Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Am J Bioeth. Author manuscript; available in PMC 2021 December 06.

Published in final edited form as:

Am J Bioeth. 2020 November ; 20(11): 18–20. doi:10.1080/15265161.2020.1820115.

It is Time for Bioethicists to Enter the Arena of Machine Learning Ethics

Michaela Hardt[a], **Marshall H. Chin**[b]

[a]Independent Scholar

[b]University of Chicago

Increasingly, data scientists are training machine-learning (ML) models for diagnosis, treatment selection, and resource allocation. The U.S. Food and Drug Administration has given regulatory approval for some models to analyze data from X-ray, CT, MRI, ECG, and other sources. Therefore, ethical guidance is urgently needed along with standards and regulation, to ensure appropriate use of ML models in healthcare.

We believe that ML models have great potential to improve health outcomes, and technical models must be integrated with human judgment and ethical principles to lead to the best results. Bioethicists can make important contributions to teams developing and evaluating novel medical technologies including ML. However, with a deeper understanding of how ML models are created and used in an application, bioethicists can be a more effective voice at the table. In our earlier work we described how to integrate health equity considerations into the model development and deployment process (Rajkomar et al. 2018). Char et al. raise additional ethical considerations including audit-ability, informed consent of study participants, and accountability (Char et al. 2020). We briefly review the basics of ML and then highlight key opportunities for bioethicists to contribute to the development and deployment of ML models in healthcare applications.

A *model* is a mapping from input *features* of an example to an output prediction. During training the mapping of a model is learned by fitting its parameters so that its predictions are close to the correct output *labels*. We *evaluate* the model based on how close its predictions are to the labels of new examples. Various metrics capturing the performance exist, e.g. for binary prediction tasks in which the output takes only two values (true and false) we can compute the overall *accuracy* (the fraction of examples for which the model makes the correct prediction) or break it down to *sensitivity* (for what fraction of true examples the model correctly predicts true), and *specificity* (for what fraction of false examples the model correctly predicts false). We *deploy* the model and obtain predictions for examples for which we do not have labels.

For example, in a previous case study, the University of Chicago built a model to predict how likely a patient is to be discharged from the hospital within the next few days

CONTACT Michaela Hardt hardtmichaela@gmail.com Berkeley, CA 94707, USA.

DISCLOSURE STATEMENT

Michaela Hardt is employed by Amazon, but her views expressed in this commentary do not represent the views of her employer.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

(Rajkomar et al. 2018). They used historical data of patients with known discharge dates to train a model to make this prediction based on patient information from the electronic health record including demographic information. The model parameters were fitted so that the model's predictions about how likely a patient is to be discharged soon were close to 1 for patients who were actually discharged soon (i.e. positive example) and low probability for patients with longer stays (i.e. negative example). The data scientists assessed model performance using the area-under-the-receiver-operating-curve (au-roc) which measures the probability that the model will correctly assign a higher score to a random patient that is being discharged soon compared to a random patient whose stay extends (Fawcett 2006). They worked to improve this performance metric by carefully choosing and representing the input features.

Bioethicists can play an important role helping to steer the course of ML in healthcare. It starts with how to use ML in healthcare. To prioritize applications, it would be extremely helpful to obtain the ethicist's perspective on questions such as: Given limited resources, what models should we build? Who is going to benefit from the application and how? These are important questions for those building models (e.g. academic researchers and industry data scientists), making deployment decisions (e.g. hospital systems) or funding research, and each of these stakeholders has multiple competing interests. Ethicists can work with data scientists who weigh in on the technical feasibility, clinicians who weigh in on the medical benefit and risk, and patients who weigh in with their preferences and values. While in practice any given stakeholder may consider and balance different factors, bioethicists are well-trained to ensure that the concerns of the broader population and society and different subgroups such as vulnerable and disempowered patients and communities are appropriately considered and addressed.

Next, comes data collection, model training, and evaluation. Bioethicists can assess these steps by working backwards from the goal of the application and its intended impact on the patients and the healthcare system. Currently, some data scientists focusing on model development may not fully understand the meaning of the data and their intended clinical impact. Similarly, healthcare providers may have limited technical understanding of the model development. Hence a multi-disciplinary effort is needed.

We can consider equal health outcomes as an option for distributive justice (Rajkomar et al. 2018). Working backwards from equal health outcomes, we seek to identify performance metrics that can help achieve those outcomes. If these performance metrics point to stark differences across groups, we may want to revise the model at this early stage to strive toward a more equal benefit across groups. In fact, we can use algorithmic fairness tools to create models with comparable performance (e.g. sensitivity and/or specificity) across groups (Agarwal et al. 2018; Hardt et al. 2016; Platt 1999; Woodworth et al. 2017; Zhang et al. 2018). The choice of the metric and the set of groups, however, still require human judgment.

While Char et al. posit that an equitable application provides "equivalent levels of accuracy … across groups," they urge us to not limit ethical considerations to algorithmic fairness (Char et al. 2020). We agree and in our article we outlined various scenarios in which

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

we may observe unequal outcomes despite achieving equal model performance through these tools (Rajkomar et al. 2018). We therefore correct Char et al.'s mischaracterization of our article to "pursue an ideal of algorithmic fairness." Algorithmic fairness tools cannot correct many biases including a biased access to an application or the informed mistrust of patients from a group that has been historically exploited (e.g. through experimentation or sterilization without consent). We firmly believe that identifying and correcting biases should be considered to advance distributive justice. Char et al caution that we "risk introducing a complex set of unintended biases in attempts to correct the initial bias." This is why we prioritize equal outcomes through clinical trials in which we can also test the effectiveness of the algorithmic fairness tools and their impact on outcomes. It is critical to monitor outcomes during the deployment of a model. We also believe that human judgment is needed to determine which biases, performance metrics, and outcomes are of importance for a given application.

For example, the stated goal of the University of Chicago application was to reduce unnecessary length-of-stay. The hope was that by using a model to identify patients who would be discharged soon, case management resources could be assigned to them to avoid unnecessary delays. During a review of the model with collaborators from the University of Chicago Medicine's Diversity and Equity Committee, team members pointed out the important different goal of equitable distribution of the case management resources. By raising this concern, the team prevented a model deployment which would have diverted case management resources away from African Americans with greater medical and social needs who had longer stays.

Notably, this example is not about whether the au-roc metric indicated similar performance across groups. The more important question was identified by understanding the impact of the model's predictions on patients. The consideration of an equitable distribution of resources was not captured in the au-roc metric nor the original priority of reducing length-of-stays. An additional concern not captured by the original model was patient preferences: Do some patients feel rushed out of the hospital too early through the algorithm and mobilization of case management services?

Bioethicists can help navigate tradeoffs in the development and deployment of ML applications. Should we use algorithmic fairness tools that equalize model performance across groups in cases when they reduce the performance for at least one group? Should we launch an application when we demonstrate that each group benefits even if overall the application widens disparities by benefitting one group more than another? Char et al. recommend requiring models to have a transparent "explainable architecture." But what if more complicated "black box" models achieve both higher accuracy overall and a more equitable allocation across groups (Kleinberg and Mullainathan 2019)?

We call for bioethicists to collaborate with companies and data scientists to develop models and with healthcare systems to deploy and monitor them. It will be important to share findings through case studies, reports, and articles. These studies can be the foundation for refined practical guidelines and standards paving the way for regulations and legislation to make sure that all patients benefit from ML applications in healthcare.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

ACKNOWLEDGMENTS

The authors thank John Fahrenbach and Alvin Rajkomar for their review of portions of an earlier version of this manuscript.

FUNDING

Dr. Chin was supported in part by the Robert Wood Johnson Foundation Advancing Health Equity: Leading Care, Payment, and Systems Transformation Program Office, the Merck Foundation Bridging the Gap: Reducing Disparities in Diabetes Care National Program Office, and the Chicago Center for Diabetes Translation Research [NIDDK P30 DK092949].

REFERENCES

Agarwal A, Beygelzimer A, Dudik M, Langford J, and Wallach H. 2018. A reductions approach to fair classification. Proceedings of Machine Learning Research 80:60–69.

Char DS, Abràmoff MD, and Feudtner C. 2020. Identifying ethical considerations for machine learning healthcare applications. The American Journal of Bioethics 20 (11):7–17. doi: 10.1080/15265161.2020.1819469.

Fawcett T 2006. An introduction to ROC analysis. Pattern Recognition Letters 27 (8):861–874.

Hardt M, Price E, and Srebro N. 2016. Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems 29:3315–3323.

Kleinberg J, and Mullainathan S. 2019. Simplicity creates inequity: Implications for fairness, stereotypes, and inter-pretability. Proceedings of the 2019 ACM Conference on Economics and Computation, 807–808.

Platt JC 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Advances in large margin classifiers, ed. Smola AJ, Bartlett Peter, Schölkopf B and Schuurmans D, 61–74. Cambridge, MA: MIT Press.

Rajkomar A, Hardt M, Howell MD, Corrado G, and Chin MH. 2018. Ensuring fairness in machine learning to advance health equity. Annals of Internal Medicine 169 (12):866–872. [PubMed: 30508424]

Woodworth B, Gunasekar S, Ohannessian MI, and Srebro N. 2017. Learning non-discriminatory predictors. Proceedings of Machine Learning Research 65:1920–1953.

Zhang BH, Lemoine B, and Mitchell M. 2018. Mitigating unwanted biases with adversarial learning. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 335–340. New Orleans, LA, USA.