Published in final edited form as:
Ann Intern Med. 2018 December 18; 169(12): 866–872. doi:10.7326/M18-1990.

Ensuring Fairness in Machine Learning to Advance Health Equity

Alvin Rajkomar, MD[*],
Google, Mountain View, and University of California, San Francisco, San Francisco, California

Michaela Hardt, PhD[*],
Google, Mountain View, California

Michael D. Howell, MD, MPH,
Google, Mountain View, California

Greg Corrado, PhD, and
Google, Mountain View, California

Marshall H. Chin, MD, MPH
University of Chicago, Chicago, Illinois

Abstract

Machine learning is used increasingly in clinical care to improve diagnosis, treatment selection, and health system efficiency. Because machine-learning models learn from historically collected data, populations that have experienced human and structural biases in the past—called *protected groups*—are vulnerable to harm by incorrect predictions or withholding of resources. This article describes how model design, biases in data, and the interactions of model predictions with clinicians and patients may exacerbate health care disparities. Rather than simply guarding against these harms passively, machine-learning systems should be used proactively to advance health equity. For that goal to be achieved, principles of distributive justice must be incorporated into model design, deployment, and evaluation. The article describes several technical implementations of distributive justice—specifically those that ensure equality in patient outcomes, performance, and resource allocation—and guides clinicians as to when they should prioritize each principle. Machine learning is providing increasingly sophisticated decision support and population-level monitoring, and it should encode principles of justice to ensure that models benefit all patients.

Corresponding Author: Alvin Rajkomar, MD, Google LLC, 1600 Amphitheatre Way, Mountain View, CA 94043; alvinrajkomar@google.com.

[*]Drs. Rajkomar and Hardt contributed equally to this work.

Current author addresses and author contributions are available at Annals.org.

Current Author Addresses: Drs. Rajkomar, Hardt, Howell, and Corrado: Google LLC, 1600 Amphitheatre Way, Mountain View, CA 94043.
Dr. Chin: University of Chicago, 5841 South Maryland Avenue, MC2007, Chicago, IL 60637.

Author Contributions: Conception and design: A. Rajkomar, M. Hardt.
Analysis and interpretation of the data: A. Rajkomar, M. Hardt.
Drafting of the article: A. Rajkomar, M. Hardt, M.D. Howell.
Critical revision for important intellectual content: A. Rajkomar, M. Hardt, M.D. Howell, G. Corrado, M.H. Chin.
Final approval of the article: A. Rajkomar, M. Hardt, M.D. Howell, G. Corrado, M.H. Chin.
Statistical expertise: A. Rajkomar, M. Hardt.
Administrative, technical, or logistic support: M.D. Howell.

Disclosures: Disclosures can be viewed at www.acponline.org/authors/icmje/ConflictOfInterestForms.do?msNum=M18-1990.

Machine learning can identify the statistical patterns of data generated by tens of thousands of physicians and billions of patients to train computers to perform specific tasks with sometimes superhuman ability, such as detecting diabetic eye disease better than retinal specialists (1). However, historical data also capture patterns of health care disparities, and machine-learning models trained on these data may perpetuate these inequities. This concern is not just academic. In a model used to predict future crime on the basis of historical arrest records, African American defendants who did not reoffend were classified as high risk at a substantially higher rate than white defendants who did not reoffend (2, 3). Similar biases have been observed in predictive policing (4) and identifying which calls to a child protective services agency required an in-person investigation (5, 6). The implications for health care led the American Medical Association to pass policy recommendations to "promote development of thoughtfully designed, high-quality, clinically validated health care AI [artificial or augmented intelligence, such as machine learning] that … identifies and takes steps to address bias and avoids introducing or exacerbating health care disparities including when testing or deploying new AI tools on vulnerable populations" (7).

We argue that health care organizations and policymakers should go beyond the American Medical Association's position of doing no harm and instead proactively design and use machine-learning systems to advance health equity. Whereas much health disparities work has focused on discriminatory decision making and implicit biases by clinicians, policymakers, organizational leaders, and researchers are increasingly focusing on the ill health effects of structural racism and classism—how systems are shaped in ways that harm the health of disempowered, marginalized populations (8). For example, the United States has a shameful history of purposive decisions by government and private businesses to segregate housing. Zoning laws, discrimination in mortgage lending, prejudicial practices by real estate agents, and the ghettoization of public housing all contributed to the concentration of urban African Americans in inferior housing that has led to poor health (9, 10). Even when the goal of decision makers is not outright discrimination against disadvantaged groups, actions may lead to inequities. For example, if the goal of a machine-learning system is to maximize efficiency, that might come at the expense of disadvantaged populations.

As a society, we value health equity. For example, the Healthy People 2020 vision statement aims for "a society in which all people live long, healthy lives," and one of the mission's goals is to "achieve health equity, eliminate disparities, and improve the health of all groups" (11). The 4 classic principles of Western clinical medical ethics are justice, autonomy, beneficence, and nonmaleficence. However, health equity will not be attained unless we purposely design our health and social systems, which increasingly will be infused with machine learning (12), to achieve this goal.

To ensure fairness in machine learning, we recommend a participatory process that involves key stakeholders, including frequently marginalized populations, and considers distributive justice within specific clinical and organizational contexts. Different technical approaches can configure the mathematical properties of machine-learning models to render predictions that are equitable in various ways. The existence of mathematical levers must be supplemented with criteria for *when* and *why* they should be used—each tool comes with tradeoffs that require ethical reasoning to decide what is best for a given application.

We propose incorporating fairness into the design, deployment, and evaluation of machine-learning models. We discuss 2 clinical applications in which machine learning might harm protected groups by being inaccurate, diverting resources, or worsening outcomes, especially if the models are built without consideration for these patients. We then describe the mechanisms by which a model's design, data, and deployment may lead to disparities; explain how different approaches to distributive justice in machine learning can advance health equity; and explore what contexts are more appropriate for different equity approaches in machine learning.

Case Study 1: Intensive Care Unit Monitoring

A common area of predictive modeling research focuses on creating a monitoring system—for example, to warn a rapid response team about inpatients at high risk for deterioration (13–15), requiring their transfer to an intensive care unit within 6 hours. How might such a system inadvertently result in harm to a protected group? In this thought experiment, we consider African Americans as a protected group.

To build the model, our hypothetical researchers collected historical records of patients who had clinical deterioration and those who did not. The model acts like a "diagnostic test" of risk for intensive care unit transfer. However, if too few African American patients were included in the training data—the data used to construct the model—the model might be inaccurate for them. For example, it might have a lower sensitivity and miss more patients at risk for deterioration. African American patients might be harmed if clinical teams started relying on alerts to identify at-risk patients without realizing that the prediction system underdetects patients in that group (automation bias) (16). If the model had a lower positive predictive value for African Americans, it might also disproportionately harm them through dismissal bias—a generalization of alert fatigue in which clinicians may learn to discount or "dismiss" alerts for African Americans because they are more likely to be false-positive (17).

Case Study 2: Reducing Length of Stay

Imagine that a hospital created a model with clinical and social variables to predict which inpatients might be discharged earliest so that it could direct limited case management resources to them to prevent delays. If residence in ZIP codes of socioeconomically depressed or predominantly African American neighborhoods predicted greater lengths of stay (18), this model might disproportionately allocate case management resources to patients from richer, predominantly white neighborhoods and away from African Americans in poorer ones.

What Is Machine Learning?

Traditionally, computer systems map inputs to outputs according to manually specified "if–then" rules. With increasingly complex tasks, such as language translation, manually specifying rules becomes infeasible, and instead the mapping (or model) is *learned* by the system given only input examples represented through a set of *features* together with their desired output, referred to as *labels*.

The quality of a model is assessed by computing evaluation metrics on data not used to build the model, such as sensitivity, specificity, or the c-statistic, which measures the ability of a model to distinguish patients with a condition from those without it (19, 20). Once the model's quality is deemed satisfactory, it can be deployed to make predictions on new examples for which the label is unknown when the prediction is made. The quality of the models on retrospective data must be followed with tests of clinical effectiveness, safety, and comparison with current practice, which may require clinical trials (21).

Traditionally, statistical models for prediction, such as the pooled-cohort equation (22), have used few variables to predict clinical outcomes, such as cardiovascular risk (23). Modern machine-learning techniques, however, can consider many more features. For example, a recent model to predict hospital readmissions examined hundreds of thousands of pieces of information, including the free text of clinical notes (24). Complex data and models can drive more personalized and accurate predictions but may also make algorithms hard to understand and trust (25).

What Can Cause a Machine-Learning System to Be Unfair?

The Glossary lists key biases in the design, data, and deployment of a machine-learning model that may perpetuate or exacerbate health care disparities if left unchecked. The Figure reveals how the various biases relate to one another and how the interactions of model predictions with clinicians and patients may exacerbate health care disparities. Biases may arise during the design of a model. For example, if the label is marred by health care disparities, such as predicting the onset of clinical depression in environments where protected groups have been systematically misdiagnosed, then the model will learn to perpetuate this disparity. This represents a generalization of test-referral bias (26) that we refer to as label bias. Moreover, the data on which the model is developed may be biased. Data on patients in the protected group might be distributed differently from those in the nonprotected group because of biological or nonbiological variation (9, 27). For example, the data may not contain enough examples from a group to properly tailor the predictions to them (minority bias) (28), or the data set of the protected group may be less informative because features are missing not at random as a result of more fragmented care (29, 30).

The immediate effect of these differences is that the model may not be as accurate for patients in the protected class, but the effects on patient outcomes and resource allocation are usually mediated through how clinicians and administrators interact with the model. For example, do clinicians trust the model even when it is wrong (automation bias) or ignore it when they should not (dismissal bias)? Will administrators use a flawed model to determine which patients are at high risk for poor outcomes and who should then receive more assistance?

Patients in the protected group may also be negatively affected by privilege bias if models are not built for diseases that disproportionately affect them or if models are disproportionately deployed to areas where they do not seek care (for example, concierge practice vs. safety-net clinic) (31). They also may be affected by informed mistrust if protected groups distrust using models for their own care (32).

Distributive Justice Options in Machine Learning

What can be done to mitigate the biases that make a model unfair? We propose using 3 central axes inspired by principles of distributive justice to understand fairness in machine learning.

Equal Outcomes

Equal patient outcomes refers to the assurance that protected groups have equal benefit in terms of patient outcomes from the deployment of machine-learning models (33). A weak form of equal outcomes is ensuring that both the protected and nonprotected groups benefit similarly from a model (equal benefit); a stronger form is making sure that both groups benefit and *any* outcome disparity is lessened (equalized outcomes). Ensuring equal outcomes is the most critical aspect of fairness and can be advanced by interventions proactively designed to reduce disparities (34, 35). It may be hard to know in advance, though, if any well-intentioned general, nontailored intervention, whether a quality improvement project or a machine-learning system, might disproportionately harm or benefit a protected group. However, besides equal outcomes, other options that might advance health equity can be analyzed and addressed prospectively.

Equal Performance

If a model systematically makes errors disproportionately for patients in the protected group, it is likely to lead to unequal outcomes. *Equal performance* refers to the assurance that a model is equally accurate for patients in the protected and nonprotected groups. Equal performance has 3 commonly discussed types: equal sensitivity (also known as equal opportunity [36]), equal sensitivity *and* specificity (also known as equalized odds), and equal positive predictive value (commonly referred to as predictive parity [37]). Not only can these metrics be calculated, but techniques exist to force models to have one of these properties (36, 38–41).

When should each type of equal performance be considered? A higher false-negative rate in the protected group in case 1 would mean African American patients were missing the opportunity to be identified; in this case, equal sensitivity is desirable. A higher false-positive rate might be especially deleterious by leading to potentially harmful interventions (such as unnecessary biopsies), motivating equal specificity. When the positive predictive value for alerts in the protected group is lower than in the nonprotected groups, clinicians may learn that the alerts are less informative for them and act on them less (a situation known as class-specific alert fatigue). Ensuring equal positive predictive value is desirable in this case.

Equal performance, however, may not necessarily translate to equal outcomes. First, the recommended treatment informed by the prediction may be less effective for patients in the protected group (for example, because of different responses to medications and a lack of research on heterogeneous treatment effects [42]). Second, even if a model is inaccurate for a group, clinicians might compensate with additional vigilance, overcoming the model's deficiencies.

Third, forcing a model's predictions to have one of the equal performance characteristics may have unexpected consequences. In case 1, ensuring that a model will detect African American and non–African American patients at equal rates (equal sensitivity) could be straightforwardly accomplished by lowering the threshold for the protected class to receive the intervention. This simultaneously increases the false-positive rate for this group, manifesting as more false alarms and subsequent class-specific alert fatigue. Likewise, equalized odds can be achieved by *lowering* accuracy for the nonprotected group, which undermines the principle of beneficence.

Equal Allocation

Predictions are often used to allocate resources, such as in case 2, in which some patients are given additional case management. The third type of equity is equal allocation (also known as demographic parity [43]), which ensures that the resources are proportionately allocated to patients in the protected group. Because the comorbidity distribution may differ across groups, the desired allocation might first be adjusted for relevant variables (44). This is distinct from equal performance, because allocation is determined by the rate of *positive* predictions (such as predictions above a threshold) without regard to their accuracy.

In some cases, judging accuracy is misleading when labels have historical bias, explaining why equal allocation may be preferable. Consider a model to identify which patients presenting emergently with chest pain should automatically activate a cardiac catheterization team. If African American women were historically sent for this procedure at inappropriately low rates compared with white men (45), then "correct" predictions (based on historical data) would underidentify these women. Equal allocation could be used to lower the threshold for African American women so that the catheterization laboratory would be activated at equal rates across groups, thereby correcting for past bias. This may not necessarily translate to equal outcomes if it leads to a higher rate of false-positive activations of the laboratory with respect to *actual clinical need* or to a continuation of lower true-positive rates if clinicians dismiss the predictions because of the underlying bias against recommending the procedure for women. Whether the net effect of the model is a reduction in health care disparities, especially compared with *not* implementing a model, is uncertain.

Tradeoffs

The computer science community was rocked when a machine-learning model used to help predict which criminal defendants were at risk for committing a future crime was found to be unfair with respect to equalized odds: African American defendants who did not reoffend were classified as high risk at a substantially higher rate than white defendants who did not reoffend. The model builders, however, asserted that the model had equal positive and negative predictive value across the groups (46). Subsequent analysis revealed that various types of fairness are sometimes incompatible: A model may be fair with respect to equal positive and negative predictive value but unfair with respect to equalized odds (or vice versa), but it is impossible for any model to satisfy both. This impossibility also holds for equalized odds and equal allocation, and for equal allocation and equal positive and negative predictive value (37). Machine-learning fairness is not just for machine-learning specialists to understand; it requires clinical and ethical reasoning to determine which type of fairness is appropriate for a given application and what level of it is satisfactory. Although no cookie-cutter solution exists, the examples and recommendations provide a starting point for this reasoning. We believe that in practice, satisfactory levels of the desired fairness types can be achieved.

Recommendations

In the Table, we present recommendations for how to incorporate fairness into machine learning. Researchers should consider how prior health care disparities may affect the design and data of a model. For example, if advanced-stage melanoma is diagnosed more frequently in patients with dark skin than in other groups, might a skin cancer detection model fail to detect early-stage disease in patients with dark skin (47, 48)? During training and evaluation, researchers should measure any deviations from equal accuracy and equal allocation, and consider mitigating them by using techniques during training (38–40) or by postprocessing a trained model (30, 36, 41). Before deployment, launch reviews should formally assess model performance and allocation of resources across groups. The reviews should determine whether a model promotes equal outcomes, broadly defined as “the patient’s care experience, functional status, and quality of life, as well as… personalization of care and resource stewardship” (49). If a model is deployed, the performance of the model and outcome measurements should be monitored, possibly through formal trial design (such as stepped-wedge trials [50]). Moreover, the model may be improved over time by collecting more representative or less biased data.

We purposefully do not recommend the commonly discussed fairness principle of “unawareness,” which states that a model should not use the membership of the group as a feature. Complex models can infer a protected attribute even if it is not explicitly coded in a data set, such as a model identifying a patient’s self-reported sex from a retinal image even though ophthalmologists cannot (51). Moreover, removing features may lead to poorer performance for all patients.

Conclusions

Consideration of fairness in machine learning allows us to reexamine historical bias and proactively promote a more equitable future. We are optimistic that machine learning can substantially improve the care delivered to patients if it is thoughtfully designed and deployed. Case 2 is based on a University of Chicago Medicine example in which data scientists from the Center for Healthcare Delivery Science and Innovation collaborated with experts from the Diversity and Equity Committee to identify the equity problem and to design a local checklist for model building and deployment that advances equity.

Machine-learning fairness is not just about preventing a model from harming a protected group; it may also help focus care where it is really needed. Models could be used to provide translation services where inperson interpreters are scarce, provide medical expertise in areas with a limited number of specialists, and even improve diagnostic accuracy for rare conditions that are often misdiagnosed. By including fairness as a central consideration in how the models are designed, deployed, and evaluated, we can ensure that *all* patients benefit from this technology.

Acknowledgment:

The authors thank Meredith Whittaker, Roxanne Pinto, Moritz Hardt, Gerardo Flores, Charina Chou, Kathryn Rough, Ashley Hayes, Jutta Williams, Katherine Chou, Andrew Smart, Alex Beutel, Valeria Espinosa, Adam Sadilek, Kaspar Molzberger, and Yoni Halpern for insightful comments on the interplay of fairness in building and deploying machine learning and for helpful comments on early versions of the manuscript. They also thank John Fahrenbach, James Williams, and their colleagues from the Chicago Center for Diabetes Translation Research for providing critical analysis of the ideas of the manuscript applied to real clinical prediction tasks.

Grant Support: Dr. Chin was supported in part by the Chicago Center for Diabetes Translation Research (grant NIDDK P30 DK092949), Robert Wood Johnson Foundation Finding Answers: Solving Disparities Through Payment and Delivery Reform Program Office, and Merck Foundation Bridging the Gap: Reducing Disparities in Diabetes Care National Program Office.

Glossary

Biases in model design

Label bias: A label that does not mean the same thing for all patients because it is an imperfect proxy that is subject to health care disparities rather than an adjudicated truth. This is a generalization of test-referral and test-interpretation bias in the statistics literature

Cohort bias: Defaulting to traditional or easily measured groups without considering other potentially protected groups or levels of granularity (e.g., whether sex is recorded as male, female, or other or more granular categories)

Biases in training data

Minority bias: The protected group may have insufficient numbers of patients for a model to learn the correct statistical patterns

Missing data bias: Data may be missing for protected groups in a nonrandom fashion, which makes an accurate prediction hard to render (e.g., a model may underdetect clinical deterioration in patients under contact isolation because they have fewer vital signs)

Informativeness bias: Features may be less informative to render a prediction in a protected group (e.g., identifying melanoma from an image of a patient with dark skin may be more difficult)

Training–serving skew: The model may be deployed on patients whose data are not similar to the data on which the model was trained. The training data may not be representative (i.e., selection bias), or the deployment data may differ from the training data (e.g., a lack of unified methods for data collection or not recording data with standardized schemas)

Biases in interactions with clinicians

Automation bias: If clinicians are unaware that a model is less accurate for a specific group, they may trust it too much and inappropriately act on inaccurate predictions

Feedback loops: If the clinician accepts the recommendation of a model even when it is incorrect to do so, the model's recommended versus administered treatments will always match. The next time the model is trained, it will learn to continue these mistakes

Dismissal bias: Conscious or unconscious desensitization to alerts that are systematically incorrect for a protected group (e.g., an early-warning score for patients with sepsis). Alert fatigue is a form of this

Allocation discrepancy: If the protected group has disproportionately fewer positive predictions, then resources allocated by the predictions (e.g., extra clinical attention or social services) are withheld from that group

Biases in interactions with patients

Privilege bias: Models may be unavailable in settings where protected groups receive care or require technology/sensors disproportionately available to the nonprotected class

Informed mistrust: Given historical exploitation and unethical practices, protected groups may believe that a model is biased against them. These patients may avoid seeking care from clinicians or systems that use the model or deliberately omit information. The protected group may be harmed by not receiving appropriate care

Agency bias: Protected groups may not have input into the development, use, and evaluation of models. They may not have the resources, education, or political influence to detect biases, protest, and force correction

Distributive justice options for machine learning

Equal patient outcomes: The model should lead to equal patient outcomes across groups

Equal performance: The model performs equally well across groups for such metrics as accuracy, sensitivity, specificity, and positive predictive value

Equal allocation: Allocation of resources as decided by the model is equal across groups, possibly after controlling for all relevant factors

References

1. Krause J, Gulshan V, Rahimy E, Karth P, Widner K, Corrado GS, et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. Ophthalmology 2018;125:1264–72. doi:10.1016/j.ophtha.2018.01.034 [PubMed: 29548646]
2. Angwin J, Larson J, Kirchner L, Mattu S. Machine bias. ProPublica 23 5 2016 Accessed at www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing on 13 December 2017.
3. Kleinberg J. Inherent trade-offs in algorithmic fairness [Abstract].. Abstracts of the 2018 Association for Computing Machinery International Conference on Measurement and Modeling of Computer Systems; Irvine, California. 18 –22 June 2018; New York: Association for Computing Machinery; 2018. 40
4. Lum K, Isaac W. To predict and serve? Significance 2016;13:14–9.

5. Chouldechova A, Benavides-Prado D, Fialko O, Vaithianathan R. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. Proc Mach Learn Res 2018: 134–48.

6. Hurley D Can an algorithm tell when kids are in danger? The New York Times Magazine 2 1 2018 Accessed at www.nytimes.com/2018/01/02/magazine/can-an-algorithm-tell-when-kids-are-in-danger.html on 2 January 2018.

7. American Medical Association. AMA passes first policy recommendations on augmented intelligence 2018 Accessed at www.ama-assn.org/ama-passes-first-policy-recommendations-augmented-intelligence on 6 July 2018.

8. Chin MH, King PT, Jones RG, Jones B, Ameratunga SN, Muramatsu N, et al. Lessons for achieving health equity comparing Aotearoa/New Zealand and the United States. Health Policy 2018; 122:837–53. doi:10.1016/j.healthpol.2018.05.001 [PubMed: 29961558]

9. Smedley BD, Stith AY, Nelson AR, eds; Institute of Medicine. Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care Washington, DC: National Academies Pr; 2003.

10. Rothstein R The Color of Law: A Forgotten History of How Our Government Segregated America New York: Liveright; 2017.

11. Healthy People 2020 About Healthy People. 2018 Accessed at www.healthypeople.gov/2020/About-Healthy-People on 9 October 2018.

12. Hinton G Deep learning—a technology with the potential to transform health care. JAMA 2018;320:1101–2. doi:10.1001/jama.2018.11100 [PubMed: 30178065]

13. Escobar GJ, Turk BJ, Ragins A, Ha J, Hoberman B, LeVine SM, et al. Piloting electronic medical record-based early detection of inpatient deterioration in community hospitals. J Hosp Med 2016;11 Suppl 1:S18–24. doi:10.1002/jhm.2652 [PubMed: 27805795]

14. Bates DW, Zimlichman E. Finding patients before they crash: the next major opportunity to improve patient safety [Editorial]. BMJ Qual Saf 2015;24:1–3. doi:10.1136/bmjqs-2014-003499

15. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff (Millwood) 2014;33:1123–31. doi:10.1377/hlthaff.2014.0041 [PubMed: 25006137]

16. Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. J Am Med Inform Assoc 2017;24:423–31. doi:10.1093/jamia/ocw105 [PubMed: 27516495]

17. Drew BJ, Harris P, Zègre-Hemsey JK, Mammone T, Schindler D, Salas-Boni R, et al. Insights into the problem of alarm fatigue with physiologic monitor devices: a comprehensive observational study of consecutive intensive care unit patients. PLoS One 2014;9: e110274. doi:10.1371/journal.pone.0110274 [PubMed: 25338067]

18. Epstein AM, Stern RS, Tognetti J, Begg CB, Hartley RM, Cumella E Jr, et al. The association of patients' socioeconomic characteristics with the length of hospital stay and hospital charges within diagnosis-related groups. N Engl J Med 1988;318:1579–85. [PubMed: 3131674]

19. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 2007;115:928–35. [PubMed: 17309939]

20. Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. Clin Chem 2008;54:17–23. [PubMed: 18024533]

21. Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. NPJ Digit Med 28 8 2018;1:40.

22. Muntner P, Colantonio LD, Cushman M, Goff DC Jr, Howard G, Howard VJ, et al. Validation of the atherosclerotic cardiovascular disease pooled cohort risk equations. JAMA 2014;311:1406–15. doi:10.1001/jama.2014.2630 [PubMed: 24682252]

23. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA 2018;319:1317–8. doi:10.1001/jama.2017.18391 [PubMed: 29532063]

24. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med 8 5 2018;1:18.

25. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. JAMA 2017;318:517–8. doi:10.1001/jama.2017.7797 [PubMed: 28727867]

26. Owens DK, Sox HC. Biomedical decision making: probabilistic clinical reasoning. In: Shortliffe EH, Cimino JJ, eds. Biomedical Informatics London: Springer-Verlag; 2014:67–107.

27. Manrai AK, Patel CJ, Ioannidis JPA. In the era of precision medicine and big data, who is normal? JAMA 2018;319:1981–2. doi:10.1001/jama.2018.2009 [PubMed: 29710130]

28. Ferryman K, Pitcan M. Fairness in precision medicine. Data & Society 2018 Accessed at https://datasociety.net/research/fairness-precision-medicine on 31 May 2018.

29. Howell EA, Brown H, Brumley J, Bryant AS, Caughey AB, Cornell AM, et al. Reduction of peripartum racial and ethnic disparities: a conceptual framework and maternal safety consensus bundle. Obstet Gynecol 2018;131:770–82. doi:10.1097/AOG.0000000000002475 [PubMed: 29683895]

30. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med 2018. doi:10.1001/jamainternmed.2018.3763

31. Veinot TC, Mitchell H, Ancker JS. Good intentions are not enough: how informatics interventions can worsen inequality. J Am Med Inform Assoc 2018;25:1080–8. doi:10.1093/jamia/ocy052 [PubMed: 29788380]

32. Insel TR. Digital phenotyping: technology for a new science of behavior. JAMA 2017;318:1215–6. doi:10.1001/jama.2017.11295 [PubMed: 28973224]

33. Cohen IG, Amarasingham R, Shah A, Xie B, Lo B. The legal and ethical concerns that arise from using complex predictive analytics in health care. Health Aff (Millwood) 2014;33:1139–47.doi: 10.1377/hlthaff.2014.0048 [PubMed: 25006139]

34. Chin MH, Clarke AR, Nocon RS, Casey AA, Goddu AP, Keesecker NM, et al. A roadmap and best practices for organizations to reduce racial and ethnic disparities in health care. J Gen Intern Med 2012; 27:992–1000. doi:10.1007/s11606-012-2082-9 [PubMed: 22798211]

35. National Academies of Sciences, Engineering, and Medicine, ed. Systems Practices for the Care of Socially At-Risk Populations Washington, DC: National Academies Pr; 2016.

36. Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning [Abstract]. In: Proceedings from the Conference on Advances in Neural Information Processing Systems 2016, Barcelona, Spain, 5–10 December 2017 La Jolla, CA: Neural Information Processing Systems; 2017:3315–23.

37. Chouldechova A Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big Data 2017;5:153–63. doi:10.1089/big.2016.0047 [PubMed: 28632438]

38. Woodworth B, Gunasekar S, Ohannessian MI, Srebro N. Learning non-discriminatory predictors. Proc Mach Learn Res 2017;65: 1920–53.

39. Beutel A, Chen J, Zhao Z, Chi EH. Data decisions and theoretical implications when adversarially learning fair representations Accessed at https://arxiv.org/pdf/1707.00075.pdf on 8 May 2018.

40. Zhang BH, Lemoine B, Mitchell M. Mitigating unwanted biases with adversarial learning. Artificial Intelligence, Ethics, and Society 2018 Accessed at http://arxiv.org/abs/1801.07593 on 8 May 2018

41. Platt JC. Probabilities for SV machines. In: Smola AJ, Bartlett PJ, Schuurmans D, Schölkopf B, eds. Advances in Large Margin Classifiers Cambridge, MA: MIT Pr; 1999:61–74.

42. Dhruva SS, Mazure CM, Ross JS, Redberg RF. Inclusion of demographic-specific information in studies supporting US Food & Drug Administration approval of high-risk medical devices. JAMA Intern Med 2017;177:1390–1. doi:10.1001/jamainternmed.2017.3148 [PubMed: 28738116]

43. Pleiss G, Raghavan M, Wu F, Kleinberg J, Weinberger KQ. On fairness and calibration. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan, et al., eds. Proceedings from the Conference on Advances in Neural Information Processing Systems 2017, Long Beach, California, 4–9 December 2017 La Jolla, CA: Neural Information Processing Systems; 2017:5680–9.

44. Kilbertus N, Rojas-Carulla M, Parascandolo G, Hardt M, Janzing D, Schölkopf B. Avoiding discrimination through causal reasoning. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., eds. Proceedings from the Conference on Advances in Neural Information Processing Systems 2017, Long Beach, CA, 4–9 December 2017 La Jolla, CA: Neural Information Processing Systems; 2017:656–66.

45. Schulman KA, Berlin JA, Harless W, Kerner JF, Sistrunk S, Gersh BJ, et al. The effect of race and sex on physicians' recommendations for cardiac catheterization. N Engl J Med 1999;340:618–26. [PubMed: 10029647]

46. Dieterich W, Mendoza C, Brennan T. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity Golden, CO: North-pointe; 7 2016.

47. Adamson AS, Smith A. Machine learning and health care disparities in dermatology. JAMA Dermatol 2018. doi:10.1001/jamadermatol.2018.2348

48. Gloster HM Jr, Neal K. Skin cancer in skin of color. J Am Acad Dermatol 2006;55:741–60. [PubMed: 17052479]

49. Goroll AH. Emerging from EHR purgatory—moving from process to outcomes. N Engl J Med 2017;376:2004–6. doi:10.1056/NEJMp1700601 [PubMed: 28538132]

50. Portela MC, Pronovost PJ, Woodcock T, Carter P, Dixon-Woods M. How to study improvement interventions: a brief overview of possible study types. BMJ Qual Saf 2015;24:325–36. doi: 10.1136/bmjqs-2014-003620

51. Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, et al. Predicting cardiovascular risk factors from retinal fundus photographs using deep learning. Nat Biomed Eng 2017;2: 158–64.

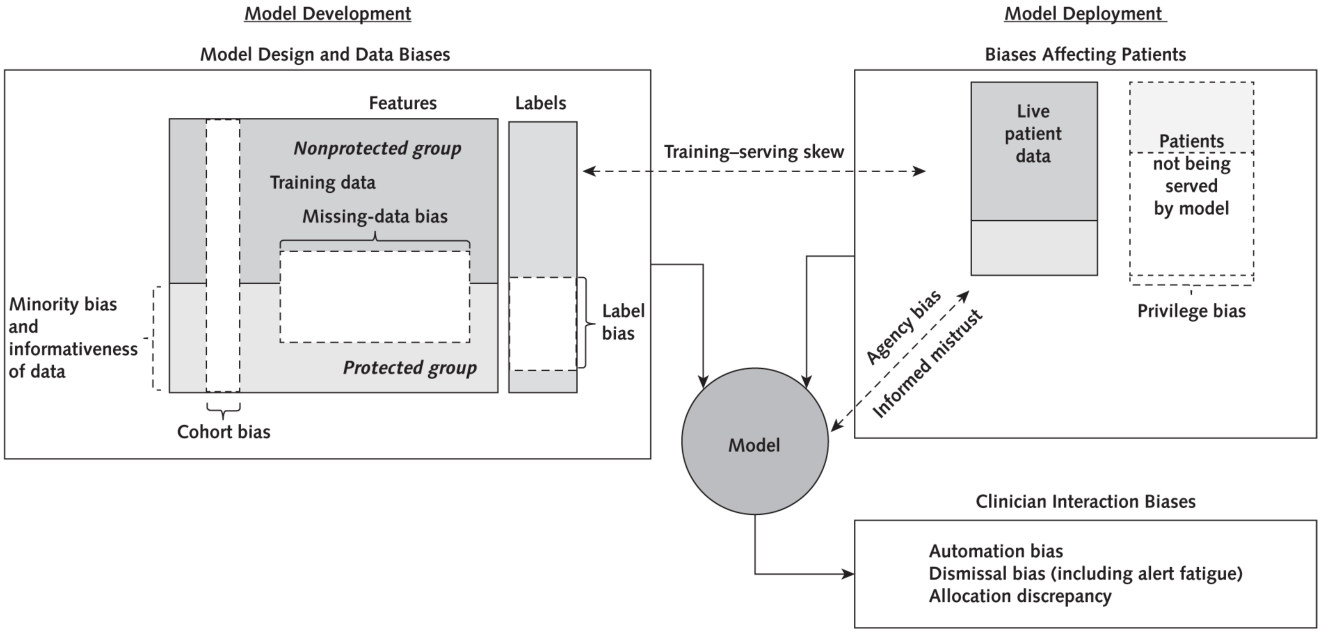

Figure. Conceptual framework of how various biases relate to one another.

During model development, differences in the distribution of features used to predict a label between the protected and nonprotected groups may bias a model to be less accurate for protected groups. Moreover, the data used to develop a model may not generalize to the data used during model deployment (training–serving skew). Biases in model design and data affect patient outcomes through the model's interaction with clinicians and patients.

Table.

Recommendations

Design
Determine the goal of a machine-learning model and review it with diverse stakeholders, including protected groups.
Ensure that the model is related to the desired patient outcome and can be integrated into clinical workflows.
Discuss ethical concerns of how the model could be used.
Decide what groups to classify as protected.
Study whether the historical data are affected by health care disparities that could lead to label bias. If so, investigate alternative labels.
Data collection
Collect and document training data to build a machine-learning model.
Ensure that patients in the protected group can be identified (weighing cohort bias against privacy concerns).
Assess whether the protected group is represented adequately in terms of numbers and features.
Training
Train a model taking into account the fairness goals.
Evaluation
Measure important metrics and allocation across groups.
Compare deployment data with training data to ensure comparability.
Assess the usefulness of predictions to clinicians initially without affecting patients.
Launch review
Evaluate whether a model should be launched with all stakeholders, including representatives from the protected group.
Monitored deployment
Systematically monitor data and important metrics throughout deployment.
Gradually launch and continuously evaluate metrics with automated alerts.
Consider a formal clinical trial design to assess patient outcomes.
Periodically collect feedback from clinicians and patients.