinicians in a new way of doing things, or acquire new forms of equipment, or otherwise alter workflows. If it were to turn out that predictions or recommendations should be different, based on experience, for one or another subgroup of patients in a particular setting, those changes could theoretically be implemented in an embedded AI system rapidly and seamlessly, showing up in results in the electronic medical record without major hurdles. This, of course, is the vision; getting there is much harder.

---

[91] *See* NATIONAL ACADEMY OF MEDICINE, *supra* note 85, at 89.

[92] INSTITUTE OF MEDICINE, THE LEARNING HEALTHCARE SYSTEM 6 (2007) [hereinafter IOM, LEARNING HEALTHCARE SYSTEM].

[93] *See* INSTITUTE OF MEDICINE, CLINICAL DATA AS THE BASIC STAPLE OF HEALTH LEARNING: CREATING AND PROTECTING A PUBLIC GOOD (2010); INSTITUTE OF MEDICINE, DIGITAL INFRASTRUCTURE FOR THE LEARNING HEALTH SYSTEM: THE FOUNDATION FOR CONTINUOUS IMPROVEMENT IN HEALTH AND HEALTH CARE (2011); INSTITUTE OF MEDICINE, BEST CARE AT LOWER COST: THE PATH TO CONTINUOUSLY LEARNING HEALTH CARE IN AMERICA (2013).

[94] *See* IOM, LEARNING HEALTHCARE SYSTEM, *supra* note 92, at 48.

[95] *See*, *e.g.*, Sarah M. Greene at al., *Implementing the Learning Health System: From Concept to Action*, 157 ANN. INT. MED. 207, 208 (2012) (discussing how learning health systems continuously collect and analyze real-time clinical data both to identify gaps in care quality and to evaluate whether implemented changes produce their intended outcomes).

[96] *Id.*

[97] Charles P. Friedman et al., *Socio-technical infrastructure for a learning health system*, 8 LEARNING HEALTH SYS., 2024, at 2.

LLMs enable one version of this system that is enticing, though frankly too risky for implementation anytime soon, if ever. That is, given that LLMs can incorporate new information almost on-the-fly through retrieval-augmented generation and on-the-fly through information included in potentially lengthy prompts, LLMs used in a health system could—potentially—adjust on a daily basis by incorporating new data.[98] This is almost certainly a bad idea—there's too much black-boxiness going on inside LLMs to treat this kind of rapid, unvalidated change, and too much likelihood of overfitting and data irregularities leading to significant problems.[99] It's worth mentioning, though, as an example of taking an AI-enabled learning health system too far, too fast.

For more judicious but still rapid AI learning to occur, more information is needed than just the negative information normally called for in oversight. Some information is well fit to the idea of oversight as avoiding failures, such as the adverse events already (if very unevenly[100]) reported to MAUDE or the information called for by those seeking greater oversight like near misses and performance drift. But more is needed for improvement. If an AI system performs particularly poorly on one subgroup, that's information worth sharing (and would probably be shared in a well-functioning oversight model)—but when an AI system performs particularly *well* for a different subgroup, that's not information that would need to be shared in a pure oversight model but absolutely should be shared in a more integrative learning health system model. And that kind of information to the extent that data-sharing infrastructure enables the sharing of the first form of data (poor subgroup performance), it should similarly enable sharing of the second (better subgroup performance); it's just a matter of choosing to report more broadly.

Not all forms of learning are so (potentially) straightforward to share. When health systems adjust workflow in a way that enables an AI system to perform more effectively and for its insights to be implemented better, that information can be a source of learning not only for that health system but also for others, and would ideally be shared. But that's also not the kind of information that can be shared readily via an automated system. Classically, this information would be shared in an academic publication or not shared at all, the latter because the information was deliberately as a trade secret or simply not worth taking the effort of publication. A better infrastructure to share this kind of information is also needed; one model might be the best practices repository being developed by the Health AI Partnership.[101]

---

[98] *See supra* notes 45-47 and accompanying text.

[99] One might think that clinicians involved in care could catch such errors, but that's a much more challenging task than one might think. *See* Price, *Clinicians in the Loop*, *supra* note 11, at 1279.

[100] *See supra* Part II.B.

[101] *Key Decision Points*, HEALTH AI PARTNERSHIP, https://healthaipartnership.org/key-decisions-in-adopting-an-ai-solution [https://perma.cc/6QZB-678P] (last visited Mar. 7, 2026); *See generally Empowering Healthcare with Responsible AI: Scaling and Fortifying a Community of Practice*, HEALTH AI PARTNERSHIP, https://healthaipartnership.org/insight/empowering-healthcare-with-responsible-ai-scaling-and-fortifying-a-community-of-practice

A. An Example of AI Learning: Epic's Cosmos System

Indeed, we have an example of something like this sort of information sharing in the real world—but confined to a specific (if important) commercial environment. Epic, the largest vendor of electronic medical records, has developed an AI product named Cosmos, "which collects de-identified patient records from participating health systems and displays predicted outcomes on similar patients based on Cosmos data directly in the EHR."[102] Cosmos, however, is available only to those customers of Epic who choose to share their data with the company.[103] Though limited information is available about exactly how this product works, there are at least two aspects here which are plausibly connected with the idea of sharing health data for a learning health system based on AI. First, the Cosmos product itself specifically uses AI and data from a combination of health systems to make predictions about the health care path of an individual patient;[104] that's a clear example of cross-contextual learning, but limited to the context of care for one patient at a time. Second, the lure of sharing data for Cosmos[105] (or with other incentives) for Epic means that the company has data from multiple health systems available to improve learning and performance for its many *other* AI health systems embedded within its EHR.

The question, then, is how to enable such within- and cross-system learning in the context of medical AI in environments that *aren't* just Epic. It's a problem for many reasons if only one company is able principally to benefit from this type of learning—not least because Epic has had some notable failures in the past with respect to cross-contextual performance of its own algorithms.[106]

B. FDA Surveillance Standards as an Opportunity

One key opportunity is FDA's current effort to develop standards for continuous performance evaluation. As described above, the agency is focused on

---

[https://perma.cc/QBY4-J8D4] (last visited Mar. 7, 2026) (discussing how the Health AI Partnership developed and continues to expand a community-sourced repository of best practice guides to help healthcare organizations adopt AI safely, effectively, and equitably).

[102] Brittany Trang, *Epic's AI overhaul promises to address EHR headaches for clinicians and patients*, STAT (Aug. 20, 2025), https://www.statnews.com/2025/08/20/epic-ehr-artificial-intelligence-microsoft/ [https://perma.cc/QY8R-Y5DX]; *Cosmos*, EPIC SYSTEMS, https://cosmos.epic.com/ [https://perma.cc/2KM2-MKSZ] (last visited Mar. 7, 2026).

[103] *See* Trang, *supra* note 102.

[104] *Id.*

[105] As one health tech reporter put it, "[Epic CEO Judy] Faulkner subtly implied that customers who don't hand over their data won't be able to give the best care to their patients: 'If you haven't signed up for Cosmos yet, please consider doing so,' she said, 'so you can get these capabilities that go with Epic artificial intelligence and [that will] eventually, we think, be important for the best patient care.'" *Id.*

[106] Andrew Wong, et al., *External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients*, 181 JAMA INTERNAL MED. 1065 (2021) https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2781307 [https://perma.cc/WJC3-KHBV].

how to monitor AI systems better with respect to their performance in the real world.[107] The default focus for this monitoring—like with MAUDE and other adverse-event reporting systems in other contexts—is to focus on problems that arise and avoiding patient-injuring errors and decrements in system performance. But sharing positive learning about improvements in system performance, effective structures surrounding the AI system and enabling its success,[108] better performance in specific subgroups, and the like, can all be part of the surveillance system.

To be sure, there's not an existing system that captures this information. Nor are there set standards for exactly what needs to be reported, unlike the death/injury/malfunction requirements of reporting to MAUDE. But proponents of better safety monitoring for AI systems already recommend broadening oversight systems.[109] If we are to capture and share real-world performance data in a form that can be readily organized and shared with FDA—as the agency clearly contemplates—now is the time to plan for capturing other information as well within the same system.

There's room for more opportunities beyond FDA's systems, of course. I've pushed in the past for robust data sharing of patient information via, for instance, federated learning so that other systems can learn directly from patient data, rather than indirectly via other systems' experiences with those data.[110] Embedding requirements for sharing such data within ongoing monitoring efforts might be outside the scope of FDA's oversight. But it would nevertheless be significant in driving the improvement of AI and its improved implementation.

Sharing specific stories of AI integration in health systems is important as well, though such sharing isn't as helpful for enabling the rapid implementation vision of AI as a driver of change in the learning health system.

V. COMPLEXITIES

In addition to the main central task—how to share information for both robust oversight and learning for improvement—a few complexities arise, which I touch on briefly here. First, what are the dynamic effects on information generation and incentives? Second, how shall this be paid for? Third, who shall own the resulting data? And fourth, how do international considerations fit in?

---

[107] WORLD HEALTH ORGANIZATION, *supra* note 73, at 24-25. *See also supra* notes 74-79, and accompanying text.

[108] Gerke et al., *supra* note 71, at 3-4.

[109] *See supra* Part III.

[110] *See* W. Nicholson Price II, *Secrecy, Health Data Infrastructure, and Medical AI*, in RESEARCH HANDBOOK ON TRADE SECRECY IN DATA AND DATA INFRASTRUCTURE (Rochelle Dreyfuss, Katherine Strandburg, & Christopher Morten, eds.) (forthcoming 2026) [hereinafter Price, *Secrecy*].

A. Dynamic Incentives

One way to look at the problems of medical AI are that developers have inadequate incentives to develop and share information about safety, and especially inadequate incentives to share information about new innovation.[111] Under this framing, I'm essentially proposing that these inadequate incentives be solved through mandates: firms must share information about safety issues (and proposals are that they share much more), and—along for the ride, I propose—they should also be required to share information about what works well and how. Fine, for the moment, but what about the dynamic impacts of those mandates? What impacts will mandates have on incentives going forward?

For safety, mandatory disclosure seems a mostly unalloyed good. Developers and deployers of AI systems are already required, by regulation and tort law, to make sure they're safe. To the extent that better disclosure increases the strength of that mandate, it should increase incentives to make sure these systems are safe.[112]

For learning for improvement, though, the incentives are a bit more complicated. Classically, the ability to free-ride off another's innovation is the central problem that intellectual property as a regime is trying to solve, because free-riders decrease the ability to capture the benefits of innovation and thus decrease the incentives to innovate in the first place.[113] Would mandating disclosure about what works, and when, and how, decrease incentives to generate that information in the first place? My intuition is no, for a couple of reasons. First, many relevant actors in this space—in particular, hospitals, health systems, and clinicians—are determining what works because they want the systems to work so that they can provide more effective or more efficient care. There may be some marginal gain to proprietary methods[114] ("come to our hospital because our sepsis prediction system is better integrated with critical care nursing teams!"), but mostly, it's about solving problems and making things better.[115] In this sense, many relevant innovators are akin to user innovators.[116] Second, tasks like figuring out how to integrate AI into a workflow effectively are likely to be relatively cheap

---

[111] My thanks to Doni Bloomfield for crisply articulating this facet of the situation.

[112] We could theoretically take this too far—I've argued in the past that a key risk of medical AI is that we don't use it *enough*, worry too much about safety, and leave a lot of potential health improvement on the table. *See* W. Nicholson Price II, Sara Gerke & I. Glenn Cohen, *Potential Liability for Physicians Using Artificial Intelligence*, 322 JAMA 1765, 1765 (2019) ("[T]he challenge is that current [tort] law incentivizes physicians to minimize the potential value of AI."). https://jamanetwork.com/journals/jama/fullarticle/2752750 [https://perma.cc/2DX7-NHQ8]. But disclosure obligations seem unlikely to do that.

[113] *See* WILLIAM M. LANDES & RICHARD A. POSNER, THE ECONOMIC STRUCTURE OF INTELLECTUAL PROPERTY LAW at 11 (2003).

[114] Some of this information may be difficult to fully exclude anyway. *See* Amy Kapczynski & Talha Syed, *The Continuum of Excludability and the Limits of Patents*, 122 YALE L.J. 1900, 1037-41 (noting the difficulty of excluding others from using information about improvements in healthcare quality).

[115] *See* Price & Sachs, *supra* note 6, at 9-16.

[116] *Id.*

compared to other biomedical innovation, like developing a new AI system from scratch (or a physical medical device or a drug).[117] Nevertheless, even if these mandates aren't likely to quash innovation, they still need support—especially if gains aren't privately appropriable—and it is to that that I turn next.

B. Funding

It's all very well and good to say that a robust AI-enabled learning health system will make things better, including how well AI performs and the care that's enabled by that performance—but who's going to pay for it? As I've argued in prior work with Mark Sendak and Suresh Balu, the reimbursement and payment structures for health systems don't provide adequate resources or incentives for robust monitoring as-is, much less a vision that encompasses a broader set of information being shared for learning.[118]

There's something to be gained from piggybacking. To the extent that FDA requires more robust real-world surveillance and postmarket monitoring,[119] health systems and developers will need to implement infrastructure to share information for the purposes of that safety monitoring. For some types of data to be shared for learning, that infrastructure should be essentially the same, making broader sharing relatively cheap, and requiring less in the way of funding.

For other forms of information—reports of what works well, and how workflows can be improved in practice, for instance—it's not so easy to report results automatically. More traditional grant funding can help support these sharing efforts, but so too can standards, like best practices for reporting these sorts of effectiveness improvements. To be sure, each place and workflow are unique—but there are commonalities in medical practice, and common variations, which can cut down on the challenge of reporting, analyzing, and learning from the shared experiences of others.[120]

We might also look for funding (coupled with a mandate) in another major driver of hospital reporting: the Centers for Medicare and Medicaid Services (CMS), which have implemented reporting and oversight requirements in the past, such as for 30-day readmission rates to hospitals.[121] Based on the assumption that

---

[117] *C.f.,* Mark P. Sendak, Suresh Balu & Kevin A. Schulman, *Barriers to Achieving Economics of Scale in Analysis of EHR Data,* 8 APPLIED CLIN. INFORMATICS 826, 828 (2017) (estimating a cost of $90,000 to validate a kidney AI system and incorporating it into a local clinical workflow, compared to over $215,000 to develop the system from scratch).

[118] Sendak et al., *supra* note 63.

[119] *See supra* notes 74-79 and accompanying text.

[120] *See* Bruno Valan et al., *Evaluating sepsis watch generalizability through multisite external validation of a sepsis machine learning model*, 8 NPJ DIGIT. MED. Art. no. 350, at 4-5 (2025) (evaluating the performance of an AI system developed in an academic medical center when implemented in a "dramatically different" clinical setting) https://www.nature.com/articles/s41746-025-01664-5 [https://perma.cc/EKT7-FN9U].

[121] *See* R. Neal Axon & Mark V. Williams, *Hospital Readmission as an Accountability Measure*, 2011 JAMA 504, 504-05 (2011) (describing CMS's implementation of public reporting and payment adjustments tied to 30-day readmission rates) https://pubmed.ncbi.nlm.nih.gov/2128

implementing AI systems can improve patient care and also increases efficiency for health systems, CMS could couple funding and reimbursement for the use of AI systems with requirements that implementers share not only quantitative performance data but also standardized accounts of how AI systems are incorporated into the delivery of care, what works, and what doesn't.

C. Ownership

One persistent issue that arises in the context of sharing health system data (or other types of data) is who owns the data.[122] I mention it here to offer a blunt response: it's deeply unhelpful to think about information on AI performance in different contexts and over time as proprietary, to be owned, kept secret, licensed, hoarded, or any variation of those concepts.[123] The health system desperately needs improvement, and keeping these types of secondary health data proprietary is a major barrier to that improvement (and, for the reasons mentioned above, not really necessary[124]).

There have been efforts to increase transparency and require data sharing about model performance at the initial phases of review and implementation. For instance, the Office of the National Coordinator for Health Information Technology (ONC) has promulgated a rule requiring transparency from vendors of AI embedded in electronic health records, including about model training and performance.[125] Such requirements should also apply to ongoing performance monitoring on behalf of developers and health systems alike. In the context, unlike other contexts where data-sharing mandates raise Takings Clause concerns about the government appropriating trade secrets,[126] almost all of the relevant sharing

---

5430/ [https://perma.cc/86RT-PCKR].

[122] There are lots of nuances here about what data ownership means—rights to control, to profit, etc. *See, e.g.*, Jorge L. Contreras, *The False Promise of Health Data Ownership*, 94 N.Y.U. L. REV. 624, 631-32 (2019) (arguing personal health information is not property) https://nyulawreview.org/issues/volume-94-number-4/the-false-promise-of-health-data-ownership/ [https://perma.cc/44D8-9QWM]. I'm not focusing on those nuances here, for reasons that will become obvious.

[123] *See* Price, *Secrecy*, *supra* note 110.

[124] *See supra* Part V.A.

[125] Health Data, Technology, and Interoperability: Certification Program Updates, Algorithm Transparency, and Information Sharing, 89 FR 1192 (Jan. 9, 2024) (codified at 45 C.F.R. pts. 170-71). As of this writing, the Trump Administration has proposed rolling back some of these requirements. Health Data, Technology, and Interoperability: ASTP/ONC Deregulatory Actions To Unleash Prosperity, 90 FR 60970 (proposed Dec. 29, 2025) (to be codified at 45 C.F.R. pts. 170-171). https://www.federalregister.gov/documents/2025/12/29/2025-23896/health-data-technology-and-interoperability-astponc-deregulatory-actions-to-unleash-prosperity [https://perma.cc/BC9Y-TMYG].

[126] *See, e.g.,* W. Nicholson Price II & Arti K. Rai, *Manufacturing Barriers to Biologics Competition and Innovation*, 101 IOWA L. REV. 1023, 1054-55 (discussing Takings Clause concerns with disclosure of biologics manufacturing information and retroactivity issues) https://ilr.law.uiowa.edu/sites/ilr.law.uiowa.edu/files/2022-10/Manufacturing%20Barriers%20to%20Biologics%20Competition%20and%20Innovation.pdf [https://perma.cc/7U5S-LKVN]; Price, *Secrecy*, *supra* note 110, Part IV.C.

would be prospective, which is acceptable in the context of conditioning funding (for health systems, via CMS) or regulatory approval (for developers by FDA or ONC certification).

D. International Harmonization

Finally, is there some way to use AI as a key tool for a truly international learning health system, as well as at a national level? To be honest, it's possible that the answer is mostly no—it could be that differences in care, resources, patient populations, and medical conditions are substantial enough that there's minimal benefit to trying to implement such a vision. But there's also the possibility for big learning steps, since the circumstances of care may differ so widely. Trying to accomplish learning at an international level will require harmonization of standards in terms of sharing performance and other data, which is a substantial hurdle.[127]

In fact, some types of internationally oriented learning may be substantially more straightforward, such as in countries where care is more centrally managed and therefore variation among different care settings comes in more predictable forms. Such settings might also enable better learning than mere observation, such as by implementing ongoing pragmatic trials of different workflow integration of AI products, which would yield more robust data about what works and what doesn't.[128]

VI. CONCLUSION

Among the many opportunities that medical AI is bringing to the health system, it brings the possibility of helping to enable a true learning health system on a scale that has been previously unattainable. Right now is an opportunity—not only for improving care by bringing in safe and effective AI products, but also for improving the infrastructure of learning and improvement so that patient care continues to improve more readily and more rapidly than in our current systems. Building systems and infrastructure for just catching errors is valuable, but stopping there would be a waste; policymakers have the chance to do substantially more to improve the health system and patient care.

---

[127] Sandeep Reddy, *Global Harmonization of Artificial Intelligence-Enabled Software as a Medical Device Regulation: Addressing Challenges and Unifying Standards*, 3 MAYO CLIN. PROC. DIGIT. HEALTH 100191 (2025), https://pubmed.ncbi.nlm.nih.gov/40207007/ [https://perma.cc/R9CM-NKQP]; Santra et al., *supra* note 43.

[128] *Cf.* Rachel E. Sachs et al., *What can policymakers learn from the UK's RECOVERY trial to improve clinical research for COVID-19 and beyond?*, WRITTEN DESCRIPTION (May 3, 2021), https://writtendescription.blogspot.com/2021/05/what-can-policymakers-learn-from-uks.html [https://perma.cc/X3PY-GGSY] (describing how the UK was able to conduct broad, distributed pragmatic trials of COVID-19 interventions during the pandemic).