3-7-2022

Distributed Governance of Medical AI

W. Nicholson Price II
University of Michigan Law School, wnp@umich.edu

Forthcoming 25 SMU Sci. & Tech. L. Rev. __ (2022)
Comments welcome to wnp@umich.edu

DISTRIBUTED GOVERNANCE OF MEDICAL AI

W. Nicholson Price II*

Artificial intelligence (AI) promises to bring substantial benefits to medicine. In addition to pushing the frontiers of what is humanly possible, like predicting kidney failure or sepsis before any human can notice, it can democratize expertise beyond the circle of highly specialized practitioners, like letting generalists diagnose diabetic degeneration of the retina. But AI doesn't always work, and it doesn't always work for everyone, and it doesn't always work in every context. AI is likely to behave differently in well-resourced hospitals where it is developed than in poorly resourced frontline health environments where it might well make the biggest difference for patient care. To make the situation even more complicated, AI is unlikely to go through the centralized review and validation process that other medical technologies undergo, like drugs and most medical devices. Even if it did go through those centralized processes, ensuring high-quality performance across a wide variety of settings, including poorly resourced settings, is especially challenging for such centralized mechanisms. What are policymakers to do? This short Essay argues that the diffusion of medical AI, with its many potential benefits, will require policy support for a process of distributed governance, where quality evaluation and oversight take place in the settings of application—but with policy assistance in developing capacities and making that oversight more straightforward to undertake. Getting governance right will not be easy (it never is), but ignoring the issue is likely to leave benefits on the table and patients at risk.

* Professor of Law, University of Michigan Law School. JD, Columbia Law School, 2011. PhD (Biology), Columbia University, 2010. This work was presented as part of the SMU SciTech Law Review's 2022 Symposium on AI & Medicine: The Emerging Legal and Ethical Frameworks for Artificial Intelligence in Medicine. I thank Mark Sendak for helpful discussions and comments on an earlier draft; Ana Bracic, Karandeep Singh, and Yindalon Aphinyanaphongs for helpful discussions; and Phoebe Roque for exemplary research assistance.

I. THE NEED FOR GOVERNANCE

Why does medical AI require governance? Put plainly, quality, safety, and efficacy are quite difficult to assess. Many health-care technologies are credence goods,[1] requiring either blind faith in quality (not a great plan) or some form of rigorous, systemic evaluation (a better plan!) because one-off evaluations in the moment don't cut it. AI is very much a credence good.[2] It is novel; it deals in probabilities rather than certainties; and it relies on algorithms that are typically quite opaque, whether because of secrecy, inherent technological limits, or merely their underlying complexity.[3] Like any technology involved in medical care, whether AI works is hard knowledge to come by.

And quality concerns are well founded; researchers have demonstrated deep flaws with AI systems, including some in wide use. Sometimes the systems just aren't useful, and sometimes they're actively harmful. Health-care vendor Epic's AI-powered system to predict the risk of sepsis, distributed and used in hospitals around the country, turns out to be a very poor predictor of risk—perhaps because the algorithm used as a predictive variable whether a physician had *already* ordered antibiotics, a typical *response* to sepsis.[4] COVID-19 prediction algorithms, developed rapidly and heralded as a success story for quick AI innovation in a global pandemic, turn out not to have performed very well at all and to have made little difference.[5] A tool for

[1] Uwe Dulleck, Rudolf Kerschbamer & Matthias Sutter, *The Economics of Credence Goods: An Experiment on the Role of Liability, Verifiability, Reputation, and Competition*, 101 AM. ECON. REV. 526, 526 (2011).

[2] W. Nicholson Price II, *Regulating Black-Box Medicine*, 116 MICH. L. REV. 421, 432 (2017).

[3] *Id.* at 432–37.

[4] Andrew Wong et. al, *External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients* 181 J. AM. MED. INTER. MED. 1065 (2021); Casey Ross, *Epic's Sepsis Algorithm Is Going Off the Rails in the Real World. The Use of These Variables May Explain Why*, STAT (Sept. 27, 2021), https://www.statnews.com/2021/09/27/epic-sepsis-algorithm-antibiotics-model/.

[5] *E.g.*, William Douglas Heaven, *Hundreds of AI Tools Have Been Built to*

analyzing X-rays for pneumonia flopped when it was tested in another hospital (it was relying on X-ray procedural clues rather than actual patient image traits).[6] And a tool used by an insurer to allocate care coordinators was shown to be strongly biased against Black patients because of a careless proxy decision made in development.[7] AI has a lot of promise—but adopters are right to be cautious, and real governance is required to make sure that the systems being considered actually work in general, work where they are used, and work for the patients and providers in that particular setting.

II. THE LIMITS OF CENTRALIZED REGULATION

The default turn for regulation of new medical technologies is to FDA. The agency regulates drugs and medical devices, and software—including AI software—is explicitly within the bounds of regulated "medical devices"[8] (or, at least, may be[9]), whether embedded in other devices (Software *in* a Medical Device or SiMD) or on its own (Software *as* a Medical Device, or SaMD).[10] FDA

Catch Covid. None of Them Helped, MIT TECH. REV. (July 30, 2021) https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/; Jordana Cepelewicz, *The Hard Lessons of Modeling the Coronavirus Pandemic*, QUANTA MAG. (Jan. 28, 2021) https://www.quantamagazine.org/the-hard-lessons-of-modeling-the-coronavirus-pandemic-20210128/.

[6] W. Nicholson Price II, Rachel E. Sachs & Rebecca S. Eisenberg, *New Innovation Models in Medical AI,* XX WASH U. L. REV (forthcoming 2022) (manuscript at 44–45) (on file with author).

[7] Ziad Obermeyer, Brian Powers, Christine Vogeli & Sendhil Mullainathan, *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, 366 SCI. 447 (2019); *see also* Jenna Wiens, W. Nicholson Price II & Michael W. Sjoding, *Diagnosing Bias in Data-Driven Algorithms for Healthcare*, 26 NATURE MED. 22 (2020) (responding to Obermeyer et al.).

[8] Barbara Evans & Frank A. Pasquale, *Product Liability Suits for FDA-Regulated AI/ML Software*, *in* THE FUTURE OF MEDICAL DEVICE REGULATION: INNOVATION AND PROTECTION (I. Glenn Cohen, Timo Minssen, W. Nicholson Price II, Christopher Robertson & Carmel Shachar eds., forthcoming 2022) (manuscript at 2–3) (on file with author).

[9] *Id.*; *see also* Nathan Cortez, *Substantiating Big Data in Health Care*, 14 I/S: J. L. & POL'Y INFO. SOC'Y, 61, 72–81 (2017).

[10] *See Software as a Medical Device (SaMD): Key Definitions*, International Medical Device Regulators Forum ("IMDFR") (Dec. 19, 2013), http://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-131209-samd-key-

regulates (or at least can regulate) software via the familiar Class I-II-III risk-based categorization of medical devices, with scrutiny titrated to risk.[11] This process is meant to ensure that medical devices are safe and effective before entering into commerce and then into practice—but the process has severe limitations on how well it can govern the quality of the broad sweep of AI products developed and used today.

A. What does FDA see?

The first problem is one of coverage: Many AI systems in development or already in use have not gone through FDA review of any kind, and many likely never will. As I have written with Rachel Sachs and Rebecca Eisenberg, there is substantial user innovation in the space of medical AI, where health systems, hospitals, and insurers are developing AI systems for use within their own walls.[12] AI innovation is well within the capacity of many well-resourced actors in this space, in a way that the development of novel drugs or complex physical medical devices, for instance, may not be.[13] Academic medical centers have developed in-house predictors for the likelihood of sepsis,[14] hospital systems have developed AI systems to model patient flow (and,

---

definitions-140901.pdf; "*Software as a Medical Device": Possible Framework for Risk Categorization and Corresponding Considerations*, IMDFR (Sept. 18, 2014), http://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-140918-samd-framework-risk-categorization-141013.pdf; *Global Approach to Software as a Medical Device*, FDA (Dec. 6, 2017) https://www.fda.gov/medical-devices/software-medical-device-samd/global-approach-software-medical-device (adopting IMDFR's regulatory framework).

[11] *Classify Your Medical Device*, FDA (Feb. 7, 2020), https://www.fda.gov/medical-devices/overview-device-regulation/classify-your-medical-device.

[12] Price, Sachs & Eisenberg, *supra* note 6.

[13] *Id.* at 7–8. Notably, health systems have begun developing the capacity to manufacture drugs, though so far their efforts are entirely focused on generic manufacturing and not developing new products. *See*, *e.g.*, CIVICARX, https://civicarx.org/ (last visited Feb. 19, 2022) (describing a collaborative generic drug manufacturing initiative undertaken by health systems).

[14] *See, e.g.*, Mark Sendak et al., *Real-World Integration of a Sepsis Deep Learning Technology into Routine Clinical Care: Implementation Study,* 8 JMIR MED. INFORM. 1 (2020) (discussing Duke Health's Sepsis Watch program).

accordingly, resource allocation),[15] and insurers have developed models to allocate care coordinators to reduce costs.[16] These user innovations are very unlikely to see any sort of formal FDA review for multiple reasons, including FDA's long-time exercise of enforcement discretion over laboratory-developed tests,[17] the exclusion of many clinical decision support software products (CDS) from the definition of medical devices under the 21st Century Cures Act,[18] and potentially other jurisdictional issues.[19] These systems are already being used. And they undoubtedly have quality problems—indeed, multiple examples of quality failures described above were in-house systems. But FDA does not review them, at least not typically.[20]

On a broader scale, AI products are being embedded within electronic health record (EHR) systems that are then distributed

---

[15] *See, e.g.*, Michael Thompson, *New Ways to Improve Hospital Flow with Predictive Analytics* (March 2019), https://www.slideshare.net/healthcatalyst1/new-ways-to-improve-hospital-flow-with-predictive-analytics (outlining Cedar-Sinai Medical Center's AI tool that aims to reduce capacity strain by predicting patient census).

[16] Obermeyer, Powers, Vogeli & Mullainathan, *supra* note 7, at 447 ("Large health systems and payers rely on this algorithm to target patients for 'high-risk care management' programs. These programs seek to improve the care of patients with complex health needs by providing additional resources, including greater attention from trained providers, to help ensure that care is well coordinated. Most health systems use these programs as the cornerstone of population health management efforts, and they are widely considered effective at improving outcomes and satisfaction while reducing costs.").

[17] Price, Sachs & Eisenberg, *supra* note 6, at 26; *see also* FDA, *Draft Guidance for Industry, Food and Drug Administration Staff, and Clinical Laboratories, Framework for Regulatory Oversight of Laboratory Developed Tests (LDTs)*, at 5–7 (Oct. 2014); FDA, *Discussion Paper on Laboratory Developed Tests (LDTs)*, at 4–5 (Jan. 2017).

[18] 21st Century Cures Act, Pub. L. 114–255 §3060, 130 Stat. 1033, 1130–33 (2016), codified at 21 U.S.C. § 360.

[19] Price, Sachs & Eisenberg, *supra* note 6, at 22 ("The FDCA only applies to products that are introduced, delivered, or received in interstate commerce, an important limitation that may exclude many user innovations.") (citing 21 U.S.C. § 331).

[20] This is not to say that FDA is totally uninvolved; at least some developers have had discussions with FDA officials about how to *avoid* FDA review by ensuring sufficient presence of a human in the algorithmic loop to stay within the CDS exclusion of the Cures Act. Price, Sachs & Eisenberg, *supra* note 6, at 25–26.

to health care providers, sometimes quite broadly.[21] Epic, the market leader for EHR systems, has developed several AI-based tools that are integrated into its EHR suites, including some addressing COVID-19 risk and deterioration probability for patients more generally.[22] It is also developing a marketplace for developers to interact with its EHR systems.[23] Other EHR vendors, such as Cerner, have also developed AI products, though Epic appears to have taken an early lead.[24] Although these products are already distributed widely, it appears that no EHR-vendor-developed AI systems have gone through FDA premarket review.[25] And they have quality issues, too; a wide-ranging review

---

[21] *See* Sehj Kashyap, Keith E. Morse, Birju Patel & Nigam H. Shah, *A Survey of Extant Organizational and Computational Setups for Deploying Predictive Models in Health Systems*, 28 J. AM. MED. INFORM. ASSOC. 2445 (2021).

[22] Epic integrated EHR systems into its COVID-19 prediction models—models that are used to determine an individual's likelihood of testing positive for COVID-19 as well as their likelihood of needing critical care after testing positive (*e.g.*, the Deterioration Index). Alicia Reale-Cooney, *COVID-19 Risk Model Developed by Cleveland Clinic Now Available to Health Systems Around the World Through Epic*, CLEVELAND CLINIC (Nov. 29, 2020), https://newsroom.clevelandclinic.org/2020/11/09/covid-19-risk-model-developed-by-cleveland-clinic-now-available-to-health-systems-around-the-world-through-epic/; *Epic AI Helps Clinicians Predict When COVID-19 Patients Might Need Intensive Care*, EPIC (May 18, 2020), https://www.epic.com/epic/post/epic-ai-helps-clinicians-predict-covid-19-patients-might-need-intensive-care. Prior to the pandemic, Epic produced SlicerDicer, an AI-based tool that "allows a provider to tap into patient data to investigate clinical conjectures or make new discoveries about patient populations." Christina DuVernay, *SlicerDicer Reveals Practice-Based Data*, JOHNS HOPKINS MEDICINE (Aug. 31, 2017), https://www.hopkinsmedicine.org/office-of-johns-hopkins-physicians/best-practice-news/slicerdicer-reveals-practice-based-data. Epic is also currently developing an EHR document assistant that would enable AI to transcribe patient-clinical interactions. Christopher Jason, *Epic in Process of Developing AI EHR Documentation Assistant*, EHR INTELLIGENCE (Feb. 21, 2021), https://ehrintelligence.com/news/epic-in-process-of-developing-ai-ehr-documentation-assistant.

[23] Epic, *Epic App Orchard*, https://apporchard.epic.com (last visited Feb. 20, 2022).

[24] Christopher Jason, *Epic Systems, Cerner Lead EHR Vendors in AI Development*, EHR INTELLIGENCE (May 12, 2020), https://ehrintelligence.com/news/epic-systems-cerner-lead-ehr-vendors-in-ai-development.

[25] In September 2021, the FDA released a list of AI and machine-learning enabled medical devices that have been reviewed and approved for the U.S.

found problems in many of Epic's AI products,[26] and another documented problems with its model reporting guidelines.[27]

Overall, then, a substantial fraction of AI products already being used in care settings do not appear to pass through centralized, national-level review for safety and efficacy.[28] And this lack of governance isn't because the products are great. But limited scope of FDA review isn't the only problem of centralized governance.

B. What can FDA do?

The second problem is more systematic: Even for those AI products that do go through FDA review, that review only addresses a subset of the issues for which governance is necessary.

A burgeoning literature addresses the limitations of FDA review for AI products writ large. For instance, the vast majority of AI products that have received some sort of FDA marketing authorization have gone through the 510(k) clearance process rather than a full approval.[29] 510(k) clearance requires

---

market. The list does not include any products by Epic or Cerner. *Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices*, FDA (Sept. 22, 2021), https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices ; *see also* Jodi K. Scott, Kristin Zielinski Duggan, Lina Kontos, Suzanne Levy Friedman & Kelliann Payne, *FDA Launches List of AI and Machine Learning-Enabled Medical Devices*, HOGAN LOVELLS (Sep. 23, 2021), https://www.engage.hoganlovells.com/knowledgeservices/news/fda-launches-list-of-ai-and-machine-learning-enabled-medical-devices.

[26] Ross, *supra* note 4.

[27] Jonathan H. Lu et al., *Low Adherence to Existing Model Reporting Guidelines by Commonly Used Clinical Prediction Models*, MEDRXIV 2021.07.21.21260282, https://www.medrxiv.org/content/10.1101/2021.07.21.21260282v1 (Jul. 21, 2021),

[28] FDA review is not the only form of such review; insurers also sometimes play a quality-review function—but appear not to be playing such a role here. Price, Sachs & Eisenberg, *supra* note 6, at 40.

[29] Charlotte Tschider, *Medical Device Artificial Intelligence: The New Tort Frontier e*, 46 BYU L. REV. 1551, 1597 (2021). The minority that have not gone through the 510(k) process have been *de novo* classified as Class I and II devices rather than undergoing the full Class III premarket approval process. Medical

demonstrating substantial equivalence to a product that is already marketed,[30] and FDA has made clear that an AI product can be demonstrated to have substantial equivalence to an already approved non-AI product.[31] But the "substantial equivalence" standard has long been criticized for insufficient rigor.[32] Such arguments apply in the context of AI systems as well.[33] Critiques also note FDA's potential deficit in terms of AI-expert personnel[34] (a deficit FDA is trying to remedy[35]) and raise the challenges of regulating products that can and potentially should be updated

---

Device De Novo Classification Process, 86 Fed. Reg. 54, 826 (Oct. 5, 2021) (to be codified at 21 C.F.R. pt. 860); *see also The De Novo Pathway: What Has Changed in 10 Years?,* GUIDED SOLUTIONS (Aug. 19, 2019), https://www.guidedsolutions.co.uk/blog/the-de-novo-pathway/.

[30] *Premarket Notification 510(k)*, FDA (March 13, 2020), https://www.fda.gov/medical-devices/premarket-submissions/premarket-notification-510k#se.

[31] *See How FDA Regulates Artificial Intelligence in Medical Products,* PEW TRUSTS (Aug. 15, 2021) https://www.pewtrusts.org/en/research-and-analysis/issue-briefs/2021/08/how-fda-regulates-artificial-intelligence-in-medical-products.

[32] Jeffrey K. Shapiro, *Substantial Equivalence Premarket Review: The Right Approach for Most Medical Devices*, 69 FOOD & DRUG L.J. 365, 365 (defending the substantial equivalence standard but acknowledging that it "is frequently compared unfavorably to [pre-market] approval as a means of establishing the safety and effectiveness of new devices, as an affront to the original intent of the Medical Device Amendments of 1976 (MDA), and as a "regulatory loophole" that should be scrapped or, if that is not practical, at least limited to the extent possible.") (internal citations omitted). Indeed, the Supreme Court has recognized the difference in rigor between approval and clearance, finding that approval preempts state tort litigation for certain types of defect, while clearance does not. *Riegel v. Medtronic, Inc.*, 552 U.S. 312 (2008); *see also* Tschider*, supra* note 27, at 1597.

[33] Soleil Sha & Abdul El-Sayed, *The FDA Should Better Regulate Medical Algorithms*, SCI. AM. (Oct. 7, 2021) https://www.scientificamerican.com/article/the-fda-should-better-regulate-medical-algorithms/ .

[34] Tschider, *supra* note 27, at 1586 ("To what degree FDA personnel or panel members actually provide expert direction in this [pre-market approval] review process is unknown, though facially it seems unlikely that personnel and panel members are equipped to review software design and anticipate real patient risks for new software technology like AI from a position of deep expertise.").

[35] Dave Muoio, *Commissioner Hahn: FDA Hiring More Data Experts to Help Healthcare 'Unleash the Power of Data*', MOBI HEALTH NEWS (Oct. 12, 2020), https://www.mobihealthnews.com/news/commissioner-hahn-fda-hiring-more-data-experts-help-healthcare-unleash-power-data.

relatively frequently as new data and performance metrics become available.[36] Some note the challenge of regulating systems incorporating AI as whole systems, rather than trying to focus on the algorithm itself.[37] And finally, the set of entities with the resources and capabilities of taking an AI system through the FDA evaluation process is itself limited—especially given uncertainty about payment and reimbursement mechanisms[38]—impacting who is able to innovate effectively if the default model runs through FDA.[39]

Even setting aside the questions about how to get centralized governance functioning well in the first place, centralized governance can only do so much for AI systems that frequently need to be adapted and responsive to local environments.[40] Some products might reasonably expect to work the same essentially irrespective of context; whether a retina shows signs of diabetic retinopathy when examined through a uniform camera system *hopefully* doesn't change depending on whether that system is used in an academic medical center in Boston or a clinic in Alabama.[41] Other systems that similarly rely on measurements expected to be universally applicable might fit well into a centralized, national (or international) regulatory paradigm, whether those systems calculate heart volumes using machine learning,[42] identify

---

[36] *Id.*

[37] *See* Sara Gerke, Boris Babic, Theodoros Evgeniou & I. Glenn Cohen, *The Need for a System View to Regulate Artificial Intelligence/Machine Learning-based Software as Medical Device*, NPJ DIGIT. MED., April 7, 2020, at 1.

[38] *See* Price, Sachs & Eisenberg, *supra* note 6, at __.

[39] *See* Price, *supra* note 2, at 452–53.

[40] *See* Mark Sendak et al., *Machine Learning in Health Care: A Critical Appraisal of Challenges and Opportunities*, 7 eGEMS 1, 2, http://doi.org/10.5334/egems.287 (2019) ("Personalized medicine will require mass customization of models that are trained and re-calibrated at the hospital and cohort level. Modern machine learning techniques focus on generalization beyond a training dataset, not on generalization to different sites.")

[41] *See* Michael D. Abràmoff, Philip T. Lavin, Michele Birch, Nilay Shah & James C. Folk, *Pivotal Trial of an Autonomous AI-based Diagnostic System for Detection of Diabetic Retinopathy in Primary Care Offices*, NPJ DIGIT. MED., Aug. 28, 2018, at 1.

[42] *See* Akhil Narang et. al, *Machine Learning Based Automated Dynamic Quantification of Left Heart Chamber Volumes*, 20 EUR. HEART J. 541 (2018)

cerebral hemorrhage on CT scans,[43] or perform other presumably generalizable tasks. (I say "might" because differences in how data are recorded and processed or underlying populations could still limit widespread applicability.[44])

On the other hand, many systems are inherently quite hard to generalize. Most obviously, systems that predict or recommend based on site workflow are closely tied to that site—a patient or staff volume predictor may only be applicable to the venue where it was developed.[45] Predictors of patient deterioration may not generalize well, whether because of differing patient populations,[46] differing care infrastructure and treatment patterns,[47] or simply differing data infrastructures so that different data are recorded and available for systems to use.[48] Predictors of sepsis, for instance, have been quite difficult to generalize across contexts, as have predictions of infection by *Clostridium difficile*.[49]

---

[43] *See* Mohammad R. Arbabshirani et. al, *Advanced Machine Learning in Action: Identification of Intracranial Hemorrhage on Computed Tomography Scans of the Head with Clinical Workflow Integration*, NPJ DIGIT. MED., April. 4, 2018, at 1.

[44] The commercially distributed IDx-DR diabetic retinopathy detection system requires a specific camera system for precisely this reason. Digital Diagnostics, *IDx-DR*, https://www.digitaldiagnostics.com/products/eye-disease/idx-dr/ (last visited Feb. 20, 2022).

[45] S*ee, e.g.*, Thompson, *supra* note 4.

[46] W. Nicholson Price II, *Contextual Bias and Medical AI*, 33 HARV. J.L. & TECH. 66, 91–94 (2019).

[47] *See id.* at 96–97; *c.f.*, M. Scottie Eliassen, Ashleigh King, Christopher Leggett, Sukdith Punjasthitkul & Jonathan Skinner, *The Dartmouth Atlas of Health Care: 2018 Data Update,* DARTMOUTH ATLAS PROJECT (2021) (providing the latest update in a series dedicated to reporting the variation in care and medical services throughout the United States).

[48] Price, *supra* note 41, at 100.

[49] Wong et al., *supra* note 4 (finding that Epic's sepsis prediction model correctly identified patients' risk of sepsis only 63% of the time due to hospitals' varying definitions of and billing codes for sepsis); Jeeheh Oh et al., *A Generalizable, Data-Driven Approach to Predict Daily Risk of Clostridium difficile Infection at Two Large Academic Health Centers* 39 INFECTION CONTROL & HOSP. EPIDEMIOLOGY 425, 425 (2018) (finding that institution-specific models for estimating risk for *Clostridium difficile* provide "earlier and more accurate identification of high-risk patients and better targeting of infection prevention strategies" than a one-size-fits-all approach).

It's not always straightforward to figure out which generalizability bucket a particular product will fit into. Looking at skin lesions to identify potential skin cancer seems like it might be generalizable—except that it turns out patient skin color makes a big difference, and dermatological image databases from different places are, you guessed it, very different in patient demographics as well.[50] Questions of generalization arise not only across high-resource contexts in the U.S. but especially in contexts with different levels of resources (and presumably different care patterns),[51] and of course are likely to be even more prominent in international contexts,[52] though international implications are largely beyond the scope of this piece.

All of this is not to say that AI systems developed specifically for a particular context can't usefully be applied to other contexts—just that that application can take substantial work and is often not straightforward; more importantly for present purposes, centralized, national-level governance is a poor fit for localized application.

III. DISTRIBUTED GOVERNANCE

Distributed, localized governance will be an essential complement to national regulators in providing robust oversight for medical AI. That recognition in itself is important; the normal approach to validating biomedical credence goods simply won't cover the gamut

[50] Veronica Rotemberg et al., *A Patient-Centric Dataset of Images and Metadata for Identifying Melanomas Using Clinical Context*, 8 SCI. DATA 34, 41 (2021); *see also* Nicole Westman, *Data Used to Build Algorithms Detecting Skin Disease is Too White*, VERGE (Sept. 23, 2021), https://www.theverge.com/2021/11/9/22770852/data-dermatology-algorithms-skin-tone-ethnicity. *But see* Angela Lashbrook, *AI-Driven Dermatology Could Leave Dark-Skinned Patients Behind*, ATLANTIC (Aug, 16, 2018), https://www.theatlantic.com/health/archive/2018/08/machine-learning-dermatology-skin-color/567619/ (citing Haenssle et al.'s study and arguing that, despite its clear racial disparities, machine-learning software could aid marginalized communities that don't have access to a dermatologist and can be improved over time).

[51] Price, *supra* note 41, at 95–97.

[52] *See* Daniel E. Weissglass, *Contextual Bias, The Democratization of Healthcare, and Medical Artificial Intelligence in Low–and Middle–Income Countries*, BIOETHICS, Aug. 2021, at 1.

of useful medical AI, at least any time in the near future.[53] With the need for local governance taken as given for the near term, then, what could such local governance look like?

A. Examples of local governance

One starting point for the normative is the positive; here, some well-resourced health environments are already engaged in organized local governance efforts to validate AI before, during, and after deployment. I focus here on academic medical centers which have the capacity both for in-house development and outsourcing models, with the recognition that governance capabilities and procedures may differ substantially in different contexts.

Duke, for instance, has devoted substantial resources to developing and deploying AI in medical practice.[54] It uses a stage-gate process, where a new AI model needs to move through various procedural steps before implementation. Outcome measures are defined with the input of clinical leaders (that is, what is the algorithm supposed to do in practice), and then performance and improvement targets are set relative to the baseline of standard practice. An algorithm is validated on retrospective data, then on current data while development continues. Once the algorithm is trained, if adequate performance metrics are met, it is integrated into Duke's electronic health records system—but in the background, and with a smaller set of beta testers, who may see

[53] Could we imagine AI engines so powerful, stable, and flexible, that they can be validated at a national level, then deployed locally to integrate into various data ecosystems, collecting and refining data and self-validating through reports to centralized regulators like FDA? Of course! But that health-algorithmic utopia is a long way off.

It's also worth noting that localized governance is the norm for much of the *rest* of medical practice, such as physician oversight or hospital safety. Biomedical technologies have typically been treated differently, as described above.

[54] All details in this paragraph are taken from a video interview with Mark Sendak, Population Health & Data Science Lead, Duke Institute for Health Innovation, November 3, 2021; *see also* Duke AI Health, *Algorithm-Based Clinical Decision Support (ABCDS) Oversight*, https://aihealth.duke.edu/algorithm-based-clinical-decision-support-abcds/ (last visited Feb. 20, 2022).

delayed results rather than real-time data and who do not use the algorithm to alter patient care. This period is used to both evaluate performance and determine what effective integration into the real clinical workflow would require. If a model passes this background beta testing, it is evaluated in a prospective study, and only after passing that can be considered for integration into routine care. At each stage, the process involves IT, individuals trained in technical evaluation and data analysis (such as the Duke Institute for Health Innovation[55]), and a set of clinical stakeholders across clinical departments: end users who will need to buy into the new technology, clinical leadership (including both nursing and physician staff), and operational leadership.

The University of Michigan has a similar governance process in substantive terms, but processes initiatives through the Clinical Informatics Committee.[56] The Committee includes academics, care providers, and other professionals. Its membership includes informatics representatives from physicians, pathology, nursing, and research; it also includes several representatives from Health IT and Learning Health Systems and adds guests and ad hoc members as needed. The Committee evaluates requests and suggestions for models to incorporate, and similarly runs them as background models for at least six months, after which decisions are made about potential integration into clinical workflow.

New York University's Langone Medical Center takes a more siloed approach: it has an in-house team, conceived of as part of the IT department but including researchers and clinical staff, that develops and deploys models, including the process of evaluating model performance and clinical workflow integration.[57]

---

[55] Duke Institute for Health Innovation, https://dihi.org (last visited Feb. 25, 2022).

[56] All details in this paragraph are taken from a video interview with Karandeep Singh, Chair, Michigan Medicine Clinical Intelligence Committee, Oct. 26, 2021.

[57] NYU Langone Health Center for Healthcare Innovation and Delivery Science, *Predictive Analytics and Machine Learning*, https://med.nyu.edu/centers-programs/healthcare-innovation-delivery-science/predictive-analytics-unit.

What do these various models tell us? Most basically, evaluating, implementing, and governing a model is a challenging, complex task. It involves substantial time, effort, and expertise across a significant timeframe and across both technical and clinical domains. Some entities treat this as a relatively enclosed process, largely entrusting it to one unit within the organization (e.g., NYU); others treat it more as a committee service requirement developed on top of other responsibilities (e.g., Michigan); and still others as a process involving both built-for-purpose subunits and broader clinical involvement (e.g., Duke). This is a complex, multimodal, demanding process of governance.

And governance doesn't stop at implementation. After the process of deciding to develop or import a model, evaluating it, and then integrating it, effective local governance also requires ongoing monitoring of performance and undertaking maintenance and updating efforts as required. Among other things, data drift within a particular environment means that an AI system will tend to lose performance over time, absent regular updating.[58]

B. Problematic capacity variations

Relying on the variety of sites that can use medical AI to create their own governance structures is, frankly, a recipe for failure. Resources vary wildly across contexts. While the sort of well-resourced academic medical center that has the capacity to develop AI may well have the capacity to evaluate, deploy, and otherwise govern that AI (with a bit of emphasis on "may," since those capacities are distinct), contexts with fewer resources generally are similarly less likely to have the governance infrastructure for evaluation and ongoing monitoring.[59]

---

[58] Adarsh Subbaswamy & Suchi Saria, *From Development to Deployment: Dataset Shift, Causality, and Shift-Stable Models in Health AI*, 21 BIOSTATISTICS 345 (arguing that datasets need to be monitored and maintained overtime to account for data shifts and their accompanying performance decay).

[59] *See, e.g.*, Tiankai Wang, Yangmei Wang & Alexander McLeod. *Do Health Information Technology Investments Impact Hospital Financial Performance and Productivity?* 28 INT'L J. ACCT. SYS. 1 (2018) (finding that health information technology investments, like electronic health records adoption, lead to positive financial performance and productivity); AM. HOSPITAL ASS'N, RURAL REPORT:

At a basic level, health entities differ substantially with respect to the information technology resources necessary for deploying and evaluating models. Running models silently in the background of systems for evaluative purposes demands more technical expertise and intervention than simply turning on a vendor-provided model and letting it run (though even that task is often fraught). And accurately collecting performance metrics brings similar challenges when deciding whether to go forward and integrate a model into the clinical workflow.

Perhaps more significantly, the human resources necessary for model evaluation are also highly disparate between settings. Michigan, for instance, relies on committee work that typically sits atop other responsibilities. Settings stretched for personnel simply may not be able to call on those human resources (and may not have staff with the relevant technical expertise in any case). Standalone units, like NYU Langone's in-house development team, are similarly out of reach for all but the most well-resourced medical environments.

Existing, intense, location-specific modes of governance are simply infeasible for many medical environments—including those that might benefit most from the ability of medical AI to democratize expertise, expand capacity, and improve care.

IV. POLICY INTERVENTIONS

CHALLENGES FACING RURAL COMMUNITIES AND THE ROADMAP TO ENSURE LOCAL ACCESS TO HIGH-QUALITY, AFFORDABLE CARE 14 (Feb. 2019) ("Rural hospitals are committed to improved care through use of HIT in order to meet past and current regulatory requirements. . . . Rural hospitals must meet the same regulatory requirements [in this area] as other hospitals, yet often do not need the additional technology functionality contained in required, expensive system upgrades; nor do they have the available infrastructure such as adequate broadband to support them."); Jordan Rau & Emmarie Huetteman, *Some Urban Hospitals Face Closure or Cutbacks as the Pandemic Adds to Fiscal Woes*, NPR (Sept. 15, 2020) https://www.npr.org/sections/health-shots/2020/09/15/912866179/some-urban-hospitals-face-closure-or-cutbacks-as-the-pandemic-adds-to-fiscal-woe.

Given the need for local, ongoing governance of medical AI and the currently unevenly distributed resources to conduct that governance, how can policymakers help enable that governance in the future? My aim here is not to solve the problem, but to sketch two directions for potential solutions: offloading and building capacity.

A. *Offloading*

One approach is to provide resources for low-resource institutions to offload whatever governance tasks can be reasonably offloaded. At a federal level, FDA review of AI tools could presumably perform a partial oversight role (for those tools that go through FDA[60])—demonstrating that a tool works in principle, for some set of assumptions—and then leaving the last-mile task of localized validation to the local entity. But FDA is not the only option. OCHIN, for instance, vets AI products and provides interfaces between those products and health system IT infrastructures to more easily adopt the products of trusted partners.[61] Other organizations, including for-profit entities that help health systems make procurement decisions, could develop similar capabilities.[62]

---

[60] *See supra* Section II.A.

[61] OCHIN collaborates with a range of technology partners with the following goal: "As leaders in the EHR space, we take great pride in being able to provide add-on functionality, services and upgrades to our members from our vast array of technological partnerships. . . . OCHIN coordinates the interface and build required to use these products so our members don't have to, and we are able to make it available for a fraction of market cost." *See Ochin's Preferred Technology Partners*, OCHIN, https://ochin.org/technology-partners (last visited Nov. 21, 2021); *see also* Christopher Jason, *eHealth Exchange Taps Electronic Case Reporting for Interoperability*, EHR INTELLIGENCE (Aug. 17, 2020), https://ehrintelligence.com/news/ehealth-exchange-taps-electronic-case-reporting-for-interoperability; Christopher Jason, *Epic Systems, OCHIN Launch COVID-19 Preparedness Screening App*, EHR INTELLIGENCE (March 30, 2020) https://ehrintelligence.com/news/epic-systems-ochin-launch-covid-19-preparedness-screening-app; OCHIN, *OCHIN Joins NIH Funded AI/ML Consortium to Advance Health Equity and Researcher Diversity,* PR Newswire (Oct. 7, 2021), https://www.prnewswire.com/news-releases/ochin-joins-nih-funded-aiml-consortium-to-advance-health-equity-and-researcher-diversity-301395495.html.

[62] *See*, *e.g.* Vizient, *Clinical Cost Management*,

B. *Building Capacity*

Offloading is unlikely to get us all the way there; policymakers should also consider investments in the infrastructure necessary to enable distributed governance, whether that infrastructure is technical, data-based, or procedural.

Technical infrastructure refers to the information technology tools necessary to process data, run AI tools, integrate those tools into the care workflow, and—crucially—monitor and evaluate the outputs to measure performance.[63] Technical infrastructure, in addition to facilitating in-house distributed governance, can also facilitate the flow of monitoring data *out* of the low-resource context, to enable the sort of outsourced monitoring described above.[64] Technical infrastructure also includes developing programs or technical tools to monitor performance, to integrate AI systems developed elsewhere into local clinical data ecosystems and workflow, and to facilitate the training of local care providers on new tools.[65]

---

https://www.vizientinc.com/our-solutions/clinical-solutions/clinical-cost-management (last visited Feb. 20, 2022).

[63] *See, e.g.*, *What is IT Infrastructure?*, IBM, https://www.ibm.com/topics/infrastructure (last visited Nov. 20, 2022); *cf.* W. Nicholson Price II, *Risk and Resilience in Health Data Infrastructure* 16 COLO. TECH. L.J. 65, 67, 77–79 (2017) ("[H]ealth data infrastructure would be infrastructure *for* health data—that is, infrastructure on which health data can be stored and transmitted (such as computer systems, shared data standards, and the like). But it should also be infrastructure *of* health data—that is, a platform of shared data on which to base further efforts to increase the efficiency or quality of care. In an infrastructure *of* data, the data themselves are a resource to enable productive downstream activity that can improve the health care system.").

[64] *Cf.* Kai Hu et al., *Federated Learning: A Distributed Shared Machine Learning Method*, COMPLEXITY, Aug. 2021, at 1, 1 (analyzing federated learning: a machine learning (ML) framework where "multiple clients collaborate to solve traditional distributed ML problems under the coordination of the central server without sharing their local private data with others.").

[65] For instance, technical infrastructure can facilitate the reporting of adverse events, itself typically left to widely distributed individuals. *See* U.S. FDA, *FDA Adverse Event Reporting System (FAERS) Electronic Submissions*,

Data infrastructure refers to the development of data resources on a broad, representative level for the development of AI tools. Why is this necessary for distributed *governance*, rather than just for the development of tools? Three reasons. First, representative datasets allow at least some types of variation to be built into the development of AI tools, whether developed for a national audience or in-house with cross-validation on large, infrastructural datasets.[66] Accordingly, local governance should be easier, because some problems will be weeded out earlier. Second, the variations in large datasets can help illuminate the quirks and complexities of application to varied subsets of data—which can correspondingly flag issues that local governance should take into account.[67] And third, infrastructural datasets can help establish performance baselines against which tools can be measured.

Procedural infrastructure refers to development of processes for governance so that each entity doing tasks need not reinvent the wheel but can instead rely on best practices and guides prepared by experts.[68] Many standards have been developed for determining

---

https://www.fda.gov/drugs/questions-and-answers-fdas-adverse-event-reporting-system-faers/fda-adverse-event-reporting-system-faers-electronic-submissions (last visited Feb. 20, 2022) (describing distributed reporting of adverse events).

[66] S*ee, e.g.*, *Core Values*, NAT'L INST. HEALTH ALL OF US RSCH. PROGRAM, https://allofus.nih.gov/about/core-values (last visited Nov. 20, 2021) ("The All of Us Research Program is guided by a set of core values: . . . Participants reflect the rich diversity of the United States. To develop individualized plans for disease prevention and treatment, researchers need more data about the differences that make each of us unique. Having a diverse group of participants can lead to important breakthroughs. These discoveries may help make health care better for everyone."); AIM-AHEAD, *Data and Research Core*, https://aim-ahead.net/home/leadership/research (last visited Feb. 20, 2022) (describing the goal of "linking and preparing multiple sources and types of research data to form an inclusive basis for AI / ML").

[67] Jeffrey Brown et al., *Data Quality Assessment for Comparative Effectiveness Research in Distributed Data Networks*, 51 MED. CARE S22, S28 (2013).

[68] *See, e.g.*, *Responsible AI Practices*, GOOGLE, https://ai.google/responsibilities/responsible-ai-practices/(last visited Nov. 20, 2021); *Good Machine Learning Practice for Medical Device Development: Guiding Principles*, FDA, https://www.fda.gov/medical-devices/software-

how best to conduct AI studies and report results,[69] including performance metrics;[70] similar tools can ease the process of governance once tools are developed and reported. What steps need to be taken to evaluate an AI tool for deployment—and if only a subset is feasible, what steps are most crucial? Correspondingly, what steps are so essential that insufficient capacity to perform them means that model deployment is too risky to go forward? (This question is a tricky one! It's easy to answer with the assumption that everything is necessary, an assumption which is only realistic from the comfortable seat of a high-resource setting. Knowing which corners can be reasonably cut is a key question for democratizing expertise in low-resource settings, in AI implementation just as much as in care.[71]) Accordingly, process-based infrastructure tools should include some way to evaluate the

---

medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles (last visited Nov. 20, 2021); Reid Blackman, *A Practical Guide to Building Ethical AI*, HARV. BUS. REV. (Oct. 15, 2020), https://hbr.org/2020/10/a-practical-guide-to-building-ethical-ai.

[69] *See, e.g.*, Gary S. Collins, Johannes B. Reitsma, Douglas G. Altman & Karel G.M. Moons, *Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement*, 162 ANNALS INTERN. MED. 55 (2015) (describing the development of the TRIPOD Statement, "a set of recommendations for the reporting of studies developing, validating, or updating a prediction model, whether for diagnostic or prognostic purposes"); Robert F. Wolff et al., *PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies*, 170 ANNALS INTERN. MED. 51, (describing PROBAST, "a tool for assessing the risk of bias and applicability of diagnostic and prognostic prediction model studies"); Gary S. Collins et al., *Protocol for Development of a Reporting Guideline (TRIPOD-AI) and Risk of Bias Tool (PROBAST-AI) for Diagnostic and Prognostic Prediction Model Studies Based on Artificial Intelligence*, BMJ OPEN, July 2021, at 1 (explaining how PROBAST and the TRIPOD statement will be extended to prediction models that utilize machine learning techniques).

[70] *See, e.g.*, Mark P. Sendak et al., *Presenting Machine Learning Model Information to Clinical End Users with Model Facts Labels*, 3 NPJ DIGIT. MED. 1 (2020).

[71] *Cf.* Price, *supra* note 41, at 113–14 ("The most straightforward way for AI algorithms to address cost issues would be to add those issues to the AI's optimization function: that is, when scoring outcomes as desirable or undesirable (for the purposes of care recommendations, at least), the cost of care could be included in the score, rather than just patient health measures. Algorithms would then prioritize not simply outcomes or duplicating the patterns prevalent in High-Resource Hospitals, but also cost-effectiveness.").

costs, not only of deploying an AI system, but of effectively governing that system going forward.

Both infrastructure development and offloading have the potential for significant scaling, whether in a domestic or an international context. For instance, tools to monitor performance, processes for implementing governance structures, and dataset-driven knowledge about places to seek performance glitches should all be at least potentially deployable broadly—ideally to environments globally (with the obvious need for tweaks). Other interventions, like simply purchasing computer systems or deploying roving integration-and-evaluation teams, will be less easy to scale.

CONCLUSION

None of these interventions will be a magic bullet. Low-resource environments will still have few resources, barring massive structural change across the health system (domestic or worldwide). The people charged with adopting new technology, AI or otherwise, will still be overworked, under-resourced, and generally charged with doing too much with too little. Adding the initial and ongoing governance of AI tools onto overfull plates seems unfair. Nevertheless, there is hope! To the extent that AI has the possibility of making that task easier in *other* domains—of allowing more to be done with less for patient care, or resource management, or whatever else—it should be distinctly worth it in the end. Making the responsible adoption and governance of AI tools as easy and straightforward as possible looks likely to pay considerable dividends down the road and should be a focus of policymakers going forward.