Machine Learning in Health Care: Ethical Considerations Tied to Privacy, Interpretability, and Bias

Thomas Hofweber, Rebecca L. Walker

Machine learning models hold great promise with medical applications, but also give rise to a series of ethical challenges. In this survey we focus on training data, model interpretability, and bias and the related issues tied to privacy, autonomy, and health equity.

Introduction

Machine learning is the technique behind much contemporary work in artificial intelligence. In machine learning, models trained on a particular dataset are then deployed, for example, to make a prediction or to identify a particular feature in a new data sample. In health care, a few examples of how machine learning can be used include in diagnosis of a specific disease, risk prediction for various health conditions, determining likely effectiveness of different medical interventions, and detecting cancer in a radiological image [1, 2]. Machine learning thus holds great promise for progress in medicine, but with this promise come notable ethical challenges.

For machine learning to generate useful information in a health care setting, models must be trained on large amounts of medical data. Access to medical data, certain features of machine learning, and existing limitations of available data each create ethical challenges for its use in medicine, which we will outline here in terms of their implications for privacy, autonomy, and health equity.

Privacy and Big Data

Machine learning is very powerful when models are trained on enough data. Well-trained models have the potential to surpass predictions and identifications made by human experts, promising significant advances in patient outcomes when used in a health care setting. Where data are publicly available, for example with information about stock market performance, anyone can attempt to build a model that can then be deployed in making future predictions, allowing for innovation and providing training data for the development of future highly useful models. In contrast, the use of medical data is significantly constrained within certain legal and oversight parameters, such as the Health Insurance Portability and Accountability Act of 1996 (HIPAA). This creates barriers to innovation and limits training data. With the help of electronic medical records, hospitals, insurance companies, and government providers have sufficient data to train valuable models to use internally. But the wider release of such models is problematic, since sometimes training data can be recovered from the model using adversarial attacks, thus risking privacy violations [3]. So-called de-identified medical information is a growing commodity for those who wish to train machine learning models, and certain types of de-identified data are publicly available through government resources such as the Department of Veterans Affairs (DVA), and the Centers for Medicare and Medicaid Services (CMS) [4–6]. However, these resources face problems including concerns that data cannot be fully de-identified and worries about sources of bias and error that are nontransparent and embedded in the data [4, 7].

Medical data is especially protected, for good reasons. The promise of confidentiality within the health care setting allows patients to fully disclose sensitive information to their provider, shoring up trust and improving patient outcomes [8]. At the same time, the lack of medical datasets for the training of machine learning models by researchers or innovative startup companies also comes at a real cost to patients in terms of less powerful medical predictions and discernments. Confidentiality is a core principle of medical ethics, appearing from the time of Hippocrates [9], but faces challenges in modern medical practice given the necessity of many different entities accessing and sharing patient data [10]. The conundrum raised by machine learning—of how to build models that benefit patients without endangering their privacy—raises these questions in a new and more pressing way, inviting a potential reevaluation of our current standards for data sharing.

Electronically published July 10, 2024.

Address correspondence to Thomas Hofweber, Department of Philosophy, Caldwell Hall, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-3125 (hofweber@unc.edu).

N C Med J. 2024;85(4):240-245. 0029-2559/2024/85406

Pencina **SIDEBAR**

Autonomy and Interpretability

One well-known feature of machine learning models is that their accurate predictions do not come with reasons for these predictions. For example, a model can accurately predict whether a patient will get heart disease, but not reveal why or what features of the patient are responsible for this outcome. These models are often described as "black boxes" that take patient data as input and spit out probabilities for specified outcomes. This description notwithstanding, such models can be fully transparent to those who have access to them while still not yielding insights into why particular predictions are made, often because of the model's complexity. Making the "black box" transparent does little to alleviate the "why" problem.

This problem is referred to as the lack of interpretability

of a model and is especially significant when it comes to basing medical decisions on model predictions. A model can be highly accurate without being interpretable. This can occur for predictions about disease as well as other predictions and identifications, such as what is likely the most successful course of treatment, or whether a scan indicates a cancer. Thus, a physician may be able to accurately indicate which course of treatment is best for a particular patient but be unable to give that patient any explanation of why that is the case. Or a patient may face a difficult prognosis without understanding what they can do to turn the tide. The lack of interpretability of a model is thus in tension with patient autonomy in two different ways. First, insofar as autonomy is a capacity for self-determination, a lack of interpretability in predictive medicine may leave patients adrift without the tools they need to deliberatively structure their own future.

Second, insofar as autonomous decision-making in medicine requires understanding [11], a lack of interpretability leaves patients and their providers selecting between treatment options without understanding why they might work.[a]

What are the possible solutions to the tensions identified between autonomy and the lack of interpretability of machine learning models? It may help to first clarify that the identified problems for autonomy aren't prohibitive of autonomous decision-making. Just as a patient can autonomously decide to put all their trust in their doctor when it comes to determining which course of treatment is best for them [11], they can similarly autonomously choose to prioritize the likelihood of accuracy of a predictive model regardless of interpretability. Yet others will demand a reason for why they should consent to a particular treatment beyond the mere prediction of its success. Exercise of their autonomy may require that they can assess what is best for themselves by considering proposed alternatives and their personal values. Indeed, the basic tenets of informed consent typically include not merely a recommendation of a particular course of treatment, but also an explanation as to its risks and benefits, information about relevant alternatives (including the risks and benefits of each), and the patient's adequate understanding of each of these factors [13]. Not being able to provide an explanation of why a particular treatment has been recommended, other than its predicted success, seems to undercut the sense in which adequate patient understanding is even relevant to informed consent.

[a] Here we are focused on problems for autonomy stemming from a lack of interpretability of some models, however it is important to point out that AI in health care generally can also enhance patient autonomy. For example, through the development of virtual assistants which can make patients more able to manage their conditions given overwhelming amounts of medical information [12].

There are several possible solutions to this problem. One is to avoid the use of non-interpretable models altogether, and only use interpretable models instead [14]. But restricting oneself outright to only interpretable models potentially loses some of the benefits and power of machine learning models, since interpretable models are generally simpler and only form a subclass of all models. Another option is to use a non-interpretable model in decision-making, and then conduct a post-hoc "explanation" for why the model made a particular prediction. This can be attempted by manipulating patient data to see how this impacts the model's output and thereby trying to gain insight into what factors are relevant to the model's prediction or identification [15].

Whether these approaches provide a reason for a recommendation or prognosis is a matter of ongoing debate, however they may be sufficiently helpful to support patient autonomy when it comes to pragmatic questions of self-determination. For example, a patient who is told not only of a significant likelihood of heart disease, but also that tweaking their cholesterol levels significantly changes that prediction, may be less adrift in terms of their capacity to make relevant choices in keeping with their values.

Health Equity and Bias

Another crucial ethical concern for machine learning in health care is bias that may perpetuate social injustice through the recommendations of these models. Bias in machine learning is a complex topic, with some aspects of the problem having more obvious solutions than others. Among the in-principle solvable problems is that of inadequate representation of specific socially identifiable groups in model training data. For example, a model trained on heart attack data from male patients only that is then deployed to predict heart attacks for patients of all sexes might perform unacceptably badly for non-male patients, since there are relevant differences between sexes in what predicts a heart attack. This problem is solvable by including sufficiently diverse training data, and then evaluating the accuracy of the model on a variety of subpopulations. While easily solvable in principle, however, this problem is deeper than it appears from a practical point of view. It is likely that members of underserved groups will not be properly represented in the training data and including them will require targeted efforts that may be underfunded or otherwise difficult to achieve [16]. Such efforts are crucial to combatting the potential for machine learning models to further entrench health disparities, since underserved populations otherwise face the risk of being further disadvantaged by having "precision" medical information generated by a model that is not accurate for them [17]. Avoiding such bias is a crucial responsibility of those building models that generate important medical information. However, health care providers using machine learning to make medical suggestions to patients also must be aware of these limitations if they are to fulfill their obligations responsibly.

Other problems with bias in machine learning are deeper than the mere lack of adequate inclusion in the relevant training data and likewise present serious health equity obstacles to the deployment of these models in health care. Machine learning models only make predictions based on what has happened in the past, as reflected in the training data. As addressed in the prior section, the model does not consider why something has or hasn't happened. However, understanding the reasons why, for example, an intervention has or hasn't worked in the past can be a crucial component of addressing problems in health equity. To illustrate, consider individuals who are living with poverty. A model might accurately predict that a particular treatment is not successful for patients in a particular income range, and thus recommend against deploying it for a particular patient living with poverty. However, the reason the treatment wasn't successful in the past might simply be that poor patients with this disease were less able to stick to the treatment because their economic situation did not allow them to follow the prescribed drug regimen [18]. The model will indicate a treatment isn't effective for the individual, but in fact it would be effective if adequate provisions could be made to provide it for the patient. The recommendation should be for ensuring adequate medical support, not for avoiding this line of care. Relying on the machine learning models in such a case would deny effective treatment for members of an already disadvantaged group.

Relying on machine learning models in decision-making perpetuates the status quo, since they are trained on data of the past and use this past data to project results into the future. The model does not evaluate why things were as they were in the past, or whether particular aspects of the past are positive or negative. Thus, the deployment of machine learning models in making decisions for the future will reflect, for better or for worse, how the values of the past impacted individual lives and outcomes. This problem has no obvious solution, but it is a crucial topic to address in moving machine learning in health care forward.

Conclusion

Machine learning models hold great promise for advances in medicine due to their power of finding patterns in vast amounts of data that are not possible for human beings to analyze unaided. At the same time, they also hold dangers that need to be properly addressed to avoid damage to patients going forward, including the potential for violating privacy, undermining autonomy, and further deepening health disparities. NCMJ

Thomas Hofweber, PhD Professor and Director of the AI Project, Department of Philosophy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.
Rebecca L. Walker, PhD Professor of Philosophy and of Social Medicine; Core Faculty, Center for Bioethics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.

Acknowledgments

The authors report no relevant conflicts of interest.

References

1. Gerke S, Minssen T, Cohen G. Ethical and legal challenges of artificial intelligence-driven healthcare. In: Bohr A, Memarzadeh K, Eds. *Artificial Intelligence in Healthcare.* pp. 295-336. Academic Press; 2020.
2. Rajpurkar P, Chen E, Banerjee O, Topol EJ. AI in health and medicine. *Nature Medicine.* 2022;28(1):31-38. https://doi.org/10.1038/s41591-021-01614-0
3. Fredrikson, Matt, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In: Ray I, Li N, Eds. *Proc. 22nd ACM SIGSAC Conference on Computer and Communications Security.* Association for Computing Machinery; 2015.
4. Alberto IRI, Alberto NRI, Ghosh AK, et al. The impact of commercial health datasets on medical research and health-care algorithms. *Lancet Digit Health.* 2023;5(5):e288–e294. doi: 10.1016/S2589-7500(23)00025-0
5. US Department of Veterans affairs. Open Data. Accessed May 10, 2024. https://www.data.va.gov/stories/s/About-Data-va-gov/mdjj-7nhd/
6. US Centers for Medicare and Medicaid Services. Datasets. Accessed May 10, 2024. https://data.cms.gov/search
7. Mandl KD, Perakslis ED. HIPAA and the leak of "deidentified" EHR data. *N Engl J Med.* 2021;384(23):2171-2173. doi: 10.1056/NEJMc2111490
8. Kipnis K. Medical confidentiality. In: Francis LP, Silvers A, Rhodes R, Eds. *The Blackwell Guide to Medical Ethics.* Blackwell; 2007.
9. Estroff SE, Walker RL. Confidentiality: concealing "things shameful to be spoken about". *Virtual Mentor.* 2012;14(9):733-737. doi: 10.1001/virtualmentor.2012.14.9.msoc1-1209
10. Siegler M. Sounding Boards. Confidentiality in medicine- a decrepit concept. *N Engl J Med.* 1982;307(24):1518-1521. doi: 10.1056/NEJM198212093072411
11. Beauchamp TL, Childress JF. Principles of Biomedical Ethics, 8th Edition. Oxford University Press; 2019.
12. Tirth D, Anirudh Athaluri S, Satyam S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell.* 2023;6:1169595. doi: 10.3389/frai.2023.1169595
13. Faden RR, Beauchamp TL. *A History and Theory of Informed Consent.* Oxford University Press; 1986
14. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell.* 2019;1(5):206-215. doi: 10.1038/s42256-019-0048-x
15. Ribeiro MT, Singh S, Guestrin C. Why should I trust you?: Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM; 2016.
16. Geneviève LD, Martani A, Shaw D, Simone Elger B, Wangmo T. Structural racism in precision medicine: leaving no one behind. *BMC Medical Ethics.* 2020;21:17. https://doi.org/10.1186/s12910-020-0457-8
17. Price WN, Cohen IG. Privacy in the age of medical big data. *Nat Med.* 2019;25(1):37-43. doi: 10.1038/s41591-018-0272-7
18. Hensley C, Heaton PC, Kahn RS, Luder HR, Frede SM, Beck AF. Poverty, transportation access, and medication nonadherence. *Pediatrics.* 2018;141(4):e20173402. doi: 10.1542/peds.2017-3402