2019

Medical AI and Contextual Bias

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

Recommended Citation

Price, W. Nicholson, II. "Medical AI and Contextual Bias." *Harvard Journal of Law & Technology* 33, no. 1 (2019): 65-116.

Harvard Journal of Law & Technology
Volume 33, Number 1 Fall 2019

MEDICAL AI AND CONTEXTUAL BIAS

*W. Nicholson Price II**

TABLE OF CONTENTS

---

* Professor of Law, University of Michigan Law School; Core Partner, Centre for Advanced Studies in Biomedical Innovation Law at the University of Copenhagen; and Co-PI, Project on Precision Medicine, AI, and the Law at the Petrie-Flom Center for Health Law Policy, Biotechnology, and Bioethics at Harvard Law School. For helpful comments and conversations, my thanks to Sam Bagenstos, Nick Bagley, Ana Bracic, Rochelle Dreyfuss, Rebecca Eisenberg, Janet Freilich, Margot Kaminski, Mark Lemley, Kyle Logue, Yuan Luo, Arti Rai, Natalie Ram, Gabriel Rauterberg, Barak Richman, Dan Rubin, David Schwartz, Kayte Spector-Bagdady, Rina Spence, Megan Stevenson, Charlotte Tschider, Jordan Woods, Christopher Yoo, and Kathryn Zeiler. This project benefited from comments at the Stanford/Penn/Northwestern Law/STEM Junior Faculty Forum; the Duke Law and Health Policy Colloquium; South NeBrooklyn; faculty workshops at the University of Michigan, Boston University, Florida State University, the Ohio State University, and the University of Wisconsin; the Health Law Professors Conference; the University of Copenhagen CeBIL Workshop; and the Yale Law Roundtable on AI, Robotics, & Telemedicine. For excellent research assistance, I thank Monica Daegele, Erin Edgerton, and Angela Theodoropoulos. This work was supported by the Cook Fund at the University of Michigan Law School and the Novo Nordisk Foundation (NNF17SA0027784). All errors are my own.

I. INTRODUCTION

Artificial intelligence is entering medical practice. The combination of medical big data and machine learning techniques allows developers to create AI usable in medical contexts — also called "black-box medicine" due to its inherent opacity — that can help improve human health and health care. Only a few years ago, black-box medicine seemed far from real-world use. Today, there are already FDA-approved devices that use AI to diagnose diabetic retinopathy or to flag radiologic images for further study.[1] Hospitals have used AI to help develop care pathways for increasingly specified groups of patients. Future uses are multiplying.

But there is a problem lurking in the development of AI in medicine.[2] A key promise of medical AI is its ability to democratize medical expertise, allowing providers of all sorts to give care that otherwise might be beyond their capacity.[3] Medical AI is typically trained in high-resource settings: academic medical centers or state-

1. *See infra* Section III.C.1.

2. Actually, there are lots of problems, including how to set proper incentives, how to regulate for safety and efficacy, how to use the tort system to encourage providers and hospitals to adopt the best medical AI products, challenges to the doctor-patient relationship, and questions of diminishing human expertise. For an initial overview on those problems and an introduction to medical AI generally, see W. Nicholson Price II, *Black-Box Medicine*, 28 HARV. J.L. & TECH. 419 (2015) [hereinafter Price, *Black-Box Medicine*] (introducing medical AI and canvassing several issues). This Article is focused on a different problem.

3. *See, e.g.*, Victoria J. Mar &. Peter H. Soyer, *Artificial Intelligence for Melanoma Diagnosis: How Can We Deliver on the Promise?*, 29 ANNALS ONCOLOGY 1625, 1625 (2018) ("[A]rtificial intelligence (AI) promises a more standardised level of diagnostic accuracy, such that all people, regardless of where they live or which doctor they see, will be able to access reliable diagnostic assessment.").

of-the-art hospitals or hospital systems.[4] These sites typically have well-trained, experienced practitioners and are most likely to have high-quality data collection systems; training medical AI in these systems makes intuitive sense. Democratizing medical expertise, though, requires deploying that medical AI in low-resource settings like community hospitals, community health centers, practitioners' offices, or rural health centers in less-developed countries.[5] This translation runs into a problem: low-resource contexts have different patient populations and different resources available for treatment than high-resource contexts, and disparities in available data make it hard for AI to account for those differences.

The translational disconnect between high-resource training environments and low-resource deployment environments will likely result in predictable decreases in the quality of algorithmic recommendations for care, limiting the promise of medical AI to actually democratize excellence. To take a simple example: at Memorial Sloan Kettering, one of the best cancer centers in the world, it may well make sense to give a patient a cocktail of powerful chemotherapeutics with potentially fatal side effects, since trained oncology nurses and other specialists are available to monitor problems and intervene if things go wrong. In a community hospital without those safeguards, though, it may be a better call to administer less drastic remedies, avoiding the chance of catastrophic failure. That danger is even more pronounced in even lower-resource settings, such as rural areas of less-developed countries. But medical AI trained only on data from Memorial Sloan Kettering would have no way of taking that resource constraint into account and would provide a poor recommendation to providers in those lower-resource settings.[6]

Contextual bias is an under-addressed kind of bias in the legal AI literature.[7] Rather than the bias arising from *problems* in the underlying data, such as when policing algorithms end up silently replicating

---

4. *See* MICHAEL E. MATHENY ET AL., NAT'L ACAD. OF MED., AI & MACHINE LEARNING IN HEALTH CARE, Section 2.E.2 (forthcoming 2019) (manuscript at 46) (on file with author) (noting that "[i]n the United States, MIT, Stanford and Carnegie Mellon pioneered AI research in the 1960s, and these, and many others, continue to do so today").

5. I focus in this Article on medical AI that is used in health-care settings, not consumer-focused devices, though some of the same issues arise in the latter context as well.

6. It is not impossible to take resource constraints into account in AI decision-making, but, as the rest of this Article demonstrates, doing so is complicated and requires more data than are available from just high-resource settings.

7. For a "whirlwind tour" of AI bias issues, see Karen Hao, *This Is How AI Bias Really Happens and Why It's So Hard to Fix*, MIT TECH. REV. (Feb. 4, 2019), https://www.technologyreview.com/s/612876/this-is-how-ai-bias-really-happensand-why-its-so-hard-to-fix [https://perma.cc/XNG3-XWY7]. In the computer science literature, see, for example, Adarsh Subbaswamy & Suchi Saria, *Counterfactual Normalization: Proactively Addressing Dataset Shift and Improving Reliability Using Causal Mechanisms*, ARXIV, Aug. 9, 2018, https://arxiv.org/pdf/1808.03253v1.pdf [https://perma.cc/CCZ5-W6JP] (revised from print version).

racial bias in underlying arrest patterns and the data they generate[8] or when health algorithms accurately mirror racial or gender biases already present in health care,[9] this bias arises in the process of *translating* algorithms from one context to another. The care provided in high-resource contexts may be superb and untinged by problematic human bias of any kind, and this bias would still arise.[10]

I do not mean to suggest that AI developers are unaware of the challenges of translating AI from one context to another, or the differences between high- and low-resource contexts. The technique of "transfer learning," for instance, focuses on taking insights from one environment and using them in another.[11] And some work, especially nonprofit work in the global health space, focuses intently on developing robust AI especially for deployment in low-resource contexts in less-developed countries.[12] But this Article places the dynamics of cross-context translation into a legal context where, particularly in the United States, incentives actively promote problematic development patterns; it also suggests why the data most useful to address problems of contextual bias are least likely to be available.

This Article analyzes how medical AI can run into problems through an otherwise reasonable process of development and deployment. It proceeds in four Parts. Part II briefly describes the promise of artificial intelligence in medicine, focusing on the idea of democratizing medical expertise. Part III explores the incentives for developing

---

8. *See, e.g.*, Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 GA. L. REV. 109, 119–43 (2018).

9. *See, e.g.*, Ziad Obermeyer et al., *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, 366 SCI. 447, 447 (2019) (finding that a widely used algorithm used to predict patient risk was biased based on race because the algorithm predicted health care *costs*, and less is spent on African-American patients than comparable white patients).

10. High-resource care may be biased. *See*, *e.g.*, David A. Ansell & Edwin K. McDonald, *Bias, Black Lives, and Academic Medicine*, 372 NEW ENG. J. MED. 1087, 1087–89 (2015). But contextual bias can occur independently from any bias in the high-resource care on which the training data are generated, as described in Parts IV and V.

11. *See, e.g.*, Jenna Wiens et al., *A Study in Transfer Learning: Leveraging Data from Multiple Hospitals to Enhance Hospital-Specific Predictions*, 21 J. AM. MED. INFORMATICS ASS'N 699, 699 (2014) (examining the transfer of learning among three hospitals); Dianbo Liu et al., *FADL: Federated-Autonomous Deep Learning for Distributed Electronic Health Record*, ARXIV, Nov. 28, 2018, https://arxiv.org/pdf/1811.11400.pdf [https://perma.cc/DHA6-R3ZV] (suggesting a federated network where generalized insights can be applied in individual contexts); Awni Y. Hannun et al., *Cardiologist-Level Arrhythmia Detection and Classification in Ambulatory Electrocardiograms Using a Deep Neural Network*, 25 NATURE MED. 65, 69 (2019) (noting need to calibrate algorithms to local populations).

12. *See, e.g.*, Valentina Bellemo et al., *Artificial Intelligence Using Deep Learning to Screen for Referable and Vision-Threatening Diabetic Retinopathy in Africa: A Clinical Validation Study*, 1 LANCET DIGITAL HEALTH e35 (2019) (validating in Zambia a model trained on diabetic patients from Singapore); ELEONORE PAUWELS, U.N. UNIV. CTR. FOR POLICY RESEARCH, THE NEW GEOPOLITICS OF CONVERGING RISKS: THE UN AND PREVENTION IN THE ERA OF AI 28 (2019), https://i.unu.edu/media/cpr.unu.edu/attachment/3472/PauwelsAIGeopolitics.pdf [https://perma.cc/2Y7S-7JAP].

medical AI in high-resource medical contexts. It explores how technological factors around data availability are buttressed by legal and economic incentives to focus AI training on high-resource contexts.

Part IV, the heart of the paper, lays out the different types of errors that can arise when medical AI trained in high-resource contexts is deployed in low-resource contexts. It notes problematic differences in patient populations, differences in recommended treatments based on the available resources of the medical environment, and systematic influences on cost.

Part V addresses a question of scope: isn't all medicine contextual? Treatments are developed and doctors are trained in one set of contexts — often high-resource — and then care occurs in a wide array of different contexts. In one sense, medical AI embodies the same type of contextual bias. But medical AI carries the illusory promise of being different because it can theoretically take into account exactly those contextual differences to tailor care and can learn from its own performance. However, this safeguard fails if medical AI lacks data from different contexts to adjust its recommendations. The resulting contextual bias is especially insidious because medical AI is typically opaque, hiding the negative effects that may result.

Part VI discusses potential solutions. It begins with two obvious but flawed solutions. First, could we rely on human doctors "in the loop" to provide common-sense checks on medical AI contextual bias errors? Unfortunately, even assuming that doctors have the knowledge, incentive, and willingness to correct AI errors — assumptions that may not be merited — in many low-resource situations where AI can bring the most benefit, well-trained human providers will simply not be present. Second, could we simply rely on labeling to inform users of its limitations? I argue that labeling is unlikely to solve the problem, since training-based labels are difficult to design, likely to be ignored, and, if followed, would eviscerate much of the promise of democratizing expertise. This Part suggests instead that a better solution requires a combination of public investment in data infrastructure and regulatory mandates of data showing that AI focuses well across different contexts. This combination would ameliorate the problem of contextual translation and help ensure that medical AI actually does provide benefits more broadly, rather than just to those who can already access high-resource care.

Part VI also notes that while the problem of contextual bias needs addressing, policymakers should not be misled by the Nirvana fallacy.[13] Some forms of even imperfect medical AI promise substantial

13. *See* Harold Demsetz, *Information and Efficiency: Another Viewpoint,* 12 J.L. & ECON. 1, 1 (1969) (defining the "nirvana approach" as seeking "to discover discrepancies

benefit to underserved patients, and the field's growth should not be strangled while we await perfection.

Before proceeding, one caveat is in order. Medical AI is on the cusp of entering practice, and a few specific examples of medical AI are already available. But it is early yet, and some of the features key to this discussion are largely in development, especially AI that recommends a particular treatment for a particular patient. The predicates of the argument made here — medical AI, training in high-resource contexts, differences in patient population and resources, and impact of resources on treatment plans — are all already present. I argue that their combination is likely to lead to problems in the process of contextual translation, barring action specifically taken to avoid those problems. But I cannot yet point to instances where such problems have happened, and it is possible that careful developers and regulators will ensure that they never do, even without explicit policy intervention.[14] Nevertheless, the risk needs to be identified and brought to the fore now. Medical AI is developing rapidly and will become increasingly embedded in medical practice; the problem of pervasively biased treatment will be easier to avoid than to fix.

II. THE PROMISE OF BLACK-BOX MEDICINE

Medical AI promises big things. Big data and machine learning can help health-care providers explore new biological relationships and new methods of treatment, automate many low-level tasks that fill providers' days, and raise the general level of care by allowing many types of providers to access expertise through the intermediary of medical AI.[15] Each of these possibilities can bring substantial changes to the world of health care. This Part briefly describes the first two, and then focuses in depth on the third, which the rest of this Article addresses.

A. Advancing Medical Knowledge

Black-box medicine's headline promise is to stretch the boundaries of medical care by uncovering and using new information about

---

between the ideal and the real" and finding "the real is inefficient" without comparing relevant choices between real institutional arrangements).

14. *Cf.* Jorge L. Contreras, *The Anticommons at 20: Concerns for Research Continue*, 361 SCI. 335, 336 (2018) (noting that concerns about innovation stagnation theorized by Michael Heller and Rebecca Eisenberg twenty years earlier had not come to pass in part due to community efforts to avoid them).

15. *See generally* W. Nicholson Price II, *Artificial Intelligence in the Medical System: Four Roles for Potential Transformation*, 18 YALE J. HEALTH POL'Y L. & ETHICS, 21 YALE J.L. & TECH., Special Issue 122 (2019). (describing these three roles and also noting the use of AI in resource allocation).

how humans work and how to care for them. Human biology is tremendously complex and our tools for understanding it are limited; artificial intelligence promises to find and use complex underlying relationships to improve care, discover new treatments, and advance scientific hypotheses even if we don't understand those underlying relationships.[16]

Medical AI is already pushing boundaries. IBM's Watson for Drug Discovery used AI to identify genes likely to be associated with Alzheimer's disease and flagged them as potential targets for new drugs.[17] AI systems can similarly allow things we can't do now; a wearable device could predict the onset of stroke by analyzing a person's gait[18] or AI software could notice the onset of Parkinson's disease by monitoring trembling of a computer mouse and the characteristics of web searches.[19] AI systems could also predict which patients might react better to a particular treatment by noticing subtle groupings among patients that are currently undetectable through standard analysis.[20] All of these possibilities promise to push past the current frontiers of medical knowledge.

B. Automating the Routine

A second, more quotidian promise of medical AI is automating medical drudgery. The problem here is that much of medical practice consists of tasks that aren't really about practicing medicine; instead, they focus on paperwork and routine tasks that often don't do much to help patients and contribute to physician burnout. Providers are deluged with data searching and data entry tasks; one study found that physicians spent almost half of their time on electronic health record work and desk work, and only a quarter of their time seeing patients.[21]

---

16. *See* Price, *Black-Box Medicine*, *supra* note 2, at 434–37.

17. *See, e.g.*, Nadine Bakkar et al., *Artificial Intelligence in Neurodegenerative Disease Research: Use of IBM Watson to Identify Additional RNA-Binding Proteins Altered in Amyotrophic Lateral Sclerosis*, 135 ACTA NEUROPATHOLOGICA 227, 229 (2018) (describing IBM Watson's processing of the scientific literature to identify new genes linked to ALS).

18. *See* Fei Jiang et al., *Artificial Intelligence in Healthcare: Past, Present, and Future*, 2 STROKE & VASCULAR NEUROLOGY 230, 240 (2017).

19. *See* Ryen W. White et al., *Detecting Neurodegenerative Disorders from Web Search Signals*, NATURE: NPJ DIGITAL MED., Apr. 23, 2018, at 1, 1.

20. Jiang et al., *supra* note 18, at 239–40 (noting proposed AI-based stroke treatment models); *id.* at 241 (describing AI-based cancer treatment prediction). AI may also enhance existing medical device usage. Charlotte A. Tschider, *Deus ex Machina: Regulating Cybersecurity and Artificial Intelligence for Patients of the Future*, 5 SAVANNAH L. REV. 177, 189 (2018) (describing the evolution of medical devices from self-executing, device-bound code to AI and distributed infrastructure models).

21. Christine Sinsky et al., *Allocation of Physician Time in Ambulatory Practice: A Time and Motion Study in 4 Specialties*, 165 ANNALS INTERNAL MED. 753, 753 (2016); *see also* Ming Tai-Seale et al., *Electronic Health Record Logs Indicate that Physicians Split Time Evenly Between Seeing Patients and Desktop Medicine*, 36 HEALTH AFF. 655, 655 (2017)

Even in the examination room, physicians spent only about half of their time interacting with patients — and about a third interacting with electronic health records and desk work.[22] The ability of AI to automate at least some of this work could have a profound effect on the provision of health care, potentially improving the doctor-patient relationship[23] and reducing the rate of provider burnout.[24]

Automation of routine tasks, though not as exciting as pushing medical frontiers, could still change medical practice for the better. Some action could happen on the front end: AI could automatically identify and highlight the most relevant medical information from patient medical records to reduce the time spent scrolling through records looking for information.[25] It could provide the most relevant medical literature to doctors based on natural-language processing.[26] And speech-recognition software based on AI could automatically transcribe patient conversations and provider notes and fill out forms afterward.[27] Assuming eventual accuracy on the part of AI, such assistance could also reduce the rate of transcriptional errors and even improve privacy as details are read by machines rather than medical scribes.

(finding that physicians "logged an average of 3.08 hours on office visits and 3.17 hours on desktop medicine each day").

22. *See* Sinsky et al., *supra* note 21, at 753.

23. *Cf.* Maria Alcocer Alkureishi et al., *Impact of Electronic Medical Record Use on the Patient–Doctor Relationship and Communication: A Systematic Review*, 31 J. GEN. INTERNAL MED. 548, 550–57 (2016) (evaluating many studies and finding both positive and negative impacts of EHRs on patient-doctor relationships).

24. In a 2018 survey, "too many bureaucratic tasks" was the most commonly cited contributor to physician burnout. Carol Peckham, *National Physician Burnout and Depression Report 2018*, MEDSCAPE (Jan. 17, 2018), https://www.medscape.com/slideshow/2018-lifestyle-burnout-depression-6009235#13.

25. *See* Kory Kreimeyer et al., *Natural Language Processing Systems for Capturing and Standardizing Unstructured Clinical Information: A Systematic Review*, 73 J. BIOMEDICAL INFORMATICS 14, 14 (2017); *cf.* Theresa A. Koleck et al., *Natural Language Processing of Symptoms Documented in Free-Text Narratives of Electronic Health Records: A Systematic Review*, 26 J. AM. MED. INFORMATICS ASS'N 364, 365 (2019) (explaining that the previously manual process of extracting symptom information from patient records could be automated through natural language processing to reduce time spent by clinical experts).

26. *Cf.* Kreimeyer et al., *supra* note 25, at 15.

27. *Id.*; *see also* Linda Dawson et al., *A Usability Framework for Speech Recognition Technologies in Clinical Handover: A Pre-Implementation Study*, 38 J. MED. SYS., June 2014, at 1, 1. Yet another potential use for medical AI comes in its use to analyze and direct medical resources: assigning scarce resources to patients based on likelihood of aiding them, improving workflow, or even finding ways to optimize medical billing. *Cf.* I. Glenn Cohen et al., *The Legal and Ethical Concerns That Arise from Using Complex Predictive Analytics in Health Care*, 33 HEALTH AFF. 1139, 1140 (2014). These interventions, focused less directly on care encounters, are outside the scope of this work.

C. Democratizing Expertise

Finally, medical AI promises to democratize medical expertise. Today, there are tremendous differences in the quality and level of care patients receive based on the context in which they receive that care.[28] This is reflected in everything from the availability of a specialist (e.g., whether a patient can see a board-certified ophthalmologist or dermatologist rather than relying on a primary care physician for more complex care) to the type of practitioner involved (e.g., physician versus nurse practitioner) to the qualifications of the provider (e.g., top-of-her-class with elite fellowships to less-exalted qualifications).[29] Medical AI promises to reduce this variation by "leveling up" — allowing a much broader swath of providers to provide care at the level of excellent specialists, which is what I mean by democratization of medical expertise. Medical AI is scalable in a way that human expertise simply is not; although gathering data, training algorithms, and validating algorithmic performance are all hard and expensive tasks,[30] duplicating an existing algorithm for use in another setting is much easier and cheaper than training new people for the same tasks.[31] It's not free or easy — information infrastructure still needs to be set up,[32] and the data the algorithm will analyze need to be properly collected and formatted on-site[33] — but an algorithm is easier to copy than an oncologist.

This Section describes how AI can democratize different types of medical expertise. It then considers where AI can bring expertise — a span that ranges from other high-resource settings like mid-level hos-

---

28. *See, e.g.*, John E. Wennberg, *Unwarranted Variations in Healthcare Delivery: Implications for Academic Medical Centres*, 325 BMJ 961, 962–63 (2002); DARTMOUTH ATLAS PROJECT, https://www.dartmouthatlas.org [https://perma.cc/YA3A-5JCY] (cataloging regional differences in care).

29. I recognize that quality of care is not uni-dimensional; for a gastrointestinal problem, a patient would likely rather see a novice nurse practitioner than an experienced neurosurgeon, and with good reason. Nevertheless, there are many situations for which the expertise of a well-trained specialist can improve care.

30. *See* W. Nicholson Price II, *Big Data, Patents, and the Future of Medicine*, 37 CARDOZO L. REV. 1401, 1411 (2016) [hereinafter Price, *Big Data*].

31. Such duplication is not costless, of course, and in some situations doing the transfer right might actually be more expensive than home-growing a solution. *See, e.g.*, JAMES E. TCHENG ET AL., NAT'L ACAD. OF MED., OPTIMIZING STRATEGIES FOR CLINICAL DECISION SUPPORT: SUMMARY OF A MEETING SERIES 28 (2017) ("While the creation of CDS [clinical decision support] content in-house is an expensive and resource-intensive endeavor, sharing CDS content, either with peers or through the licensing of vendor content, is presently perceived to be equally or more expensive; thus this duplication of effort at each site has persisted.").

32. *See* W. Nicholson Price II, *Risk and Resilience in Health Data Infrastructure,* 16 COLO. TECH. L.J. 65, 71 (2017) [hereinafter Price, *Risk and Resilience*].

33. Further down the road, AI can help here too; natural language processing will make it easier to accept unstructured data about patients rather than requiring data to be in a certain format. *See supra* Section II.B.

pitals to very low-resource settings like rural providers in less-developed countries.

1. Diagnostics and Treatment Recommendations

AI can democratize different types of medical expertise. Though medical expertise comes in many flavors, with many interconnections, we can usefully consider two rough classes: diagnostics and treatment recommendations.[34]

a. Diagnostics

Diagnosis is the process of figuring out what's wrong with a patient.[35] If a patient comes in complaining of an exceptionally bad headache, is she suffering from tension headache, a migraine, or a subdural hematoma? One demands over-the-counter painkillers, another a large set of unpredictable medications, and the last an immediate trip to the emergency department to avoid death or severe brain injury. Fans of the television series *House* will readily recognize the recurrent problem of finding out what malady (or combination of maladies) underlies a collection of symptoms. Diagnosis is hard (though Dr. House makes it look easy); it depends on recognizing the right symptoms and using them to identify underlying problems from a vast realm of possibilities. Excellent diagnosticians, when available, are tremendously valuable to medical care — but not everyone can be an excellent diagnostician. Providers may reach incorrect diagnoses because they never acquired the relevant medical knowledge, the knowledge they acquired is outdated, they lack time to conduct the relevant research, they suffer from heuristic biases such as recalling

34. Both of these forms of expertise can also be advanced by AI. *See supra* Section II.A. This section focuses on their democratization. There are more things AI can do in medicine. Prognostics, for instance, are an area of active development; it is good to be able to predict what will happen to a patient, to know how long they might live, and who may become sicker. *Cf.* Ziad Obermeyer & Ezekiel J. Emanuel, *Predicting the Future — Big Data, Machine Learning, and Clinical Medicine*, 375 NEW ENG. J. MED. 1216, 1217 (2016) (discussing machine learning and prognostics generally); Alvin Rajkomar et al., *Scalable and Accurate Deep Learning with Electronic Health Records*, NATURE: NPJ DIGITAL MED., May 8, 2018, at 1, 1 (presenting a model with high accuracy predicting patient mortality, unplanned readmission, and prolonged length of stay). This Article focuses on diagnostics and treatment recommendations as two possibilities for medical AI focused most closely on direct patient care.

35. Diagnosis is not always entirely separable from treatment. In many circumstances, the mere provision of a correct diagnosis can provide relief to patients who know more about what is happening to them and can enable useful self-care. *See, e.g.*, Sumi Sexton & Robert Loflin III, *The Relief of Getting a Diagnosis*, 80 AM. FAM. PHYSICIAN 1223, 1223 (2009); *see also* racheldoesstuff, *A Diagnosis*, YOUTUBE (Nov. 17, 2017), https://www.youtube.com/watch?v=uic_3vlI5BE [https://perma.cc/F4BS-U2LW].

memorable rare diseases rather than common ones,[36] or, most simply, they are unfamiliar with the area of care. Artificial intelligence, based largely on pattern recognition, can help democratize diagnostic expertise, allowing access to this expertise even when an excellent human diagnostician is not available.[37]

EyeDiagnosis's IDx-DR software for diabetic retinopathy is an example of leveling-up that AI can bring to medical diagnosis.[38] Diabetic retinopathy is a condition wherein diabetes causes loss of small blood vessels in the retina; new blood vessels that grow to replace them can cause vision problems.[39] The current standard of care is for patients with diabetes to visit an ophthalmologist yearly to check for signs of retinopathy, so that treatment can begin before the retina worsens.[40] But this requires regularly visiting an ophthalmologist, which is not easy or even possible for many patients.

EyeDiagnosis has developed a system that enables primary care physicians (or other non-specialist practitioners) to use an essentially automated camera to take images of the retina; those images are then analyzed by a machine-learning algorithm trained on a gold-standard dataset of retina images (annotated by expert ophthalmologists).[41] The algorithm returns a diagnosis of more-than-mild diabetic retinopathy, in which case the patient should seek further care, or not, in which case the patient should ideally be retested in a year.[42] IDx-DR is approved by FDA for this level of autonomous diagnosis and performs at a level comparable to ophthalmologists, even when operated by novices.[43] In this scenario, the diagnostic expertise is that possessed by most ophthalmologists (and by their supporting camera technicians). IDx-DR brings that level of diagnostic expertise to primary care physicians without the relevant experience.[44]

---

36. *See, e.g.*, Jill G Klein, *Five Pitfalls in Decisions About Diagnosis and Prescribing*, 330 BMJ 781, 782 (2005).

37. AI diagnosis is not *just* about democratizing expertise. AI could also replace very easy, routine diagnostics (automating drudgery) or point us to disease variants previously unrecognized (advancing medical knowledge). But to the extent that many maladies are diagnosable by expert diagnosticians but not by those with less experience or expertise, AI can help bridge that gap.

38. *IDx-DR*, IDX, https://www.eyediagnosis.net/idx-dr [https://perma.cc/9GZR-MTUB].

39. AM. ACAD. OF OPHTHALMOLOGY, QUALITY OF CARE SECRETARIAT, INFORMATION STATEMENT: SCREENING FOR DIABETIC RETINOPATHY 1, 1 (2014), https://www.aao.org/clinical-statement/screening-diabetic-retinopathy [https://perma.cc/XW88-VZ27].

40. *Id.* at 2.

41. *IDx-DR*, *supra* note 38.

42. *Id.*

43. *See Performance*, IDX, https://www.eyediagnosis.net/performance [https://perma.cc/3Q83-6QLB].

44. In clinical trials for the IDx-DR, the developer specifically sought out technicians who had not been trained on any retinal imaging system — the opposite of an imaging expert. *Id.*

b. Treatment Recommendations

After diagnosis comes treatment. Once providers have determined what ails the patient, they must select from a menu of possibilities to determine the best option for improvement.[45] Consider a well-trained and experienced oncologist; knowing that a patient has a certain type of cancer, she also (hopefully) knows what the best course of treatment is: surgery, radiotherapy, chemotherapy, or some combination — and within each class, which drugs or protocols are likely most effective, given what she knows about the patient. That expertise, like diagnostic expertise, is hard-won and hard to apply; becoming a skilled oncologist takes time, money, and practice. Such expertise is accordingly hard to come by, especially outside specialized cancer centers like Memorial Sloan Kettering or MD Anderson.

Medical AI offers possibilities of democratizing expertise here as well. Indeed, one well-known example of medicine, IBM's Watson for Oncology ("Watson Oncology"), addresses exactly this challenge.[46] I should note that this example is in some ways a problematic one. There appears to be some discrepancy between how IBM says Watson Oncology works and how it actually works in practice, though details are scarce.[47] I will analyze the program as described by IBM, on the basis that this description is at least aspirational; where others offer critiques of this account, I'll note them in footnotes. Whatever the precise contours of Watson Oncology, it is by far the highest-profile example of using AI to democratize medical expertise existing today.

Watson Oncology uses machine-learning-based natural language processing to analyze patient medical records to determine cancer type and then provides recommendations for treatment.[48] The system is an AI/decision-rule hybrid: AI is involved in the initial stages, but the treatment recommendation is based not on any particular machine-learning approach, but instead on what oncologists at Memorial Sloan Kettering would do when faced with a similar patient.[49] IBM aims

45. This picture is naturally somewhat stylized; sometimes, for instance, providers may need to jump straight from symptoms to treatment without knowing the underlying problem, as when treating severe dehydration without first determining the cause.

46. *IBM Watson Health*, IBM, https://www.ibm.com/watson/health/oncology-and-genomics [https://perma.cc/F39K-UF64].

47. Casey Ross & Ike Swetlitz, *IBM's Watson Supercomputer Recommended 'Unsafe and Incorrect' Cancer Treatments, Internal Documents Show*, STAT (July 25, 2018), https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments [https://perma.cc/EVP7-AB72].

48. *IBM Watson Health*, *supra* note 46. According to STAT, Watson Oncology is actually trained on synthetic patient records (that is, records created by doctors to match typical patient patterns) rather than actual patient records. Ross & Swetlitz, *supra* note 47.

49. A. Michael Froomkin et al., *When AIs Outperform Doctors: Confronting the Challenges of a Tort-Induced Over-Reliance on Machine Learning*, 61 ARIZ. L. REV. 33, 43

explicitly to use Watson Oncology to democratize medical expertise (for a price, of course). It licenses Watson Oncology for use at hundreds of hospitals worldwide[50] and has evaluated its performance at hospitals (relatively high-resource hospitals, to be sure) in Thailand,[51] India,[52] and Mexico.[53] Results from initial trials in Mexico conclude that "Oncologists . . . felt [Watson Oncology] would be particularly beneficial in clinics that lack subspecialist expertise."[54] Watson Oncology is thus especially targeted at contexts that lack existing oncologist expertise.

Democratization of the expertise needed to recommend treatments, though, does not always translate to democratization of the expertise needed to actually treat. If medical AI recommends taking tablet A over tablet B for a particular patient, as long as both tablets are available, that recommendation may be easy to follow. But if medical AI recommends a complicated surgery, successful implementation depends on the presence of a skilled surgeon. This problem is explored in more detail below, and is a key challenge for democratizing medical expertise.[55]

2. Contexts of Application

AI has the potential to democratize medical expertise to many medical contexts, ranging from other high-resource contexts like ma-

---

(2019) ("[Watson Oncology] is really a decision-support tool enhanced with preprogrammed suggestions based on what a committee of doctors at Sloan Kettering said they would do when presented with various symptoms and scenarios.").

50. *See Watson Health: Get the Facts*, IBM, https://www.ibm.com/blogs/watson-health/watson-health-get-facts [https://perma.cc/KU84-MEBD]; *see also Manipal Hospitals Adopts Watson for Oncology to Help Physicians Identify Options for Individualized, Evidence-Based Cancer Care Across India*, IBM (Dec. 2, 2015), https://www-03.ibm.com/press/us/en/pressrelease/48189.wss [https://perma.cc/5JJP-KWXD].

51. Suthida Suwanvecho et al., *Concordance Assessment of a Cognitive Computing System in Thailand*, 35 J. CLINICAL ONCOLOGY SUPPLEMENT 6589 (2017), https://meetinglibrary.asco.org/record/150478/abstract [https://perma.cc/Y3AU-4W9T] (abstract presented at the 2017 American Society for Clinical Oncology Annual Meeting). Notably, some disagreement between Thai oncologist recommendations and Watson Oncology's recommendations was attributed to "local oncologist preferences." *Id.*

52. S.P. Somashekhar et al., *Early Experience with IBM Watson for Oncology (WFO) Cognitive Computing System for Lung and Colorectal Cancer Treatment*, 35 J. CLINICAL ONCOLOGY SUPPLEMENT 8527 (2017), http://ascopubs.org/doi/abs/10.1200/JCO.2017.35.15_suppl.8527 [https://perma.cc/DS38-VFW5] (abstract presented at the 2017 American Society for Clinical Oncology Annual Meeting).

53. Catherine Sarre-Lazcano et al., *Cognitive Computing in Oncology: A Qualitative Assessment of IBM Watson for Oncology in Mexico*, 35 J. CLINICAL ONCOLOGY e18166 (2017), https://meetinglibrary.asco.org/record/152386/abstract [https://perma.cc/G7B9-LW6B] (abstract presented at the 2017 American Society for Clinical Oncology Annual Meeting).

54. *Id*. Other appraisals are less complimentary. *See* Ross & Swetlitz, *supra* note 47 (quoting a doctor from Jupiter Florida hospital describing the product as "a piece of s—").

55. *See infra* Section IV.A.

jor hospitals, to medium-resource contexts like community hospitals or community health centers, to low-resource contexts like rural providers in less-developed countries. The higher-resource the destination context, the easier the translation — but the smaller the potential for transforming medical care.

The most straightforward translation is from the absolute top-notch, very high-resource hospitals to other slightly-less-high-resource hospitals — taking the expertise of Memorial Sloan Kettering's cancer center, for instance, and making it accessible to smaller hospitals with less specialized or less experienced oncologists. IBM is already doing this; it's the easiest path, because those settings already have the basic resources and infrastructure in place. The information technology is in place, and oncologists are already on hand who can take AI recommendations and use them to change — ideally to improve — their own practice (or reject them, as the case may be).[56] This is democratization of expertise on a small scale; very valuable, but perhaps not transformative. But this is not the only potential context.

Close to the other end of the spectrum, medical AI could be deployed to genuinely low-resource contexts: small rural hospitals, community health centers or clinics, solo practitioners' offices or small doctors' practices. Where specialists are unavailable — to say nothing of highly skilled, experienced specialists — medical AI could make a tremendous difference in the type and level of care that could be offered. IDx-DR provides exactly this sort of potential: in places without available ophthalmologists, the AI/camera combination allows providers to check patients with diabetes for diabetic retinopathy, availing themselves of ophthalmologic expertise through the AI system.[57] Deploying AI in these contexts demands resources, but almost certainly far fewer resources than improving care by training and employing new medical specialists.

AI could truly transform care in the lowest-resource contexts. In Liberia, as of 2016, there were 298 doctors for a population of 4.5 million, including only fifteen pediatricians and six ophthalmologists.[58] In rural India, a single doctor can be responsible for as many

---

56. *See supra* note 51 (noting that some Thai oncologists rejected Watson Oncology recommendations based on local preferences); *see also* Ross & Swetlitz, *supra* note 47 (noting that some Watson Oncology recommendations, based on Memorial Sloan Kettering practice, differed from national guidelines); *infra* Section VI.A (noting difficulties with using human-in-the-loop safeguards for medical AI generally).

57. *See* discussion *supra* Section II.C.1.

58. Al-Varney Rogers, *Liberian Doctors Threaten Go-Slow over Salary Arrears*, FRONT PAGE AFR. (Nov. 15, 2016), https://frontpageafricaonline.com/health/liberian-doctors-threaten-go-slow-over-salary-arrears [https://perma.cc/MGA3-LLCG] (citing a July 2016 report by the Liberia Medical and Dental Council).

as 30,000 residents in the rural health system.[59] In such low-resource environments, medical AI could provide front-line access for simple diagnostics and treatment recommendations, triaging patients who need to seek further help, as well as the more complex tasks that AI can facilitate in higher-resource contexts. In India, where the doctor shortage extends to ophthalmologists, the Google AI team is already deploying its own AI system to diagnose diabetic retinopathy for patients who cannot access ophthalmologists for recommended yearly screenings.[60] Further work has suggested that smartphones may be suitable for such machine-learning diagnoses, which could lower the barriers to AI-mediated care even further.[61] Overall, while AI has the potential to incrementally improve care in relatively high-resource settings, it could revolutionize care in very low-resource contexts.

* * *

Medical AI can make a difference in many areas of medicine, but one of the most exciting is democratizing medical expertise, especially by bringing diagnostic and treatment recommendation expertise to lower-resource settings where they are otherwise unavailable. The next Part explores the first part of that process: developing algorithms that incorporate medical expertise.

III. WHERE MEDICAL AI IS DEVELOPED — AND WHY

Black-box medical algorithms are predominantly developed in partnership with high-resource medical settings. These are often academic medical systems, but I also include high-resource standalone hospitals. I'll refer to the group collectively as "High-Resource Hospitals." In a typical arrangement, the AI system developer partners with the High-Resource Hospital with an agreement to use the High-Resource Hospital's data to train and develop a new medical algorithm. In the examples above, IBM's Watson Oncology partners with

---

59. Devarsetty Praveen et al., *SMARTHealth India: Development and Field Evaluation of a Mobile Clinical Decision Support System for Cardiovascular Diseases in Rural India*, 2 JMIR MHEALTH & UHEALTH, Oct.–Dec. 2014, at e54, https://mhealth.jmir.org/2014/4/e54/pdf [https://perma.cc/73M5-2RV4].

60. Kamala Thiagarajan, *The AI Program That Can Tell Whether You May Go Blind*, GUARDIAN: THE UPSIDE (Feb. 8, 2019, 1:00 AM), https://www.theguardian.com/world/2019/feb/08/the-ai-program-that-can-tell-whether-you-are-going-blind-algorithm-eye-disease-india-diabetes [https://perma.cc/Y6AY-CW2X]; *see also* Varun Gulshan et al., *Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs*, 316 J. AM. MED. ASS'N 2402, 2402 (2016) (describing an AI diagnostic system).

61. Ramachandran Rajalakshmi et al., *Automated Diabetic Retinopathy Detection in Smartphone-Based Fundus Photography Using Artificial Intelligence*, 32 EYE 1138, 1138 (2018).

Memorial Sloan Kettering[62] and EyeDiagnosis partners with the University of Iowa Health System and the University of Arizona.[63] At the academic/pre-development level, similar patterns manifest: over 500 medical AI studies have been based on the MIMIC dataset, the most-used publicly available health dataset for AI — which includes data only from patients seen at Beth Israel Deaconess Medical Center, a high-resource Harvard-affiliated hospital in Boston.[64]

Developer focus on High-Resource Hospitals does not reflect the delivery of medical care, either nationally or worldwide. Academic medical centers, for instance, make up only a small fraction of all hospitals, and deliver a small (though larger) fraction of care.[65] Many more medical encounters take place in practitioner offices, community health centers, or community hospitals than in High-Resource Hospitals of various flavors.

Algorithm developers partner with High-Resource Hospitals for a varying combination of technical, legal, and business reasons.[66] First and most importantly, High-Resource Hospitals are more likely to have large, high-quality data sets. Second, training algorithms on data from High-Resource Hospitals may facilitate convincing potential clients or insurers that the algorithm is high-quality and worth paying for. Third, training algorithms on High-Resource Hospital data decreases the risk of adverse outcomes from three legal processes: receiving regulatory approval, avoiding tort liability for potential problems once the algorithm is in use, and winning reimbursement from payers.

---

62. *Memorial Sloan Kettering Trains IBM Watson to Help Doctors Make Better Cancer Treatment Choices*, MEMORIAL SLOAN KETTERING CANCER CTR. (Apr. 11, 2014), https://www.mskcc.org/blog/msk-trains-ibm-watson-help-doctors-make-better-treatment-choices [https://perma.cc/HB8U-36VT].

63. *Pipeline*, IDX, https://www.eyediagnosis.net/pipeline [https://perma.cc/L5W4-S2UZ].

64. Rebecca Robbins, *How Patient Records from One Boston Hospital Fueled an Explosion in AI Research in Medicine*, STAT (July 12, 2019), https://www.statnews.com/2019/07/12/boston-hospital-records-fuel-artificial-intelligence-research [https://perma.cc/YGF6-AKWZ].

65. *See, e.g.*, Joanna Bisgaier et al., *Academic Medical Centers and Equity in Specialty Care Access for Children*, 166 ARCHIVES PEDIATRICS & ADOLESCENT MED. 304, 304 (2012) (observing that academic medical centers were "only 6% of the nation's hospitals [yet] provide 28% of all discharges of Medicaid enrollees"); *Academic Medical Centers: Shaping the Future of Healthcare*, UCI HEALTH (June 23, 2016), http://www.ucihealth.org/news/2016/06/academic-medical-centers-future-of-healthcare [https://perma.cc/F89N-BY65] (noting that "[a]cademic medical centers make up 2 to 2.5 percent of all hospitals in the country").

66. I do not claim that all of these reasons apply in each case, and they may be of varying strength; one anonymous industry insider, for instance, described data availability as a "need to have" and potential easing of FDA review as "nice to have."

A. That's Where the Data Are

The first reason for developer focus on High-Resource Hospitals is fundamental: High-Resource Hospitals have more data. Indeed, they may be the only places that actually have high-volume, high-quality data. To take one simple example: health data are hard to use or access unless they are in electronic format. In health care settings, that typically means that the data are recorded in an electronic health record.[67] By now, electronic health records are almost universal; by 2017, essentially all hospitals had adopted electronic health records systems, as had about 90% of office-based practices.[68] However, if a developer wants longitudinal data, or the ability to track results over time, adoption one or two years ago is insufficient — and in 2008, only about 10% of hospitals had EHR systems in place.[69] Which hospitals were those? High-Resource Hospitals.[70]

The mere presence of electronic health records is not enough. For a health-care provider to collect data that can be used to develop medical AI, the provider needs the right infrastructure.[71] This includes developing (1) systems so that providers input the right data, in the right format; (2) databases to ensure that data are collected, categorized, and made available for future use; and (3) quality checks to ensure that the data collected are correct.[72] This infrastructure can be

---

67. Other health data that can be used for training medical AI include pharmacy records or insurance claims data — or non-medical data such as internet search histories or personal health trackers. *See generally* SHARONA HOFFMAN, ELECTRONIC HEALTH RECORDS AND MEDICAL BIG DATA: LAW AND POLICY (2016). These, too, need to be in electronic format. However, electronic health records are the most direct source of data about health-care encounters in particular. *Id.* at 9.

68. Vindell Washington et al., *The HITECH Era and the Path Forward*, 377 NEW ENG. J. MED. 904, 904–05 (2017). Electronic health record adoption received a substantial push in the HITECH Act, which largely mandated their adoption. *See What is the HITECH Act?*, HIPAA J., https://www.hipaajournal.com/what-is-the-hitech-act [https://perma.cc/8W63-5BJ2].

69. Washington et al., *supra* note 68, at 905 (showing data for nonfederal acute care hospitals). Older, paper-based records may be digitized by scanning, but such data migration creates a complicated hybrid system. *See, e.g.*, Diane Dolezel & Jackie Moczygemba, *Implementing EHRs: An Exploratory Study to Examine Current Practices in Migrating Physician Practice*, 12 PERSP. HEALTH INFO. MGMT., Winter 2015, at 1e, 1e, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4700870/pdf/phim0012-0001e.pdf [https://perma.cc/39DX-RP7D]. Eventually, this problem will lessen as even new EHR systems acquire longitudinal data over time — but that will take substantial time, by definition, and if developers are to take advantage of that eventual broadening, policymakers must ensure that the *current* system is not locked in as the default, legal and otherwise.

70. *See, e.g.*, John D. Halamka et al., *Early Experiences with Personal Health Records*, 15 J. AM. MED. INFORMATICS ASS'N 1, 1 (2008) (describing early EHR systems at the Palo Alto Medical Foundation, Beth Israel Deaconess Medical Center, and Boston Children's Hospital).

71. Price, *Big Data*, *supra* note 30, at 1413.

72. *Id.* at 1411–15; *see also* Sharona Hoffman & Andy Podgurski, *The Use and Misuse of Biomedical Data: Is Bigger Really Better?*, 39 AM. J.L. & MED. 497, 515–20 (2013) (describing pitfalls and precautions for biomedical database development).

complex, challenging, and expensive.[73] It demands information technology resources, data scientists or data managers (themselves in short supply), and attention from management.[74] These requirements are *a priori* harder to meet for low-resource health-care providers than for High-Resource Hospitals — they have fewer resources, by definition — skewing the distribution of health-record data to the latter context.

Law also creates hurdles to the collection and use of big health data for research purposes, especially through the Health Insurance Portability and Accountability Act's[75] Privacy Rule (the HIPAA Privacy Rule)[76] and requirements for informed consent.[77] Under the HIPAA Privacy Rule, "covered entities" — including essentially all health-care providers and hospitals[78] — are prohibited from using or disclosing individually identifiable health information without authorization, except for a list of specifically identified purposes.[79] Research is not one of those specifically identified purposes.[80] Providers wishing to use patient data for research purposes must therefore either obtain individual authorization[81] (a closely prescribed and potentially sample-biasing process)[82] or remove identifying information from the sample (which makes linking different data together difficult).[83]

The requirement to obtain informed consent and research approval for use of patient data similarly imposes costs on that use.[84] In-

---

73. *See* Hoffman & Podgurski, *supra* note 72; Price, *Big Data*, *supra* note 30, at 1411–15.

74. HOFFMAN, *supra* note 67, at 152–68; *see also* Hoffman & Podgurski, *supra* note 72, at 527–32; Price, *Big Data*, *supra* note 30, at 1414–15.

75. Health Insurance Portability and Accountability Act of 1996, Pub. L. No. 104-191, 110 Stat. 1936 (codified as amended in scattered sections of 26, 29, and 42 U.S.C.).

76. Standards for Privacy of Individually Identifiable Health Information, 45 C.F.R. §§ 160, 164 (2018). State privacy rules also come into play and further complicate the situation. *See* Barbara J. Evans, *The Ethics of Postmarketing Observational Studies of Drug Safety Under 505(o)(3) of the Food, Drug, and Cosmetic Act*, 38 AM. J.L. & MED. 577, 594 (2012).

77. *See* W. Nicholson Price II, *Drug Approval in a Learning Health System*, 102 MINN. L. REV. 2413, 2446–48 (2018) [hereinafter Price, *Drug Approval*].

78. 45 C.F.R. § 160.103 (2018).

79. *Id.* § 164.502.

80. *See* Rebecca S. Eisenberg & W. Nicholson Price II, *Promoting Healthcare Innovation on the Demand Side*, 4 J.L. & BIOSCIENCES 3, 35–36 (2017) (describing the lack of a research exemption, and noting that the "operations" and "quality improvement" exemptions do not cover research).

81. *See* Kayte Spector-Bagdady & Andrew G. Shuman, *Reg-ent Within the Learning Health System*, 158 OTOLARYNGOLOGY — HEAD & NECK SURGERY 405, 405 (2018).

82. *See, e.g.*, Sharona Hoffman & Andy Podgurski, *Balancing Privacy, Autonomy, and Scientific Needs in Electronic Health Records Research*, 65 SMU L. REV. 85, 114–19 (2012) (describing this bias).

83. *See* Eisenberg & Price, *supra* note 80, at 36–37.

84. *See, e.g.*, Hoffman & Podgurski, *supra* note 82, at 123 (describing empirical evidence on informed consent costs); Mark J. Pletcher et al., *Informed Consent in Randomized Quality Improvement Trials: A Critical Barrier for Learning Health Systems*, 174 JAMA

formed consent requirements are part of a suite of oversight and ethical requirements,[85] typically enforced by institutional review boards that review research.[86] Obtaining informed consent for research use of patient data can be an arduous and costly process.[87]

These legal hurdles tend to concentrate the collection and use of patient health data for research purposes in High-Resource Hospitals. The hurdles may be justified — though that claim has been questioned[88] — and certainly were put in place to serve laudable aims.[89] Nevertheless, the costs imposed by these legal hurdles weigh especially heavily in low-resource contexts, like small hospitals, community health centers, or solo practitioners in rural areas, which have fewer resources to start with. Even de-identifying patient data to comply with the HIPAA Privacy Rule and informed consent requirements may impose its own costs.[90] Those low-resource settings are unlikely to have the resources to spend on addressing legal compliance issues, just as they are unlikely to have spare resources to meet the technological requirements for a useful data infrastructure that can support future research.[91] These resource constraints help drive the concentration of medical big data — and the concomitant ability to develop black-box medical algorithms — in high-resource contexts.

B. Reputational Effects

Reputational effects also push algorithm developers to partner with High-Resource Hospitals. Developers of black-box medical algorithms must persuade potential clients that these algorithms will provide excellent results, whether diagnoses or treatment

---

INTERNAL MED. 668, 668 (2014) (describing how informed consent requirements make large-scale clinical trials and data collection more challenging).

85. *See, e.g.*, Ruth R. Faden et al., *Informed Consent, Comparative Effectiveness, and Learning Health Care*, 370 NEW ENG. J. MED. 766, 768 (2014).

86. *See* Price, *Drug Approval*, *supra* note 77, at 2446 n.208.

87. *See id.* at 2457; Charlotte A. Tschider, *The Consent Myth: Improving Choice for Patients of the Future*, 96 WASH. U. L. REV. 1505, 1507 (2019) (finding HIPAA's informed consent process largely incompatible with health AI).

88. *See, e.g.*, Price, *Drug Approval*, *supra* note 77, at 2449–52. *See generally* CARL E. SCHNEIDER, THE CENSOR'S HAND (2015) (critiquing research oversight by institutional review boards, including the procedural requirements of informed consent).

89. *See, e.g.*, Nancy E. Kass et al., *The Research-Treatment Distinction: A Problematic Approach for Determining Which Activities Should Have Ethical Oversight*, 43 HASTINGS CTR. REP., S4, S5 (Jan.–Feb. 2013).

90. *See, e.g.*, Elizabeth Ford et al., *Extracting Information from the Text of Electronic Medical Records to Improve Case Detection: A Systematic Review*, J. AM. MED. INFORMATICS ASS'N 1007, 1013 (2016).

91. Interview with Researcher, Univ. of Mich. (Feb. 2018) (describing the process of developing a learning health system at a low-resource Michigan health system); Interview with Medical AI Researcher, Vanderbilt Univ. (July 2018).

recommendations.[92] Making that pitch is likely easier when the developer can state that the algorithm is trained on data from presumably expert doctors at High-Resource Hospitals, rather than a more run-of-the-mill medical practice.[93] IBM, for instance, notes that "Watson for Oncology can provide clinicians with evidence-based treatment options based on expert training by Memorial Sloan Kettering (MSK) physicians."[94]

C. Legal Influences

Finally, three legal regimes also suggest the utility of training algorithms with data from practitioners at the top of their profession: FDA approval, tort liability, and insurer reimbursement. In no case does the legal regime require high-resource context training, but in each case risk-averse developers may find that such training decreases the possibility of unexpected problems.

1. FDA Approval

Many forms of medical AI will require FDA approval to be marketed. The FDA regulates "medical devices" under the Federal Food, Drug, and Cosmetics Act and defines "device" quite broadly so that many forms of medical AI will qualify.[95] The FDA has released guidance on regulating Software as a Medical Device ("SaMD") generally[96] and has also released guidance on regulation of clinical decision support software under the 21st Century Cures Act ("Cures Act").[97] Both suggest that FDA will regulate medical AI.[98] And indeed, a few

92. *See* Price, *Black-Box Medicine*, *supra* note 2, at 465–66. Empirical studies on the challenge of provider adoption present an interesting avenue for future work. To my knowledge, none yet exist.

93. Interview with Lawyer for a Major Medical AI Developer (May 2018).

94. *IBM Watson Health*, *supra* note 46; *see supra* Section II.C.1.b.

95. *See* 21 U.S.C. § 321(h) (2012).

96. *See* FDA, SOFTWARE AS A MEDICAL DEVICE (SAMD): CLINICAL EVALUATION — GUIDANCE FOR INDUSTRY AND FOOD AND DRUG ADMINISTRATION STAFF (2017), https://www.fda.gov/regulatory-information/search-fda-guidance-documents/software-medical-device-samd-clinical-evaluation-guidance-industry-and-food-and-drug-administration [https://perma.cc/VCH3-CJUG].

97 . *See* FDA, CLINICAL DECISION SUPPORT SOFTWARE: DRAFT GUIDANCE FOR INDUSTRY AND FOOD AND DRUG ADMINISTRATION STAFF (2019), https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software [https://perma.cc/GP6D-CXEY]; 21st Century Cures Act, Pub. L. No. 114-255, 130 Stat. 1033, 1130–33 (2016) (amending 21 U.S.C. § 360j (2012)).

98. *See* W. Nicholson Price II, *Regulating Black-Box Medicine*, 116 MICH. L. REV. 421, 439–42 (2017). Under the 21st Century Cures Act, many forms of clinical decision support software are excluded from the definition of medical devices. 21st Century Cures Act, Pub. L. No 114-255, 130 Stat. at 1130–33 (2016). That is, software that merely informs doctors of treatment options or that makes recommendations may not be regulated as a medical device — but only if the software provides an explanation of its recommendations and al-

devices using medical AI have already been allowed on the market by FDA, including one product that uses machine learning to estimate cardiac volume;[99] one product that identifies radiological images of breast abnormalities for further review;[100] and one product, the IDx-DR mentioned above, that analyzes retinal images to autonomously diagnose diabetic retinopathy.[101] Many more medical AI devices are likely to come through FDA's approval or clearance pathways.[102]

Training medical AI with high-quality data from high-resource contexts may ease the path to FDA approval. The FDA does not yet have any explicit standards or rules about the quality or source of data used in training medical AI.[103] In a sense, the agency is learning as it goes along in this area of very new technology.[104] Nevertheless, all

---

lows the provider "to independently review the basis for such recommendations . . . so that it is not the intent that such [providers] rely primarily on any of such recommendations to make a clinical diagnosis or treatment decision regarding an individual patient." 21st Century Cures Act, Pub. L. No. 114-255, 130 Stat. at 1131 (2016). Medical AI — at least the type of medical diagnosis and treatment AI discussed here — will rarely meet this description because it will typically be unable to provide reasoning sufficient for independent review. *See* Price, *supra*, at 440. This will not be the case for all medical AI; some systems at least make claims to explain the reasoning behind their decisions, though this is a contested area and there may be tradeoffs between algorithmic performance and explainability requirements. Other types of AI are not medical devices because they do not inform or direct the care of individual patients; AI used in billing, or to provide medical literature references to doctors, would be excluded. *See, e.g.*, 21st Century Cures Act, Pub. L. No. 114-255, 130 Stat. at 1131 (2016). Nevertheless, as described below, even when FDA approval is not required, such as for devices with sufficient explainability to sit within § 3060's exemption, FDA approval brings other benefits. *See infra* Section III.C.3.

99. Letter from FDA to Arterys, Inc. (Jan. 5, 2017), https://www.accessdata.fda.gov/cdrh_docs/pdf16/K163253.pdf [https://perma.cc/H6M4-6QQH] (determining that the Arterys Cardio DL system is substantially equivalent to legally marketed predicate devices).

100 . Letter from FDA to Quantitative Insights, Inc. (July 19, 2017), https://www.accessdata.fda.gov/cdrh_docs/pdf17/DEN170022.pdf [https://perma.cc/UX5U-T2X5] (classifying QuantX as a Class II medical device under the de novo pathway).

101. Press Release, FDA, FDA Permits Marketing of Artificial Intelligence-Based Device to Detect Certain Diabetes-Related Eye Problems (Apr. 11, 2018), https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm604357.htm [https://perma.cc/5K8X-JMBU].

102. The extent to which FDA will evaluate AI medical devices as components of a larger system or holistically is also unclear. *See* Tschider, *supra* note 20, at 207 (describing limitations of classifying health-care AI systems as components when they may be used for differing diagnostic purposes).

103. Interview with Senior FDA Official (June 2018).

104. Interviews with Regulatory Affairs Personnel at Medical AI Developers (May and June 2018); *see* FDA, *Challenge Questions*, https://www.fda.gov/downloads/MedicalDevices/DigitalHealth/DigitalHealthPreCertProgram/UCM605686.pdf [https://perma.cc/SA26-5BWV]; FDA, *Digital Health Software Precertification (Pre-Cert) Program*, https://www.fda.gov/medicaldevices/digitalhealth/digitalhealthprecertprogram/default.htm [https://perma.cc/AB4S-5VST].

things being equal, training algorithms on the highest quality data available removes at least one cause for risk and uncertainty.[105]

2. Tort Liability

The tort regime also creates incentives for medical AI development.[106] The tort landscape for medical AI is largely theoretical, as the technology is just entering practice. Even the limited scholarly work on the topic has focused more on medical malpractice liability of providers using medical AI, rather than on liability for the developers of the AI products.[107] If a patient is injured through the failure of medical AI, however, liability could be found for developers on theories of negligence or strict liability, alleging design or manufacturing defect.[108] Successful tort claims against a medical AI developer face numerous challenges, including identifying the injury, demonstrating causation within the context of opaque recommendations, overcoming the reluctance of courts to find liability for software generally,[109] and avoiding the doctrine of the learned intermediary.[110] But there is still a risk of liability.

High-quality data from high-resource settings could serve as potential insulation from possible tort liability. It is not clear whether design defect or manufacturing defect would more accurately encapsulate a failure to train an algorithm correctly. But training algorithms on data from excellent doctors in high-resource settings creates an easier case that the developer exercised due care in the development process. To the extent that developing algorithms based on high-

105. All things may not be equal. For instance, if a lower-resource setting also provides lower-quality care (not a certainty), then an AI system could more easily show a greater effect in the lower-resource setting.

106. Tort law shapes other aspects of the development of black-box medicine. *See, e.g.*, W. Nicholson Price II, *Medical Malpractice and Black-Box Medicine, in* BIG DATA, HEALTH LAW, AND BIOETHICS 295, 295–96 (I. Glenn Cohen et al. eds., 2018) [hereinafter Price, *Medical Malpractice*]; Froomkin et al., *supra* note 49, at 35. Among other aspects, to the extent that tort law relies on demonstrating causation, and to the extent that demonstrating causation is hampered by essentially opaque machine-learning algorithms, we might expect that tort law pushes medical AI *away* from explainability and reliance on explicit factors, and *toward* models that are harder to interrogate — and therefore harder landscapes to pinpoint causation.

107. *See, e.g.*, Froomkin et al., *supra* note 49, at 55; Price, *Medical Malpractice*, *supra* note 106, at 295; Nicolas P. Terry & Lindsay F. Wiley, *Liability for Mobile Health and Wearable Technologies*, 25 ANNALS HEALTH L. 62, 81 (2016).

108. *Cf.* Daniel A. Crane, et al., *A Survey of Legal Issues Arising from the Development of Autonomous and Connected Vehicles*, 23 MICH. TELECOMM. & TECH. L. REV. 191, 261 (2017) (discussing potential products liability claims against software providers in the event of an autonomous vehicle crash).

109. *Cf.* Frances E. Zollers et al., *No More Soft Landings for Software: Liability for Defects in an Industry that Has Come of Age*, 21 SANTA CLARA COMPUTER & HIGH TECH. L.J. 745, 766 (2004).

110. *See* Timothy S. Hall, *Reimagining the Learned Intermediary Rule for the New Pharmaceutical Marketplace*, 35 SETON HALL L. REV. 193, 195 (2004).

resource data becomes standard practice, failure to do so — if injuries result — could potentially be viewed as a failure to exercise due care in development.[111]

Although tort law seems unlikely to be a principal source of incentives to develop algorithms on high-resource data, it likely reinforces existing pressures in that direction.

3. Insurer Reimbursement

Finally, convincing insurers that these new technologies should be reimbursed could easily follow a similar pattern.[112] Training algorithms on the data from highly skilled doctors is at least a proxy signal of quality. All things being equal, it suggests that the algorithms are likely to be higher quality, and therefore worthier of reimbursement. While the source of the training data is unlikely completely to replace other quality metrics (e.g., patient outcomes, decreased costs), linking an algorithm to the reputation of its training data may supplement those metrics on the path to reimbursement by payers.[113]

* * *

These issues are not totally distinct. The link between FDA approval and reimbursement by the Centers for Medicare and Medicaid Services ("CMS"), the largest payer in the United States, is substantial for drugs, though less so for medical devices in general.[114] However, the link is prominent in the development of new diagnostic tests, including those reliant on big data.[115] Foundation Medicine pursued FDA approval of its Foundation One biomarker test simultaneously with CMS review for payment in a prominent example of joint FDA approval/CMS coverage determination.[116] CMS suggested that the

111. *See* Froomkin et al.*, supra* note 49, at 36–37, 49 (arguing that as medical AI improves, it will become standard of care to use it and a failure to do so might result in liability).

112. *See, e.g.*, Rachel E. Sachs, *Prizing Insurance: Prescription Drug Insurance as Innovation Incentive*, 30 HARV. J.L. & TECH. 153, 178–79 (2016) (describing insurance reimbursement process); *cf.* Rebecca S. Eisenberg & Harold Varmus, *Insurance for Broad Genomic Tests in Oncology*, 358 SCI. 1133, 1133 (2017) (describing the practice of insurers declining to cover new next-generation sequencing techniques because of lack of clinical evidence).

113. *See* Price, *Black-Box Medicine*, *supra* note 2, at 462–64 (discussing reimbursement challenges for black-box medicine).

114. *See* Rachel E. Sachs, *Delinking Reimbursement*, 102 MINN. L. REV. 2307, 2309, 2311, 2342 (2018).

115. *See* Rachel E. Sachs, *Innovation Law and Policy: Preserving the Future of Personalized Medicine*, 49 U.C. DAVIS L. REV. 1881, 1885 (2016).

116. *See* Eisenberg & Varmus, *supra* note 112, at 1134; Press Release, FDA, FDA Announces Approval, CMS Proposes Coverage of First Breakthrough-Designated Test to Detect Extensive Number of Cancer Biomarkers (Nov. 30, 2017), https://www.fda.gov/

level of reimbursement available for such next-generation diagnostics would be linked to the type of FDA review sought.[117] Devices taken through FDA's more stringent Class III approval pathway would receive full reimbursement, while devices that were only "cleared" through FDA's less-stringent 510(k) clearance pathway would receive lower reimbursement rates.[118] This stringent difference did not make it into final policy.[119]

FDA approval is also linked to tort liability. State tort lawsuits alleging negligent design of medical devices are preempted if the device was approved by FDA through the premarket approval process (but not if the device was cleared under the 510(k) pathway).[120] Thus, to the extent that the FDA approval pathway is smoothed by the demonstration of high-quality data reliance, that also has indirect impacts on the ease of obtaining reimbursement for the product and on reducing tort liability.

D. Caveats

The reliance on medical data from High-Resource Hospitals is both over-determined and under-determined. In many situations, firms will rely on High-Resource Hospital data for multiple reasons, any combination of which may be independently sufficient. By contrast, in two notable exceptions, algorithms may be trained on data from different sources.

First, some types of medical data are so highly standardized that the particular setting from which they arise does not matter very much. For instance, because ophthalmological examinations are highly standardized, the field has developed gold-standards for images and

---

news-events/press-announcements/fda-announces-approval-cms-proposes-coverage-first-breakthrough-designated-test-detect-extensive [https://perma.cc/ZBL2-A7SM].

117. *See* CTRS. MEDICARE & MEDICAID SERVS., PROPOSED DECISION MEMO FOR NEXT GENERATION SEQUENCING (NGS) FOR MEDICARE BENEFICIARIES WITH ADVANCED CANCER (CAG-00450N) (Nov. 30, 2017), https://www.cms.gov/medicare-coverage-database/details/nca-proposed-decision-memo.aspx?NCAId=290 [https://perma.cc/LP8R-PVFS]; Rebecca S. Eisenberg, *Opting into Device Regulation in the Face of Uncertain Patentability*, MARQ. INTELL. PROP. L. REV. (forthcoming) (draft on file with author) [hereinafter Eisenberg, *Device Regulation*].

118. *See* Eisenberg, *Device Regulation*, *supra* note 117, at 20–21; Price, *Regulating Black-Box Medicine*, *supra* note 98, at 438 (describing the approval and clearance pathways).

119. *See* CTRS. MEDICARE & MEDICAID SERVS., DECISION MEMO FOR NEXT GENERATION SEQUENCING (NGS) FOR MEDICARE BENEFICIARIES WITH ADVANCED CANCER (CAG-00450N) (Mar. 16, 2018), https://www.cms.gov/medicare-coverage-database/details/nca-decision-memo.aspx?NCAId=290 [https://perma.cc/WB86-AXWW].

120. *See* Riegel v. Medtronic, Inc., 552 U.S. 312, 323, 326–27 (2008) (finding preemption for devices that underwent premarket approval); Medtronic, Inc. v. Lohr, 518 U.S. 470, 496–98 (1996) (finding no preemption for devices cleared through 510(k)).

diagnoses for use in training ophthalmologists, and these data can similarly be used to train medical AI.[121]

Second, the picture changes drastically in an international context. The particular patterns of health data acquisition, storage, and use — especially legal, but also technical — are artifacts of the peculiar American health system. For instance, when the federal government mandated that hospitals and other providers adopt EHRs, it left the choice of system to the market.[122] As a result, different providers and hospitals use different EHR systems, which makes it hard to assemble data from different medical environments.[123] In China and some other international contexts, on the other hand, the central government mandates specific EHRs and their adoption, and data collection is thus more widespread and uniform across different medical contexts, though those contexts may bring other challenges.[124] Algorithms trained on foreign data, then, may be less likely to rely on data from High-Resource Hospitals.[125]

On the other hand, reliance on U.S. data from High-Resource Hospitals may be over-determined in some cases. The concentration of those data at High-Resource Hospitals may be a sufficient condition to drive company reliance on High-Resource Hospital data, because without those data, there is nothing on which to train the algorithms. However, the other factors mentioned — reputation and avoidance of legal risks — might themselves be independently sufficient were the data to become available from more contexts.[126] This matters because if use of data from High-Resource Hospitals is indeed over-determined for a subset of algorithm developers, then fixing merely one problem — availability of data or a broader path to FDA

---

121 . *See* WISCONSIN FUNDUS PHOTOGRAPH READING CENTER, https://www.ophth.wisc.edu/research/fprc [https://perma.cc/S8W3-RF88].

122. *See* HOFFMAN, *supra* note 67, at 1–2.

123. *See* Price, *Risk and Resilience*, *supra* note 32, at 70.

124. *See, e.g.*, Luxia Zhang et al., *Big Data and Medical Research in China*, 360 BMJ, Feb. 5, 2018, at 1–2. Of course, there may be other concerns with centrally-mandated EHR; for instance, although the United Kingdom developed a plan to centralize health data for biomedical research, that process was halted amid intense controversy. *See* Siobhan Fenton, *Controversial Mega-database of Medical Records Scrapped Over Privacy Concerns*, INDEPENDENT (July 6, 2016), https://www.independent.co.uk/life-style/health-and-families/health-news/nhs-database-medical-records-care-data-scrapped-privacy-concerns-chilcot-report-a7123126.html [https://perma.cc/7FAR-33V9].

125. *See, e.g.*, Kasumi Widner & Sunny Virmani, *New Milestones in Helping Prevent Eye Disease with Verily*, GOOGLE (Feb. 25, 2019), https://www.blog.google/technology/health/new-milestones-helping-prevent-eye-disease-verily [https://perma.cc/5BCK-7XU9] (describing Google working with data from a chain of Indian eye hospitals). *But see* Corinne Abrams, *Google's Effort to Prevent Blindness Shows AI Challenges; Company's AI Can Detect a Condition That Causes Blindness in Diabetes Patients, But in Rural India It Doesn't Always Work*, WALL ST. J. (Jan. 26, 2019), https://www.wsj.com/articles/googles-effort-to-prevent-blindness-hits-roadblock-11548504004 [https://perma.cc/V3GU-WZDJ] (describing challenges using Google's algorithms in field clinics).

126. *See supra* Sections III.B–C.

approval, for instance — will not actually result in medical AI being trained on data from different contexts. Instead, solving just part of the problem may result in developers continuing to train principally on data from High-Resource Hospitals.

But what is the impact of training medical AI on data from High-Resource Hospitals? The reasons listed above all seem reasonable justifications for training algorithms on those data. What's the problem? The next Part explores the challenges that arise in translating algorithms trained on data from High-Resource Hospitals into less-elite health-care settings.

IV. TRANSLATIONAL CHALLENGES

The promise of black-box medicine — at least, the promise that is the focus of this work — is that it can help democratize medical expertise, raising the level of run-of-the-mill practitioners and improving medical care. Achieving those goals requires that algorithms actually be deployed in those run-of-the-mill settings. How will algorithms trained on data from High-Resource Hospitals fare outside those settings? This Part argues that problems are likely to arise in translation in two principal areas: quality of care and cost of care.

One preliminary note: other technical challenges arise in the process of translation itself, which are not the focus of this Part. For example, it can be difficult to ensure that algorithms trained on data from one electronic health record system can accurately analyze data within the context of another electronic health record system.[127] One study found that an algorithm developed in Washington state to identify lung cancer patients who would likely respond to targeted therapy performed well in Washington, but quite poorly in Kentucky, based in part on different language used in electronic health records.[128] Such technical issues may be particularly likely when the deployment context is relatively under-resourced; community health centers may be ill-equipped to deal with EHR incompatibility issues, for instance. Nevertheless, even if these more straightforward technical hurdles are overcome, less visible challenges of decreased patient care quality and increased cost may remain.[129]

---

127. *See* Price, *Risk and Resilience*, *supra* note 32, at 71.

128. *See* Bernardo Haddock Lobo Goulart et al., *Validity of Natural Language Processing for Ascertainment of* EGFR *and* ALK *Test Results in SEER Cases of Stage IV Non-Small-Cell Lung Cancer*, JCO CLINICAL CANCER INFORMATICS 1, 7 (2019).

129. In fact, overcoming technical challenges may give a false sense of security, thus obscuring the other problems that arise in translation.

A. Treatment Quality

The most significant problem with applying algorithms developed in High-Resource Hospitals in lower-resource settings is that those algorithms are likely to make diagnoses and treatment recommendations that are systematically suboptimal in those lower-resource settings. These can arise in at least two different ways: differences in diagnoses and treatment recommendations based on systematically different patient populations, and differences in recommended treatments based on treatment rankings whose order shifts with available medical resources. This distinction is a bit abstract, so the next sections will illustrate with examples from current care and then describe how these examples could become embedded in relatively opaque black-box algorithms and negatively impact the quality of care.

1. Patient Population Differences

Algorithmic translation can cause problems in care when there are systematic differences between the patient populations used to train the algorithm and those where the algorithm is later used. If the patients in the training data — the High-Resource Hospital — differ systematically from the patients in low-resource settings where the algorithm is deployed as part of an AI system, the system won't do a good job dealing with those patients.

Patient population differences, including ancestral origin/genetic variation, socioeconomic status,[130] or general health status,[131] can influence recommendations for treatment in many ways. These differences can influence both proper diagnosis and proper treatment. Consider two examples, one on the prediction side and one on the treatment side.

A prominent example on the diagnosis/prophylactic front comes from hypertrophic cardiomyopathy.[132] In this condition, the wall of the heart thickens abnormally, potentially leading to abnormal rhythms and even sudden death; it is particularly dangerous for young athletes who can be asymptomatic and then die during strenuous exer-

---

130. *See, e.g.*, Dhruv Khullar, *A.I. Could Worsen Health Disparities*, N.Y. TIMES (Jan. 31 2019), https://www.nytimes.com/2019/01/31/opinion/ai-bias-healthcare.html [https://perma.cc/6QXJ-SQZS] (noting AI may entrench current inequities in health: "If, for example, poorer patients do worse after organ transplantation or after receiving chemotherapy for end-stage cancer, machine-learning algorithms may conclude such patients are less likely to benefit from further treatment — and recommend against it.").

131. For instance, of all patients with a particular disease, those with the most severe symptoms might disproportionately choose to go to High-Resource Hospitals, which would skew the data from which an algorithm could learn.

132. *See* Arjun K. Manrai et al., *Genetic Misdiagnoses and the Potential for Health Disparities*, 375 NEW ENG. J. MED. 655, 655 (2016).

tion.[133] Genetic tests are used to identify the disorder — but a 2016 study found that black Americans were underrepresented in the initial data, and as a result many black patients were told they were at risk based on a mutation that does not in fact predict a higher risk for them.[134] Medical AI could easily use this type of genetic information, especially once genetic sequencing becomes more common, to drive preliminary diagnoses and recommendations for further screening — and unless that medical AI was trained on more representative data, it would provide poor results for underrepresented groups.[135]

On the treatment side, consider clopidogrel, marketed in the United States as Plavix for preventing heart attacks and stroke.[136] The gene CYP2C19 is related to the efficacy of clopidogrel. One particular CYP2C19 allele reduces how well clopidogrel works — but only appears in those of European ancestry 10–20% of the time, as opposed to those of Pacific Islander descent (40–77%) or East Asian descent (23–45%).[137] Unfortunately, 95% of participants in the initial clinical studies were of European descent — leading to the conclusion that the drug is much more broadly effective than it actually is.[138] The state of Hawaii sued Bristol-Myers Squibb and Sanofi-Aventis, the makers of Plavix, for false, unfair, and deceptive marketing based on the failure to disclose that treatment efficacy differed based on patient populations.[139]

Differences between patients are well-recognized. Those differences drive the development of precision medicine: the idea that medical treatment should take into account the characteristics of each individual patient.[140] For drugs, that means getting the right drug to the right patient, at the right time.[141] For medicine to take those differences into account, though, especially AI, medical technologies need to be developed in environments that actually show representative variation. If, as posited here, certain types of variation are not reflected in development environments, those potential benefits are lost. That is to say, if High-Resource Hospitals have notably different pa-

133. *See id.* at 656.

134. *See id.* at 659–60.

135. *See* Lucia A. Hindorff, et al., *Prioritizing Diversity in Human Genomics Research*, 19 NATURE REVS. GENETICS 175, 175 (2018) ("Increased attention to diversity will increase the accuracy, utility and acceptability of using genomic information for clinical care.").

136. *See* Alan H.B. Wu et al., *The Hawaii Clopidogrel Lawsuit: The Possible Effect on Clinical Laboratory Testing*, 12 PERSONALIZED MED. 179, 179 (2015).

137. *See id.* at 180.

138. *See* Rachel Huddart et al., *Are Randomized Controlled Trials Necessary to Establish the Value of Implementing Pharmacogenomics in the Clinic?*, 106 CLINICAL PHARMACOLOGY & THERAPEUTICS 284, 285 (2019).

139. *See* Vence L. Bonham et al., *Will Precision Medicine Move Us Beyond Race?*, 374 NEW ENG. J. MED. 2003, 2004 (2016).

140. *See, e.g.*, Margaret A. Hamburg & Francis S. Collins, *The Path to Personalized Medicine*, 363 NEW. ENG. J. MED. 301, 301 (2010).

141. *See, e.g.*, *id.*

tient populations, then we should expect that medical AI trained on data from those populations and then deployed in different settings should encounter problems based on those patient population differences.

And in fact, at least some High-Resource Hospitals show substantially skewed patient populations. Roosa Tikkanen and her colleagues found that white patients were three times as likely as black patients to be admitted to academic medical centers in New York City in 2009, controlling for insurance status, age, and gender.[142] Even after the Affordable Care Act went into effect, the ratio was still more than two to one.[143] Thus, the data collected in those High-Resource Hospitals would substantially underrepresent black patients. This pattern is not universal among High-Resource Hospitals; Boston academic medical centers did not show the same underrepresentation as in New York.[144] Other studies have found similar results in terms of minority representation in academic medical centers.[145]

Genomic data provide a useful example of the underrepresentation of diverse populations in big health data. To be sure, genomic data differ from electronic health records — EHRs are records of patient care that *may* be used for research, while genome sequences are frequently generated specifically for research purposes. Nevertheless, genomic data are key elements of big health data, especially those that push boundaries to increase the precision of medicine, and are important for medical AI. And genomic sequence databases are tremendously non-representative. In 2009, 96% of participants in genome-wide association studies were of European descent.[146] More recently, the diversity of those databases has increased — but almost exclusively because of increased genomic sequencing efforts by Asian cen-

142. Roosa Sofia Tikkanen et al., *Hospital Payer and Racial/Ethnic Mix at Private Academic Medical Centers in Boston and New York City*, 47 INT'L J. HEALTH SERVS. 460, 464 (2017).

143. *Id.*

144. *See id.* High-Resource Hospitals whose patient populations are more generally representative will tend to produce algorithms with fewer translational problems — at least on the dimension of patient population differences.

145. *See, e.g.*, Neil S. Calman et al., *Separate and Unequal Care in New York City*, 9 J. HEALTH CARE L. & POL'Y 105, 107 (2006); Romana Hasnain-Wynia et al., *Disparities in Health Care Are Driven by Where Minority Patients Seek Care: Examination of the Hospital Quality Alliance Measures*, 167 ARCHIVES INTERNAL MED. 1233, 1237–38 (2007); Ashish K. Jha et al., *The Characteristics and Performance of Hospitals that Care for Elderly Hispanic Americans*, 27 HEALTH AFF. 528, 533–35 (2008).

146. *See* Anna C. Need & David B. Goldstein, *Next Generation Disparities in Human Genomics: Concerns and Remedies*, 25 TRENDS GENETICS 489, 490 (2009).

ters.[147] Patients of Latin-American and African descent remain rare in these databases.[148]

More generally, researchers are increasingly realizing that the data used to train medical AI are not representative of the populations in which those AI may be used. Voice recognition AI often performs poorly when analyzing accented voices.[149] And the databases of skin lesions used to train dermatological AI to recognize melanomas are largely missing images from patients with darker skin.[150]

Overall, differences in patient populations may limit the generalizability of medical AI. Where AI is trained on data including only a limited and non-representative set of patients, it will work less well for patients outside that set. This problem has a familiar flavor; other forms of medical intervention, such as drugs, are also developed in particular patient contexts, and generalizability is an ongoing challenge.[151] Some instances will matter more than others. It might be the case that retinal images look pretty much the same from any population of patients in the world, so that contextual bias in retinal-image-based diagnoses is a minimal concern — but skin images look very different depending on whether the skin is fair or not. The problem will vary, unsurprisingly, depending on the context.

But a second set of translational challenges also exists, more dependent on the pattern of medical AI's development: challenges that arise from the differences in resource capacity between High-Resource Hospitals and other settings where black-box medical algorithms will be deployed.

---

147. *See* Alice B. Popejoy & Stephanie M. Fullerton, *Genomics Is Failing on Diversity*, 538 NATURE 161, 163 (2016).

148. *See id.* at 162. All of Us, the NIH-led initiative to obtain health records and genomic sequences for more than a million Americans, is a notable effort to reflect patient diversity and is discussed in detail below. *See infra* Section VI.C.

149. *See* Sonia Paul, *Voice is the Next Big Platform, Unless You Have an Accent*, WIRED, (March 20, 2017, 12:00 AM), https://www.wired.com/2017/03/voice-is-the-next-big-platform-unless-you-have-an-accent [https://perma.cc/9TXN-RBXH] ("AI can only recognize what it's been trained to hear. Its flexibility depends on the diversity of the accents to which it's been introduced."); *see also* Will Knight, *AI Programs are Learning to Exclude Some African-American Voices*, MIT TECH. REV. (Aug. 16, 2017), https://www.technologyreview.com/s/608619/ai-programs-are-learning-to-exclude-some-african-american-voices [https://perma.cc/M53G-PLCF] (noting similar problems for both voice and text recognition).

150. *See* Adewole S. Adamson & Avery Smith, *Machine Learning and Health Care Disparities in Dermatology*, 154 JAMA DERMATOLOGY 1247, 1247 (2018).

151. For a small sampling of the extensive literature on pharmacogenomics, a field based on this reality, see, for example, Mary V. Relling & William E. Evans, *Pharmacogenomics in the Clinic*, 526 NATURE 343 (2015), and Simona Volpi et al., *Research Directions in the Clinical Implementation of Pharmacogenomics: An Overview of US Programs and Projects*, 103 CLINICAL PHARMACOLOGY & THERAPEUTICS 778 (2018).

2. Resource Capacity Differences

A second major source of contextual bias occurs because differences in resources change which option is better — that is, which treatment option an algorithm *should* recommend. This problem arises with treatment recommendations in a way that it doesn't with diagnostic expertise. When medical AI gives a particular diagnosis, it simply provides that information, which is either accurate or not, whatever the situation. Whether a patient actually has a subdural hematoma does not depend on whether the patient presents at Mass General or in a rural Nigerian clinic. Whether AI gets that diagnosis *right* may change based on the contexts of training and application.[152] But the right diagnosis — the ground truth — does not change. Treatment recommendations are different, because they need to be put into practice — the provider and patient must actually undertake the treatment, and that process differs in different contexts.[153]

Given a menu of treatment options for a given ailment, the "best" or most appropriate option in a high-resource setting may well be quite different than the best option in a low-resource setting. The most straightforward version of this dichotomy is when recommended treatment options are simply unavailable. In lower-resource settings, patients and providers may not have access to machines necessary for certain types of care (e.g., directed radiotherapy or laparoscopic surgery) or certain drugs, either because they are too expensive or because they require specific conditions for transport and storage. In the very lowest-resource settings, drugs that require refrigeration may not be available if reliable cold-chain transport is absent. But these types of context disparities, while troubling, are at least easy to see; if AI says to do X, but X isn't possible, that's an easy recommendation to ignore. Algorithms with lots of those unhelpful recommendations won't improve care very much in lower-resource contexts, but at least those algorithms won't actively compromise care.

More problematically, some treatments work very well when performed by experts with excellent support structures, but poorly if performed without those resources. Algorithms trained in high-resource settings may learn to prefer treatments that are only the best treatments when performed in those same high-resource settings. When those algorithms are applied in lower-resource settings, lower-quality care may result, and that drop in quality may be tough to observe. Some examples may clarify the pattern.

Gallbladder cancer and inflamed gallbladders demonstrate the tremendous difference in optimal choice based on the resources of the

152. *See supra* Section IV.A.1.
153. *See supra* Section II.C.2.

medical setting.[154] Gallbladder cancer is both extremely rare and extremely aggressive; if it metastasizes beyond the gallbladder, patients have among the worse outcomes of any cancer. Cholecystitis, or inflammation of the gallbladder, on the other hand, is common. Cholecystitis is treated with a low-risk, technically straightforward surgery wherein the surgeon laparoscopically removes the inflamed gallbladder. Often, though, a patient will present to a doctor with what appears to be cholecystitis, but is actually gallbladder cancer. When that happens, the surgeon needs to notice — in the middle of the laparoscopic surgery — the signs of likely cancer and then decide — again, mid-surgery — whether to try to remove the cancer or stop the surgery and send the patient to a higher-resource hospital.[155] Doing the surgery (that is, removing the cancer and some surrounding tissue)

> requires significantly more surgeon skill, as well as surgeon education/understanding of the anatomy of the liver, the gallbladder, and the blood vessels and ducts. It also requires different, more specialized operative instruments. It will take longer, and it can be much harder. But if it's done correctly, the patient has their appropriate, necessary cancer operation at the time of (suspected) diagnoses. They may be cured at that point, or they might need chemo, but it gives them the best treatment and the best long-term survival.
>
> The problem with this option is that if there's an error, the surgeon can seriously injure the liver itself, the blood vessels and ducts to/from the liver, or, much worse, tear the gallbladder and spill cancer throughout the abdomen. Any of those has severe consequences that will require significant resources to address (or . . . advance the cancer and kill the patient).[156]

On the other hand, stopping the surgery and sending the patient to a better-trained surgeon with better equipment has essentially no risk (except the time elapsed), and no immediate benefit. Choosing to pur-

---

154. E-mail from Dr. Clare French, General Surgeon, SurgOne, P.C. (on file with author).

155. *Not* noticing the cancer is itself highly problematic (and a situation where AI could help); if the surgeon does not notice the cancer, common techniques such as opening the gallbladder to drain it before removal could be disastrous, spilling cancer throughout the abdomen and dooming the patient to a rapid death. E-mail exchange with Dr. Clare French, *supra* note 154.

156. *Id.*

sue the surgery may well be the right option in a setting with trained surgeons and better equipment — but is often a poor option in settings without those resources.[157] AI helping make that choice could easily make the wrong choice if trained only in environments with highly skilled surgeons.

Interventions do not have to be surgical. Choices among drugs may also be resource-dependent. Consider the example of a chemotherapeutic raised above: while a powerful drug may be most efficacious against a cancer, it may also carry high risks of serious side effects that require highly skilled monitoring to avoid. In a high-resource setting, the stronger chemotherapeutic may be the right call; outside such a setting, it may be catastrophic.

Overall, diagnostics and interventions that are the best options in high-resource settings will frequently not be the best options in low-resource contexts. When black-box algorithms are trained exclusively in high-resource settings, we should expect them to perform worse in low-resource settings where both patient populations and available resources are different.

B. Cost

Training medical AI in High-Resource Hospitals may also bias the resulting algorithms toward selecting more costly procedures. High-Resource Hospitals are on the cutting edge of medical treatment; they are where the most sophisticated and up-to-date techniques and technologies are developed and used. Academic medical centers also tend to treat patients more intensively than do other medical settings.[158] These treatments are often excellent — some researchers find that academic medical centers do better by patients than other hospitals[159] — but they are also more expensive.[160]

---

157. Nevertheless, surgeons in community hospitals certainly do sometimes think that attempting to remove the gallbladder cancer on the spot is the right option, and make that choice — despite the higher risk and potential failure. *Id.*

158. *See, e.g.*, Teryl Nuckols et al., *What Value-Based Payment Means for Academic Medical Centers*, NEJM CATALYST (May 30, 2019), https://catalyst.nejm.org/value-based-payment-academic-medical-centers [https://perma.cc/J8G4-T39Z].

159. *See, e.g.*, Laura G. Burke et al., *Association Between Teaching Status and Mortality in US Hospitals*, 317 JAMA 2105, 2107–10 (2017) (finding lower mortality rates at teaching hospitals for a range of common medical conditions, and finding that major teaching hospitals did better than minor teaching hospitals); John Z. Ayanian & Joel S. Weissman, *Teaching Hospitals and Quality of Care: A Review of the Literature*, 80 MILBANK Q. 569, 574–77 (2002) (reviewing the literature). *But see* Andrew M. Ibrahim et al., *Association of Hospital Critical Access Status with Surgical Outcomes and Expenditures Among Medicare Beneficiaries*, 315 JAMA 2095, 2096–99 (2016) (finding critical access rural hospitals performed as well as noncritical access non-rural hospitals on common surgeries).

160. *See, e.g.*, Lisa Rapaport, *Teaching Hospitals in U.S. Are Expensive, But Have Lower Death Rates*, REUTERS (May 23, 2017 3:59 PM), https://www.reuters.com/article/us-health-hospitals-usa-mortality-idUSKBN18J2UG [https://perma.cc/6VZD-JED3]; Nuckols et al.,

To the extent that providers in high-resource settings tend to choose more intense, more costly interventions, medical AI will learn those patterns and recommend them when translated to low-resource settings. Such effects may develop over time. If an algorithm continually suggests that patients get PET scans, for instance, and a community hospital does not have a PET scanner, obviously the patient cannot get that scan at that time. But providers in that hospital may note the continued recommendations and push for the hospital to buy a PET scanner to comport with the algorithm, resulting in higher costs over time. Such higher costs may be warranted — perhaps the hospital really needs a PET scanner to provide appropriate care efficiently and effectively. But in other cases, the AI may just suggest the more expensive option because that option is more prevalent in high-resource contexts, when a lower-cost option may be more appropriate for the lower-resource context.

This pattern is likely to result in anti-frugal effects. Although medical AI may reduce some costs — presumably, software is cheaper to run than an additional diagnostician is to hire — it may increase other costs by systematically changing preferred patterns of care to more closely match those at high-resource, more expensive care settings.[161] Overall, translation between contexts looks to have problematic effects on both the quality of care and the cost of care.

V. ISN'T ALL MEDICINE CONTEXTUAL?

On being presented with these translational challenges, one might ask: isn't all medicine contextual anyway? That is to say, isn't it the case that all medicine depends on the particular patient in front of the particular provider, the evidence upon which the intervention is based and in what populations that evidence was developed, and the resources available to the provider in the immediate medical encounter?[162] This Part gives three replies to this question: first, even if medical AI is just contextual like other medicine, that is worth noting; second, the opacity of medical AI may hide contextual changes that would otherwise be noticed; and third, the rhetoric and development

---

supra note 158. *But see* Laura G. Burke et al., *Comparison of Costs of Care for Medicare Patients Hospitalized in Teaching and Nonteaching Hospitals*, JAMA NETWORK OPEN (June 7, 2019), https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2735462 [https://perma.cc/M2EL-Y677] (presenting recent findings that indicate teaching hospitals may not necessarily be as expensive as previously thought).

161. These increases in care complexity and intensity may have differential effects on different patient populations, as well. Poorer patients are less able to undergo frequent follow-up visits, for instance, so a shift to more intensive follow-up care may disproportionately affect those patients.

162. In an even more direct comparison, doctors who work in low-resource contexts are themselves typically trained in high-resource academic medical centers (that's what makes them academic medical centers). But they adapt their practice to their new contexts.

of medical AI suggests that contextual dependence will not be a problem, making its existence more striking. The next Part addresses a different set of issues, focusing on how medicine already deals with contextual knowledge, and why those tools work less well for medical AI.

First, even if the ultimate conclusion from this work is only that medical AI is contextual like other types of medicine, and that its contextual effects may be concentrated across the gradient of resources, that is worth knowing. Medical AI has the apparent promise of addressing contextual challenges in medicine by distributing expertise and by taking account of patient variations to make care especially precise — and hopefully, it will fulfill that promise! But unless the status quo is changed, medical AI is subject to its own set of contextual biases. It is not an automatic panacea.

Second, opacity makes it potentially harder to spot problems that may arise from contextual bias than to spot parallel problems in well-understood systems. Medical AI is black-box medicine; it is difficult to know how it makes its recommendations.[163] This opacity makes it hard to spot the problems of contextual bias when they appear. If a provider consults a Physician's Desk Reference and sees that a particular treatment option is generally preferred but requires more resources to perform well than are available in a low-resource setting, she can decide to pursue a less effective but more practical option. But if this recommendation comes from an algorithm with no reasoning given — it might be based on specific patient characteristics, or the particulars of the diagnosis, or something else — it is harder to know whether that recommendation or some alternative is the better choice.[164] Thus, medical AI's contextual bias may be harder to understand and to rectify than in other medical situations.

Third and most importantly, the nature of machine-learning systems and the possibility of self-improvement provides an illusory safety rail. This requires some unpacking. For other medical technologies, we recognize (or, at least, we're starting to recognize) that the technology is developed in a specific context and might not work so well when deployed in other contexts. When the clinical trials used to approve a drug include no pregnant women, we recognize — or should — that evidence for its safe and effective use in pregnant women is lacking.[165] Perhaps more pointedly, whether or not the drug works safely for pregnant women, *we don't expect the drug to change*.

---

163. *See, e.g.*, Price, *Regulating Black-Box Medicine*, *supra* note 98, at 429–31.

164. *See, e.g.*, Jenna Burrell, *How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms*, BIG DATA & SOC'Y, Jan.–June 2016, at 1, 1.

165. *See* Jeanne S. Sheffield et al., *Designing Drug Trials: Considerations for Pregnant Women*, 59 CLINICAL INFECTIOUS DISEASES S437, S437 (2014) (supplement article) ("The study of therapeutic agents in pregnant women has been virtually nonexistent for decades.").

But that's not true for medical AI. Machine-learning systems hold out the possibility of improvement; it's right there in the name. Data from deployment and use can be used to improve the algorithms so that they get better over time.[166] And so we might reasonably be more optimistic about context-specificity in medical AI: while the algorithm isn't perfect when it starts, it will learn from deployment contexts and improve. But that safeguard is illusory. Contexts where medical AI is likely to run into problems are precisely the contexts where we lack the data needed to improve its performance: low-resource environments that lack the data infrastructure to train, improve, or even evaluate algorithms.[167] That lack of infrastructure doesn't end with the deployment of some black-box medicine implementations. Rather, unless that deployment is embedded within new data infrastructure that itself returns data to the algorithm's development, we should expect that any contextual problems will remain unaddressed even as the algorithm is used in the new context. Thus, medical AI holds out the promise of improvement over time, but that promise will do little to solve the problem of contextual bias in low-resource contexts.

So yes — all medicine is contextual, and black-box medicine is as well. But given black-box medicine's capacity for democratizing expertise, opacity, and capability for self-improvement in aspects other than contextual bias, bias in black-box medicine demands special attention.

VI. SOLUTIONS

The problem of contextual bias in medical AI is likely to dampen the potential benefits of democratizing medical expertise. Reducing this problem will be tricky. This Part discusses several possible solutions and closes with a discussion of traps to avoid in implementing them.

Two fairly obvious solutions for the quality problem both have real challenges. First, for several reasons, "humans-in-the-loop" — providers who can review and implement care options — won't prevent the problems above, though they may sometimes ameliorate them. Second, labeling of medical AI based on how and where it was trained faces substantial difficulties in implementation, and even if it works as intended, will not solve the problem.

Two quality solutions have more promise. First, public investment in data infrastructure can help tackle the problem at the front-

166. *See* Burrell, *supra* note 164, at 5; *see also* Ariel Bleicher, *Demystifying the Black Box that is AI*, SCI. AM. (Aug. 9, 2017), https://www.scientificamerican.com/article/demystifying-the-black-box-that-is-ai [https://perma.cc/3946-VCYY].

167. *See supra* Section III.A; *see also* Price, *Risk and Resilience*, *supra* note 32, at 74–75.

end by increasing the representativeness of data on which medical AI is trained. Second, regulatory requirements of at least some evidence of cross-contextual efficacy will reduce the incentives developers face to develop medical AI only in high-resource contexts.

Cost is its own tremendous tangle of issues; I briefly discuss how we might think through addressing it.

Finally, a coda considers three traps to avoid: too much contextualization, too little contextualization, and the innovation-stunting paralysis of the Nirvana fallacy.

A. Provider Safeguards and Humans-in-the-loop

Why doesn't the presence of human providers in care settings resolve the concerns described above? Medical practice already incorporates variation between different contexts; different doctors have different preferred strategies, and patients are different everywhere. The way we tend to resolve this is by relying on providers at the point of care to make the decision that is most appropriate for the patient in front of them — the original version of "personalized medicine." Why doesn't that work here to avoid these problems? I posit four reasons of increasing force: (1) present provider ignorance; (2) reliance on algorithms; (3) future provider ignorance; and (4) provider absence. Each reduces the force providers can bring to bear to correct translational errors of medical AI — and, more generally, should decrease our confidence in relying on "human-in-the-loop" safety mechanisms for medical AI.

1. Present Provider Ignorance

First, providers often don't know what the best options are, and therefore may not be suited to exercise independent corrective judgment on the decisions of algorithms. Famously, a large fraction of medicine as practiced is not evidence-based.[168] Providers may not know which option is preferable in general among a menu of options, much less what treatment is preferable for the specific patient in front of them or in the specific resource context of the medical encounter. The examples from current practice listed above exemplify this pattern; it may be the wrong call to undertake surgical removal of

168. *See, e.g.*, Diana Herrera-Perez et al., *A Comprehensive Review of Randomized Clinical Trials in Three Medical Journals Reveals 396 Medical Reversals*, ELIFE, June 11, 2019, at 1, 5–12, https://doi.org/10.7554/eLife.45183.001 [https://perma.cc/WS22-LCSQ] (cataloging medical reversals); Kayte Spector-Bagdady et al., *Stemming the Standard-of-Care Sprawl: Clinician Self-Interest and the Case of Electronic Fetal Monitoring*, 47 HASTINGS CTR. REP. 16, 16–17 (2017) (describing the persistence of electronic fetal monitoring and identifying the role of the legal system in that persistence).

gallbladder cancer in low-resource settings, but some providers still choose those options. Why would we assume that providers would somehow acquire the knowledge to correct the errors of medical AI when they currently make at least some similar errors in practice without AI present?[169]

2. Reliance on Algorithms

Second, even if providers currently know what the ideal diagnostic or treatment pathway is, they may not actually exercise independent judgment when confronted with an algorithm providing a different conclusion. Automation bias refers to a phenomenon where individuals rely on the results of automation even when they know or should know that the automation is wrong.[170] Sometimes, the individuals follow incorrect recommendations (commission errors), and sometimes they fail to notice problems when the software does not flag them for review (omission errors).[171] Both types of errors have been observed in the context of clinical decision support software in areas including prescriptions, mammogram interpretation, EKG interpretation, and clinical scenario management.[172] Overall, we probably want at least some level of automation bias, because good software still improves the level of care, even if it occasionally makes mistakes.[173] If providers are constantly second-guessing medical AI, we lose the benefits of increased performance and efficiency that they promise.[174] Nevertheless, the presence of automation bias decreases our ability to rely on humans-in-the-loop to correct problems of medical AI, whether based on problems with contextual translation or not.

169. In fact, sometimes this ignorance may lead to the opposite of the desired result: a provider may override the correct decision of an algorithm because it does not accord with her own knowledge or intuition, but that knowledge or intuition may be precisely wrong. *See* Price, *Medical Malpractice*, *supra* note 106, at 296–98.

170. *See* David Lyell et al., *Automation Bias in Electronic Prescribing*, BMC MED. INFORMATICS DECISION MAKING, Mar. 16, 2017, at 1, 1, https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-017-0425-5 [https://perma.cc/Q3RL-JULR].

171. *Id.*

172. *Id.* at 1–2 (reviewing the literature and describing the results of a study on prescription automation bias).

173. Software that performs worse than humans is another story; avoiding software like that is a big part of FDA's role in this picture. *See* Price, *Regulating Black-Box Medicine*, *supra* note 98, at 455–57.

174. *See* Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085, 1110 (2018); Price, *Medical Malpractice*, *supra* note 106, at 299–305.

3. Future Provider Ignorance

Third, even if providers *currently* know enough to correct the errors of medical AI, that knowledge base may decrease over time as medical AI becomes more commonplace. Michael Froomkin, Ian Kerr, and Joelle Pineau have painted a picture of what might happen to medical practice as medical AI becomes better and more available.[175] They argue that as medical AI becomes more competent, doctors will be pushed to rely on it by, among other things, medical malpractice, and that over time doctors will lose the knowledge necessary to practice medicine well and to know how well medical AI is performing.[176] One need not accept their argument wholesale to believe that as medical AI comes to perform at at-or-above human levels on some medical tasks, and to be widely available, the incentives for providers to train in those tasks will substantially decrease. This tendency will likely interact with the automation bias described immediately above, with the result that providers will be less able to catch errors resulting from problems in medical AI.

4. Provider Absence

Fourth and finally, all the models of humans-in-the-loop to reduce errors from medical AI, including contextual translation errors, rely on humans actually being present to take their place in the loop. But this is quite an assumption and will often be wrong. Consider again the benefit of medical AI on which this work focuses: the possibility of democratizing expertise, making high-level medical expertise available to those who otherwise might not have it. When we assume the presence of a skilled provider who can oversee the recommendations of medical AI, in a partnership rather than a replacement model, and can correct errors of the sort discussed above, we assume away this problem that medical AI can help us fix. Of course, a skilled surgeon can recognize the problems of trying to remove gallbladder cancer without the right operating tools, and could countermand the recommendation of medical AI to proceed — but what about the common situation where there is no skilled surgeon present? An excellent pathologist may recognize the mistakes of an AI-provided diagnosis of a particular pathology slide, but often there will be no excellent pathologist available, especially in the type of lower-resource settings on which this work focuses.

---

175. *See generally* Froomkin et al., *supra* note 49 (describing changes to existing medical liability rules to maintain physician participation and to avoid over-reliance on medical AI).

176. *Id.*; *see also* Federico Cabitza et al., *Unintended Consequences of Machine Learning in Medicine*, 318 J. AM. MED. ASS'N 517, 517 (2017).

In low-resource medical settings, we simply cannot assume the presence of practitioners with the right set of knowledge to recognize and fix the suboptimal recommendations medical AI may provide when its insights translate poorly to that exact low-resource context. Whether we are talking about community health centers, community hospitals, solo practitioners, or rural health settings with very limited provider availability, those settings will lack many types of expertise — and again, that's precisely the point of medical AI.[177] This is not to say that human-in-the-loop is not a laudable model; there are reasons to prefer rich provider involvement[178] (though there are also reasons to circumscribe that involvement[179]), and reasons to suspect that in high-resource contexts skilled providers will be unwilling to cede responsibility to medical AI.[180] But relying on humans to catch AI errors will not work in many contexts where medical AI promises to do a tremendous amount of good.

B. Labeling

Labeling medical AI to provide more information to users provides a solution that is both obvious and problematic. It is obvious because labeling is a common and straightforward way to recognize the limitations of technology, especially medical technology. It is problematic for three reasons. First, it is unclear how to label medical AI appropriately to recognize the problem of contextual bias. Second, providers often ignore medical labels and use technologies "off-label." Third, even if providers follow labeling restrictions about where to use medical AI, such a path hobbles the goal of democratizing medical expertise.

Labeling could mean two distinct things in this context: the more general labels that give instructions for any product, or FDA-mandated labeling. Labels are familiar in many regulated contexts;[181]

---

177. *See supra* Section II.C.2.

178. *See* Kiel Brennan-Marquez & Stephen E. Henderson, *Artificial Intelligence and Role-Reversible Judgment*, 109 J. CRIM. L. & CRIMINOLOGY 137, 146–48 (2019); Selbst & Barocas, *supra* note 174, at 1138–39.

179. For instance, if providers consistently second-guess the recommendations of algorithms, and the algorithms perform at a higher level than the average provider, provider reversal will on average lessen the quality of recommendations. *See* Price, *Medical Malpractice*, *supra* note 106, at 299–305 (discussing this dynamic); Selbst & Barocas, *supra* note 174, at 1129.

180. Among other things, for the foreseeable future, providers are likely to bear ultimate responsibility for final medical decisions. *See* Price, *Medical Malpractice*, *supra* note 106, at 303. *But see* Cabitza et al., *supra* note 176, at 517 (noting provider willingness to defer to algorithms).

181. *See generally* Omri Ben-Shahar & Carl E. Schneider, *The Failure of Mandated Disclosure*, 159 U. PA. L. REV. 647 (2011) (surveying mandatory disclosure regimes).

cigarettes carry warnings about potential health risks[182] and food carries labels stating nutrition content.[183] Labeling relies on a combination of disclosure and choice: product users should be able to choose how and whether to use a particular product, but they should be informed about the salient facts before making that choice (particularly if those facts are hard for users to discern on their own).[184]

The FDA mandates its own specific form of labeling for products it regulates, including drugs and medical devices — which as described above, will typically include medical AI.[185] Labels for medical devices must include adequate direction for use, including "[s]tatements of all conditions, purposes, or uses for which such device is intended."[186] Use outside those conditions, just like for drugs, is "off-label use."[187] The rest of this Section will assume the existence of an FDA-approved label for medical AI as a medical device, but similar arguments apply to non-FDA-approved labeling that just discloses information about a product to inform users.

Determining what information should go on a medical AI label will be hard. Ideally, a label would provide enough information such that those choosing to deploy it in a new context would know how well to expect the algorithm to perform in that context, and what types of failure or errors might be expected — and we don't know that yet. This Article has sought to open that conversation — mentioning, among other things, patient composition and resource availability (broadly defined) of the setting in which the algorithm was trained. But to really know how to impose labeling requirements that contain enough information to inform use meaningfully, we need to know a lot more about the relevant sources of patient and provider variation

182. *See, e.g.*, Kristin M. Sempeles, Note, *The FDA's Attempt to Scare the Smoke Out of You: Has the FDA Gone Too Far with the Nine New Cigarette Warning Labels?*, 117 PA. ST. L. REV. 223, 232–35 (2012) (describing the labeling regime); *see also* Sara C. Hitchman et al., *Changes in Effectiveness of Cigarette Health Warnings Over Time in Canada and the United States, 2002–2011*, 16 NICOTINE & TOBACCO RES. 536, 536 (2013) (evaluating the effectiveness of warning labels).

183. Nutrition Labeling and Education Act, Pub. L. No. 101-535, 104 Stat. 2353 (1990).

184. Ben-Shahar & Schneider, *supra* note 181, at 649–50.

185. *See supra* Section III.C.1.

186. 21 C.F.R. § 801.5 (2019); *see also* FDA, *General Device Labeling Requirements*, https://www.fda.gov/medical-devices/device-labeling/general-device-labeling-requirements [https://perma.cc/8U4W-SH69] (citing 21 C.F.R. § 801.5 (2019)).

187. *See, e.g.*, Randall S. Stafford, *Off-Label Use of Drugs and Medical Devices: A Review of Policy Implications*, 91 CLINICAL PHARMACOLOGY & THERAPEUTICS 920, 920 (2012) (providing an overview of off-label use of drugs and devices); Jamie S. Sutherell et al., *Pediatric Interventional Cardiology in the United States is Dependent on the Off-Label Use of Medical Devices*, 5 CONGENITAL HEART DISEASE 2, 2–3 (2010) (finding that half of all pediatric cardiac interventions involved an off-label use and that 99% of stent implantations were off-label).

than we know now.[188] This is not to say that such labeling is a futile enterprise, but it will be difficult to get right.

More problematic is how labels are actually used or not used in clinical care. In general, transparency is of limited efficacy in shaping the use of technology, though it is a commonly-prescribed solution.[189] Off-label use of drugs is famously common.[190] Perhaps unsurprisingly, some of the most common off-label uses, such as pediatric use without trials to support pediatric approval, [191] mimic common gaps in health big data. And although drug labels rarely specify that they are principally tested in relatively ancestrally homogeneous populations, we might think of the widespread use of drugs or other treatments in ancestral minorities in whom the treatments were not originally tested as a sort of ersatz off-label use. So, too, we should expect that medical AI would be used off-label just as other medical treatments are.[192] If an algorithm were trained in a relatively limited population, then using it in another population would be unsurprising — especially if the algorithm otherwise seems to be a good tool, trained on data from doctors in a high-resource setting.[193]

Third, finally, and most importantly: even if labels are well-designed and even if providers actually follow them — *that just gets us back to the original problem.* Recall the key goal of medical AI that drives the issue of translation across contexts (and the rest of this work): democratizing medical expertise, and allowing the provision of excellent medical care in settings where it might otherwise be outside the capabilities of providers in that setting. If labels state that medical AI is developed in high-resource settings with relatively limited patient populations and should be limited to similar situations, and if providers follow those labels to avoid using the medical AI in low-resource settings with different patient populations, then the medical AI doesn't actually democratize expertise at all. Respecting limita-

188. One parallel solution is to just let the algorithms sort all of this variation out, such that rather than labels noting variation, the algorithms themselves take all relevant variation into account. But this begs the question — that solution requires that algorithms be developed with enough data to see that range of variation, which by hypothesis throughout this piece is not the case.

189. *See, e.g.*, Ben-Shahar & Schneider, *supra* note 181, at 679. *But see* Ryan Bubb, *TMI? Why the Optimal Architecture of Disclosure Remains TBD*, 113 MICH. L. REV. 1021, 1042 (2015) (arguing for the effectiveness of some disclosures).

190. *See, e.g.*, Rebecca S. Eisenberg, *The Problem of New Uses*, 2 YALE J. HEALTH POL'Y L. & ETHICS 717, 731 nn.62–63 (2005) [hereinafter Eisenberg, *Problem of New Uses*].

191. *See* Sutherell et al., *supra* note 187, at 2–3.

192. One could imagine technological limitations built into algorithms, such that an algorithm trained only on adults, for instance, would simply not provide a recommendation in a pediatric case. This would be challenging to implement and would also not solve the immediately following problem.

193. *See supra* Section III.B (describing the reputational benefit of training in high-resource settings).

tions of algorithmic development by avoiding potentially problematic contexts is like responding to the problem of biased policing by keeping police out of minority neighborhoods entirely: it may decrease the problem of bias, but it also loses any potential benefit that the technology or intervention might give to those in the second context.[194]

Labeling may still have *some* benefit. If, for instance, providers or other purchasers of medical AI actually do follow restrictive labeling, then developers would face incentives to demonstrate cross-context efficacy to obtain a broader label and therefore broader use.[195] But this assumes that labels are closely followed, and also that the resources available in low-resource settings are sufficient to outweigh incentives to focus development on high-resource settings — assumptions that are easy to challenge. On the other hand, training-based labels might be of more use when combined with two other interventions: investment in data infrastructure and regulatory mandates for cross-context efficacy data.

C. Representative Datasets

A third way to ameliorate problems in contextual translation involves addressing the root of the issue: the initial training data. If the existing dynamic is principally driven by data location — High-Resource Hospitals are where the data are[196] — then policymakers could push to generate and collect more data to change that initial condition. The public — and by public, here I largely mean the government, whether state or federal[197] — can invest in two types of data infrastructure.[198] First, it can invest in infrastructure *for* data: resources like computer servers, personnel, standards, and procedures that let data be collected, controlled for quality, and made available at lower-resource settings such as community health centers.[199] The public can also invest in the infrastructure *of* data: large collections of

194. *See, e.g.*, Solon Barocas & Andrew D. Sebst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 689–90 (2016); I. Glenn Cohen & Harry S. Graver, *Cops, Docs, and Code: A Dialogue Between Big Data in Health Care and Predictive Policing*, 51 U.C. DAVIS L. REV. 437, 443–44 (2018).

195. *See* Eisenberg, *Problem of New Uses*, *supra* note 190, at 734–35.

196. *See supra* Section III.A.

197. Private investors could also invest in funding such data infrastructure, but private investment in infrastructure tends to be socially suboptimal because private actors cannot adequately capture the spillover benefits that infrastructural goods provide for other innovators and downstream users. *See* BRETT M. FRISCHMANN, INFRASTRUCTURE: THE SOCIAL VALUE OF SHARED RESOURCES 66 (2012); Price, *Risk and Resilience*, *supra* note 32, at 77–78. Private firms also have incentives to keep data collected as trade secrets to maintain competitive advantage, which causes problems both for data aggregation across firms and for external validation of medical AI models. *See* Price, *Big Data*, *supra* note 71, at 1432–35.

198. Price, *Risk and Resilience*, *supra* note 32, at 78–79.

199. *Id.*

data that can enable broad sets of future innovation and economic activity, such as developing black-box medical algorithms.[200]

Investing in infrastructure *for* data could take several forms. Most obviously, grants specifically directed to support the purchase of computer systems or the hiring of data personnel could ameliorate the data acquisition problem of low-resource settings.[201] Less obviously, governments can set standards, which themselves are a sort of infrastructure. Electronic health records currently use a hodgepodge of formats due to an early decision not to federally mandate a centralized format; this situation has resulted in problems of interoperability and data fragmentation.[202] The federal government is moving to address interoperability problems, but could and probably should go further to require standards for electronic health records.[203] Government efforts could also ease the burden of developing data infrastructure: adding a research exemption to the HIPAA Privacy Rule, for instance, would make it easier for low-resource settings to collect data by removing one set of legal concerns.[204]

The advantage of public funding of infrastructure *for* and *of* data, whether through grant funding or direct spending, is that such infrastructure does not have to follow the pre-existing patterns of what is collected and where. Instead, data collected through a public infrastructure effort can better represent the care that many patients actually experience.[205] If high-quality data are collected about a wide variety of patients, the concerns about effects from different patient populations decrease.[206] And if data are collected about a wider range of care settings — not just High-Resource Hospitals, but community hospitals, community health centers, practitioner's offices, and the like — those data can more accurately reflect the resources available in the care setting, the range of practices followed, the types of treatment implemented, and the outcomes that result.[207]

---

200. *Id.*; *see also* Price, *Big Data*, *supra* note 71, at 1439–44 (proposing an infrastructure model for gathering data to promote the development of black-box medical algorithms); MATHENY ET AL., *supra* note 4, at 169–71 (arguing that, for medical AI to reach its potential, datasets must be conceptualized as a "public good").

201. Even here, there may be backlash along the lines of, "Why fund data when we have insufficient funding for care?" The awkward answer is that better, broader data make future care better and cheaper — but that may be a difficult sell to those facing resource gaps.

202. Julia Adler-Milstein, *Moving Past the EHR Interoperability Blame Game*, NEJM CATALYST (July 18, 2017), http://catalyst.nejm.org/ehr-interoperability-blame-game [https://perma.cc/BPT7-4SLM].

203. *Id.*

204. *See* Price, *Drug Approval*, *supra* note 77, at 2460–61. In fact, an earlier draft of the 21st Century Cures Act included such a provision, but it was removed in the final text. *Id*.

205. Indeed, grant funding involving data collection could be conditioned on a requirement that data be more representative.

206. *See supra* Section IV.A.1.

207. This data collection goal is essential to the development of a learning health system more broadly. *See* Elizabeth A. McGlynn et al., *Developing a Data Infrastructure for a*

The difficulty of this endeavor should also not be minimized. There are reasons that data collection practices today are as they are.[208] Gathering data well is hard and can be expensive.[209] Privacy concerns are also implicated in the gathering, use, and sharing of large amounts of sensitive health data.[210] Nevertheless, investment in those data-gathering capacities — across many contexts — is likely to pay substantial dividends down the line, including in those same low-resource contexts.

The NIH's All of Us initiative is a prominent example of exactly this type of infrastructural investment in data.[211] All of Us (formerly the Precision Medicine Cohort) is a part of the Precision Medicine Initiative.[212] Through All of Us, the NIH plans to gather detailed health information — including genetic sequences, treatment information, and outcome data — from over one million Americans. Crucially, the sample population for All of Us is meant to be nationally representative.[213] According to Francis Collins, the Director of the NIH, the program has a goal that half of its participants come from traditionally underrepresented groups.[214] If the definition of diversity is broadened to include socioeconomic status and rural status, then the

---

Learning Health System: The PORTAL Network, 21 J. AM. INFORMATICS ASS'N 596, 598–600 (2014). Increased data infrastructure also allows other types of innovation and measurements of health system quality more generally. *Id.*

208. *See supra* Section III.A.

209. *Id.* Secondary questions also arise as to the best allocation of resources. One might argue that any new resources allocated to low-resource medical contexts should be aimed directly at improving care rather than improving data infrastructure. That calculus is complex. I argue here only that investment in data infrastructure in low-resource contexts will benefit patients in those contexts down the road, not that such investment is the best use of scarce resources. However, infrastructure is often a useful investment, considering the amount by which it can increase innovation and future welfare. *See, e.g.*, W. Nicholson Price II, *Grants*, 34 BERKELEY TECH. L.J. 1, 59 (2019) (discussing the role of government grants in providing infrastructure for future innovation).

210. I have examined the privacy concerns of medical big data and medical AI in some detail elsewhere. *See* Roger Allan Ford & W. Nicholson Price II, *Privacy and Accountability in Black-Box Medicine*, 23 MICH. TELECOMM. & TECH. L. REV. 1, 19–20 (2016) (discussing the privacy challenges of medical AI in general and noting the tension between third-party validation and privacy protections); Price, *Drug Approval, supra* note 77, at 2458 (describing the limitations HIPAA and other privacy rules place on the innovation of a learning health system where patient data are constantly used to improve medical knowledge); W. Nicholson Price II & I. Glenn Cohen, *Privacy in the Age of Medical Big Data*, 25 NATURE MED. 37, 42 (2019) (surveying the privacy landscape for medical big data and arguing against a "privacy maximalist" approach).

211. *See* NAT'L INST. OF HEALTH, *All of Us Research Program*, https://allofus.nih.gov [https://perma.cc/T2B9-6Z3D]. Other examples include the UK Biobank. *See* Editorial, *UK Biobank Data on 500,000 People Paves Way to Precision Medicine*, NATURE (Oct. 10, 2018), https://www.nature.com/articles/d41586-018-06950-9 [https://perma.cc/WK9U-GJZW].

212. NAT'L INST. OF HEALTH, *supra* note 211.

213. *Id.*

214 Francis Collins, *Keynote: An Update on All of Us,* PROC. PRECISION MED. WORLD CONF. (June 6, 2018).

NIH plans to draw 75% of participants from diverse groups more generally.[215] This goal speaks directly to diversity and representativeness of patient population, and at least indirectly to the diversity of medical contexts, given that many of these participants are likely to seek medical care in low-resource contexts.[216] Efforts like All of Us should be supported, continued, and expanded.

To be clear, more representative datasets do not need to be publicly funded. The dynamics described in Part III make non-representative data an easy default for private parties, but private parties could also seek to address it (especially if required to as described immediately below). One approach could blend private spending on infrastructure for data with private acquisition of data. If developers sought data from low-resource contexts but recognized that those contexts lacked the resources to generate high-quality data, those developers could themselves provide the resources — technological or personnel-based — in exchange for access to the data generated, which would then fuel better performance down the road.

D. FDA Regulation and Concordance

The FDA could also play a role in decreasing problems of contextual translation. As described above, the reliance on data from High-Resource Hospitals may be over-determined: not only are data currently found in high-resource contexts, but using data from those contexts also helps avoid risks arising from FDA regulation, tort liability, and insurance reimbursement pathways. To help reduce these pressures to focus medical AI training on data from a limited set of medical contexts, the FDA approval process for medical AI products could be shifted to require explicit concordance data and demonstration of cross-context performance.

The FDA could explicitly require that developers seeking clearance or approval for medical devices using AI or machine learning include concordance data demonstrating performance in contexts outside the original development context.[217] More specifically, if an algorithm proposes to recommend treatment pathways based on patient characteristics, FDA could require that the validation of those pathways consider not only the high-resource contexts where the algo-

215. *Id.*

216. *See supra* Section IV.A.1.

217. The FDA could implement such requirements for other technology, including other algorithms; however, as described in this Article, black-box algorithms are particularly worrisome and therefore merit special attention.

rithm was developed but also low-resource contexts where it is likely to be deployed.[218]

A requirement for low-resource concordance demonstrations would not be trivial. In the current state of the world, low-resource contexts will often have insufficient data to allow purely data-based validation.[219] Thus, demonstrating concordance now might require extra clinical trials, which are costly and don't always match well with the development of black-box medicine.[220] Implementation of better data infrastructure — ideally, of a data-based learning health system more broadly — should eventually decrease the difficulty of validation of performance in different environments.[221] The near term is likely to be messy. But FDA requirements and infrastructural investments could interact in a self-sustaining cycle: infrastructural investments in data can help support the ability to demonstrate concordance to FDA, while FDA requirements to demonstrate concordance would encourage data infrastructure investment.

While FDA requirements for concordance would be unusual, such requirements have some precedent. The FDA already encourages greater gender, racial, and ethnic diversity among clinical trial participants, though it does not require it.[222] Clinical research funded by the NIH has even stronger diversity requirements; in 1993, Congress passed the National Institutes of Health Revitalization Act, which required the NIH Director to ensure inclusion of women and minorities in clinical research.[223] These requirements are not squarely on point — the NIH policy stems from grant funding of clinical trials, and FDA encouragement is voluntary — but they demonstrate a similar commitment to ensuring that clinical trials show that treatments work in different groups.

---

218. *See, e.g.*, *Performance*, *supra* note 43, at 4–5 (describing clinical trial testing IDx-DR in ten primary care clinics across the United States).

219. *See supra* Section III.A; *see also* Ford & Price, *supra* note 210, at 18–21 (2016) (discussing the idea of data-based validation); Price, *Regulating Black-Box Medicine*, *supra* note 98, at 432–37 (same).

220. *See* Price, *Regulating Black-Box Medicine*, *supra* note 98, at 434–35.

221. This infrastructure, as described above, would also weaken the primary motivator for focus on High-Resource Hospitals, their predominance in possessing relevant data. *See supra* Section VI.C.

222. FDA, *FDA Encourages More Participation, Diversity in Clinical Trials* (Jan. 16, 2018), https://www.fda.gov/ForConsumers/ConsumerUpdates/ucm535306.htm [https://perma.cc/V24R-8TN2]; FDA, *Racial and Ethical Minorities in Clinical Trials* (Aug. 6, 2018), https://www.fda.gov/forconsumers/byaudience/minorityhealth/ucm472295.htm [https://perma.cc/QVT7-3A68].

223. Pub. L. No. 103-43, § 131, 107 Stat. 122, 133–35 (1993) (codified at 42 U.S.C. § 492B (1988)); *see also* NAT'L INST. OF HEALTH, *NIH Policy and Guidelines on the Inclusion of Women and Minorities as Subjects in Clinical Research* (Oct. 9, 2001), https://grants.nih.gov/grants/funding/women_min/guidelines.htm [https://perma.cc/QVT7-3A68].

FDA involvement in demonstrating concordance and applicability across contexts could also help resolve the other two legal incentives currently pushing for development based on high-resource data: tort liability and insurance reimbursement.[224] When FDA approves a medical device as a Class III device (i.e., a higher-risk device) under its premarket approval pathway, state tort liability is largely preempted for that device under *Riegel v. Medtronic*.[225] Thus, at least for companies that pursue Class III premarket approvals — and assuming that concordance data helps persuade FDA to grant such approvals — tort liability concerns should also largely be resolved by that process.[226] This doesn't resolve all liability concerns. Bringing a device to market through the 510(k) preclearance pathway (i.e., a finding that the device has an already-approved predicate device on the market) does not preempt state tort lawsuits,[227] and so far developers have been able to use that pathway (or *de novo* classification[228]) and to bring devices to market as Class I or Class II devices rather than undergoing the costlier premarket approval pathway for Class III devices.[229]

Finally, FDA approval, especially if that process includes concordance data, should help resolve issues of insurer reimbursement. As described above, FDA approval and CMS reimbursement decisions are frequently linked,[230] and private payers frequently follow CMS's lead.[231] An FDA-approved demonstration that an algorithm works in different contexts could similarly support payer determinations that the technology is worth reimbursing across those different contexts, even in the absence of the current quality proxy of training

224. *See supra* Sections III.C.2–3.

225. *See* Riegel v. Medtronic, Inc., 552 U.S. 312, 345 (2008).

226. This analysis assumes that *Riegel*'s bright-line rule — Class III premarket approval preempts state tort suits — remains. The possibility of medical AI changing over time with new data might suggest that this rule should be revisited because a prior approval might no longer serve as evidence of current safety in the same way as for a relatively static medical device.

227. *See* Medtronic, Inc. v. Lohr, 518 U.S. 470, 471 (1996).

228. As far as I know, no court has yet determined whether a determination by FDA under the *de novo* pathway that a device is Class II is sufficient to preempt state tort liability. Predicting the result is outside the scope of this work.

229. *See supra* notes 99–101. I have argued elsewhere that a rigid premarket approval process is likely to stifle innovation in black-box medicine. Price, *Regulating Black-Box Medicine*, *supra* note 98, at 451–54. The FDA is currently engaged in efforts to ease the path to market for digital health generally, including medical AI. *See* FDA, *Digital Health Innovation Action Plan*, https://www.fda.gov/downloads/MedicalDevices/DigitalHealth/UCM568735.pdf [https://perma.cc/ZLP9-G77U]. It remains to be seen how much FDA's efforts at more flexible market pathways will end up making premarket approval itself a more attractive option for developers. *See* Eisenberg, *Device Regulation*, *supra* note 117 (discussing this dynamic in the context of diagnostics more generally).

230. *See supra* Section III.C.3; *see also* Sachs, *Delinking Reimbursement*, *supra* note 114, at 2311.

231. Eisenberg & Price, *supra* note 80, at 31–32.

on data from high-resource medical contexts. Such FDA approval incentives could be even stronger if reimbursement is greater for devices that go through full premarket approval rather than 510(k) clearance — a difference that CMS currently is pressing.[232]

FDA approval modifications will not be a panacea, of course; insurers may still prefer the cachet of high-resource settings and developers may also seek the prima facie liability reduction that could come with name-brand training data. Persuading providers and health systems to adopt black-box medical systems may also still be easier while partnering prominently with High-Resource Hospitals.[233] But linking FDA approval, with its concomitant benefits, to a development process that at least attempts to ensure validity across contextual translation may help ease the path to broadly useful medical AI.

E. Incorporating Cost

The cost problem is extremely tough. As described above, algorithms that learn about the right kind of care in high-resource settings may simultaneously learn that the right kind of care is an expensive form of care, with many interventions and fancy, costly tools.[234] This may sometimes even be correct; sometimes, high-intervention care is the right way to go, and some costly interventions rightly spread from high-resource contexts to low-resource contexts. But it also creates the possibility for AI acting as a vector in increasing costs in a system which sorely needs to reduce costs, and in which AI has at least the potential to contribute to that reduction.[235]

A potential solution is easy to state but hard to implement. 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.

Implementing such measures could be quite challenging, especially since in the U.S. rationing health dollars is a hot-button issue.[236] And, at least in a fee-for-service system, which still exists in many contexts, provider and health system incentives typically push for more care, and costlier care, rather than efficient and cost-effective

232. *See* CTRS. MEDICARE & MEDICAID SERVS., *supra* note 119 and accompanying text.
233. *See supra* Section III.C.3.
234. *See supra* Section IV.B.
235. *Id.*
236. *See, e.g.*, Elizabeth Weeks Leonard, *Death Panels and the Rhetoric of Rationing*, 13 NEV. L.J. 872, 873–74, 886 (2013).

care. It seems more likely that the first medical AI systems focused on cost will aim to promote revenue maximization rather than efficiency.[237] But building costs into the initial models could eventually help some AI systems reduce system costs, assuming that cost reduction becomes a goal of system developers.

F. Traps to Avoid

Figuring out how best to develop and to deploy AI to democratize medical expertise is hard. It's made harder because contextual bias is not the only challenge that medical AI faces. If we try too hard to eliminate contextual bias, we could wind up with any of three related problems: too much contextualization, insufficient contextualization, or inadequate adoption.[238]

First, pushing too hard to ensure that AI is trained for each context could result in too much contextualization. One potential solution to the problem of contextual bias is to train AI in a wide variety of contexts so that every context has its own AI matched specifically to it. But the health system is rife with disparity, and AI might replicate or enhance those disparities.[239] Not only do many low-resource contexts lack the capacity to generate the data to train medical AI, or to support the training and validation necessary once those data are gathered, any AI that might result would be trained on a context with, by definition, a lack of resource-based expertise. Medical AI trained in health centers in rural India would avoid any problems of contextual bias when translated to other rural Indian health centers (or perhaps to other developing-world rural health centers), but it would lack the benefit of being trained on providers with the most extensive training, tools, and experience in high-resource settings. Such an approach would democratize only limited forms of medical expertise and would leave much of the benefit of medical AI on the table.[240]

---

237. AI developed by *payers*, on the other hand, might prioritize efficiency. *See* Eisenberg & Price, *supra* note 80, at 16–18.

238. A separate and complex set of issues concerns the distributional effects of efforts to ensure broad applicability — why not just allow medical AI to be developed for those in high-resource contexts, and perhaps those benefits will eventually trickle down to those in low-resource contexts? One preliminary answer might be that to the extent that medical AI is touted for its benefit in democratizing expertise, this Article takes that goal as a given and focuses on how to actually achieve it successfully. Other answers about the ideal path for development and spread of new technology more generally are outside the scope of this work.

239. *See, e.g.*, Khullar, *supra* note 130; Charlotte A. Tschider, *Regulating the Internet of Things: Discrimination, Privacy, and Cybersecurity in the Artificial Intelligence Age*, 96 DENV. U. L. REV. 87, 98–100 (2018) (describing concerns of AI systems codifying existing disparities).

240. *See, e.g.*, Alvin Rajkomar et al., *Ensuring Fairness in Machine Learning to Advance Health Equity*, 169 ANNALS INTERNAL MED. 866, 866–868 (2018).

Second, trying to avoid variation in algorithmic performance in different contexts could result in too little contextualization. In an ideal world, medical AI would be able to take advantage of differences in resources. If a hospital has a top-notch PET scanner and very experienced surgeons, AI algorithms that make recommendations should consider those options within the set of possibilities. In an ideal world, everyone would have access to the best care, but that is not our world, and not all hospitals have such resources. We don't want medical AI never to suggest using a PET scanner or undertaking a risky surgery just because those are unhelpful or actively harmful suggestions in some medical contexts. And patient populations *do* differ, both as groups and as individuals; medical AI should be able to take account of those differences as well.[241] Part of the appeal of black-box medicine is the possibility of intensely personalized analysis and recommendations for care; requiring too stringent replicability across contexts might sacrifice some of that precision. Those designing concordance policies need to tread a middle path.

Third, focusing too much on these problems — contextual bias, too much contextualization, and insufficient contextualization — could result in decisionmakers throwing up their hands and avoiding the new problems that come with medical AI, preferring the status quo. This is the Nirvana fallacy, where new options are compared to perfection rather than a flawed status quo.[242] But the status quo itself already has lots of problems, some of which form the impetus for medical AI in the first place.[243] The promise of democratizing expertise is enticing precisely because we have too few experts, and many patients face barriers to accessing high-quality care in all but the highest-resource settings. Avoiding the adoption of medical AI because it might not work as well in low-resource contexts does nothing to aid patients who already lack options because of the lack of resources.[244] Ultimately, even flawed medical AI may prove transformative for millions of patients, and we should endeavor to see that promise even while we try to avoid the pitfalls of cross-context translation.

VII. CONCLUSION

Medical AI has tremendous promise to bring excellent medical care to those that might not otherwise see such care. Translating

241. *See* Price, *Black-Box Medicine*, *supra* note 2, at 425–30.

242. Demsetz, *supra* note 13, at 1.

243. *See* Price, *Black-Box Medicine*, *supra* note 2, at 434–37.

244. Adoption of medical AI is already likely to face barriers from providers and potentially patients; too much focus on contextual bias is likely to increase already-existing hurdles. *See, e.g.*, Price, *Black-Box Medicine*, *supra* note 2, at 437–42 (discussing barriers to adoption).

black-box algorithms from high-resource contexts to low-resource contexts, though, brings the risk of problems; what works well in one context may not in another. If we are to avoid the risks of compromising care for those in low-resource settings, now is the time to consider how medical AI can be developed not just for those who already have access to excellent care, but for those who can benefit most from the advent of this new technology.