SCHOOL OF LAW

CASE WESTERN RESERVE UNIVERSITY

Artificial Intelligence and Discrimination in Health Care

Sharona Hoffman
and Andy Podgurski

Case Research Paper Series in Legal Studies
Paper 2020-29
December 2020

Artificial Intelligence and Discrimination in Health Care

Sharona Hoffman & Andy Podgurski*

Abstract:

Artificial intelligence (AI) holds great promise for improved health-care outcomes. It has been used to analyze tumor images, to help doctors choose among different treatment options, and to combat the COVID-19 pandemic. But AI also poses substantial new hazards. This Article focuses on a particular type of health-care harm that has thus far evaded significant legal scrutiny. The harm is algorithmic discrimination.

Algorithmic discrimination in health care occurs with surprising frequency. A well-known example is an algorithm used to identify candidates for "high risk care management" programs that routinely failed to refer racial minorities for these beneficial services. Furthermore, some algorithms deliberately adjust for race in ways that hurt minority patients. For example, according to a 2020 *New England Journal of Medicine* article, algorithms have regularly underestimated African Americans' risks of kidney stones, death from heart failure, and other medical problems.

This Article argues that algorithmic discrimination in medicine can violate civil rights laws such as Title VI and Section 1557 of the Affordable Care Act when it exacerbates health disparities or perpetuates inequities. It urges that algorithmic fairness constitute a key element in designing, implementing, and validating AI and that both legal and technical tools be deployed to promote fairness. To that end, we call for the reintroduction of the disparate impact theory as a robust litigation tool in the health-care arena and for the passage of an algorithmic accountability act. We also detail technical measures that AI developers and users should implement.

* Sharona Hoffman, Edgar A. Hahn Professor of Law and Professor of Bioethics, Co-Director of Law-Medicine Center, Case Western Reserve University School of Law; B.A., Wellesley College; J.D., Harvard Law School; LL.M. in Health Law, University of Houston; S.J.D. in Health Law, Case Western Reserve University. Author of ELECTRONIC HEALTH RECORDS AND MEDICAL BIG DATA: LAW AND POLICY (Cambridge University Press 2016). For more information, see https://sharonahoffman.com. Andy Podgurski, Professor of Computer and Data Sciences, Case Western Reserve University; B.S., M.S., Ph.D., University of Massachusetts. The authors thank Peter Gerhart, Jessie Hill, Katharine Van Tassel, and all participants in the Case Western Reserve University School of Law summer faculty workshop for their very helpful comments on earlier drafts. We also thank Mariah Dick for her invaluable research assistance.

TABLE OF CONTENTS

INTRODUCTION

Artificial intelligence (AI) is no longer a novelty in the medical field, and its use is increasingly prevalent.[1] According to a 2020 *Washington Post* article, "From diagnosing patients to policing drug theft in hospitals, AI has crept into nearly every facet of the health-care system, eclipsing the use of machine intelligence in other industries."[2] A KPMG survey of hundreds of business decision makers found that eighty-nine percent of respondents from the health-care industry believed that AI has already generated efficiencies in medical care, and ninety-one percent believe it has enhanced patients' access to care.[3]

AI, which does its work through learning algorithms and models,[4] thus holds great promise for improved health-care outcomes, but it also poses substantial new risks and hazards.[5] This article focuses on a particular type of health-care harm that has thus far evaded significant legal scrutiny. The harm is algorithmic discrimination.

In a June 2019 statement, the American Medical Informatics Association urged the Food and Drug Administration to address AI biases related to ethnicity, gender, age, socioeconomic status, and disability.[6] It suggested that the agency

---

1. MELANIE MITCHELL, ARTIFICIAL INTELLIGENCE: A GUIDE FOR THINKING HUMANS 119 (2019) (noting that AI will soon become widespread in medicine, "assisting physicians in diagnosing diseases and in suggesting treatments; discovering new drugs; and monitoring the health and safety of the elderly in their homes"); Amisha, Paras Malik, Monika Pathania & Vyas Kumar Rathaur, *Overview of Artificial Intelligence in Medicine*, 8 J. FAM. MED. & PRIMARY CARE 2328, 2328 (2019); W. Nicholson Price II, *Risks and Remedies for Artificial Intelligence in Health Care*, BROOKINGS (Nov. 14, 2019), https://www.brookings.edu/research/risks-and-remedies-for-artificial-intelligence-in-health-care.

2. Meryl Kornfield, *The Health 202: Artificial Intelligence Use Is Growing in the U.S. Health-Care System*, WASH. POST (Feb. 24, 2020, 7:41 AM EST), https://www.washingtonpost.com/news/powerpost/paloma/the-health-202/2020/02/24/the-health-202-artificial-intelligence-use-is-growing-in-the-u-s-health-care-system/5e52f13188e0fa632ba81ec7.

3. *Living in an AI World: Achievements and Challenges in Artificial Intelligence Across Five Industries*, KPMG 5 (2020), https://advisory.kpmg.us/content/dam/advisory/en/pdfs/2020/living-in-ai-world.pdf. This study surveyed 751 business decision-makers from five industries who had "at least a moderate knowledge of AI. *Id.* at 2.

4. *See infra* notes 33-34 and accompanying text. Researchers sometimes use the terms "learning algorithm" and "model" interchangeably. More accurately, however, the term "model" suggests a representation of knowledge that is created by an algorithm. MAX KUHN & KJELL JOHNSON, APPLIED PREDICTIVE MODELING 2 (2013); SHAI SHALEV-SHWARTZ & SHAI BEN-DAVID, UNDERSTANDING MACHINE LEARNING: FROM THEORY TO ALGORITHMS 13-14 (2014).

5. Michael J. Rigby, *Ethical Dimensions of Using Artificial Intelligence in Health Care*, 21 AMA J. ETHICS E121, E121-23 (2019); *The Dangers of AI in the Healthcare Industry*, THOMAS (May 7, 2019), https://www.thomasnet.com/insights/the-challenges-and-dangers-of-ai-in-the-health-care-industry-report.

6. *AMIA Supports, Encourages Further Refinement of FDA AI/Machine Learning Regulatory Framework*, AMIA (June 5, 2019), https://www.amia.org/news-and-publications/press-release/amia-supports-encourages-further-refinement-fda-aimachine-learning.

issue guidance about testing and adjustment of algorithms.[7]

There are many examples of algorithmic discrimination that have become infamous outside of the medical field. An algorithm designed to predict criminal recidivism exhibited bias against Black defendants.[8] It incorrectly labeled Black defendants as likely to reoffend almost twice as often as in the case of White defendants, and it mislabeled White defendants as low-risk more frequently than Black defendants.[9] In the employment arena, Amazon developed artificial intelligence-driven software to identify its best job candidates.[10] It turned out, however, that the algorithm was biased against women and routinely concluded that men were preferable candidates.[11] As a third example, in March of 2019, the Department of Housing and Urban Development sued Facebook, asserting that it kept certain users from seeing housing ads based on machine-learning algorithms' inferences about their race.[12]

Algorithmic discrimination in employment, criminal law, housing, and other fields has garnered attention in the legal literature.[13] Surprisingly, however, the

---

7. *Id.*

8. Julia Angwin, Jeff Larson, Surya Mattu & Lauren Kirchner, *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

9. *Id.*; *see also* Melissa Hamilton, *Debating Algorithmic Fairness*, 52 UC DAVIS L. REV. 261, 264 (2019) (reporting that the risk tool's corporate owner denied the allegation and stated that its reanalysis of the data led it to conclude that "the tool was unbiased as blacks and whites had similar positive predictive values for recidivism"); Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218, 2221-22 (2019) (discussing algorithmic risk assessment in the criminal justice system and its racial impact).

10. MICHAEL KEARNS & AARON ROTH, THE ETHICAL ALGORITHM 60-61 (2020) (relating that Amazon's algorithm "was found to be explicitly penalizing resumes that contained the word *women's*, as in "women's chess club captain," and downgraded candidates who listed the names of two particular all-women colleges"); Katherine Maher, Opinion, *Without Humans, A.I. Can Wreak Havoc*, N.Y. TIMES (Mar. 12, 2019), https://www.nytimes.com/2019/03/12/opinion/artificial-intelligence-wikipedia.html.

11. *Id.*; *see infra* Sections II.B-C for a discussion of bias.

12. Margot E. Kaminski & Andrew D. Selbst, Opinion, *The Legislation That Targets the Racist Impacts of Tech*, N.Y. TIMES (May 7, 2019), https://www.nytimes.com/2019/05/07/opinion/tech-racism-algorithms.html.

13. Ifeoma Ajunwa, *The Paradox of Automation as Anti-Bias Intervention*, 41 CARDOZO L. REV. 1671, 1692-96 (2020) (discussing automated decision-making in employment); Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 675 (2016) (focusing on Title VII's prohibition of employment discrimination); Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE. L.J. 1043, 1053-54 (2019) (discussing the discriminatory effects of implementing AI in the criminal justice setting); Sonia K. Katyal, *Private Accountability in the Age of Artificial Intelligence*, 66 UCLA L. REV. 54, 56 (2019) (discussing various applications of artificial intelligence that lead to discrimination, including in the criminal justice, housing, and employment realms); Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857, 860 (2017) (discussing employers' use of data analytics to make workplace decisions); Gerhard Wagner & Horst Eidenmüller, *Down by Algorithms? Siphoning Rents, Exploiting Biases, and Shaping*

legal literature has not focused on AI-related discrimination in health care, even though it clearly occurs.[14] A well-known example is an algorithm used to identify candidates for "high risk care management" programs that routinely failed to refer racial minorities for these beneficial services.[15] Other algorithms explicitly adjust for race, adding or subtracting risk points based on patients' ancestral background.[16] This Article, therefore, fills a noticeable gap in the treatment of AI in legal scholarship.

Learning algorithms[17] are trained on data, which means that the quality of the data is vital to the reliability of the AI algorithm.[18] Data sources such as electronic health records (EHR) or insurance claims can be rife with errors, systemic biases, and data gaps that might be particularly pronounced for minorities who do not receive optimal care.[19] In addition, datasets may be too small or not diverse enough because disadvantaged populations face health-care access barriers.[20] Moreover, if datasets capture historical health disparities, AI could learn to perpetuate patterns of discrimination.[21] These defects and others can make algorithms work poorly when they are deployed in the real world.[22]

This Article argues that algorithmic discrimination may violate Title VI of the Civil Rights Act and Section 1557 of the Affordable Care Act.[23] It further argues that algorithmic fairness must be a key element in designing, implementing, and validating AI. To that end, AI experts and policy makers must employ both technical and legal tools to promote algorithmic fairness. Among other recommendations, the Article calls for the reintroduction of the disparate impact

---

Preferences: Regulating the Dark Side of Personalized Transactions, 86 U. CHI. L. REV. 581, 583 (2019) (discussing the ways in which big data and artificial intelligence exploit human bias in online marketing and purchases).

14. *See infra* Section II.E (providing examples of algorithmic bias that generate discriminatory outcomes).

15. *See infra* notes 114-118 and accompanying text.

16. *See infra* notes 146-148 and accompanying text.

17. Strictly speaking, the algorithms at issue are called "supervised learning algorithms." Danilo Bzdok, Martin Krzywinski & Naomi Altman, *Machine Learning: Supervised Methods*, 15 NATURE METHODS 5, 5 (2018). For purposes of brevity, we will use the term "learning algorithm."

18. *See* Ignacio Cofone, *Algorithmic Discrimination Is an Information Problem*, 70 HASTINGS L.J. 1389, 1410 (2019) ("[A]n algorithmic decision-making process can only be as good as the data that it uses."); Ravi B. Parikh, Stephanie Teeple & Amol S. Navathe, *Addressing Bias in Artificial Intelligence in Health Care*, 322 JAMA 2377, 2377 (2019); *A.I. Bias in Healthcare: Human Pride, Machine Prejudice*, MED. FUTURIST (Sept. 19, 2019), https://medicalfuturist.com/a-i-bias-in-healthcare. See *infra* Part I, for a discussion of how AI works.

19. Parikh et al., *supra* note 18, at 2377.

20. *A.I. Bias in Healthcare: Human Pride, Machine Prejudice*, *supra* note 18.

21. Alvin Rajkomar, Michaela Hardt, Michael D. Howell, Greg Corrado & Marshall H. Chin, *Ensuring Fairness in Machine Learning to Advance Health Equity*, 18 ANNALS INTERNAL MED. 866, 866 (2018).

22. *Id.*

23. *See infra* Part III.

theory as a robust litigation tool in the health-care arena.[24]

Fairness is a complicated concept with no comprehensive or universally accepted definition in the AI context,[25] or for that matter, even in philosophy.[26] For the purposes of this Article, a useful conception includes three elements: equal outcomes, equal performance, and equal allocation.[27] More specifically, fairness requires that minority and majority groups benefit equally from AI in terms of patient outcomes, that AI is equally accurate for minority and non-minority patients, and that AI allocate resources proportionately to all groups.[28] We use the term "minority" broadly to include all individuals whom the civil rights laws aim to protect, including women, older people, and individuals with disabilities.[29] It is further important to understand that there are frequently competing notions of fairness that cannot all be fulfilled simultaneously.[30] For example, group fairness may be inconsistent with individual fairness.[31]

The remainder of this Article proceeds as follows. Part I discusses the use of AI in medicine and describes its benefits. Part II analyzes the discrimination-related pitfalls of AI. It explains measurement error, selection bias, and feedback loop bias and provides numerous examples of algorithmic discrimination in health care. It also discusses other discrimination risks associated with AI, including inequitable deployment of AI and the development of racially tailored medicine by which AI potentially recommends different treatments for members of different populations. Part III focuses on theories of discrimination that may apply to health-care inequities. These include intentional discrimination and disparate impact under Title VI of the Civil Rights Act of 1964 and Section 1557 of the Affordable Care Act. Under existing law, however, plaintiffs face many hurdles and may well

24. *See infra* Section III.A.

25. KEARNS & ROTH, *supra* note 10, at 69-72; Deborah Hellman, *Measuring Algorithmic Fairness*, 106 VA. L. REV. 811, 820-28 (2020); Alexandra Chouldechova & Aaron Roth, *A Snapshot of the Frontiers of Fairness in Machine Learning*, 63 COMM. ACM 82 (2020).

26. Reuben Binns, *Fairness in Machine Learning: Lessons from Political Philosophy*, 81 PROC. MACHINE LEARNING RES. 1, 1 (2018) ("Various definitions proposed in recent literature make different assumptions about what terms like discrimination and fairness mean and how they can be defined in mathematical terms.").

27. Rajkomar et al., *supra* note 21, at 868-69.

28. *Id.*

29. *See infra* notes 202-203 and accompanying text (describing protected classes under the civil rights statutes and listing relevant laws).

30. KEARNS & ROTH, *supra* note 10, at 84-86 (discussing "fairness fighting fairness" (capitalization in title omitted)); Hellman, *supra* note 25, at 827 (discussing circumstances in which it is "impossible to have parity between . . . groups along all the possible dimensions of fairness").

31. *See infra* text accompanying notes 378-387; *see also* Doaa Abu-Elyounes, *Contextual Fairness: A Legal and Policy Analysis of Algorithmic Fairness*, 2020 U. ILL. J.L. TECH. & POL'Y 1, 38 ("[F]rom both a policy and technical perspective, satisfying several notions of fairness simultaneously is mutually incompatible."); Jason R. Bent, *Is Algorithmic Affirmative Action Legal?*, 108 GEO. L.J. 803, 817-20 (2020) (discussing group fairness and individual fairness).

eschew litigation. Consequently, many discriminatory algorithms could be left unchallenged.

The last part of the paper transitions to formulating a series of recommendations. Part IV addresses legal intervention. First, it suggests adding an explicit private cause of action for disparate impact to Title VI and Section 1557. Second, it discusses and critiques the proposed Algorithmic Accountability Act. Third, it briefly addresses regulation by the Food and Drug Administration. Part V develops recommendations for improving algorithm design, validation, and monitoring processes. These include steps that both algorithm designers and algorithm users can implement. This section also cautions that AI experts, health-care providers, and patients must have realistic expectations about the degree of fairness they can achieve and may often need to prioritize among competing fairness goals. Part VI concludes.

I. ARTIFICIAL INTELLIGENCE IN MEDICINE

A. How AI Works

The term "artificial intelligence," (AI) refers to computers' ability to mimic human behavior and learn.[32] Learning is carried out with the aid of algorithms. An algorithm is a "computational procedure that takes some value, or set of values, as input and produces some value, or set of values, as output."[33] It is thus "a sequence of computational steps that transform the input into the output."[34] Users often rely on AI to help them make decisions or to make decisions for them.[35] They may input information about a patient's symptoms, medical history, and demographics and obtain a likely diagnosis or recommended treatment as the AI output.[36]

A large subfield of AI is machine learning (ML), which enables computers to "automatically detect patterns in data, and then use the uncovered patterns to predict future data or to perform decision-making tasks under uncertainty."[37]

32. IAN GOODFELLOW, YOSHUA BENGIO & AARON COURVILLE, DEEP LEARNING 1-2 (2016).
33. THOMAS H. CORMEN ET AL., INTRODUCTION TO ALGORITHMS 5 (3d ed. 2009).
34. *Id*.
35. *See infra* Section I.B. (discussing the benefits of AI).
36. Xiaoxuan Liu, *A Comparison of Deep Learning Performance against Health-Care Professionals in Detecting Diseases from Medical Imaging: a Systematic Review and Meta-Analysis*, 1 LANCET DIGITAL HEALTH E271, E271 (2019); *AI System Works with Physicians to Identify the Most Helpful Treatments for People Diagnosed with Depression*, MAYO CLINIC MAG., Fall 2019, https://mayomagazine.mayoclinic.org/2019/11/ai-system-works-with-physicians-to-identify-the-most-helpful-treatments-for-people-diagnosed-with-depression ("AI methodologies can discover patterns in a patient's data . . . that can explain unique characteristics of the specific patient, allowing for the right treatment to be chosen at the right time and right dose to achieve the therapeutic benefit.").
37. KEVIN P. MURPHY, MACHINE LEARNING: A PROBABILISTIC PERSPECTIVE 1 (2012); *see also* David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn about Machine*

Scientists train machine-learning algorithms to do analytical work by feeding them information, known as training data.[38] For example, scientists might show a learning algorithm a large number of tumor x-rays or scans, indicating which ones are and are not cancerous.[39] These designations of input data are known as labels.[40] The algorithm then learns to distinguish between benign and malignant masses based on patterns in the tumor images, so that it can identify cancerous tumors when shown new images.[41] Once data scientists determine that the algorithm's performance is satisfactory, it can be deployed to classify images with unknown labels.[42]

Some machine-learning algorithms are trained only once, and others continuously learn and adapt over time.[43] If an algorithm is adaptive and perpetually learns based on its real-world experience, the outputs it generates for particular inputs may change over time .[44]

Algorithms often examine large collections of information, known as "big data," from sources such as EHR databases or the Internet in order to unearth hidden knowledge or patterns.[45] "Big data" can be defined as data that is of high volume, variety, and velocity, the last referring to the speed with which it is generated.[46] In medicine, big data can come from a myriad of sources, including patients, health-care providers, insurers, manufacturers, the government, and even mobile devices such as smartphones and wearables.[47]

Learning, 51 UC DAVIS L. REV. 653, 671 (2017) ("Fundamentally, machine learning refers to an automated process of discovering correlations (sometimes alternatively referred to as relationships or patterns) between variables in a dataset, often to make predictions or estimates of some outcome."); Alvin Rajkomar, Jeffrey Dean & Isaac Kohane, *Machine Learning in Medicine*, 380 NEW ENG. J. MED. 1347, 1348 (2019) ("[I]n machine learning, a model learns from examples rather than being programmed with rules.").

38. *See* SHALEV-SHWARTZ & BEN-DAVID, *supra* note 4, at 13-14 (discussing "the statistical learning framework"); *see, e.g.*, Niha Beig et al., *Perinodular and Intranodular Radiomic Features on Lung CT Images Distinguish Adenocarcinomas from Granulomas*, 290 RADIOLOGY 783, 784 (2019) ("A machine classifier was trained on a cohort of 145 patients . . . .").

39. Beig et al., *supra* note 38, at 784.

40. Rajkomar et al., *supra* note 21, at 867.

41. Beig et al., *supra* note 38, at 792.

42. Rajkomar et al., *supra* note 21, at 867.

43. U.S. FOOD & DRUG ADMIN., PROPOSED REGULATORY FRAMEWORK FOR MODIFICATIONS TO ARTIFICIAL INTELLIGENCE/MACHINE LEARNING (AI/ML)-BASED SOFTWARE AS A MEDICAL DEVICE (SAMD) 3 (2019), https://www.fda.gov/files/medical%20devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf; *AMIA Supports, Encourages Further Refinement of FDA AI/Machine Learning Regulatory Framework*, *supra* note 6.

44. U.S. FOOD & DRUG ADMIN., *supra* note 43, at 3.

45. JIAWEI HAN, MICHELINE KAMBER & JIAN PEI, DATA MINING: CONCEPTS AND TECHNIQUES 8 (3d ed. 2012).

46. SHARONA HOFFMAN, ELECTRONIC HEALTH RECORDS AND MEDICAL BIG DATA: LAW AND POLICY 111 (2016).

47. Nathan Cortez, *Substantiating Big Data in Health Care*, 14 I/S: J.L. & POL'Y FOR INFO.

Algorithms have different degrees of transparency and explainability.[48] In some cases, they are opaque because they rely on extremely complex rules, and even their programmers are unsure of exactly how they work in particular instances.[49] Some experts describe clinician reliance on nontransparent, noninterpretable algorithms as "black-box medicine."[50]

B. The Benefits of AI in Medicine

AI can generate many benefits by allowing experts to analyze very large data sets quickly and efficiently, potentially delivering improved health care at a lower cost.[51] If computers rather than humans do some of the work, health-care providers can lower staffing costs and accomplish tasks more quickly.[52]

AI is valuable for physicians, researchers, and policy makers.[53] Learning algorithms can help doctors predict which patients are likely to have either poor or successful treatment outcomes and to adjust medical decisions accordingly.[54] AI may also help identify high-risk individuals whom doctors should screen regularly for specific illnesses.[55] Likewise, AI can analyze EHRs in order to determine which patients are good candidates for clinical trials so that researchers can recruit them.[56] AI can further expedite medical discoveries as learning algorithms examine big data and discern previously unknown patterns, connections, and causal effects.[57]

SOC'Y 61, 63-65 (2017) (discussing the breadth of big data sources).

48. Milena A. Gianfrancesco, Suzanne Tamang, Jinoos Yazdany & Gabriela Schmajuk, *Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data*, 178 JAMA INTERNAL MED. 1544, 1546 (2018) ("Certain machine learning models . . . are less transparent than others . . . and therefore may be harder to interpret."); W. Nicholson Price II, *Artificial Intelligence in the Medical System: Four Roles for Potential Transformation*, 21 YALE J.L. & TECH. (SPECIAL ISSUE) 122, 124 (2019) (referring to "explainable algorithms versus black-box methods").

49. Tokio Matsuzaki, *Ethical Issues of Artificial Intelligence in Medicine*, 55 CAL. W. L. REV. 255, 269 (2018) ("One concern is that AI decision-making . . . often has no transparency. This means that doctors and patients are not able to know how the Al system reached the decision."); W. Nicholson Price II, *Regulating Black-Box Medicine*, 116 MICH. L. REV. 421, 430 (2017).

50. Price, *supra* note 49, at 429; *see* Eric J. Topol, *High-Performance Medicine: The Convergence of Human and Artificial Intelligence*, 25 NATURE MED. 44, 51 (2019); Effy Vayena, Alessandro Blasimme & I. Glenn Cohen, *Machine Learning in Medicine: Addressing Ethical Challenges*, PLOS MED., Nov. 2018, art. no. e1002689, at 3.

51. Alicia Phaneuf, *Use of AI in Healthcare & Medicine Is Booming – Here's How the Market Is Benefiting from AI in 2020 and Beyond*, BUS. INSIDER (July 31, 2019, 10:48 AM), https://www.businessinsider.com/artificial-intelligence-healthcare.

52. *Id.* (noting that "30% of healthcare costs are associated with administrative tasks").

53. EWOUT W. STEYERBERG, CLINICAL PREDICTION MODELS 1-3, 11 (2009).

54. *Id.* at 11.

55. *Id.*

56. Stefan Harrer, Pratik Shah, Bhavna Antony & Jianying Hu, *Artificial Intelligence for Clinical Trial Design*, 40 TRENDS PHARMACOLOGICAL SCI. 577, 580 (2019).

57. W. Nicholson Price II, *Black Box Medicine*, 28 HARV. J.L. & TECH. 419, 421 (2015).

Public health authorities and health-care providers are now using AI to address the COVID-19 pandemic.[58] Researchers hope that AI will facilitate tracking the disease and predicting how and where it will spread.[59] They are also undertaking initiatives to develop and understand the potential of AI tools for the diagnosis of patients and prediction of their disease course.[60] To that end, experts are training AI models to diagnose COVID-19 using chest images and are developing AI tools to predict which COVID-19 patients will become severely ill.[61] Likewise, a large Israeli health maintenance organization is using AI to help identify which of its participants is most at risk of severe COVID-19 symptoms.[62]

Many hope that AI will also accelerate the development of a vaccine and the discovery of effective treatments.[63] To illustrate, machine learning led researchers to conclude that the drugs atazanavir and baricitinib could possibly be repurposed to treat COVID-19.[64]

Finally, AI has been harnessed to enforce public health orders. According to one report, "At airports and train stations across China, infrared cameras are used to scan crowds for people with high temperatures. They are sometimes used with a facial recognition system, which can pinpoint the individual with a high temperature and whether he or she is wearing a surgical mask."[65]

Experts acknowledge, however, that AI has been of limited efficacy in the COVID-19 battle thus far. One reason is that AI algorithms require large amounts of data for training purposes, and obtaining adequate data can be costly and work-

58. Marcello Ienca & Effy Vayena, *On the Responsible Use of Digital Data to Tackle the COVID-19 Pandemic*, 26 NATURE MED., 463, 463 (2020).

59. Wim Naudé, *Artificial Intelligence vs. COVID-19: Limitations, Constraints and Pitfalls*, 35 AI & SOC'Y 761, 761-62 (2020).

60. *Id.*

61. Xiangao Jiang et al., *Towards an Artificial Intelligence Framework for Data-Driven Prediction of Coronavirus Clinical Severity*, 63 COMPUTERS, MATERIALS & CONTINUA 537 (2020); Naudé, *supra* note 59, at 762-63.

62. Will Douglas Heaven, *Israel Is Using AI to Flag High-Risk COVID-19 Patients*, MIT TECH. REV. (Apr. 24, 2020), https://www.technologyreview.com/2020/04/24/1000543/israel-ai-prediction-medical-testing-data-high-risk-covid-19-patients.

63. Naudé, *supra* note 59.

64. Bo Ram Beck, Bonggun Shin, Yoonjung Choi, Sungsoo Park & Keunsoo Kang, *Predicting Commercially Available Antiviral Drugs that May Act on the Novel Coronavirus (SARS-CoV-2) Through a Drug-Target Interaction Deep Learning Model*, 18 COMPUTATIONAL & STRUCTURAL BIOTECHNOLOGY J. 784 (2020); Justin Stebbing, Anne Phelan, Ivan Griffin, Catherine Tucker, Olly Oechsle, Dan Smith & Peter Richardson, *COVID-19: Combining Antiviral and Anti-Inflammatory Treatments*, 20 LANCET 400, 400-01 (2020).

65. Andy Chun, *In a Time of Coronavirus, China's Investment in AI Is Paying Off in a Big Way*, S. CHINA MORNING POST (Mar. 18, 2020, 10:00 AM), https://www.scmp.com/comment/opinion/article/3075553/time-coronavirus-chinas-investment-ai-paying-big-way.

intensive.[66] Most studies to date have drawn information from small datasets.[67] In addition, in the United States, patients' records are often fragmented and located at different facilities that do not have interoperable[68] EHRs.[69] Thus, it could be impossible to obtain a sufficiently large and representative patient dataset to allow for accurate predictions about disease prognosis.[70] In the area of surveillance, thermal scanning can be hampered by people wearing eyeglasses "because scanning the inner tear duct gives the most reliable indication" of fever from a distance.[71]

II. DISCRIMINATION-RELATED PITFALLS OF AI

The above-described problems with employing AI to combat COVID-19 provide a preview of the shortcomings of AI more generally. AI can often generate incorrect results. In some instances, AI defects can have discriminatory effects and can severely disadvantage certain groups of patients.[72] Flawed outcomes can stem from a number of problems. This part focuses on three key problems. First, the data themselves can be incomplete or incorrect,[73] thus causing measurement error.[74] Second, the data set that trains the algorithm may be under-inclusive or otherwise skewed (e.g., containing records of only White males) so that AI outcomes are not generalizable to the population as a whole.[75] Third, the training data may capture historical patterns of discrimination, causing the algorithm to perpetuate the inequitable treatment. This problem is called feedback loop bias.[76] The section also briefly discusses other sources of uncertainty.

---

66. Naudé, *supra* note 59, at 761-63; Don Roedner, *Why 96% of Enterprises Face AI Training Data Issues*, DATACONOMY (July 30, 2019), https://dataconomy.com/2019/07/why-96-of-enterprises-face-ai-training-data-issues.

67. *Id.*

68. Interoperability means "the ability for systems to exchange data and operate in a coordinated, seamless manner." BIOMEDICAL INFORMATICS: COMPUTER APPLICATIONS IN HEALTH CARE AND BIOMEDICINE 952 (Edward H. Shortliffe & James J. Cimino eds., 3d ed. 2006).

69. HOFFMAN, *supra* note 46, at 54-55; *see also* Heaven, *supra* note 62.

70. Heaven, *supra* note 62.

71. Naudé, *supra* note 59.

72. Ian A. Scott, *Hope, Hype and Harms of Big Data*, 49 INTERNAL MED. J. 126, 127 (2019).

73. Vayena et al., *supra* note 50, at 2 (discussing "cases in which the data sources themselves do not reflect true epidemiology within a given demographic, as for instance in population data biased by the entrenched overdiagnosis of schizophrenia in African Americans").

74. Timo B. Brakenhoff, Maarten van Smeden, Frank L.J. Visseren & Rolf H.H. Groenwold, *Random Measurement Error: Why Worry? An Example of Cardiovascular Risk Factors*, PLOS ONE, Feb. 2018, art. no. e0192298.

75. Vayena et al., *supra* note 73, at 2 ("Such an algorithm would make poor predictions, for example, among younger black women.").

76. David Casacuberta, *Bias in a Feedback Loop: Fuelling Algorithmic Injustice*, CCCB LAB, (May 9, 2018) http://lab.cccb.org/en/bias-in-a-feedback-loop-fuelling-algorithmic-injustice.

A. Measurement Errors

Big data that is used to train machine-learning algorithms can have missing and incorrect information.[77] Indeed, some patients' records contain a plethora of erroneous and misleading data.[78] Measurement errors can be defined as "the difference between the [actual] quantity of interest and the measured value."[79] Poor data quality inevitably leads to poor AI algorithm performance, sometimes expressed as the "garbage in-garbage out" principle.[80]

EHRs of minorities and economically disadvantaged individuals might be particularly vulnerable to missing data.[81] Members of vulnerable populations may receive health care infrequently because they are uninsured, have no transportation or childcare, or face other barriers.[82] They also often lack a primary care physician and visit multiple facilities when they do seek medical attention, so that their records are fragmented and do not contain comprehensive information.[83] Because of data gaps, AI may not recognize such patients as having the diseases or health risks that the algorithm is designed to identify.[84]

Furthermore, low-income individuals may seek care at teaching clinics where practitioners are less meticulous about recordkeeping.[85] Data gathered from these facilities may have more errors than data from facilities frequented by higher-income patients.[86]

B. Selection Bias

The word "bias" has different meanings in different contexts. Human bias is

77. Scott, *supra* note 72, at 127 (discussing numerous potential shortcomings of big data); Nilay D. Shah, Ewout W. Steyerberg & David M. Kent, *Big Data and Predictive Analytics: Recalibrating Expectations*, 320 JAMA 27, 28 (2018); Topol, *supra* note 50, at 51.

78. HOFFMAN, *supra* note 46, at 23-28.

79. Jessie K. Edwards & Alexander P. Keil, *Measurement Error and Environmental Epidemiology: A Policy Perspective*, 4 CURRENT ENVTL. HEALTH REP. 79, 79 (2017).

80. P. Elliott Miller et al., *Predictive Abilities of Machine Learning Techniques May Be Limited by Dataset Characteristics: Insights from the UNOS Database*, 25 J. CARDIAC FAILURE 479, 482 (2019) ("Our results raise the notion that large clinical datasets might lack the accuracy and granularity needed for machine learning methodologies to uncover unique associations."); Rajkomar et al., *supra* note 37, at 1355; Kun-Hsing Yu & Isaac S. Kohane, *Framing the Challenges of Artificial Intelligence in Medicine*, 28 BMJ QUALITY & SAFETY 238, 239 (2019).

81. Gianfrancesco et al., *supra* note 48, at 1545.

82. *Id.*

83. *Id.*

84. *Id.*

85. *Id.* at 1546.

86. *Id.*; Rajkomar et al., *supra* note 21, at 867 (providing the example of "predicting the onset of clinical depression in environments where protected groups have been systematically misdiagnosed").

prejudice or "unreasonably hostile feelings or opinions about a social group."[87] By contrast, algorithmic bias is present when an AI model produces results that are unintended by its creators because of its training data's shortcomings or because it is applied to an unanticipated patient population.[88]

One reason for enthusiasm about AI is the hope that it will diminish human bias in health care.[89] It is natural for human beings to have certain prejudices rooted in their background and upbringing, and this may at times influence diagnosis and treatment decisions.[90] Objective algorithmic analysis should ideally diminish or eliminate human bias.[91] However, AI algorithms are subject to their own bias problems.[92]

Big data can be subject to selection bias. Selection bias can occur if the subset of individuals represented in the training data is not representative of the patient population of interest.[93] If the data used to train a learning algorithm comes from a health system that serves particular populations (e.g., disproportionately wealthy or low-income people) but not others, the algorithm's predictions may not be generalizable to all patients of interest.[94] Several scholars have noted the following:

> Big Data has not captured certain marginalized demographics. Particularly concerning are racial minorities, people with low socioeconomic status, and immigrants. Many of the people missing from the data that come from sources such as Internet history, social media presence, and credit-card use are also missing from other sources of Big Data, such as electronic health records (EHRs) and genomic databases. The factors responsible for these gaps are diverse and include lack of insurance and the

87. *Bias*, DICTIONARY.COM, https://www.dictionary.com/browse/bias (last visited May 16, 2020); *see also* Parikh et al., *supra* note 18, at 2377 ("An AI algorithm that learns from historical electronic health record (EHR) data and existing practice patterns may not recommend testing for cardiac ischemia for an older woman, delaying potentially life-saving treatment.").

88. Irene Y. Chen, Peter Szolovits & Marzyeh Ghassemi, *Can AI Help Reduce Disparities in General Medical and Mental Health Care?*, 21 AMA J. ETHICS E167, E168 (2019); *see also* Jessica K. Paulus & David M. Kent, *Predictably Unequal: Understanding and Addressing Concerns that Algorithmic Clinical Prediction May Increase Health Disparities*, 3 NPJ DIGITAL MED., art. no. 99, 2020, at 4 (defining algorithmic bias in terms of "issues related to model design, data and sampling that may disproportionately affect model performance in a certain subgroup").

89. Gianfrancesco et al., *supra* note 48, at 1544.

90. *Id.*

91. *Id.*

92. KEARNS & ROTH, *supra* note 10, at 57-63.

93. Sharona Hoffman & Andy Podgurski, *The Use and Misuse of Biomedical Data: Is Bigger Really Better?*, 39 AM. J.L. & MED. 497, 521-23 (2013) (discussing selection bias).

94. Craig Konnoth, *Health Information Equity*, 165 U. PA. L. REV. 1317, 1361 (2017) (asserting that "relying on data that is biased towards certain social groups can have problematic effects").

> inability to access health care, to name just two . . . .[95]

Sadly, many examples of selection bias exist in the health-care field. An analysis of 2,511 genome-mapping studies from around the world found that eighty-one percent of participants were of European descent.[96] A 2014 study found that over the prior twenty years the cancer survival gap between White and African American patients did not shrink, and the researchers attributed the persistent disparity in part to the relative dearth of information about the efficacy of treatment in the Black population.[97] Unfortunately, African Americans are thirty percent less likely than Whites to participate in clinical trials.[98]

Selection bias may be particularly acute if the size of the study sample is small.[99] The sample may contain few if any data subjects who belong to particular disadvantaged groups.[100] An algorithm may misinterpret a lack of information about minorities as a lack of disease burden and consequently generate inaccurate predictions for the affected groups.[101]

C. Feedback Loop Bias

Bias can be rooted in historical patterns of discrimination. For example, police forces may send more officers to minority neighborhoods because they assume that these neighborhoods are crime-ridden.[102] With more officers present, the police will discover more crimes and make more arrests than in other areas, even if there are other locations with an equal or larger amount of crime.[103] If the arrest figures are fed into an algorithm designed to determine optimal police force allocation, the algorithm may learn that it is advisable to send more police to the minority neighborhoods because they have more crime than elsewhere. The

---

95. Sarah E. Malanga, Jonathan D. Loe, Christopher T. Robertson & Kenneth S. Ramos, *Who's Left Out of Big Data? How Big Data Collection, Analysis, and Use Neglect Populations Most in Need of Medical and Public Health Research and Interventions*, *in* BIG DATA, HEALTH LAW, AND BIOETHICS 98, 98-99 (I. Glenn Cohen, Holly Fernandez Lynch, Effy Vayena & Urs Gasser eds., 2018) (footnote omitted).

96. Alice B. Popejoy & Stephanie M. Fullerton, *Genomics Is Failing on Diversity*, 538 NATURE 161, 162 (2016).

97. Ayal A. Aizer et al., *Lack of Reduction in Racial Disparities in Cancer-Specific Mortality over a 20-Year Period*, 120 CANCER 1532, 1538 (2014).

98. *Id.*

99. Gianfrancesco et al., *supra* note 48, at 1545-46.

100. Rajkomar et al., *supra* note 21, at 867.

101. Gianfrancesco et al., *supra* note 48, at 1545-46; *A.I. Bias in Healthcare: Human Pride, Machine Prejudice*, *supra* note 18 ("[T]hese distorted datasets would be the starting points for A.I. development.").

102. KEARNS & ROTH, *supra* note 10, at 92; Chouldechova & Roth, *supra* note 25, at 84.

103. KEARNS & ROTH, *supra* note 10, at 92.

algorithm may thus make a recommendation that will perpetuate discrimination.[104]

Likewise, some patients may receive less intensive care because of their demographic characteristics rather than because of their medical needs.[105] For example, one study concluded that women are less likely than men to receive lipid-lowering medications, in-hospital procedures, and optimal care at hospital discharge, even though they are more likely to suffer hypertension and heart failure.[106] The training data used to develop algorithms relating to these conditions typically do not indicate that women received inadequate treatment compared to men and should have had additional interventions. Consequently, the algorithm will likely learn to recommend less intensive care for women thereby perpetuating and exacerbating the undertreatment problem.

D. Algorithmic Uncertainty

Medical AI users must accept that AI involves a degree of uncertainty.[107] At times, the data available for prediction will not completely characterize the class of interest.[108] Learning algorithms may be affected by incomplete observability of relevant data or incomplete modeling because not all observed information is considered in the algorithmic analysis.[109]

It is often more efficient and practical to use a simple rule with a degree of uncertainty rather than a complex one with more certainty. For example, the rule "most birds fly" is uncomplicated and highly functional. By contrast, the rule "birds fly, except for very young birds that have not yet learned to fly, sick or injured birds that have lost the ability to fly, flightless species of birds including the cassowary, ostrich and kiwi . . . " is costly to develop, maintain, and convey and will still be vulnerable to failures.[110]

A machine-learning algorithm may adopt a simple rule for a given problem and data set if it performs adequately on the training data.[111] Discrimination may occur if all or part of a minority group is mishandled by the rule, which is more likely if that group or subgroup is small.[112] In the example above, ostriches would

104. Chouldechova & Roth, *supra* note 25, at 87 ("[S]ince police are likely to make more arrests in more heavily policed areas, using arrest data to predict crime hotspots will disproportionately concentrate policing efforts on already over-policed communities.").

105. Gianfrancesco et al., *supra* note 48, at 1546.

106. Shanshan Li et al., *Sex and Race/Ethnicity-Related Disparities in Care and Outcomes After Hospitalization for Coronary Artery Disease Among Older Adults*, 9 CIRCULATION: CARDIOVASCULAR QUALITY & OUTCOMES S36, S38 (2016).

107. GOODFELLOW ET AL., *supra* note 32, at 52.

108. *Id.*

109. *Id.* at 52-53.

110. *Id.* at 53.

111. *Id.*

112. *See supra* notes 99-101 and accompanying text.

potentially suffer discrimination as a result of the rule "most birds fly" because their special circumstances would not be addressed.[113]

E. Examples of Algorithmic Bias and the Risk of Discrimination in Health Care

Algorithmic bias can function in unanticipated ways that lead to discrimination against particular groups. This concern is not merely hypothetical.

A widely publicized example is an algorithm commonly used by health systems to identify patients who could benefit from "high risk management" and who should thus receive special attention.[114] The algorithm exhibited significant racial bias, and the problem was rooted in its use of past health-care costs as a proxy for medical risks or conditions.[115] Because racial minorities often face health-care access barriers, they frequently spend less money on health care than others. Thus, their history of expenditures may not reflect their true health status or indicate the care they should have obtained if it were available to them. Economically disadvantaged individuals who utilize medical services infrequently and at low cost often have acute medical problems such as severe hypertension, diabetes, renal failure, anemia, and high cholesterol, which are prevalent in African American communities.[116] Yet, when the algorithm was deployed, its risk scores failed to reveal that African Americans were often sicker than their White counterparts who received referrals for special services.[117] Thus, the algorithm favored Whites over African Americans with greater needs. Flawed algorithms were likely used by health systems that served up to 200 million Americans.[118]

Winterlight Labs, a Toronto-based startup, built a machine-learning tool to distinguish individuals with Alzheimer's disease from those without the ailment based on short samples of their speech in response to a picture-description task.[119] It turned out that the technology was effective only for native English speakers of a specific Canadian dialect and that it misdiagnosed others.[120] It misinterpreted

---

113. GOODFELLOW ET AL., *supra* note 32, at 53.

114. Ziad Obermeyer, Brian Powers, Christine Vogeli & Sendhil Mullainathan, *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, 366 SCI. 447, 447, 449 (2019); Charlotte Jee, *A Biased Medical Algorithm Favored White People for Health-Care Programs*, MIT TECH. REV. (Oct. 25, 2019), https://www.technologyreview.com/f/614626/a-biased-medical-algorithm-favored-white-people-for-healthcare-programs.

115. Obermeyer et al., *supra* note 114, at 447; Jenna Wiens et al., *Diagnosing Bias in Data-Driven Algorithms for Healthcare*, 26 NATURE MED. 25, 25-26 (2020).

116. Obermeyer et al., *supra* note 114, at 447-50.

117. *Id.* at 447, 449.

118. *Id.* at 447.

119. Kathleen C. Fraser, Jed A. Meltzer & Frank Rudzicz, *Linguistic Features Identify Alzheimer's Disease in Narrative Speech*, 49 J. ALZHEIMER'S DISEASE 407, 407 (2016) (asserting that the researchers "obtain[ed] state-of-the-art classification accuracies of over 81% in distinguishing individuals with [Alzheimer's disease] from those without").

120. Dave Gershgorn, *If AI Is Going to be the World's Doctor, It Needs Better Textbooks*,

pauses, mispronunciations, and uncertainty rooted in language barriers as indicators of cognitive decline.[121]

Two commentators focused on machine learning that created programs to analyze images of skin lesions and to distinguish between malignant and benign moles.[122] They noted that the "patient data are heavily collected from fair-skinned populations in the United States, Europe, and Australia."[123] Consequently, they worry that the algorithms will not perform well on images of people of color, which could lead to misdiagnoses.[124]

Even algorithms that learn from accurate, fully representative data can inadvertently perpetuate discrimination. Epic, a major vendor of health information systems, released an AI tool to help medical practices identify patients who are likely to miss appointments.[125] The tool, which was built into Epic's EHRs, provided a numerical estimate of no-show likelihood, thereby encouraging clinicians to book a second patient into certain slots.[126] Because one of the input variables was prior no-shows, researchers found that the scores correlated to socio-economic status.[127] People living in poverty tend more often to have transportation or childcare problems or difficulty taking time off from work.[128] Therefore, when they did arrive at appointments, they were more likely to find a second patient booked at the same time and to receive rushed and inadequate care regardless of the complexity of their health problems.[129]

As AI technology comes into even greater use in health care, bias problems may well proliferate. Commentators have contemplated numerous other potential AI initiatives that could be tainted by bias and perpetuate discrimination.[130] To illustrate, because African American patients receive, on average, less pain treatment than Caucasians, an AI system trained on EHRs might learn to recommend lower doses of pain drugs to African American patients regardless of their need for relief.[131] As a second example, research has shown that African

---

QUARTZ (Sept. 6, 2018), https://qz.com/1367177/if-ai-is-going-to-be-the-worlds-doctor-it-needs-better-textbooks.

121. *Id.*

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

123. *Id.*

124. *Id.*

125. Sara G. Murray, Robert M. Wachter & Russell J. Cucina, *Discrimination by Artificial Intelligence in a Commercial Electronic Health Record—A Case Study*, HEALTH AFF. (Jan. 31, 2020), https://www.healthaffairs.org/do/10.1377/hblog20200128.626576.

126. *Id.*

127. *Id.*

128. *Id.*

129. *Id.*

130. *See, e.g.*, Rajkomar et al., *supra* note 21, at 867.

131. Price, *supra* note 1.

American women with chest pain are less likely to have cardiac catheterizations than are White men with the same symptoms.[132] An algorithm designed to identify patients who should undergo the procedure may well recommend the treatment for African American women at an inappropriately low rate.[133] Likewise, transgender individuals may suffer discrimination if algorithms require a binary sex input that accepts only male or female designations.[134] Algorithms may generate treatment recommendations that are incorrect or a poor fit for their needs and circumstances.

F. Other Discrimination Risks Associated with AI

1. Inequitable Deployment of AI

AI algorithms could perpetuate discrimination in other ways as well. Despite the concerns articulated above, AI is beneficial for many patients.[135] Sound learning algorithms that are free of bias can help doctors make accurate diagnostic and treatment decisions.[136] For example, they can identify patients at risk of complications or poor outcomes so that doctors can tailor their therapies accordingly.[137]

Yet resource-poor health-care providers that serve largely disadvantaged populations may not have the means to obtain and use sophisticated AI technology.[138] Commentators have noted that "informatics interventions are disproportionately available to well-off, educated, young, and urban patients and to urban and academic medical centers."[139] Health disparities will be exacerbated if low-income, minority, and rural populations are deprived of the benefits of AI technology that improve outcomes in other communities.[140]

2. Racially Tailored Medicine

Some learning algorithms deliberately adjust outputs on the basis of race in

132. Kevin A. Schulman et al., *The Effects of Race and Sex on Physicians' Recommendations for Cardiac Catheterization*, 340 NEW ENG. J. MED. 618, 618 (1999).

133. Rajkomar et al., *supra* note 21, at 869.

134. Rachel Metz, *AI Software Defines People as Male or Female. That's a Problem*, CNN (Nov. 21, 2019, 11:32 AM ET), https://www.cnn.com/2019/11/21/tech/ai-gender-recognition-problem/index.html.

135. *See supra* Section I.B.

136. *See supra* notes 51-55 and accompanying text.

137. *Id.*

138. Rajkomar et al., *supra* note 21, at 868.

139. Tiffany C. Veinot, Hannah Mitchell & Jessica S. Ancker, *Good Intentions Are Not Enough: How Informatics Interventions Can Worsen Inequality*, 25 J. AM. MED. INFORMATICS ASS'N 1080, 1081 (2018).

140. *See supra* text accompanying note 27 (including equal allocation of resources in the definition of AI fairness).

an effort to better tailor therapies to particular populations.[141] For example, a recent prostate cancer study showed that AI analysis of digital images can detect differences in the appearance of cancer between African American and White patients.[142] Researchers employed a learning algorithm to look for patterns in images of both the tumor itself and the tissue outside the tumor, known as the stroma.[143] They believe that "considering population-specific information . . . has the potential to substantially improve accuracy of prognosis and risk stratification in . . . [African American] patients with prostate cancer."[144] Similar studies are planned with respect to breast cancer.[145]

A 2020 *New England Journal of Medicine* article revealed that many clinical algorithms include "race corrections."[146] They do so because their developers believe that adjustments are justified by analyses of historical data about patient attributes and clinical outcomes.[147] The article provides the following examples:

- An American Heart Association heart failure risk score algorithm assigns three extra points to patients identified as "nonblack" so that Black patients are categorized as being at lower risk of death.
- An algorithm designed to assess kidney function reports higher estimated glomerular filtration rates for patients identified as Black, suggesting that they have better kidney function.
- The Kidney Donor Risk Index indicates a higher risk of graft failure for donors identified as Black, thus marking Black individuals as less suitable donors.
- The Vaginal Birth after Cesarean algorithm predicts a lower likelihood of vaginal birth success for African American and Hispanic women who have had a previous Cesarean, making it more likely that they will undergo further surgeries.
- An algorithm that predicts the likelihood of kidney stones in emergency

141. Hersh K. Bhargava et al., *Computationally Derived Image Signature of Stromal Morphology Is Prognostic of Prostate Cancer Recurrence Following Prostatectomy in African American Patients*, 26 CLINICAL CANCER RES. 1915, 1915 (2020); Darshali A. Vyas, Leo G. Eisenstein & David S. Jones, *Hidden in Plain Sight—Reconsidering the Use of Race Correction in Clinical Algorithms*, 383 NEW ENG. J. MED. 874 (2020).

142. Bhargava et al., *supra* note 141.

143. *Id.* at 1921 ("[T]his study is the first to show the role of stromal features in prostate cancer . . . .").

144. *Id.* at 1915.

145. Case Western Reserve University, *AI Reveals Differences in Appearance of Cancer Tissue between Racial Populations*, EUREKALERT (Mar. 5, 2020), https://www.eurekalert.org/pub_releases/2020-03/cwru-ard030520.php.

146. Vyas et al., *supra* note 141, at 874. *See also* Jessica P. Cerdeña, Marie V. Plaisime & Jennifer Tsai, *From Race-Based to Race-Conscious Medicine: How Anti-Racist Uprisings Call Us to Act*, 396 LANCET 1125, 1125-27 (2020).

147. Vyas et al., *supra* note 141, at 879.

> room patients with flank pain adds three points out of a possible thirteen to nonblack patients, thus assessing Black patients as less likely to have kidney stones.[148]

All of these algorithmic outcomes could divert resources away from African American patients or otherwise disadvantage them.[149]

Paying attention to population differences can potentially enable physicians to treat patients more effectively. The prostate cancer researchers discussed above aim to predict cancer recurrence more accurately and thus to determine which patients should receive aggressive therapies.[150] The developers of the other algorithms listed above believe that they are enhancing the accuracy of diagnoses and treatment recommendations based on empirical evidence.[151] Indeed, renowned studies, such as the Framingham Heart Study, which established now widely accepted risk factors for heart disease, have been criticized for lacking diverse study populations.[152] The Framingham Heart study derived its data from a small, middle-class town in Massachusetts with a predominantly White population of Western European descent.[153] Subsequent studies have explored racial/ethnic differences in cardiovascular disease and its risk factors and found that population-specific insights are informative for purposes of implementing preventive care.[154]

Nevertheless, racially tailored medicine carries its own serious risks,[155] and

---

148. *Id.* at 874-79; *see also* Neil R. Powe, *Black Kidney Function Matters: Use or Misuse of Race?*, 324 JAMA 737, 737 (2020); Keith Churchwell et al., *Call to Action: Structural Racism as a Fundamental Driver of Health Disparities: A Presidential Advisory from the American Heart Association*, 142 CIRCULATION e1, e11 (2020) (urging the American Heart Association to "reconsider when and how to include race/ethnicity and social determinants measures in risk calculators"); James A. Diao et al., *Clinical Implications of Removing Race From Estimates of Kidney Function*, JAMA (Dec. 2, 2020), doi:10.1001/jama.2020.22124 (noting that many U.S. medical centers are abandoning the algorithmic race adjustment for kidney function and that doing so may increase chronic kidney disease diagnoses among Black adults and improve access to care but may also exclude certain kidney donors and impact drug therapies).

149. Vyas et al., *supra* note 141, at 874 ("Many of these race-adjusted algorithms guide decisions in ways that may direct more attention or resources to white patients than to members of racial and ethnic minorities").

150. Case Western Reserve University, *supra* note 145.

151. Vyas et al., *supra* note 141, at 879 (explaining that "researchers followed defensible empirical logic," adjusting for race in their models after performing regression analyses on clinical data sets and finding that "minority patients routinely have different health outcomes from white patients").

152. Sandeep Jauhar, Opinion, *The Heart Disease Conundrum*, N.Y. TIMES (Nov. 28, 2015), https://www.nytimes.com/2015/11/29/opinion/sunday/the-heart-disease-conundrum.html ("Framingham risk models do not tell the whole story for nonwhite ethnic groups.").

153. *Id.*

154. *See* Crystel M. Gijsberts et al., *Race/Ethnic Differences in the Associations of the Framingham Risk Factors with Carotid IMT and Cardiovascular Events*, PLOS ONE, July, 2015, art. no. e0132321, at 2.

155. *See generally* Sharona Hoffman, *"Racially-Tailored" Medicine Unraveled*, 55 AM. U. L.

some institutions have ceased using algorithms that adjust for race.[156] First, race[157] in scientific studies is generally determined through subjects' self-reported identification.[158] Yet, millions of Americans are of mixed race.[159] They currently constitute up to 6.9 percent of the population,[160] and experts project that their number will triple by 2060.[161] Individuals may identify as being of a particular race but have a multi-racial background or even appear to be of different ancestry.[162] Counting such persons as members of a single race could skew research results.

Second, treating physicians attempting to apply algorithmically generated diagnostic or treatment recommendations may face a conundrum when their patients are of mixed background.[163] If the guidelines are different depending on ancestry, which ones should a doctor use for a patient who is multiracial?[164]

Third, differences that are perceived as "racial" in truth are sometimes socioeconomic.[165] For example, the health status of some (but certainly not all) African American patients might be affected by poverty or stress.[166] It would thus be inappropriate to make generalizations about all African Americans, and instead, researchers should focus on the impact of financial resources or emotional wellbeing.[167]

Fourth, so-called racial distinctions may in reality be genetic differences.[168] A

REV. 395 (2005).

156. Powe, *supra* note 148, at 737 ("A number of institutions have taken steps to remove the use of race in equations involving estimated glomerular filtration rates (eGFRs).").

157. Race is in itself a problematic term and is widely perceived as a social construct. *See* Sharona Hoffman, *Is There a Place for "Race" as a Legal Concept*, 36 ARIZ. ST. L.J. 1093, 1093 (2004). We prefer more precise terms such as color, ancestry, national origin, and others. *Id.* at 1159. We refer to race here because that is the language used in the relevant scientific studies.

158. Bhargava et al., *supra* note 141, at 1916 ("Patient race was self-reported.").

159. Kim Parker, Juliana Menasce Horowitz, Rich Morin & Mark Hugo Lopez, *Chapter 2: Counting Multiracial Americans*, PEW RES. CTR. (June 11, 2015), https://www.pewsocialtrends.org/2015/06/11/chapter-2-counting-multiracial-americans (finding that "6.9% of Americans 18 or older have a multiracial background" but noting that only "2.1% of adult Americans . . . said they were [of] two or more races in the Census Bureau's 2013 American Community Survey").

160. *Id.*

161. Marisa Franco, *What Racial Discrimination Will Look Like in 2060*, SCI. AM. BLOGS (Nov. 29, 2019), https://blogs.scientificamerican.com/voices/what-racial-discrimination-will-look-like-in-2060.

162. Nicholas Vargas & Kevin Stainback, *Documenting Contested Racial Identities Among Self-Identified Latina/os, Asians, Blacks, and Whites*, 60 AM. BEHAV. SCIENTIST 442, 442 (2016).

163. Vyas et al., *supra* note 141, at 880 ("Guidelines are silent on such issues—an indication of their inadequacy.").

164. *Id.*

165. Vyas et al., *supra* note 141 at 879-80.

166. *Id.*

167. *Id.*

168. *See* Hoffman, *supra* note 155, at 419-21 (providing the examples of cystic fibrosis, sickle cell anemia, and the BRCA1 and BRCA2 mutations that are associated with breast and ovarian

particular genetic mutation that affects disease vulnerability or treatment response might be more common in one racial group than in others.[169] Nevertheless, many members of the race in question will not have the genetic abnormality while some people with different ancestries will.[170] For example, sickle cell anemia affects not only African Americans, but also people with ancestors from Greece, Sicily, and the Arabian Peninsula, and it is not prevalent among Black South Africans.[171] Indeed, experts note that there are more genetic variations within racial groups than among them.[172] Consequently, algorithms that treat all patients identified as being of a particular race the same could provide numerous individuals with inadequate and inappropriate care and severely exacerbate health disparities.[173]

Fifth, racially tailored medicine raises concerns about stigmatization and discrimination.[174] Research findings that emphasize biological differences among racial populations may convey the message that some racial groups are biologically inferior to others.[175] For example, minorities might be seen as more diseased than non-minority patients if they are deemed more vulnerable to the recurrence of certain cancers.[176] Publicity about racially tailored research in the popular press could fuel the fires of prejudice and discrimination.

III. LITIGATING DISCRIMINATION CLAIMS

Algorithmic discrimination can hurt patients and exacerbate health disparities. Aggrieved individuals may seek compensation through litigation. Patients who suffer harm during the course of their diagnosis or treatment can turn to tort theories, regardless of whether AI was involved.[177] For example, they might sue

---

cancer, all of which are common in particular populations but not exclusive to them).

169. *Id.*

170. *Id.*

171. Hoffman, *supra* note 155, at 419; Ambroise Wonkam et al., *The Burden of Sickle Cell Disease in Cape Town*, 102 S. AFR. MED. J. 752, 752 (2012) (South Africa has a low incidence of sickle cell disease").

172. Vyas et al., *supra* note 141 at 879.

173. *Id.* at 879-80 (urging clinicians who employ race-adjusting algorithms to "be thoughtful and deliberate users").

174. Hoffman, *supra* note 155, at 421-24.

175. *Id.*

176. Alex Tsodikov et al., *Is Prostate Cancer Different in Black Men? Answers from Three Natural History Models*, 123 CANCER 2312, 2312 (2017) ("Black race has been identified as an independent prognostic factor for disease recurrence in multiple reports . . . ."); Case Western Reserve University, *supra* note 145 ("This new work on prostate cancer builds on mounting evidence that clear biological differences between races can be discovered at a cellular level").

177. Megan Sword, *To Err is Both Human and Non-Human*, 88 UMKC L. REV. 211, 219-21 (2019); Shailin Thomas, *Artificial Intelligence, Medical Malpractice, and the End of Defensive Medicine*, PETRIE-FLOM CTR.: BILL OF HEALTH (Jan. 26, 2017), http://blogs.harvard.edu/billofhealth/2017/01/26/artificial-intelligence-medical-malpractice-and-the-end-of-defensive-medicine ("As algorithms improve and doctors use them more for diagnosing and decision-making,

physicians and hospitals for medical malpractice or vendors for a device's design defects.[178] The topic of AI and tort litigation has been addressed elsewhere and is beyond the scope of this Article.[179]

This work's contribution is to focus specifically on discrimination claims. If plaintiffs wish to challenge discriminatory algorithms and to have them eliminated or corrected, their most direct route is discrimination theory.

Presumably, health-care providers will use AI in good faith and trust that the technology will improve health-care outcomes. If they do not or they act with deliberate indifference to AI's discriminatory effects, they could face intentional discrimination claims. However, as demonstrated in Part II, AI can sometimes lead to unintentional discrimination when seemingly neutral algorithms disadvantage particular groups. In such cases, the applicable discrimination principle is disparate impact. This Part explores the theory of disparate impact and its significant limitations in the health-care field. It explains why disparate impact is unlikely to be a fruitful litigation path for plaintiffs aggrieved by AI outcomes. It also addresses potential litigation alleging intentional discrimination.

A. Disparate Impact

The disparate impact theory has developed most fully in the employment arena. We therefore begin with a discussion of employment discrimination litigation under Title VII of the Civil Rights Act (Title VII) and briefly address housing discrimination caselaw before tackling disparate impact as applied to health care.

1. What Is Disparate Impact?

The disparate impact theory enables plaintiffs to prove discrimination without

---

the traditional malpractice notions of physician negligence and recklessness may become harder to apply.").

178. W. Nicholson Price II, *Medical Malpractice and Black Box Medicine*, *in* BIG DATA, HEALTH LAW, AND BIOETHICS, *supra* note 95, at 295, 300 ("Providers . . . could be held liable for harmful use of black-box medical algorithms depending on the prevailing customary practice and the extent that custom is considered dispositive."); Nicolas Terry, *Of Regulating Healthcare AI and Robots*, 18 YALE J. HEALTH POL'Y, L. & ETHICS 133, 162-63 (2019) (describing several "very difficult" questions relating to potential product liability litigation involving AI); Saurabh Jha, *Can You Sue an Algorithm for Malpractice? It Depends*, STAT (Mar. 9, 2020), https://www.statnews.com/2020/03/09/can-you-sue-artificial-intelligence-algorithm-for-malpractice.

179. *See* A. Michael Froomkin, Ian Kerr & Joelle Pineau, *When AIs Outperform Doctors: Confronting the Challenges of a Tort-Induced Over-Reliance on Machine Learning*, 61 ARIZ. L. REV. 33, 35-36 (2019); Efthimios Parasidis, *Clinical Decision Support: Elements of a Sensible Legal Framework*, 20 J. HEALTH CARE L. & POL'Y 183, 218-25 (2018); Price, *supra* note 178.

proving intent to discriminate.[180] Title VII, which prohibits employment discrimination based on race, color, religion, sex, and national origin, empowers aggrieved parties to bring disparate impact cases against employers.[181] The seminal Supreme Court disparate impact ruling came in the 1971 *Griggs v. Duke Power Co.* case.[182] *Griggs* was a class action in which African American plaintiffs successfully challenged an employer's requirement of a high school diploma or passing a standardized general intelligence test for purposes of being hired or transferring to a better job.[183] The employer could not prove that the two requirements were related to satisfactory job performance, and both disproportionately disqualified African Americans.[184]

Underlying the Title VII disparate impact theory is the premise that "some employment practices, adopted without a deliberately discriminatory motive, may in operation be functionally equivalent to intentional discrimination."[185] Advocates can use the disparate impact theory to challenge not only standardized testing by employers, but also other practices that are not job-related and systematically disadvantage members of a class that is protected under the civil rights laws.[186] Examples are employers' exclusion of workers with criminal records, which adversely affect African Americans and Hispanics,[187] and strength tests, which have an adverse impact on women.[188]

The Fair Housing Act, which prohibits housing discrimination based on color, disability, familial status, national origin, race, religion, and sex, also enables private parties to litigate disparate impact cases.[189] In the 2015 case of *Texas Department of Housing and Community Affairs v. Inclusive Communities Project,*

---

180. Michael Selmi, *Was the Disparate Impact Theory a Mistake?*, 53 UCLA L. REV. 701, 702 (2006).

181. 42 U.S.C. § 2000e-2(k) (2018).

182. 401 U.S. 424 (1971) (citing 42 U.S.C. § 2000e-2).

183. *Id.* at 425-26 (1971).

184. *Id.*

185. Pippin v. Burlington Res. Oil & Gas Co., 440 F.3d 1186, 1199 (10th Cir. 2006) (quoting Ortega v. Safeway Stores, Inc., 943 F.2d 1230, 1242 (10th Cir. 1991)).

186. *Griggs*, 401 U.S. at 430 (referring to any "practices, procedures, or tests neutral on their face, and even neutral in terms of intent" that "operate to 'freeze' the status quo of prior discriminatory employment practices").

187. U.S. EQUAL EMP'T OPPORTUNITY COMM'N, EEOC-CVG-2012-1, ENFORCEMENT GUIDANCE ON THE CONSIDERATION OF ARREST AND CONVICTION RECORDS IN EMPLOYMENT DECISIONS UNDER TITLE VII OF THE CIVIL RIGHTS ACT (2012), https://www.eeoc.gov/laws/guidance/enforcement-guidance-consideration-arrest-and-conviction-records-employment-decisions#V (We are referring specifically to Part V, entitled "Disparate Impact Discrimination and Criminal Records.").

188. U.S. EQUAL EMP'T OPPORTUNITY COMM'N, EEOC-NVTA-2007-2, EMPLOYMENT TESTS AND SELECTION PROCEDURES (2007), https://www.eeoc.gov/laws/guidance/employment-tests-and-selection-procedures.

189. 42 U.S.C. § 3604 (2018).

Inc., the plaintiff asserted that the Department's allocation of low income housing tax credits had a disparate impact on African American residents.[190] The Supreme Court confirmed that disparate impact claims are cognizable under the Fair Housing Act.[191]

One would think that plaintiffs would likewise be able to apply the disparate impact theory to health-care practices, such as AI use, that disproportionately disadvantage women or racial minority groups. An algorithm is typically facially neutral but it could affect various populations differently because of design defects or flawed training data.[192] Under current law, however, the disparate impact theory does not furnish the majority of private parties with a suitable litigation tool in health-care cases.

2. Title VI

Title VI of the Civil Rights Act of 1964 prohibits programs receiving federal financial assistance from engaging in discrimination based on race, color, or national origin.[193] Title VI regulations clarify that covered entities may not use "criteria or methods of administration which have the effect of subjecting individuals to discrimination."[194] The regulations thus forbid practices that have a disparate impact on protected groups.[195] Health-care entities such as hospitals and nursing homes receiving payments from the federal programs Medicare and Medicaid, as most do, are covered by Title VI.[196]

Title VI is enforced both by the Department of Health and Human Services' (HHS) Office of Civil Rights (OCR) and by private litigation, but to limited effect.[197] Civil rights advocates have criticized OCR for not enforcing Title VI aggressively enough.[198] In addition, in 2001, the Supreme Court foreclosed the possibility of disparate impact litigation by private parties.[199] In *Alexander v. Sandoval*, the Court held that there is no private right of action to enforce the disparate impact regulations promulgated under Title VI.[200] Consequently, private parties can pursue only claims of intentional discrimination associated with AI, and OCR has sole authority to handle AI-related disparate impact violations

190. 576 U.S. 519, 519 (2015).
191. *Id.*
192. *See supra* Sections II.A-E.
193. 42 U.S.C. § 2000d (2018).
194. 28 C.F.R. § 42.104(b)(2) (2020); 45 C.F.R. § 80.3(b)(2) (2020).
195. 28 C.F.R. § 42.104(b)(2) (2020); 45 C.F.R. § 80.3(b)(2) (2020).
196. BARRY R. FURROW ET AL., LAW AND HEALTH CARE QUALITY, PATIENT SAFETY, AND LIABILITY 385 (8th ed. 2018).
197. *Id.*
198. *Id.*
199. Alexander v. Sandoval, 532 U.S. 275, 275 (2001).
200. *Id.*

relating to race, color, or national origin.[201]

3. Section 1557 of the Affordable Care Act

Section 1557 of the Patient Protection and Affordable Care Act (ACA) prohibits discrimination based on race, color, national origin, sex, age, or disability in particular health programs or activities.[202] In describing the protected classes, the statute refers to individuals protected by Title VI of the Civil Rights Act of 1964, Title IX of the Education Amendments of 1972 (addressing sex discrimination), Section 504 of the Rehabilitation Act of 1973 (addressing disability discrimination), and the Age Discrimination Act of 1975.[203]

The provision covers health programs or activities that receive federal financial assistance or that the federal government administers.[204] These generally include "hospitals, health clinics, health insurance issuers, state Medicaid agencies, community health centers, physician's practices and home health care agencies."[205] Note that HHS maintains that funds provided under Medicare Part B (which pays for physicians' services) do not constitute federal financial assistance, so some physicians may not be bound by the Section 1557 antidiscrimination mandate.[206] However, the statute applies to doctors receiving Medicaid payments and other forms of financial support, so the majority of physicians are covered.[207]

For purposes of this Article, a particularly important question is whether Section 1557 allows for disparate impact claims. The relevant statutory language is, "The enforcement mechanisms provided for and available under such title VI, title IX, section 794, or such Age Discrimination Act shall apply for purposes of violations of this subsection."[208] Could racial minorities who are disproportionately disadvantaged by an AI algorithm assert disparate impact claims under Section 1557 while the theory is unavailable under Title VI? The

201. Our research did not reveal any AI-related disparate impact cases that were pursued by OCR thus far.

202. 42 U.S.C. § 18116(a) (2018).

203. *Id.* ("[A]n individual shall not, on the ground prohibited under title VI of the Civil Rights Act of 1964 (42 U.S.C. 2000d et seq.), title IX of the Education Amendments of 1972 (20 U.S.C. 1681 et seq.), the Age Discrimination Act of 1975 (42 U.S.C. 6101 et seq.), or section 794 of title 29, be excluded from participation in, be denied the benefits of, or be subjected to discrimination under, any health program or activity, any part of which is receiving Federal financial assistance . . . .").

204. *Id.*

205. *Section 1557: Frequently Asked Questions*, U.S. DEP'T HEALTH & HUMAN SERVICES, https://www.hhs.gov/civil-rights/for-individuals/section-1557/1557faqs/index.html (last reviewed May 18, 2017).

206. FURROW ET AL., *supra* note 196, at 416; *Section 1557: Frequently Asked Questions*, *supra* note 205.

207. FURROW ET AL., *supra* note 196, at 416.

208. 42 U.S.C. § 18116(a) (2018).

question of private litigation of disparate impact allegations under Section 1557 has generated considerable controversy.

A former HHS regulation establishes that aggrieved individuals have a private right of action under Section 1557.[209] Under the Obama administration, HHS stated that it "interprets Section 1557 as authorizing a private right of action for claims of disparate impact discrimination . . . ."[210]

In *Rumble v. Fairview Health Services*, the plaintiff alleged that he received inferior care because he was a transgender man, in violation of Section 1557.[211] A district court ruled that Congress intended to create a new cause of action for discrimination in health care that is independent of the enforcement mechanisms for the statutes listed in Section 1557 (Title VI, Title IX, the Age Discrimination Act, and the Rehabilitation Act).[212] Based on this holding, Section 1557 plaintiffs could bring both disparate treatment and disparate impact claims.[213] According to the *Rumble* court, the fact that Title VI or Title IX is understood to ban disparate impact cases would not constitute an obstacle for plaintiffs bringing disparate impact claims under Section 1557.[214]

Other courts, however, have disagreed. In *Southeastern Pennsylvania Transportation Authority v. Gilead Sciences, Inc.*, a district court held that Section 1557 does not permit private litigation of disparate impact claims related to race.[215] The case involved allegations that Gilead's pricing scheme for its Hepatitis C drugs disproportionately disadvantaged racial minorities and low-income patients in violation of Section 1557.[216] The court emphasized the statute's incorporation of "the enforcement mechanisms" of the other civil rights statutes.[217] It thus concluded that the plain language of the law reveals that Congress adopted Title VI's exclusion of disparate impact claims in Section 1557.[218]

Several district courts have held that Section 1557 also precludes individuals' disparate impact claims for sex discrimination claimants.[219] This is because Title

209. Nondiscrimination in Health Programs and Activities, 81 Fed. Reg. 31,375, 31,472 (May 18, 2016) (codified at 45 C.F.R. § 92.302(d)). 45 C.F.R. § 92.302 was later repealed by Nondiscrimination in Health and Health Education Programs or Activities, Delegation of Authority, 85 Fed. Reg. 37,160, 37,201-04 (June 19, 2020).

210. Nondiscrimination in Health Programs and Activities, 81 Fed. Reg. at 31,440.

211. No. 14-CV-2037 (SRN/FLN), 2015 WL 1197415, at *1 (D. Minn. Mar. 16, 2015).

212. *Id.* at *11.

213. *Id.*

214. *Id.*; *see infra* notes 219-220 and accompanying text (discussing Title IX).

215. 102 F. Supp. 3d 688, 698-701 (E.D. Pa. 2015).

216. *Id*. at 693, 695.

217. *Id.* at 698; *see* 42 U.S.C. § 18116(a) (2018) ("The enforcement mechanisms provided for and available under such title VI, title IX, section 794, or such Age Discrimination Act shall apply for purposes of violations of this subsection.").

218. *Gilead*, 102 F.Supp.3d at 701.

219. *See* Weinreb v. Xerox Bus. Servs., LLC Health & Welfare Plan, 323 F. Supp. 3d 501, 521

IX of the Education Amendments of 1972 does not permit private litigation of sex discrimination claims based on disparate impact.[220]

To date, there appears to have been no Section 1557 disparate impact cases filed for age discrimination.[221] However, as in the case of Title VI and Title IX, private litigation of disparate impact claims is precluded by the Age Discrimination Act of 1975, which is referenced in Section 1557.[222] Thus, most courts would likely reject age-related disparate impact claims under Section 1557.

With respect to disability, there is less certainty. The Sixth Circuit held that Section 1557 prohibits disparate impact claims by disability discrimination litigants because it has interpreted the Rehabilitation Act of 1973, which Section 1557 incorporates, as barring such claims.[223] By contrast, other circuits have found that disparate impact claims are viable under the Rehabilitation Act and thus would likely hold that the same is true for Section 1557.[224]

The Supreme Court has yet to speak on the matter of disparate impact claims under Section 1557. However, in June 2020, the Trump administration enacted a regulation explicitly establishing that Section 1557 adopts the enforcement mechanisms of each of the statutes that it incorporates.[225] This rule prevents almost all plaintiffs from pursuing disparate impact challenges under Section 1557.

(S.D.N.Y. 2018), *appeal filed*, No. 18-2809 (2d Cir. Sept. 21, 2018); Condry v. UnitedHealth Grp., Inc., No. 17-cv-00183-VC, 2018 WL 3203046, at *4 (N.D. Cal. June 27, 2018) ("[D]isparate impact claims on the basis of sex are not cognizable under section 1557."), *appeal filed*, No. 20-16857 (9th Cir. Sept. 24, 2020); Briscoe v. Health Care Serv. Corp., 281 F. Supp. 3d 725, 738 (N.D. Ill. 2017); York v. Wellmark, Inc., No. 4:16-cv-00627-RGE-CFB, 2017 WL 11261026, at *15-16 (S.D. Iowa Sept. 6, 2017), *aff'd*, 965 F.3d 633 (8th Cir. 2020).

220. *Weinreb*, 323 F. Supp. 3d at 521; *Briscoe*, 281 F. Supp. 3d at 739.

221. Nondiscrimination in Health and Health Education Programs or Activities, 84 Fed. Reg. 27,846, 27,851 n.22 (proposed June 14, 2019) ("To the Department's knowledge, no disparate impact claims on the basis of age have been filed under Section 1557 in a Federal court.").

222. Kamps v. Baylor Univ., 592 F. App'x. 282, 285-86 (5th Cir. 2014).

223. Doe v. BlueCross BlueShield of Tenn., Inc., 926 F.3d 235, 242 (6th Cir. 2019).

224. *See* Ga. State Conf. of Branches of NAACP v. Georgia, 775 F.2d 1403, 1428 (11th Cir. 1985) (citing 34 C.F.R. § 104.4); Prewitt v. U.S. Postal Serv., 662 F.2d 292, 305 (5th Cir. Unit A Nov. 1981); *see also* Alexander v. Choate, 469 U.S. 287, 299 (1985) ("[W]e assume without deciding that § 504 reaches at least some conduct that has an unjustifiable disparate impact upon the handicapped.").

225. *HHS Finalizes Rule on Section 1557 Protecting Civil Rights in Healthcare, Restoring the Rule of Law, and Relieving Americans of Billions in Excessive Costs*, U.S. DEP'T HEALTH & HUMAN SERVICES (June 12, 2020), https://www.hhs.gov/about/news/2020/06/12/hhs-finalizes-rule-section-1557-protecting-civil-rights-healthcare.html. This rule further asserted that the government will interpret the term "sex" in the Section 1557 context as encompassing only male or female "as determined by biology." *Id.* However, in June of 2020, in *Bostock v. Clayton County*, the Supreme Court held that for purposes of Title VII, the term "sex" covers sexual orientation and gender identity. 140 S. Ct. 1731 (2020). This decision may well impact other areas of the law and change future interpretations of Section 1557.

B. Intentional Discrimination

In extreme cases, plaintiffs may sue health-care providers for intentional discrimination that is related to AI.[226] For example, if malevolent health-care providers deliberately create algorithms that will disadvantage minority patients and then use them as justifications to undertreat those individuals, they may be liable for intentional discrimination.

In addition, courts have determined that deliberate indifference can constitute intentional discrimination under the civil rights laws.[227] In order to prove deliberate indifference, the plaintiff must show that the defendant had actual knowledge of the alleged discrimination and the ability to redress it but failed to do so.[228] Thus, for example, if health-care providers become aware that their AI disproportionately deprives minority patients of referrals to high-risk management programs or underestimates their risk of contracting serious diseases and do not intervene to rectify the problem,[229] they could face intentional discrimination claims under Title VI or Section 1557.

IV. IMPLEMENTING LEGAL INTERVENTIONS

AI oversight requires a multi-faceted approach that involves many stakeholders.[230] Private litigants, AI developers, AI users, and the government all have a role to play in promoting algorithmic fairness.[231] This Part recommends three forms of legal interventions to address AI discrimination problems. The first is a private cause of action for disparate impact.[232] The second is a quality control mandate in the form of an algorithmic accountability act.[233] The third, addressed

226. *See supra* notes 201 and 209 and accompanying text (discussing litigation rights under Title VI and Section 1557).

227. Sunderland v. Bethesda Hosp., Inc., 686 F. Appx. 807, 815 (11th Cir. 2017) (concluding that a jury could find that the defendant-hospital acted with deliberate indifference in violation of the Rehabilitation Act when it relied on a malfunctioning video-remote-interpreting device to communicate with a deaf patient despite the patient's complaints and requests for an alternative method of accommodation); Blunt v. Lower Merion Sch. Dist., 767 F.3d 247, 273 (3d Cir. 2014) ("[D]eliberate indifference may, in certain circumstances, establish intentional discrimination for the purposes of a Title VI claim."); S.H. *ex rel.* Durrell v. Lower Merion Sch. Dist., 729 F.3d 248, 262 (3d Cir. 2013) (noting that appellate courts have "held that deliberate indifference satisfies the requisite showing of intentional discrimination").

228. *Blunt*, 767 F.3d at 273 (citing Davis *ex rel.* LaShonda D. v. Monroe Cty. Bd. of Educ., 526 U.S. 629, 645-49 (1999)).

229. *See supra* text accompanying notes 114-118, 146-148.

230. MITCHELL, *supra* note 1, at 124.

231. *Id.*

232. *See infra* Section IV.A.

233. *See infra* Section IV.B.

briefly, is FDA regulation.[234]

A. Private Cause of Action for Disparate Impact Discrimination in Health Care

Most if not all medical AI algorithm developers are well-intentioned and strive in good faith to improve human health through their work.[235] Nevertheless, algorithms can generate discriminatory outcomes.[236] This is a classic example of disparate impact, or unintentional discrimination.[237] Assume a physician applies an algorithm to help diagnose all patients with particular symptoms. The algorithm is thus a facially neutral mechanism, and the physician has no intention of discriminating against any patients. However, if the algorithm nevertheless disproportionately disadvantage a particular population, its use may be unlawful.[238]

As in the case of other disparate impact claims, defendants would not be liable for discrimination if their use of an algorithm is justified by business necessity, such as when an algorithm truly helps doctors make sound treatment decisions.[239] Thus, if an algorithm is shown consistently to improve the accuracy of disease prognosis and treatment choice, its use is permissible. This is true even if the algorithm leads clinicians to make different decisions for people with different demographics.[240]

The Fair Housing Act, Title VII, and other employment discrimination laws permit private litigants to pursue disparate impact claims in the areas of housing and the workplace.[241] For example, in *DeHoyos v. Allstate Corp*., the plaintiffs brought a class action to challenge Allstate's credit-scoring system under the Fair Housing Act and other laws because it caused African American and Hispanic customers to pay higher insurance premiums than White customers.[242] In *Muñoz*

234. *See infra* Section IV.C.

235. *See supra* Section I.B (discussing the benefits of AI).

236. *See supra* Section II.E (providing examples of algorithmic bias).

237. *See supra* Section III.A.

238. *Id.*

239. *See supra* notes 182-186 and accompanying text (discussing employment discrimination litigation).

240. *See supra* notes 142-144 and accompanying text (discussing a cancer study that focused on differences between African American and White patients).

241. *See* Kelly Cahill Timmons, *Accommodating Misconduct under the Americans with Disabilities Act*, 57 FLA. L. REV. 187, 200-05 (2005) (discussing disparate impact under the Americans with Disabilities Act); *Questions and Answers on EEOC Final Rule on Disparate Impact and "Reasonable Factors Other Than Age" Under the Age Discrimination in Employment Act of 1967*, U.S. EQUAL EMP. OPPORTUNITY COMMISSION, https://www.eeoc.gov/regulations/questions-and-answers-eeoc-final-rule-disparate-impact-and-reasonable-factors-other-age (last visited July 7, 2020).

242. 240 F.R.D. 269, 275 (W.D. Tex. 2007) (seeking final approval of a proposed settlement); *see also* Rodriguez v. Bear Stearns Cos., No. 07-cv-1816 (JCH), 2009 WL 995865, at *7 (D. Conn.

v. Orr, a class of Hispanic males sued the U.S. Air Force under Title VII to challenge its civilian employee promotion system, which involved an algorithm.[243]

In the era of AI and "black-box medicine," it is irrational to prohibit plaintiffs from pursuing such claims in the health-care arena. Government enforcement of disparate impact cases alone is inadequate because it depends on political priorities, which may disfavor civil rights cases, and on resources, which are often scarce.[244]

Consequently, it is useful to adopt private enforcement as an adjunct to government oversight and an incentive for statutory compliance. To that end, Congress should amend existing civil rights legislation to explicitly bar disparate impact discrimination and add private rights of action for aggrieved individuals. While we are not the first to suggest it,[245] this approach is now ripe for reconsideration.

1. Amending Title VI and Other Long-Standing Civil Rights Statutes

In 2008, the late Congressman John Lewis (D-GA) and Senator Edward Kennedy (D-MA) proposed the Civil Rights Act of 2008.[246] The findings section of the bill states that "[t]he Sandoval decision contradicts settled expectations created by title VI of the Civil Rights Act of 1964, title IX of the Education Amendments of 1972 . . . , the Age Discrimination Act of 1975 . . . , and section 504 of the Rehabilitation Act of 1973 . . . ."[247] The findings further state, emphatically, that administrative enforcement alone could not achieve compliance with the antidiscrimination laws and that enforcement by "private attorneys general" is vital.[248]

The Civil Rights Act of 2008 would have amended Title VI, Title IX, and the Age Discrimination Act of 1975 to prohibit "[d]iscrimination (including exclusion from participation and denial of benefits) based on disparate impact."[249] The bill noted that the Rehabilitation Act of 1973 already covers disparate impact and

---

Apr. 14, 2009) (denying defendants' motion to dismiss plaintiffs' claims "that defendants' predatory servicing practices disproportionately harmed minority borrowers").

243. 200 F.3d 291, 292 (5th Cir. 2000) (addressing a discovery dispute regarding plaintiffs' access to the algorithm).

244. *See supra* notes 202-203 and accompanying text; *see also* Dayna Bowen Matthew, *Health Care, Title VI, and Racism's New Normal*, 6 GEO. J.L. & MOD. CRITICAL RACE PERSP. 3, 56 (2014) ("The public-private litigation model has historically proved to be an indispensable weapon in the attack against subtle and complex racial discrimination.").

245. *See infra* Section IV.A.1.

246. *See* Civil Rights Act of 2008, S. 2554, 110th Cong. (2008); Civil Rights Act of 2008, H.R. 5129, 110th Cong. (2008).

247. S. 2554 § 101(2).

248. *Id.* § 101(3).

249. *Id.* § 102(a)(2), (b)(2), (c)(2).

allows private parties to litigate disparate impact claims.[250]

The proposed bill also added an explicit right of action for any violation of the statute, including the disparate impact provisions.[251] However, the bill specified that in disparate impact cases, aggrieved individuals could recover only equitable relief, attorney's fees, and costs.[252] A finding of liability would thus require defendants to correct the AI problem but inflict limited financial pain.

The bill did not pass,[253] but its aspirations were not forgotten. Professor Dayna Bowen Matthew renewed the call for a Title VI amendment in a 2014 article.[254] Professor Matthew emphasizes the importance of combined private and governmental enforcement efforts and of empowering victims of implicit bias to seek redress for the harms they have suffered.[255] The only vehicle for doing so is a private right of action for disparate impact claims. Under Matthew's proposal, as under the proposed 2008 Civil Rights Act, plaintiffs would be able to recover only equitable remedies, including attorneys' fees and costs in disparate impact cases.[256]

Professor Matthew asserts that legislative history reveals that "[f]rom its inception, health care equity has been at the core of the legislative purpose for Title VI."[257] A private disparate impact cause of action would thus restore the law to its original purpose.[258] Now algorithmic bias threatens to exacerbate health disparities as clinicians increasingly rely on AI. This is an opportune time to reinvigorate efforts to promote health equity and bolster civil rights enforcement.

2. Amending Section 1557 of the ACA

The Civil Rights Act of 2008 would have ensured that Section 1557 would allow private litigants to assert disparate impact claims.[259] Objections to such a right of action are based on Section 1557's reference to Title VI, Title IX, and the Age Discrimination Act, which have been deemed to preclude disparate impact litigation by private parties.[260] A new law explicitly adding such a right of action to those civil rights statutes would sweep away arguments about Section 1557's limited scope of litigation rights.

Admittedly, however, amending Title VI would dramatically impact all

250. *Id.* § 101(9).
251. *Id.* § 103(a)(2), (b)(2), (c).
252. *Id.* § 104.
253. Govtrack, *H.R. 5129 (110th): Civil Rights Act of 2008*, https://www.govtrack.us/congress/bills/110/hr5129 (last visited Dec. 9, 2020)..
254. *See* Matthew, *supra* note 244, at 54-58.
255. *Id.*
256. *Id.* at 55.
257. *Id.* at 12.
258. *Id.* at 61.
259. *See supra* Section IV.A.1.
260. *See supra* Section III.A.

programs receiving federal financial assistance and thus reach well beyond health care.[261] If Congress wishes to implement a more modest legislative intervention than the Civil Rights Act of 2008, it could amend Section 1557.[262] Congress could add language that plainly states that aggrieved individuals can assert disparate impact claims under the statute.[263] This would limit the scope of reform to health-care cases only, whereas the Civil Rights Act of 2008 would have been much broader.[264] In the absence of such an amendment, civil rights advocates can urge the Biden administration to reverse the Trump administration rule[265] and hope that more courts will follow *Rumble v. Fairview Health Services* in interpreting Section 1557.[266]

B. The Algorithmic Accountability Act

A different legislative pathway is the enactment of a law that establishes oversight for algorithms and promotes AI integrity. To that end, Senators Cory Booker (D-NJ) and Ron Wyden (D-OR) and Representative Yvette Clarke (D-NY) introduced the "Algorithmic Accountability Act" in the 116th Congress on April 10, 2019.[267]

The bill is rooted in concern about discrimination. Its sponsors issued a press release in which Senator Wyden stated that "[I]nstead of eliminating bias, too often . . . algorithms depend on biased assumptions or data that can actually reinforce discrimination against women and people of color."[268] Accordingly, the purpose of the bill is to "require[] companies to study the algorithms they use, identify bias in these systems and fix any discrimination or bias they find."[269]

1. The Statutory Requirements

The bill would do the following:

- Authorize the Federal Trade Commission (FTC) to formulate regulations requiring covered entities to conduct impact assessments of highly

261. *See supra* note 193 and accompanying text.
262. 42 U.S.C. § 18116(a) (2018).
263. *Id.*
264. *See supra* Section IV.A.1.
265. *See supra* note 225 and accompanying text.
266. No. 14-CV-2037 (SRN/FLN), 2015 WL 1197415, at *1 (D. Minn. Mar. 16, 2015); *see supra* text accompanying notes 211-214.
267. Algorithmic Accountability Act of 2019, S. 1108, 116th Cong. (2019); Algorithmic Accountability Act of 2019, H.R. 2231, 116th Cong. (2019); Press Release, U.S. Senator Cory Booker of N.J., Booker, Wyden, Clarke Introduce Bill Requiring Companies to Target Bias in Corporate Algorithms, (Apr. 10, 2019), https://www.booker.senate.gov/news/press/booker-wyden-clarke-introduce-bill-requiring-companies-to-target-bias-in-corporate-algorithms.
268. Press Release, U.S. Senator Cory Booker of N.J., *supra* note 267.
269. *Id.*

sensitive automated decision systems.

- Require covered entities to evaluate their use of automated decision systems and their training data in order to determine if there are problems related to accuracy, fairness, bias, discrimination, privacy or security.
- Require covered entities to assess the extent to which their information systems protect data subjects' privacy and ensure data security.
- Require covered entities to address any problems they discover during the impact assessments.[270]

A covered entity is any person, partnership, or corporation that is subject to FTC regulations and earns more than $50 million annually, possesses or controls personal information from at least one million people or consumer devices, or primarily acts as a data broker that acquires, processes, and sells consumer data.[271] In its current form, the bill therefore would not reach many health-care providers.[272]

2. Critique of the Bill

Many hailed the Algorithmic Accountability Act as a positive first step in promoting algorithmic fairness.[273] But others voiced opposition to the bill and highlighted several shortcomings.[274]

First, the bill applies only to large or high-revenue companies, and thus smaller companies would remain unregulated with respect to AI use.[275] Second, the bill relies exclusively on the FTC for enforcement, and consumer advocates argue that the agency's enforcement activities are often anemic.[276] Third, it does not require input from diverse stakeholders for purposes of impact assessment.[277] In fact, it states that companies should consult with external third parties, such as

270. S. 1108 §§ 2(2), 2(6), 3(b); Press Release, U.S. Senator Cory Booker of N.J., *supra* note 267.

271. S. 1108 § 2(5).

272. *Id.*; MARKUS H. MEIER, BRADLEY S. ALBERT & KARA MONAHAN, FED. TRADE COMM'N, OVERVIEW OF FTC ACTIONS IN HEALTH CARE SERVICES AND PRODUCTS 1 (2019), https://www.ftc.gov/system/files/attachments/competition-policy-guidance/overview_health_care_june_2019.pdf (explaining that the FTC's "Health Care Division consists of approximately 40 lawyers and investigators who work exclusively on health care antitrust matters," implying that the FTC has regulatory power over health-care entities).

273. Kaminski & Selbst, *supra* note 12.

274. *Id.*; Joshua New, *How to Fix the Algorithmic Accountability Act*, CTR. FOR DATA INNOVATION (Sept. 23, 2019), https://www.datainnovation.org/2019/09/how-to-fix-the-algorithmic-accountability-act.

275. S. 1108 § 2(5); New, *supra* note 274.

276. Kaminski & Selbst, *supra* note 12.

277. *Id.*

"independent auditors or technology experts," only "if reasonably possible."[278] Fourth, the bill does not mandate that the public have any access to impact assessment outcomes.[279] If the proposal directed the FTC to produce annual summary reports with de-identified assessment information, it could potentially provide the public with valuable data while safeguarding industry interests in proprietary information.[280] Other criticisms include regulatory overreach, lack of definitional clarity, and insufficient guidance, among other alleged shortcomings.[281]

3. Moving Forward

The proposed Algorithmic Accountability Act did not become law.[282] However, at least a couple of local jurisdictions have begun to focus attention on the integrity of AI practices. In 2017, the New York City Council established a task force to formulate recommendations for promoting public accountability with respect to the city's algorithm use.[283] The task force issued its report in November of 2019.[284] The report emphasizes the importance of "[p]romoting fairness, equity, accountability, and transparency in the use" of automated-decision systems.[285] In 2019, legislators in Washington State held a hearing on an algorithmic accountability bill that would establish guidelines for the state government's "procurement and use of automated decision systems."[286]

In order to establish a national standard for algorithmic fairness, Congress should persist in its efforts to pass AI-oversight legislation. A national solution would be preferable to local solutions because AI use is widespread and crosses state borders.[287] Both health-care providers and AI vendors often operate in

278. *Id.* (quoting S. 1108 § 3(b)(1)(C)).

279. Kaminski & Selbst, *supra* note 12; New, *supra* note 274.

280. Kaminski & Selbst, *supra* note 12.

281. New, *supra* note 274.

282. *S. 1108: Algorithmic Accountability Act of 2019*, GOVTRACK.US, https://www.govtrack.us/congress/bills/116/s1108 (last visited June 3, 2020).

283. Press Release, N.Y. Civil Liberties Union, City Council Passes First Bill in Nation to Address Transparency, Bias in Government Use of Algorithms (Dec. 11, 2017), https://www.nyclu.org/en/press-releases/city-council-passes-first-bill-nation-address-transparency-bias-government-use.

284. N.Y.C. AUTOMATED DECISION SYS. TASK FORCE, NEW YORK CITY AUTOMATED DECISION SYSTEMS TASK FORCE REPORT (2019), https://www1.nyc.gov/assets/adstaskforce/downloads/pdf/ADS-Report-11192019.pdf.

285. *Id.* at 18-19.

286. H.B. 1655, 66th Leg., Reg. Sess. 1 (Wash. 2019); S.B. 5527, 66th Leg., Reg. Sess. 1 (Wash. 2019); DJ Pangburn, *Washington Could Be the First State to Rein in Automated Decision-Making*, FAST COMPANY (Feb. 8, 2019), https://www.fastcompany.com/90302465/washington-introduces-landmark-algorithmic-accountability-laws.

287. *See, e.g.*, Vyas et al., *supra* note 141, at 1-6 (describing a variety of race-adjusting

multiple states.[288] For purposes of this Article, the law should provide HHS with jurisdiction to regulate algorithmic use by all health-care providers. To the extent possible, any future proposal should consider and address the critiques of the existing Algorithmic Accountability Act bill.[289]

An algorithmic quality-control mandate should be a supplement to and not a replacement for litigation rights. The law might also include a private cause of action for individuals harmed by biased or flawed algorithms. Thus, if Congress does not amend the anti-discrimination laws,[290] the Algorithmic Accountability Act could serve as an alternative pathway for relief for aggrieved patients.

C. FDA Regulation

At this time, it is unclear how and to what extent the FDA will ultimately regulate AI.[291] FDA regulation is currently a patchwork and is continuously evolving.[292]

The FDA acknowledges that its "traditional paradigm of medical device regulation was not designed for adaptive artificial intelligence and machine learning technologies."[293] In 2019, the FDA published a discussion paper detailing its "foundation for a potential approach to premarket review for artificial intelligence and machine learning-driven software modifications."[294] But the FDA has not enacted a clear set of AI regulations to date.[295] The FDA typically does not regulate algorithms that are developed and employed in-house by health-care

---

algorithms that are commonly used in a variety of specialties).

288. Christian D. Becker, Katherine Dandy, Max Gaujean, Mario Fusaro & Corey Scurlock, Commentary, *Legal Perspectives on Telemedicine Part 1: Legal and Regulatory Issues*, PERMANENTE J., Summer 2019, at 93, 94 (discussing cross-state licensure for telemedicine practitioners that enables them to practice in multiple states); *About Mayo Clinic*, MAYO CLINIC, https://www.mayoclinic.org/about-mayo-clinic (last visited July 27, 2020), (stating that the Mayo Clinic has campuses in Minnesota, Arizona, and Florida); *Top Artificial Intelligence Companies in Healthcare to Keep an Eye on*, MED. FUTURIST (Jan. 21, 2020), https://medicalfuturist.com/top-artificial-intelligence-companies-in-healthcare (naming national companies such as Google Health and IBM Watson Health as key players).

289. *See supra* Section IV.B.2.

290. *See supra* Section IV.A.

291. Bradley Merrill Thompson, *New Developments in FDA Regulation of AI*, MED. DEVICE & DIAGNOSTIC INDUSTRY (Apr. 9, 2020), https://www.mddionline.com/new-developments-fda-regulation-ai.

292. *See* Sharona Hoffman, *What Genetic Testing Teaches About Predictive Health Analytics Regulation*, 98 N.C. L. REV. 123, 154-56 (2019) (discussing regulatory uncertainty).

293. *Artificial Intelligence and Machine Learning in Software as a Medical Device*, U.S. FOOD & DRUG ADMIN., https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device (current as of Oct. 5, 2020).

294. U.S. FOOD & DRUG ADMIN., *supra* note 43.

295. Murray et al., *supra* note 125 (describing the FDA's "evolving regulatory landscape").

entities.[296] The agency has clarified that it intends to regulate certain types of software, such as software that analyzes "physiological signals" for diagnosis or therapeutic purposes,[297] and it has approved many algorithms used in the field of radiology.[298] The FDA also intends to focus attention on tools that are opaque and do not allow clinicians to review the basis of recommendations independently (i.e., black-box algorithms).[299]

Determining the proper scope of FDA regulation in the realm of AI is beyond the scope of this article. However, to the extent that the agency does regulate AI algorithms, it should include requirements of algorithmic fairness in its oversight standards.[300]

V. IMPROVING ALGORITHM DESIGN, VALIDATION, AND MONITORING PROCESSES

It is appropriate and necessary to legislate quality control mandates for medical AI algorithms.[301] But how can AI developers and users realistically ensure that these algorithms do not exacerbate health disparities and perpetuate discrimination? There is already a robust literature about promoting fairness in algorithms.[302] Doing so requires deliberate action. As Professors Michael Kearns and Aaron Roth explain,

> [A]lgorithms . . . are good at optimizing what you ask them to optimize, but they cannot be counted on to do things you'd like them to do but didn't ask for, nor to avoid doing things you don't want but didn't tell them not to do. Thus if we ask for accuracy

296. Price, *supra* note 1.

297. U.S. FOOD & DRUG ADMIN., CLINICAL DECISION SUPPORT SOFTWARE: DRAFT GUIDANCE FOR INDUSTRY AND FOOD AND DRUG ADMINISTRATION STAFF 10-11, 25 (2019), https://www.fda.gov/media/109618/download; Thompson, *supra* note 291.

298. Data Sci. Inst., *FDA Cleared AI Algorithms*, AM. C. RADIOLOGY, https://www.acrdsi.org/DSI-Services/FDA-Cleared-AI-Algorithms (last visited July 7, 2020).

299. Murray et al., *supra* note 125.

300. See *infra* Section V.A, for recommendations as to how vendors can promote algorithmic fairness.

301. *See supra* Section IV.B.3.

302. *See generally* KEARNS & ROTH, *supra* note 10; Chouldechova & Roth, *supra* note 25, at 82 ("[T]he last two years have seen an unprecedented explosion in interest from the academic community in studying fairness and machine learning."); Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miroslav Dudík & Hanna Wallach, *Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?*, CHI CONF. ON HUM. FACTORS COMPUTING SYSTEMS PROC., paper no. 600, at 1 (2019) ("The potential for machine learning (ML) systems to amplify social inequities and unfairness is receiving increasing popular and academic attention."); Paulus & Kent, *supra* note 88 (proposing a provisional framework for evaluating clinical prediction models for bias and fairness).

> but don't mention fairness, we won't get fairness. If we ask for one kind of fairness, we'll get that kind but not others.[303]

This Article's purpose is not to develop a comprehensive blueprint for eliminating algorithmic bias and discriminatory AI outcomes. Instead, we want to show only that experts can take a large number of steps to protect patients. Some of these steps can be mandated in the Algorithmic Accountability Act or its regulations, and others will be best practices that developers and users implement as appropriate.[304]

This Part outlines a variety of interventions that both AI designers and users can implement to promote fairness. It also addresses ambiguities in the concept of algorithmic fairness and the need for further research in the field.

A. Algorithm Developers

Developers of medical AI algorithms should focus on fairness concerns during the requirements, design, implementation, and validation processes.[305] Developers must recognize the potential for discrimination with respect to AI that relies on population-specific identity[306] and AI that could have a disparate impact on disadvantaged populations.[307]

Since developing AI algorithms is a form of software engineering, ensuring their fairness and overall quality calls for applying software engineering best practices with special attention to fairness.[308] Well-managed software development projects typically involve a series of phases, including requirements analysis and specification, design, implementation, testing, deployment, and operation.[309]

1. Requirements Analysis

Requirements analysis and specification involves determining and documenting the *requirements* for the software: what functionality and other attributes it must have to meet the needs of its users and other stakeholders.[310] To help ensure that the requirements are complete, developers should elicit input from

303. KEARNS & ROTH, *supra* note 10, at 87.

304. *See supra* Section IV.B (discussing the Algorithmic Accountability Act).

305. *See generally* IAN SOMMERVILLE, SOFTWARE ENGINEERING 66 (8th ed. 2007) (detailing the lifecycle of software).

306. *See supra* Sections II.E and IV.A.2.

307. *See supra* Section III.A.

308. Yuriy Brun & Alexandra Meliou, *Software Fairness*, PROC. 26TH ACM JOINT EUR. SOFTWARE ENGINEERING CONF. & SYMP. ON FOUND. SOFTWARE ENGINEERING 754, 754 (2018).

309. SOMMERVILLE, *supra* note 305, at 66-67.

310. KARL E. WIEGERS, SOFTWARE REQUIREMENTS 7 (2d ed. 2003).

each distinct class of potential users and other stakeholders.[311] In the case of medical AI algorithms, relevant stakeholders include: representatives of the protected group(s) and other patient groups, doctors, other caregivers, health informaticians, data scientists, and experts on discrimination.[312] Requirements analysis should determine the fairness requirements and other ethical requirements for the algorithm, along with its medical purpose, the circumstances under which it will be used, its inputs and outputs, and its reliability, safety, performance, usability, and security requirements.[313] Developers should select specific measures for assessing achievement of these properties.[314] The requirements specifications should be validated by having them reviewed and critiqued by stakeholders, and, possibly, by implementing a prototype with which users can interact and which they can evaluate well before the production version is ready.[315]

2. Software Design

Software design involves creating a high-level description of a solution to the problem of satisfying the software requirements.[316] The description includes the software's components, their required functionality and constraints, their interfaces and their interactions, the flow of data and control between components, and the application's user interface.[317] In the case of medical AI algorithms, data scientists must additionally determine the type of learning algorithm or predictive model that will be employed (e.g., deep neural network), the specific inputs to the algorithm, and the specific output(s).[318]

3. Software Implementation

Software implementation involves programming the solution, typically by a

311. *Id.* at 101.

312. Rajkomar et al., *supra* note 21, at 866.

313. *Id.*; *see* Johan Ordish, Hannah Murfet & Alison Hall, *Algorithms as Medical Devices*, PHG FOUND. 41-44 (2019), https://www.phgfoundation.org/documents/algorithms-as-medical-devices.pdf (discussing requirements for software and manufacturers).

314. WIEGERS, *supra* note 310, at 342 ("Software measurements provide insights into your projects, products, and processes that are more accurate than subjective impressions or vague recollections of what happened in the past.").

315. *Id.* at 53-54.

316. SOMMERVILLE, *supra* note 305, at 245-46.

317. *Id.*

318. SHALEV-SHWARTZ & BEN-DAVID, *supra* note 4, at 13-14. Deep neural networks, or deep learning, is a type of machine learning that allows computers "to learn from experience and understand the world in terms of a hierarchy of concepts, with each concept defined through its relation to simpler concepts." GOODFELLOW ET AL., *supra* note 32, at 1. Therefore, computers learn more complicated concepts by building on simpler ones. *Id.*

combination of writing new program code and exploiting existing code.[319] In the case of AI applications, high-quality implementations of learning algorithms are usually already available in various machine-learning code libraries.[320] Exploiting them requires making specific choices about data representations, parameters, settings, and other details.[321] In addition, to make their software usable by health-care workers, developers must implement an intuitive user interface to guide users in invoking the algorithm appropriately to help solve a particular medical problem.[322] As software components are acquired or developed, they are integrated with other components into working versions of the overall system, which have increasingly complete functionality.[323] Fairness issues could, in principle, arise at any point as the result of design or implementation choices.[324] It stands to reason that these problems are more likely to become evident to developers and users, and thus fixable, if fairness receives special attention during design reviews and during users' evaluation of design prototypes.

Medical AI algorithms have an additional stage of implementation that non-AI software does not have: training the algorithm with data from real patients, including both individuals exhibiting the conditions of interest and individuals not exhibiting them.[325] It is critically important that the training data be representative of the larger patient population to which a medical AI algorithm will be applied, including with respect to protected classes.[326]

The main method for achieving representativeness is random sampling; that is, using a random mechanism, such as a pseudorandom number generator, to select individuals from the larger population, with every individual having a nonzero probability of selection.[327] However, simple random sampling may be inadequate if a protected class or other important class of patients is rare because then it is likely that the class will be under-sampled.[328] Alternative sampling

319. SOMMERVILLE, *supra* note 305, at 67, 447-49 (discussing component reuse).

320. *See, e.g.*, *An End-to-End Open Source Machine Learning Platform*, TENSORFLOW, https://www.tensorflow.org (last visited June 30, 2020).

321. SOMMERVILLE, *supra* note 305, at 76-79 (discussing software design and implementation).

322. *Id.* at 363-66.

323. *Id.* at 33.

324. Rajkomar et al., *supra* note 21, at 870 (emphasizing the need to focus on fairness at all stages of AI development and implementation).

325. *See supra* notes 38-41 and accompanying text.

326. *See supra* Section II.B (discussing selection bias).

327. CARL-ERIK SÄRNDAL, BENGT SWENSSON & JAN WRETMAN, MODEL ASSISTED SURVEY SAMPLING 21 (1992) (stating that random sampling protects against selection bias and is viewed as objective); Yaron Ilan, *Generating Randomness: Making the Most Out of Disordering a False Order into a Real One*, 17 J. TRANSLATIONAL MED. 49, 49 (2019) (discussing pseudorandom-number generators).

328. Lyman L. McDonald, *Sampling Rare Populations*, *in* SAMPLING RARE OR ELUSIVE SPECIES: CONCEPTS, DESIGNS, AND TECHNIQUES FOR ESTIMATING POPULATION PARAMETERS 11, 16-17 (William L. Thompson ed., 2004).

designs such as stratified sampling and adaptive sampling can be used to adequately sample such rare classes.[329]

4. Testing

For virtually all software, the final and most important form of validation is testing.[330] At the testing stage, the software is executed on a set of test cases that developers created or an automated tool generated.[331] The algorithm's behavior and output are checked for conformance to requirements and to developer and user expectations.[332] Typically, developers test the final application in-house and end users test it in the field.[333]

Medical AI algorithms require additional testing that goes beyond that applied to other kinds of software.[334] We recommend that prior to general release of a medical AI algorithm, developers evaluate it for safety, efficacy, and fairness on a large, representative sample of patients that is different from the sample from which they obtained training data. Admittedly, it may sometimes be very difficult to obtain a sizeable and appropriate sample of the relevant patient population.[335] However, researchers have developed techniques to reduce data bias.[336]

Developers should collect the following during this evaluation: (1) measures of the outcome of interest (e.g., the proportion of patients correctly diagnosed as a result of applying the algorithm), (2) general measures of predictive performance, such as sensitivity and specificity,[337] and (3) measures relating to the fairness and

329. *Id.* at 18-19 (discussing stratification of population). "In stratified sampling, the population is divided into nonoverlapping subpopulations called strata. A probability sample is selected in each stratum." SÄRNDAL ET AL., *supra* note 327, at 100. Scientists who adaptively sample search for a population of interest at predetermined locations, and if appropriate subjects are found, they continue to search nearby. David R. Smith, Jennifer A. Brown & Nancy C.H. Lo, *Application of Adaptive Sampling to Biological Populations*, *in* SAMPLING RARE OR ELUSIVE SPECIES: CONCEPTS, DESIGNS, AND TECHNIQUES FOR ESTIMATING POPULATION PARAMETERS, *supra* note 328, at 77, 77.

330. RON PATTON, SOFTWARE TESTING 21 (2001).

331. PAUL AMMANN & JEFF OFFUTT, INTRODUCTION TO SOFTWARE TESTING 21-22, 67 (2d ed. 2016).

332. *Id.* at 5-6.

333. SOMMERVILLE, *supra* note 305, at 540 ("For most systems, programmers take responsibility for testing the components that they have developed."); *see* PATTON, *supra* note 330, at 244 (discussing beta testing).

334. Sara Gerke, Boris Babic, Theodoros Evgeniou & I. Glenn Cohen, *The Need for a System View to Regulate Artificial Intelligence/Machine Learning-Based Software as Medical Device*, 3 NPJ DIGITAL MED., art. no. 53, 2020, at 1, 4.

335. *See supra* Section II.B (discussing selection bias).

336. *See* Faisal Kamiran, Indrė Žliobaitė & Toon Calders, *Quantifying Explainable Discrimination and Removing Illegal Discrimination in Automated Decision Making*, 35 KNOWLEDGE & INFO. SYSTEMS 613, 615-16 (2013) (discussing local massaging, local preferential sampling, and local direct classification).

337. XIAO-HUA ZHOU, NANCY A. OBUCHOWSKI & DONNA K. MCCLISH, STATISTICAL METHODS

proportionality of the allocation of health-care resources.[338] We recommend that, when possible, developers compute these measures for the whole sampled population and for the protected and non-protected subgroup(s) separately in order to enable comparisons between groups.

5. Deployment and Operation

Health-care providers should decide whether to deploy a medical AI algorithm only after all stakeholder groups have carefully evaluated testing results.[339] Even when a medical AI algorithm is deemed fit for general use and is deployed, its evaluation should not stop.[340] Rather, developers and users should monitor and evaluate the software continuously for reliability, safety, and fairness over its entire operational life. In between changes to the algorithm or its usage, evaluation could be less intensive (e.g., experts can review records of randomly sampled uses of the algorithm). However, if the algorithm is changed, the software should be evaluated as rigorously as it was before it was first deployed to ensure that changes did not accidentally introduce software defects.[341] Finally, the developers should also provide a mechanism by which users can report discrimination or other problems they encounter.

Proper validation, auditing, and monitoring can detect fairness problems, and appropriate interventions can often fix them.[342] If an algorithm cannot be repaired, it should be abandoned or used selectively in a manner that avoids harm to protected groups. In the case of the algorithm that predicted which patients would miss appointments,[343] experts redesigned the algorithm to omit personal attributes such as ethnicity, religion, financial status, and body mass index and left only prior history of health-care use and information about appointments in order to reduce (though not eliminate) its discriminatory impact.[344] In the case of the algorithm used to identify candidates for high-risk management care programs,[345] designers addressed its disparate impact by replacing the future cost variable with a variable

---

IN DIAGNOSTIC MEDICINE 14 (2d ed. 2011) (explaining that sensitivity is a test's "ability to detect the condition when it is present" and specificity is a test's "ability to exclude the condition in patients without the condition").

338. Rajkomar et al., *supra* note 21, at 870.

339. *Id.*

340. *Id.*

341. AMMANN & OFFUTT, *supra* note 331, at 304 (discussing regression testing and explaining that it is "the process of re-testing software that has been modified").

342. Abu-Elyounes, *supra* note 31, at 52 (emphasizing the importance of auditing); Rajkomar et al., *supra* note 21, at 870.

343. *See supra* text accompanying notes 125-129.

344. Murray et al., *supra* note 125. See *infra* text accompanying notes 370-371, for additional steps taken to eliminate the algorithm's harmful consequences.

345. *See supra* text accompanying notes 114-118.

"that combined health prediction with cost prediction."[346]

Developers (and users) should apply special scrutiny to algorithms that correct for race.[347] Experts suggest that they focus on three specific questions.[348] First, do strong evidence and statistical analyses support the need for race correction?[349] Second, is the race correction justified by a "plausible causal mechanism for the racial difference"?[350] Third, does the race correction diminish or intensify health inequities?[351]

Experts are developing a growing number of tools to promote fairness within the AI industry.[352] One example is IBM's AI Fairness 360.[353] This is an open-source software toolkit that "enables developers to use state-of-the-art algorithms to regularly check for unwanted biases . . . and to mitigate any biases that are discovered."[354] Such tools, in combination with other interventions discussed in this Article, have the potential to mitigate algorithmic biases and enhance fairness in meaningful ways.[355]

B. Algorithm Users

Some AI users develop algorithms themselves, and some employ AI that third parties develop with or without supplying their own training data.[356] Clinicians who use AI obtained from outside vendors can be responsible for discriminatory outcomes that it generates, and thus they would do well to engage in their own assessment of the technology and its impacts.[357] Like developers, AI users should

346. Obermeyer et al., *supra* note 114, at 453.

347. *See supra* Section II.F.2.

348. Vyas et al., *supra* note 141, at 880.

349. *Id*.

350. *Id*.

351. *Id.* ("In many cases, this appraisal will require further research into the complex interactions among ancestry, race, racism, socioeconomic status, and environment.").

352. Holstein et al., *supra* note 302, at 1 ("A surge of recent work has focused on the development of algorithmic tools to assess and mitigate . . . unfairness.").

353. *AI Fairness 360*, IBM DEVELOPER, https://developer.ibm.com/technologies/artificial-intelligence/projects/ai-fairness-360 (last updated Mar. 9, 2020).

354. *Id.*

355. Abu-Elyounes, *supra* note 31, at 44-45 ("While these tools could be useful and might be able to point out potential problematic behavior of algorithms, they cannot be used alone, and should be taken with a grain of salt because mitigating bias cannot be fixed by a miracle." *Id.* at 45.); Holstein et al., *supra* note 302, at 1-2 ("If such tools are to have a positive and meaningful impact on industry practice, however, it is crucial that their design be informed by an understanding of practitioners' actual challenges and needs for support in developing fairer ML systems." *Id.* at 2 (citation omitted).).

356. Emily J. Tait, Robert W. Kantner, Hilda C. Galvan & Jonathan M. Linas, *Proposed Algorithmic Accountability Act Targets Bias in Artificial Intelligence*, JONES DAY (June 2019), https://www.jonesday.com/en/insights/2019/06/proposed-algorithmic-accountability-act.

357. *See supra* Part III.

be vigilant about discrimination when implementing AI that adjusts for race[358] and AI that could have a disparate impact on disadvantaged populations.[359]

The FTC issued AI guidance to parties under its jurisdiction in April of 2020.[360] Relevant recommendations include the following:

- If you deny consumers something of value based on algorithmic decision-making, explain why.
- If you use algorithms to assign risk scores to consumers, also disclose the key factors that affected the score, rank ordered for importance.
- Don't discriminate based on protected classes.
- Focus on inputs, but also on outcomes.
- Make sure that your AI models are validated and revalidated to ensure that they work as intended, and do not illegally discriminate.[361]

Much of the FTC's advice applies to health-care providers.

1. Transparency

Health-care providers should consider discussing their use of AI with patients. Patients would likely appreciate knowing that clinicians are trying to use state-of-the-art technology for their benefit and would value an explanation of any anticipated AI limitations.

Professor I. Glenn Cohen has analyzed whether failure to disclose AI use constitutes a violation of the informed consent doctrine.[362] He concludes that it does not, with a few possible but uncertain exceptions, "such as when patients inquire about the involvement of AI/ML, when the medical AI/ML is more opaque, when it is given an outsized role in the final decision-making, or when the AI/ML is used to reduce costs rather than improve patient health."[363] Indeed, if physicians research medical literature or query colleagues in the process of making a medical decision, they are not obligated to disclose to patients that they did so.[364] Arguably, AI is an analogous source of input.[365] Nevertheless, in some cases, as Professor

358. *See supra* Section II.F.2.

359. *See supra* Section II.E.

360. Andrew Smith, *Using Artificial Intelligence and Algorithms*, FED. TRADE COMMISSION (Apr. 8, 2020, 9:58 AM), https://www.ftc.gov/news-events/blogs/business-blog/2020/04/using-artificial-intelligence-algorithms.

361. *Id.*

362. I. Glenn Cohen, *Informed Consent and Medical Artificial Intelligence: What to Tell the Patient?*, 108 GEO. L.J. 1425, 1432 (2020) (explaining that the informed consent doctrine provides that "liability could attach if a physician did not inform the patient of the risk and benefits of proposed treatment or nontreatment").

363. *Id.* at 1428-29.

364. *Id.* at 1443-44.

365. *Id.*

Cohen notes, clinicians might protect themselves from liability through disclosure and obtaining the patient's consent (e.g., if the doctor intends to rely exclusively on AI in making an important decision).[366] Even if there is no danger of liability, discussing AI use might be the right thing to do in order to be candid with patients and keep them fully informed about their care.[367]

2. Monitoring and Assessing AI Use

Health-care providers should always remain vigilant about AI outcomes and do their best to detect any discriminatory outcomes. Jones Day, a prominent law firm, advises clients using externally-developed AI to investigate the developers' mechanisms for eliminating bias and to assess whether their AI has a disparate impact on any class protected by the civil rights laws.[368] Likewise, a group of Stanford University researchers advises that doctors using machine-learning systems educate themselves "about their construction, the data sets they are built on, and their limitations" in order to avoid "ethically problematic outcomes."[369]

Clinicians using AI must be prepared to intervene as soon as discrimination problems become apparent. For example, when users realized that an algorithm designed to predict appointment no-shows had an adverse impact on disadvantaged populations, they decided it was inappropriate to double-book the appointments in question and divert resources away from vulnerable individuals.[370] Instead, they implemented "patient-positive" actions, such as appointment reminders and outreach to the identified people.[371] It is also possible that a health-care providers' patient mix will change over time, and an algorithm that was not problematic when initially deployed will generate discriminatory outcomes for a new patient population.

In time, the health-care community may develop clinical practice guidelines and educational materials about best practices that minimize AI-related discrimination. For now, providers should recognize that they should not blindly

366. *Id.* at 1466.

367. Laura M. Cascella, *Artificial Intelligence and Informed Consent*, MEDPRO GROUP, https://www.medpro.com/artificial-intelligence-informedconsent (last visited June 13, 2020) (describing what clinicians should disclose to patients about their AI use); Emily Sokol, *Artificial Intelligence's Impact on Patient Safety, Outcomes*, HEALTHITANALYTICS (Aug. 19, 2019), https://healthitanalytics.com/news/artificial-intelligences-impact-on-patient-safety-outcomes ("Patients should also be informed of the potential for inaccurate diagnosis, whether that be over-diagnosis or misdiagnosing as the result of AI technologies.").

368. Tait et al., *supra* note 356.

369. Danton S. Char, Nigam H. Shah & David Magnus, *Implementing Machine Learning in Health Care—Addressing Ethical Challenges*, 378 NEW ENG. J. MED. 981, 983 (2018).

370. Murray et al., *supra* note 125; *see supra* text accompanying notes 125-129.

371. Murray et al., *supra* note 125.

trust their AI and leave it entirely unchecked.[372]

C. Having Realistic Expectations

Improving algorithmic fairness is hard work, and fully achieving fairness is likely impossible.[373] In one study, researchers interviewed and surveyed 267 machine-learning practitioners about fairness-related challenges that they face, and respondents identified numerous difficulties.[374] For example, many AI teams lack a process to collect and curate balanced and representative training datasets.[375] Respondents stated that they struggled to determine which subpopulations they should consider to guard against selection bias in particular applications. To illustrate, while it is natural to think about ethnicity and gender when worrying about inclusivity, the relevant attribute that may skew algorithmic outcomes could be being a native English speaker.[376] In addition, teams often strain to discern the causes of unanticipated fairness problems, especially in the case of black-box AI.[377]

In some instances, there are competing fairness goals, and they cannot all be fulfilled simultaneously.[378] Imagine that an algorithm is designed to decide which applicants should receive loans and to promote fairness with respect to race.[379] The algorithm's developers will have to make some choices. They could emphasize group fairness, that is, that the same percentage of applicants of all races should get loans.[380] In the alternative, they could emphasize individual fairness, meaning that two applicants who are identical in all ways except for race should always be

372. Price, *supra* note 178, at 295-96 ("[W]hile providers and facilities are ill suited to evaluate the substantive accuracy of black-box medical algorithms, they could and perhaps should be required to exercise due care to evaluate procedural quality—the expertise of the developer and the availability of independent external validation . . . .").

373. MITCHELL, *supra* note 1, at 108 ("[I]t is often hard to tease out subtle biases and their effects."); Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns & Aaron Roth, *Fairness in Criminal Justice Risk Assessments: The State of the Art*, SOC. METHODS & RES. (forthcoming, first published July 2018) (manuscript at 1), https://journals.sagepub.com/doi/pdf/10.1177/0049124118782533 ("[T]here are at least six kinds of fairness, some of which are incompatible with one another . . . .").

374. Holstein et al., *supra* note 302, at 3-5, 6-12.

375. *Id.* at 6 ("A software engineer . . . described their team's current data collection practices as '*almost like the wild west*'.").

376. *Id.*; *see also supra* notes 119-121 and accompanying text (describing a speech-analysis machine-learning tool that misdiagnosed non-native speakers as having Alzheimer's disease because it misinterpreted pauses and mispronounciations).

377. Holstein et al., *supra* note 302, at 7.

378. KEARNS & ROTH, *supra* note 10, at 84-86 (discussing "fairness fighting fairness" (capitalization in title omitted)); Brun & Meliou, *supra* note 308, at 755.

379. Brun & Meliou, *supra* note 308, at 755; *see supra* note 30.

380. *Id.*

treated the same in terms of loan approval.[381] Imagine further that there is a significant correlation between race and income, with Whites generally having higher incomes.[382] If so, it will be impossible both to give the same percentage of applicants of all races loans and to treat all pairs of applicants that are identical in every way but race the same.[383] If applicants need to earn at least $75,000 to obtain a loan, the algorithm could safeguard individual fairness, but group fairness will be unattainable because Whites will receive loans at a higher rate than African Americans.[384] By contrast, if the lender emphasizes equalizing the percentage of applicants of all races who obtain approval for loans, it will sacrifice individual fairness.[385] Some minorities will receive loans without having an adequate income, but the same will not be true for Whites.[386] In this hypothetical, consequently, it is impossible to achieve the dual goal of group fairness and individual fairness.[387]

The AI community, therefore, will have to be realistic about the degree and types of fairness that it can achieve. It may sometimes need to identify and prioritize conflicting fairness goals. Achieving comprehensive equality of outcomes, performance, and allocation is likely impossible.[388] In addition, the government and industry must remain committed to funding and pursuing research regarding algorithmic fairness. Experts have identified a variety of vital research directions.[389] These relate to collecting and curating high quality and appropriately diverse training datasets, fairness-oriented debugging tools, auditing methods, and educational resources.[390]

CONCLUSION

The health-care community is justifiably enthusiastic about the many possible advantages of AI. But not everyone consistently benefits from the introduction of this innovative technology, and algorithms are raising growing concerns about fairness and bias.

As AI use proliferates in medicine, it is important that providers recognize its hazards and understand that some of these can lead to ethical challenges and liability exposure. AI algorithms adopt biases that are embedded in training data or that result from training data that is not sufficiently diverse and representative.[391]

381. *Id.*
382. *Id.*
383. *Id.*
384. *Id.*
385. *Id.*
386. *Id.*
387. See *supra* note 31, for other sources discussing the tension among different fairness goals.
388. *See supra* text accompanying note 27.
389. Holstein et al., *supra* note 302, at 12.
390. *Id.*
391. *See* Ben Dickson, *Healthcare Algorithms Are Biased, and the Results Can Be Deadly*, PC

In addition, some deliberately adjust for race without adequate justification for doing so. These problems can lead to patient harm and unlawful discrimination.

Private plaintiffs face very difficult terrain in attempting to litigate disparate impact discrimination claims in the health-care arena. Nevertheless, as Representative Yvette Clarke stated, "Algorithms shouldn't have an exemption from our anti-discrimination laws."[392] Consequently, this Article argues that it is necessary to reinstate disparate impact litigation as a private enforcement tool in the AI era. It also recommends that Congress legislate AI-oversight requirements through an algorithmic accountability act and that the FDA consider the potential for discrimination in its algorithmic approval processes.

It is true that many algorithms constitute black-box medicine and that even their developers often cannot fully explain how they function.[393] Nevertheless, both developers and users must make every effort to determine whether AI exacerbates health disparities and perpetuates discrimination. To that end, the Article describes a variety of interventions that both developers and users should implement while designing, validating, using, and monitoring AI in order to bolster fairness. At the same time, the health-care community must accept that it is difficult to define fairness and that it may need to prioritize among conflicting fairness goals.

As alluring as AI is and as tempting as it may be to trust it wholeheartedly, combatting discrimination requires human oversight. In the words of Dr. Steven Goodman and colleagues, "the only solution is to apply to artificial intelligence algorithms the very thing they are designed to supersede—human intelligence."[394]

With proper fairness-oriented oversight, AI can fulfill its promise of improving overall human health. Moreover, AI could in fact help combat discrimination by identifying those in greatest need and promoting more equitable allocation of health resources.[395]

MAG. (Jan. 23, 2020), https://www.pcmag.com/opinions/healthcare-algorithms-are-biased-and-the-results-can-be-deadly.

392. Press Release, U.S. Senator Cory Booker of N.J., *supra* note 267.

393. *See supra* notes 48-50 and accompanying text.

394. Steven N. Goodman, Sharad Goel & Mark R. Cullen, Editorial, *Machine Learning, Health Disparities, and Causal Reasoning*, 169 ANNALS INTERNAL MED. 883, 884 (2018).

395. Rajkomar et al., *supra* note 21, at 870.

"*The Ethics of AI in Biomedical Research, Patient Care and Public Health*"[1]

By Alessandro Blasimme and Effy Vayena

Health Ethics and Policy Lab, ETH Zurich - Switzerland

Abstract

This chapter focuses on ethical issues raised by the use of artificial intelligence (AI) in the domain of health. In particular we discuss specific issues in biomedical research, healthcare provision and public health. We devote particular attention to ethical concerns about safety and evidence standards in biomedical AI; to informed consent and the impact of automation on both professional caregivers and patients; to fairness and discrimination in algorithmic decision making on treatment and disease prevention for individuals and populations; to equity and social justice in the distribution of AI-driven health care and public health. We argue that the litany of ethical challenges that AI in medicine raises cannot be addressed sufficiently by current regulatory and ethical frameworks. We thus propose relevant governance approaches that can help address this gap.

Keywords: Artificial intelligence, machine learning, algorithms, ethics, medicine, biomedical research, public health, bioethics, medical ethics, governance.

[1] This paper is a draft version of a chapter that will appear in Markus Dubber, Frank Pasquale and Sunit Das (eds.) 'The Oxford Handbook of Ethics of Artificial Intelligence' (OUP).

1) Introduction

In March 2019 the World Health Organization announced amid a number of key reforms, the establishment of a new department of Digital Health with the aim to harness "the power of digital health and innovation by supporting countries to assess, integrate, regulate and maximize the opportunities of digital technologies and artificial intelligence"[2]. This commitment at the global level is in the same vein with several national plans announced over the last couple of years[3] as governments began to grabble with AI in health. Numerous examples of AI enabled digital health applications are available today, some have received market authorization, and if the private investment in digital health is anything to go by, the pipeline of future digital health products is going to be full. Certainly, the so-called big data revolution has been instrumental to this development.

In this chapter we discuss ethical challenges linked to the use of AI in biomedical research, patient care and public health. We then draw on a systemic oversight model for the governance of AI innovation in the health sector[4] and discuss possible ways to address emerging ethical challenges in this rapidly evolving domain. Our aim is to lay the groundwork for an ethically responsible development of AI in the domains of health research, clinical practice and

[2] See https://www.who.int/news-room/detail/06-03-2019-who-unveils-sweeping-reforms-in-drive-towards-triple-billion-targets. Accessed: 4 April 2019.
[3] Lynne E Parker, "Creation of the National Artificial Intelligence Research and Development Strategic Plan.," *AI Magazine* 39, no. 2 (2018); Corinne Cath et al., "Artificial Intelligence and the 'Good Society': The US, EU, and UK Approach," *Science and Engineering Ethics* 24, no. 2 (2018): 505–28; Sophie-Charlotte Fischer, "Artificial Intelligence: China's High-Tech Ambitions," *CSS Analyses in Security Policy* 220 (2018).
[4] Effy Vayena and Alessandro Blasimme, "Health Research with Big Data: Time for Systemic Oversight," *The Journal of Law, Medicine & Ethics* 46, no. 1 (2018): 119–29; Alessandro Blasimme and Effy Vayena, "Towards Systemic Oversight in Digital Health: Implementation of the AFIRRM Principles.," in *Cambridge Handbook of Health Research Regulation*, ed. Graeme Laurie (Cambridge University Press, forthcoming).

public health.

2) AI in Biomedical Research

In the last decade, biomedical research has become a data-centric activity[5] enabled by novel material and experimental practices linked to data collection, distribution and use.

In the burgeoning field of precision medicine[6], for instance, 'omic' data are now routinely being collected alongside clinical data, phenotypic data, lifestyle and socio-economic data to form bigger-than-ever research cohorts. Artificial intelligence is predicted to enable the simultaneous computation of such diverse arrays of data thus contributing to the promise of precision medicine to bring about more targeted approaches to diagnosis and treatment of individual patients[7]. As far as translational medicine is concerned, artificial intelligence is being employed in drug discovery to screen massive libraries of potentially therapeutic molecules, to automate searches in the biomedical literature through natural language processing techniques, to predict experimental dosage and so on[8].

Machine learning is also deployed to generate predictive models that could help doctors in prognostic assessment and in personalizing therapy and rehabilitation for individual patients, for instance in the aftermath of a stroke[9]. Electronic health records (EHR) for example offer

[5] Sabina Leonelli, *Data-Centric Biology: A Philosophical Study* (University of Chicago Press, 2016).
[6] Francis S. Collins and Harold Varmus, "A New Initiative on Precision Medicine," *New England Journal of Medicine* 372, no. 9 (February 26, 2015): 793–95; Alessandro Blasimme and Effy Vayena, "Becoming Partners, Retaining Autonomy: Ethical Considerations on the Development of Precision Medicine," *BMC Medical Ethics* 17 (2016): 67,; Alessandro Blasimme and Effy Vayena, "'Tailored-to-You': Public Engagement and the Political Legitimation of Precision Medicine," *Perspectives in Biology and Medicine* 59, no. 2 (2017): 172–188.
[7] Bertalan Mesko, "The Role of Artificial Intelligence in Precision Medicine," 2017; Jia Xu et al., "Translating Cancer Genomics into Precision Medicine with Artificial Intelligence: Applications, Challenges and Future Perspectives," *Human Genetics* 138, no. 2 (February 1, 2019): 109–24.
[8] Eric J Topol, "High-Performance Medicine: The Convergence of Human and Artificial Intelligence," *Nature Medicine* 25, no. 1 (2019): 51.
[9] See https://precise4q.eu. Accessed: 4 April 2019.

the opportunity to use real-world data to generate knowledge about the outcomes of a given medical procedure (be it a diagnosis, a prognosis, a therapy or a rehabilitation plan)[10]. AI can be employed to mine EHR to discover disease familiarity or people at risk for a given chronic disease, but also to improve the organization of health systems by providing support in triage and patient management[11]. In a recent study, deep learning was employed to create predictive modeling with EHR to accurately gauge in-hospital mortality, readmission odds, length of stay and final discharge diagnoses[12]. In another study, a machine learning algorithm identified cancer patients at high risk of 30-day mortality before they start chemotherapy (both palliative and curative)[13]. Such an algorithm can help decisions about chemotherapy initiation enabling more rational allocation of resources.

Facial recognition technologies based on machine learning are also being developed to streamline patient identification, to detect genetic disorders that correspond to specific facial traits[14] or to diagnose mood disorders such as depression[15]. Recently, researchers validated a system that, based on human-computer interaction patterns using data from a smartphone app, is able to recognize what the authors of the study call digital biomarkers of cognitive function[16]. Lately, there is increasing interest in voice analysis algorithms for health-related

[10] Institute of Medicine, *The Learning Healthcare System: Workshop Summary (IOM Roundtable on Evidence-Based Medicine)*, 2007, https://www.nap.edu/catalog/11903/the-learning-healthcare-system-workshop-summary-iom-roundtable-on-evidence.
[11] Pavel Hamet and Johanne Tremblay, "Artificial Intelligence in Medicine," *Metabolism* 69 (2017): S36–40.
[12] Alvin Rajkomar et al., "Scalable and Accurate Deep Learning with Electronic Health Records," *NPJ Digital Medicine* 1, no. 1 (2018): 18.
[13] Aymen A Elfiky et al., "Development and Application of a Machine Learning Approach to Assess Short-Term Mortality Risk among Patients with Cancer Starting Chemotherapy," *JAMA Network Open* 1, no. 3 (2018): e180926–e180926.
[14] Yaron Gurovich et al., "Identifying Facial Phenotypes of Genetic Disorders Using Deep Learning," *Nature Medicine* 25, no. 1 (2019): 60.
[15] Yu Zhu et al., "Automated Depression Diagnosis Based on Deep Networks to Encode Facial Appearance and Dynamics," *IEEE Transactions on Affective Computing* 9, no. 4 (2018): 578–84; Albert Haque et al., "Measuring Depression Symptom Severity from Spoken Language and 3D Facial Expressions," *ArXiv Preprint ArXiv:1811.08592*, 2018.
[16] Paul Dagum, "Digital Biomarkers of Cognitive Function," *Npj Digital Medicine* 1, no. 1 (2018): 10.

purposes with research concentrating on mental health[17].

The main concern raised by AI in the above context is the quality and representativeness of data used to train machine learning algorithms. In the existing medical datasets adult males of Caucasian origin are strongly overrepresented[18]. This lack of diversity is likely to result in biased algorithms trained on biased data. Similarly, EHR data used to train algorithms may suffer from issues such as missing data and misclassification[19]. For example, people of lower socioeconomic levels may be less represented in certain diagnostic categories, or may be overrepresented in categories of emergency care. Such patients maybe more concentrated to an institution than to others making research results of potential medical relevance more meaningful to overrepresented populations than minorities or socially emarginated groups.

Another concern relates to the sufficiency of informed consent as an ethical safeguard in research involving algorithmic processing. The traditional concept of informed consent is already challenged in cases of data collected in more conventional research settings, as it is increasingly hard to predict who will be accessing the data in the future, for which purposes and under which conditions[20]. The infinite uses of data and the linkage of disparate data sets, makes even the notion of broad consent – a typical safeguard of autonomy when future uses of human data and samples are hard to anticipate – weak. In the case of AI, it is still not clear whether research participants shall be specifically informed about the intention to use this

[17] Nicholas Cummins, Alice Baird, and Björn W. Schuller, "Speech Analysis for Health: Current State-of-the-Art and the Increasing Impact of Deep Learning," *Health Informatics and Translational Data Analytics* 151 (December 1, 2018): 41–54.
[18] Latrice G Landry et al., "Lack of Diversity in Genomic Databases Is a Barrier to Translating Precision Medicine Research into Practice," *Health Affairs* 37, no. 5 (2018): 780–85.
[19] Milena A Gianfrancesco et al., "Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data," *JAMA Internal Medicine* 178, no. 11 (2018): 1544–47.
[20] Effy Vayena and Alessandro Blasimme, "Biomedical Big Data: New Models of Control Over Access, Use and Governance," *Journal of Bioethical Inquiry* 14, no. 4 (2017).

form of data analysis, and whether informed consent for automated processing of personal data should reflect a heightened level of protection and, for instance, offer the option to opt out.

Issues of data privacy and security loom large on the horizon of biomedical big data research[21].

The creation of large cohorts of deeply phenotyped participants raises doubts about the huge amounts of information that such initiatives put in the hands of governments or private organizations. The latter include healthcare organizations, BigTech and companies active in the field of smart technologies that stipulate agreements with national governments to collect and analyze data from millions of citizens.

AI adds a layer of ethical complexity in that it uses data to extract fine-grained information about individuals. It is an ethical responsibility of researchers to securely protect this information from unauthorized access in order to avoid privacy-related harms to data subjects in the course of research projects. The unwanted leak of health-relevant information can lead to discriminative uses of such information in domains such as employment, education and insurance. This problem applies both to information generated and stored by researchers and to information that researchers feed back to research participants as primary, secondary or incidental findings. Return of research results enjoys widespread support as a way to show respect for the interests and the welfare of research participants[22]. In particular, precision medicine initiatives such as the US All of Us Research Program endorse a model of empowerment

[21] Omer Tene and Jules Polonetsky, "Privacy in the Age of Big Data: A Time for Big Decisions," *Stan. L. Rev. Online* 64 (2011): 63.
[22] Susan M. Wolf, "Return of Individual Research Results and Incidental Findings: Facing the Challenges of Translational Science," *Annual Review of Genomics and Human Genetics* 14, no. 1 (2013): 557–77, https://doi.org/10.1146/annurev-genom-091212-153506.

that is premised on the release of medically relevant information to research participants. This model, while laudable, can have consequences for instance for those research data subjects who intend to buy a life insurance policy. [23].

The criteria that are being employed in the evaluation of research involving human data and human subjects (including clinical trials) have been developed in the post war period and formalized in most countries since the late Seventies. Such criteria – e.g. social or scientific value, scientific validity, fair selection of participants, acceptable risk-benefit ratio, informed consent and consideration for participants' welfare and rights[24] - while being still valid at a formal level, do not adequately capture the specificities of research involving the use of AI to analyze vast amounts of personal data[25]. Consider the case of a recent study that utilizing deep neural networks analyzed the association of facial traits and self-declared sexual orientation in order to understand whether homosexuals have distinct facial characteristics[26]. Besides the technical validity of this study, its aim is highly dubitable from an ethical point of view because it lends support to stereotyped views about homosexuality – namely, the idea that male homosexuals are effeminate and that female homosexuals look boyish or anyway too manly. Moreover, while it is hard to imagine any socially beneficial use of such technique, it can be expected that stigmatization and discrimination would likely result from either intentional or unintentional misuses of it. This study exemplifies how AI can power new forms of classification based on the association between biological, personal, behavioral and social characteristics. The unprecedented classificatory power of AI can obviously produce both tangible and

[23] Alessandro Blasimme, Effy Vayena, and Ine Van Hoyweghen, "Big Data, Precision Medicine and Private Insurance: A Delicate Balancing Act," *Big Data & Society* 6, no. 1 (2019): 2053951719830111.
[24] Wendler D Emanuel EJ, "WHat Makes Clinical Research Ethical?," *JAMA* 283, no. 20 (Maggio 2000): 2701–11.
[25] Marcello Ienca et al., "Considerations for Ethics Review of Big Data Health Research: A Scoping Review," *PloS One* 13, no. 10 (2018): e0204937.
[26] Yilun Wang and Michal Kosinski, "Deep Neural Networks Are More Accurate than Humans at Detecting Sexual Orientation from Facial Images.," *Journal of Personality and Social Psychology* 114, no. 2 (2018): 246.

intangible harms[27]. Notably this particular study was reviewed by an institutional review board, it passed peer-review and was eventually published. The heated controversy that followed its publication brought to light the difficulty in assessing societal-wide effects when reviewing research as well as the lack of agreed upon criteria as how to do such an assessment.

Another issue of ethical relevance in the context of health research has emerged from collaborations between corporations with advanced capabilities in AI and health care institutions in control of health data sets. While such collaborations can be mutually beneficial, several examples to date have raised more concern than enthusiasm. The case of Deep Mind accessing 1.6 million health records from the Royal Free London NHS in order to test a kidney safety app, ended with the Information Commissioner finding a number of shortcomings in the contractual agreements. The Italian government's decision to grant an IBM research unit access to citizens' health records has been questioned by both data protection and fair competition officials[28]. Beyond the question of whether such data are used with adequate consent, or whether social benefit will be accrued from their use, the further question is how such benefit will be distributed. If for-profit entities have exclusive deals with national health data organization how will this affect access and distribution of subsequent AI products? We are still in the early days of understanding the implications of such arrangements and of articulating fair

[27] Vanessa K Ing, "Spokeo, Inc. v. Robins: Determining What Makes an Intangible Harm Concrete," *Berkeley Tech. LJ* 32 (2017): 503.
[28] See https://www.repubblica.it/economia/2017/12/05/news/dati_sanitari_alle_multinazionali_senza_consenso_passa_la_norma-183005262/. Accessed: 4 April 2019.

agreements despite the fact that there is a litany of cases that seem to raise the questions.

3) AI in Patient care

AI-driven diagnosis is certainly one of the most promising fields of application for AI in patient care. AI has largely demonstrated its ability to interpret various types of medical images, such as X-ray scans, magnetic resonance and also photographic images of body parts (such as skin or eye fundus) and digitalized pathology slides. Image interpretation and visual pattern recognition are therefore among the major drivers in this space. An obviously limited list of examples includes the use of deep learning techniques to train algorithms to detect wrist fractures in x-ray scans[29]; to help cardiologists interpret magnetic resonance images[30]; and a machine learning software that detects diabetic retinopathy by automatically interpreting images from the back of the patient's eye[31]. All the three above-mentioned applications received FDA clearance for marketing. Many more have appeared in the literature, the most promising of which may become or be embedded in approved medical devices in the near future, including algorithms that can compute cardiovascular risk factors based on retinal images[32]. In all those studies, the performance of the algorithms was tested against the benchmark of certified specialists' assessments, revealing equal or superior outcomes for AI system as compared to human physicians. This criterion is widely used in research settings but it is not yet established as a sufficient one to use of AI applications in clinical care. The issue of evidence standards has obvious implications in terms of safety and efficacy. As a consequence, a major issue

[29] Food and Drug Administration. FDA permits marketing of artificial intelligence algorithm for aiding providers in detecting wrist fractures [Internet]. Available from: https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm608833.htm. Accessed: 4 April 2019.
[30] Marr B. First FDA Approval For Clinical Cloud-Based Deep Learning In Healthcare. Forbes. 2017 Jan 20. Available from: https://www.forbes.com/sites/bernardmarr/2017/01/20/first-fda-approval-for-clinical-cloud-based-deep-learning-in-healthcare/#6af6ceef161c
[31] See https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm604357. Accessed: 4 April 2019.
[32] Ryan Poplin et al., "Prediction of Cardiovascular Risk Factors from Retinal Fundus Photographs via Deep Learning," *Nature Biomedical Engineering* 2, no. 3 (2018): 158.

with clear ethical implications is the reliability of the evidence in favor of AI clinical applications.

Some AI-driven diagnostic applications can also be operated directly by the patient on portable devices outside the clinical setting. One can imagine for example that smartphone apps could incorporate already existing AI-powered algorithms to inspect nevi and detect the presence of skin cancer[33]. Similarly, the first smart pill was approved by the FDA in 2017 and included an ingestible sensor that sends a signal to the patient's device once the pill is taken in order to help him or her adhere to prescription[34]. Commentators have highlighted that, from a patient perspective, ethical issues for this type of devices include concerns for autonomy, privacy and dependability in case of technical failures[35].

Ethical issues in the use of AI for patient care depend on specific uses and applications. It is intuitively plausible to think that ethical stakes correlate with the severity of the condition at hand or with the degree of reliance on AI for serious medical tasks such as diagnosis or treatment. It would be wrong, however, to assume that automation in health system services is less likely to have ethically relevant implications. Consider the case of triage. AI-driven decisions such as which patient is treated first or which one is offered chemotherapy (see *supra*, note 13 above) should certainly follow cost-effectiveness considerations. But exclusive reliance on algorithms may rule out that necessary degree of flexibility that allows healthcare operators to calibrate objective criteria with the reality of each individual case[36]. For instance, a

[33] Andre Esteva et al., "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks," *Nature* 542, no. 7639 (2017): 115.
[34] https://www.fda.gov/newsevents/newsroom/pressannouncements/ucm584933.htm
[35] Craig M Klugman et al., "The Ethics of Smart Pills and Self-Acting Devices: Autonomy, Truth-Telling, and Trust at the Dawn of Digital Medicine," *The American Journal of Bioethics* 18, no. 9 (2018): 38–47.
[36] Effy Vayena, Alessandro Blasimme, and I Glenn Cohen, "Machine Learning in Medicine: Addressing Ethical Challenges," *PLoS Medicine* 15, no. 11 (2018): e1002689.

system that factors the risk of longer stays into decisions about hospital admission may discriminate against the most vulnerable patients, that is, arguably, those that are more in need of care. While it is premature to say that these unfair outcomes will be the case, such ethically relevant aspects of automating clinical workflow have not yet received sufficient attention.

As to the use of AI for diagnostic purposes, the already mentioned problem of biased training dataset that lead to sub-optimal performance for underrepresented social groups creates an ethical bottleneck. In the current ethical debate about AI in medicine, the issue of whether and why the use of AI should be disclosed to patients during informed consent procedures is still in its infancy. However, a bigger discussion is ongoing as to whether black-box algorithms – that is, algorithms whose self-learned rules are too complex to reconstruct and explain – should be used in medicine[37]. Some have called for a duty to transparency in order to dispel the opacity of black-box algorithms[38]. Others, however, have highlighted that more limited requirements are sufficient to adequately protect the morally relevant interests of patients when machine learning algorithms are employed to provide them with care[39].

An important issue concerns the shift of medical authority from human physicians to algorithms – the problem of the so-called 'collective medical mind'[40]. The risk here is that AI-systems introduced as decision support tools become central nodes of medical decision-making. In this scenario, it is uncertain how the established principles of medical ethics (beneficence,

[37] W. Nicholson II Price, "Black-Box Medicine," *Harvard Journal of Law & Technology* 28 (2015 2014): 419.
[38] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi, "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation," *International Data Privacy Law* 7, no. 2 (2017): 76–99.
[39] Andrew D Selbst and Julia Powles, "Meaningful Information and the Right to Explanation," *International Data Privacy Law* 7, no. 4 (2017): 233–42; Agata Ferretti, Manuel Schneider, and Alessandro Blasimme, "Machine Learning in Medicine: Opening the New Data Protection Black Box," *European Data Protection Law Review* 4, no. 3 (2018): 320–32.
[40] Danton S Char, Nigam H Shah, and David Magnus, "Implementing Machine Learning in Health Care - Addressing Ethical Challenges," *The New England Journal of Medicine* 378, no. 11 (March 15, 2018): 981–83, https://doi.org/10.1056/NEJMp1714229.

non-maleficence, respect for patients) can still be expected to play the central role in the patient-doctor relationship that they have – or at least can be expected to have – now. The mediation of AI-powered tools can also fundamentally alter the doctor-patient relationship. AI, especially as it enables remote care or communication via robotic assistants, may create interpersonal distance between patients and their physicians. An incentive to use such tools could be the need to streamline patient care, but the downside of this phenomenon is that the patient becomes more isolated, with potentially negative repercussions on health outcomes. The same considerations can be made about AI-based home assistance platforms. In principle, these systems can be extremely useful to, for instance, providing better care to elderly people with limited mobility. However, they can also increase their social isolation.

The easiness with which an AI system can keep track of a person's health and perform accurate diagnostic has been discussed as a potential source of overdiagnosis and non-actionable diagnoses. For instance, employing deep learning to infer cardiovascular risk factors from retinal fundus pictures[41] is warranted by the fact that it could lead to lifestyle adaptations that may actually improve a patients' condition. But the use of images of retinal structures as biomarkers of dementia[42], are more problematic in the absence of concluding evidence regarding the efficacy of preventive interventions to delay or slow down dementia[43].

Finally, the use of algorithms for mood detection promises to revolutionize mental health[44]. However, privacy issues acquire particular ethical relevance in this context. Tools like DeepMood, that allow the detection of mood based on mobile phone typing dynamics, are

[41] Poplin et al., "Prediction of Cardiovascular Risk Factors from Retinal Fundus Photographs via Deep Learning."
[42] Unal Mutlu et al., "Association of Retinal Neurodegeneration on Optical Coherence Tomography with Dementia: A Population-Based Study," *JAMA Neurology* 75, no. 10 (2018): 1256–63.
[43] Engineering National Academies of Sciences and Medicine, *Preventing Cognitive Decline and Dementia: A Way Forward* (National Academies Press, 2017).
[44] David C Mohr, Heleen Riper, and Stephen M Schueller, "A Solution-Focused Research Approach to Achieve an Implementable Revolution in Digital Mental Health," *JAMA Psychiatry* 75, no. 2 (2018): 113–14.

certainly promising[45]. Yet pervasive tracking of one's emotional state is at least intrusive and may affect the legitimate interest of any individual to keep control over information about his or her mood. Mood and mental health can now be digitally tracked through sensors that capture anything from breathing patterns, to galvanic skin response, from the tone of our voice, to sleep patterns, facial expressions, our whereabouts and social media traces[46]. The possibility of being constantly monitorable as to our emotional states and mental health is certainly problematic from an ethical viewpoint as it sets the conditions for a form of granular psychological surveillance that is at odds with the values of pluralistic liberal societies. Even if these tools are employed in the context of a therapeutic relationship, their excessive use undermines a patients' capacity to remain autonomous and to maintain a sense of self-determination vis-à-vis his or her doctor.

4) AI in public health

Uses of algorithms in public health research and practice can have significant impact on population health[47]. Health is affected by several social parameters (e.g. income, education, dietary habits, environmental factors, community context), that are not confined in the health care systems. Understanding specific effects and interactions between health and various social conditions can lead to the development of more effective and efficient public health programs. Examples from AI-enabled multi-level modeling using sociomarkers have already demonstrated such potential[48]. A particular area of AI application in public health is disease

[45] Bokai Cao et al., "DeepMood: Modeling Mobile Phone Typing Dynamics for Mood Detection" (Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2017), 747–55.
[46] Paddy M Barrett et al., "Digitising the Mind," *The Lancet* 389, no. 10082 (2017): 1877.
[47] Arash Shaban-Nejad, Martin Michalowski, and David L. Buckeridge, "Health Intelligence: How Artificial Intelligence Transforms Population and Personalized Health," *Npj Digital Medicine* 1, no. 1 (October 2, 2018): 53.
[48] Eun Kyong Shin et al., "Sociomarkers and Biomarkers: Predictive Modeling in Identifying Pediatric Asthma Patients at Risk of Hospital Revisits," *Npj Digital Medicine* 1, no. 1 (October 2, 2018): 50.

surveillance. Surveillance systems monitor disease incidence, outbreaks and health behaviors. Typically these systems are state-funded and state-operated. Their purpose is to monitor the health of populations and subsequently to support decision making for allocation of resources and types of interventions necessary to improve health. As a data-driven activity, surveillance can benefit substantially from algorithmic uses. Algorithms can sort through variables that are relevant for specific health outcomes, they can recognize patterns and signals at a much faster pace and they can be used to forecast epidemics and to model their trajectories. Such algorithms have been used to mine not only standard health data collected for surveillance by state institutions, but also real-world data through social media. This seemingly unconventional approach suffered an early blow when Google Flu Trend algorithms failed to show their promised predictive power[49]. Since then however, AI-enabled analysis of social media data has produced several successful examples including better prediction of epidemics[50], detection of food poisoning cases[51]. The broader field of digital epidemiology, is a rapidly evolving field focused on epidemiological models based on content posted online by social network users[52]. Forms of AI like natural language processing obviously play a crucial role for the further development of this field. Ethical challenges in this domain revolve mainly around consent. Many commentators have stressed that the terms of use for social media fall short of complying with the rigorous requirements for informed consent in the domain of health-

[49] Declan Butler, "When Google Got Flu Wrong," *Nature News* 494, no. 7436 (2013): 155.
[50] Mohammed Ali Al-garadi et al., "Using Online Social Networks to Track a Pandemic: A Systematic Review," *Journal of Biomedical Informatics* 62 (August 1, 2016): 1–11.
[51] Jenine K. Harris et al., "Using Twitter to Identify and Respond to Food Poisoning: The Food Safety STL Project.," *Journal of Public Health Management and Practice : JPHMP* 23, no. 6 (December 2017): 577–80,.
[52] Marcel Salathé et al., "Digital Epidemiology," *PLOS Computational Biology* 8, no. 7 (lug 2012): e1002616, https://doi.org/10.1371/journal.pcbi.1002616; Antoine Flahault et al., "Precision Global Health in the Digital Age," *Swiss Medical Weekly* 147 (April 19, 2017): w14423, https://doi.org/smw.2017.14423.

related research[53].

AI combined with mobile health applications also offers a new avenue for delivering public health intervention to populations. Of relevance here are expectation for health promotion to reach populations that are marginalized by targeting them with tailored interventions[54]. An area of contest in public health ethics has been the ethical legitimacy of nudging personal behavior for health-related purposes. This is an issue that in an AI-enabled public health will generate significant concern. Continuous surveillance, tailored nudging and paternalistic interventions can generate an Orwelian form of individual control and constrained personal freedoms[55]. States and corporations with access to tools that can monitor and alter health-related behaviors, can exercise significant power over large numbers of people to further their specific interests. While in a democratic and accountable state such policies can be vetted, be transparent and revised as necessary, that is not necessarily the case everywhere nor is it the case when such behavioral manipulation occurs in arenas that are controlled entirely by institutions without public accountability.

There is significant enthusiasm for the use of AI in global health with funding agencies and international organizations investing already in public health activities in low and middle income countries. The World Health Organization, has recently committed to promote AI to achieve universal health coverage and many governments have been interested in taking stock of digital technologies to improve health care systems as they stated in a 2018

[53] Jeffrey P. Kahn, Effy Vayena, and Anna C. Mastroianni, "Opinion: Learning as We Go: Lessons from the Publication of Facebook's Social-Computing Research," *Proceedings of the National Academy of Sciences* 111, no. 38 (September 23, 2014): 13677–79.
[54] Brian Wahl et al., "Artificial Intelligence (AI) and Global Health: How Can AI Contribute to Health in Resource-Poor Settings?," *BMJ Global Health* 3, no. 4 (2018): e000798.
[55] Sarah Nettleton and Robin Bunton, "Sociological Critiques of Health Promotion," *The Sociology of Health Promotion*, 1995, 41–58.

resolution in digital health that was adopted by the 71st world Health Assembly[56]. This commitment increases the likelihood of AI entering rapidly the domain of health, adding urgency to the need of identifying and addressing the ethical tensions that AI generates[57]. The most pertinent are those related to the potential exacerbation of health disparities through biases that are perpetuated or reinforced by AI-enabled interventions. We discussed the problem of misrepresentation of certain populations in the data sets already. Several methods are currently under development to remedy bias problems in algorithms but in the meantime the problem remains and requires attention[58]. Underserved populations present certain negative health outcomes due to well-known social deficits. Algorithms that spit out decisions based on health outcomes alone, without factoring in their social causes can result in significant harm and increased health inequalities. For example, if poor, or less educated people have performed worse after certain health interventions (due to poor access to care, working schedules etc.) an algorithm can determine that people with these characteristics will always perform worse and recommend that they are not offered the intervention in the first place. This will exacerbate disparity in access to care and attainment of good health outcomes. More importantly, it will make such disparity less visible because the decision will bear the authoritative objectivity often attributed to numbers and that it typically expected from automated decision-making tools.

5) Addressing the ethical challenges

The novelty represented by AI, and machine learning in particular, might be on the verge of pushing medical research, patient care and public health into as yet uncharted ethical

[56] See http://apps.who.int/gb/ebwha/pdf_files/WHA71/A71_R7-en.pdf . Accessed: 4 April 2019.
[57] Effy Vayena and Lawrence Madoff, "Navigating the Ethics of Big Data in Public Health," in *The Oxford Handbook of Public Health Ethics*, n.d.
[58] Robert Challen et al., "Artificial Intelligence, Bias and Clinical Safety," *BMJ Quality & Safety* 28, no. 3 (March 1, 2019): 231.

territories. The impact of AI in these three domains is particularly challenging to anticipate, and in it is hard to predict whether expected benefits will offset emerging risks. In this scenario neither a precautionary approach nor a wait-and-see attitude is compatible with the widely accepted need to ensure ethically sustainable, socially robust and responsible innovation in this domain. A precautionary approach implies erring on the side of containing possible risks when evidence about how a given phenomenon will evolve is scarce and the stakes are high in terms of potential harms[59]. As far as the use of AI in medicine is concerned, a precautionary approach would likely result in disproportionate constraints that might undermine the development of promising technologies. On the other hand, a more permissive "wait-and-see" approach, while being more favorable to the development and rapid uptake of AI-driven solutions, would necessarily have to rely on existing ethical safeguards. But such safeguards, as we have seen, fall short of covering the rapidly expanding catalogue of ethical issues that AI poses in the domain of biomedicine. The collection, use, and re-use of increasingly large amounts of personal data, for instance, calls into question the adequacy of key components of the existing regulatory toolkit, such evidence standards, ethics review and informed consent[60].

What is needed to ensure responsible AI innovation is a governance approach that co-evolves with the field itself, incorporating new governance actors and experimenting with new oversight mechanisms to cope with ethical challenges as they arise from practice. Such a governance model should primarily drive attention to the ethically controversial aspects of AI-driven innovation in biomedicine, in order to ensure that emerging risks do not pass unnoticed. A second aim of an ideal governance frame would by that of channeling innovation

[59] Elizabeth Charlotte Fisher, Judith S Jones, and René von Schomberg, *Implementing the Precautionary Principle: Perspectives and Prospects* (Edward Elgar Publishing, 2006).
[60] Effy Vayena et al., "Digital Health: Meeting the Ethical and Policy Challenges," *Swiss Medical Weekly* 148 (2018): w14571.

toward socially beneficial outcomes. Finally, good governance should promote public trust in, and accountability of the innovation process. These objectives demand a specific *systemic* approach to governing a complex phenomenon whose outcomes are still largely unpredictable.

In the last two decades, scholarship on governance of controversial areas of science and innovation has given substantial consideration to so-called adaptive governance, as a model to cope with uncertainty in public policy[61]. Adaptive governance centers around constant monitoring of both the phenomenon at stake and the policy measures deployed to control it. In practical terms, this model invites oversight and regulation to take stock of evidence as it becomes available and promoting social learning among a variety of different governance stakeholders[62]. Drawing on the broad frame of adaptive governance, we have proposed a governance model for data-driven innovation in biomedicine called 'systemic oversight'[63]. Systemic oversight is specifically designed to address what gives rise to ethical issues in the use of big data and AI in biomedicine, that is, as we have seen, novel data sources, novel data uses, increased capacity to draw connections between disparate data points, and uncertainty about downstream effects of such increased classificatory powers. The systemic oversight approach is based on six components offering guidance as to the desirable features of oversight structures and processes in the domain of data-intense biomedicine: **a**daptivity, **f**lexibility **i**nclusiveness, **r**eflexivity, **r**esponsiveness and **m**onitoring (the first letters of the components

[61] Carl Folke et al., "Adaptive Governance of Social-Ecological Systems," *Annual Review of Environment and Resources* 30, no. 1 (2005): 441–473.
[62] Brian Chaffin, Hannah Gosnell, and Barbara A Cosens, "A Decade of Adaptive Governance Scholarship: Synthesis and Future Directions," *Chaffin BC, Gosnell H, Cosens BA*, 2014.
[63] Vayena and Blasimme, "Health Research with Big Data: Time for Systemic Oversight"; Blasimme and Vayena, "Towards Systemic Oversight in Digital Health: Implementation of the AFIRRM Principles."

form the acronym AFIRRM).

Adaptivity refers to the capacity of governance bodies and mechanisms to guarantee appropriate forms of oversight for new data sources and new data analytics that get incorporated in research, patient care or public health activities. Flexibility is the capacity to treat different data types depending both on their source and on their actual use, and it is based on the consideration that data acquire specific ethical meaning in different contexts of use. Inclusiveness – one of the key notions in adaptive governance – stresses the need to include all affected parties in deliberations and decision-making practices about the use of data and algorithms in specific ambits. This component refers in particular to communities and actors that are historically marginalized, vulnerable or otherwise excluded from the circuits of power, such as minorities and patient constituencies. Reflexivity requires careful scrutiny and assessment of emerging risks in the short as well as in the long run in terms of the downstream effects of big data and AI on interests, rights and values, for example in terms of fair access to healthcare services, discrimination, stigmatization, medicalization, overdiagnosis and so on. Responsiveness refers instead to the need for adequate mechanisms to mitigate the effects of unintended issues such as unauthorized access to personal health-related information. We saw above that AI is a powerful generator of such information and thus exposes research participants, patients and data subjects in general to unwanted leaks of personal data and information. Finally, monitoring expresses the need to predispose regular scrutiny of data-related activities and their effects on health-related practices in order to anticipate the emergence on new vulnerabilities and undesirable outcomes.

The implementation of the AFIRRM frame will require consideration for the well-characterized obstacles to adaptive governance in other policy domains. Particular attention needs to

be paid to the composition of oversight bodies. The demands of inclusiveness, for example, can only be appropriately fulfilled if diverse stakeholders share at least a common understanding of the intended advantages and potential risks of using AI in biomedicine. It is possible, for instance, that automating hospital services through AI-driven triage systems caters to the financial interests of hospitals (by rationalizing resource allocation), while failing to meet the expectations of severely ill patients in terms of access to care. As a consequence, the inclusion of patients' perspectives into decisions about the adoption of such systems both requires and fosters the existence of shared visions about fairness in access to health services. Along similar lines, oversight mechanisms on the use and effects of AI in clinical practice must escape purely technical considerations about the safety and efficacy of automated clinical decisions. Downstream effects on the patient-doctor relationship, or on the right of patients to decide whether they are open or not to highly automated decisions need to be considered. To this aim new review processes for clinical validation, as well as novel communication and consent requirements will have to be established. The same applies in the research domain when researchers interested in using large amounts of phenotypic data, need to negotiate the terms of use with data subjects, some of which may have value-laden views about the ethical legitimacy of certain types of research.

With the advent of AI, the agenda of academic disciplines like clinical research ethics, medical ethics and public health ethics is rapidly adapting to incorporate new issues and new controversies. Given its theoretical and thematic specificity, one may characterize this area as a separate sub-area of study in applied ethics, and call it digital bioethics. Whether and how this scholarship will inform the emergence of new oversight tools remains to be seen. In the meantime, practical proposals, criteria and best practices about the governance of AI-driven innovation in biomedicine are just starting to emerge. The UK National Institute for Clinical

Excellence (NICE), the body advising the National Health Service (NHS) on matters related to health technology assessment, has just released guidance on clinical validation of digital health technologies (DHTs)[64]. This guidance establishes evidence standards (grouped in four evidence tiers) according to the function that a given DHT is intended to perform. Such standards are going to be applied also to DHTs harboring an AI component or to stand-alone AI software. In February 2019 the NHS released an updated version of its Code of Conduct for Data-driven Health and Care Technologies[65]. The principles proposed by this code include understanding users' needs, clearly defining the expected outcomes and benefits, lawful data processing, transparency and evidence of safety and effectiveness (based on the NICE criteria). The NHS frame has been criticized for its lack of attention to the risk that AI in the healthcare space may widen social inequalities[66]. Still in the UK, The Wellcome Trust – a major funder of biomedical research in the country – has recently proposed a model called 'dynamic oversight' for emerging science and technologies that partially resembles our own systemic oversight approach and the AFIRRM principles[67].

In the US the American Medical Association (AMA) released its policy on AI in 2018[68]. This document highlights the transformative potential of AI in the clinical domain and recommends that clinically validated AI should be aligned to best clinical practices, be transparent, be reproducible, be immune to data biases, and protect patients' privacy as well as the integrity of their personal information. In the US, the FDA is the gatekeeper of AI-driven health innovation as it has statutory oversight power on medical devices and software as medical

[64] See https://www.nice.org.uk/Media/Default/About/what-we-do/our-programmes/evidence-standards-framework/digital-evidence-standards-framework.pdf. Accessed: 4 April 2019.
[65] See https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology. Accessed: 4 April 2019.
[66] Melanie Smallman, "Policies Designed for Drugs Won't Work for AI.," *Nature* 567, no. 7746 (2019): 7.
[67] See https://wellcome.ac.uk/sites/default/files/blueprint-for-dynamic-oversight.pdf. Accessed: 4 April 2019.
[68] See https://www.ama-assn.org/system/files/2019-01/augmented-intelligence-policy-report.pdf. Accessed: 4 April 2019.

device. In Europe, instead, the new 2017 Regulation on Medical Devices[69] relies on third-parties (called notified bodies) issuing conformity certificates for medical devices. The FDA is piloting a pre-certification program to identify "manufacturers who have demonstrated a robust culture of quality and organizational excellence, and who are committed to monitoring real-world performance of their products once they reach the U.S. market"[70]. In April 2019, the FDA also released a proposed regulatory framework for AI and machine learning medical software addressing the specific issue of algorithms that keep on training themselves based on new data acquired during clinical use[71].

6) Conclusions

The current proliferation of guidelines and codes of conduct demonstrates the need for ethical and technical points of reference for this rapidly evolving field. Considering the broad scope of potential applications for research, clinical use and public health, it is likely that some specific uses of AI will not be covered by existing oversight mechanisms. But reliance on existing regulatory tools alone will likely fail to ensure adequate levels of public trust and accountability. For this reason, we have advanced the systemic oversight/AFIRRM approach as a governance blueprint. Looking at the nature of ethical issues illustrated in this chapter in light of the AFIRRM principles, it seems at least advisable that certain measure be implemented in the short term. In the research domain, ethical review committees will have to incorporate reflexive assessment of the scientific and social merits of AI-driven research and, as a consequence, will have to open their ranks to new professional figures such as social scientists. Research funders, on the other hand, can require monitoring and responsiveness mechanisms to be part of research plans and could set up multi-disciplinary committees to

[69] See https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32017R0745. Accessed: 4 April 2019.
[70] See https://www.fda.gov/MedicalDevices/DigitalHealth/UCM567265. Accessed: 4 April 2019.
[71] See https://www.regulations.gov/document?D=FDA-2019-N-1185-0001 . Accessed: 4 April 2019.

periodically assess data from such activities in order to adjust their funding policies in the future. When AI-driven research amounts to large-scale project claiming data from entire communities or populations, adequate forms of inclusion must be experimented with in order to ensure social learning across different epistemic communities – including lay publics and non-academic actors.

In the domain of patient care, clinical validation is a crucial issue. Ad hoc evidence standards are a necessary condition for responsible clinical innovation, but they are not sufficient to cover the breath of potential ethical issues we saw in this area. Hospitals could equip themselves with 'clinical AI oversight bodies' charged with the task of advising clinical administrators regarding the adoption of a given AI technology, and the periodic monitoring of its effects on patient journeys and patients' engagement throughout the continuum of care. Moreover, consent requirements will need to be adapted to the presence of highly automated data-processing, for instance in the domain of diagnostics.

In the public health sphere, the new level of granularity enabled by AI in disease surveillance or health promotion will have to be negotiated at the level of targeted communities or it will result in a sense of disempowerment and, as a consequence, in a lack of public trust. The acceptable limits of data collection and algorithmic analysis, in other words, will have to result from community-wide inclusive deliberation, especially as to who is collecting and processing data and for which exact purposes.

These are just a few examples of initiatives that, if adopted, will contribute to the development AI into a socially robust technology. It is clear that we are at the very beginning of a foreseen transformation. Should this transformation occur, its real effects may be different

from those that we are able to anticipate now. This level of uncertainty, however, shall not deter societal stakeholders – including scientific and clinical institutions – from experimenting with governance arrangements aimed at reaping the benefits of AI for human knowledge and health, while at the same time paying sufficient attention to emerging ethical challenges.

Bibliography

Char, Danton S, Nigam H Shah, and David Magnus. "Implementing Machine Learning in Health Care - Addressing Ethical Challenges." *The New England Journal of Medicine* 378, no. 11 (March 15, 2018): 981–83.

He, Jianxing, Sally L. Baxter, Jie Xu, Jiming Xu, Xingtao Zhou, and Kang Zhang. "The Practical Implementation of Artificial Intelligence Technologies in Medicine." *Nature Medicine* 25, no. 1 (January 2019): 30.

Price, W. Nicholson II. "Black-Box Medicine." *Harvard Journal of Law & Technology* 28 (2015 2014): 419.

Smallman, Melanie. "Policies Designed for Drugs Won't Work for AI." *Nature* 567, no. 7746 (2019): 7.

Topol, Eric J. "High-Performance Medicine: The Convergence of Human and Artificial Intelligence." *Nature Medicine* 25, no. 1 (2019): 44.

Vayena, Effy, Alessandro Blasimme, and I Glenn Cohen. "Machine Learning in Medicine: Addressing Ethical Challenges." *PLoS Medicine* 15, no. 11 (2018): e1002689.

Vayena, Effy, and Alessandro Blasimme. "Health Research with Big Data: Time for Systemic Oversight." *The Journal of Law, Medicine & Ethics* 46, no. 1 (2018): 119–29.

Yu, Kun-Hsing, Andrew L. Beam, and Isaac S. Kohane. "Artificial Intelligence in Healthcare." *Nature Biomedical Engineering* 2, no. 10 (October 1, 2018): 719–31.