Assessing Bias in Medical AI

Melanie Ganz [1 2] **Sune H. Holm** [3] **Aasa Feragen** [4 2]

Abstract

Machine learning and artificial intelligence are increasingly deployed in critical societal functions such as finance, media and healthcare. Along with their deployment come increasing reports of their failure when viewed through the lens of ethical principles such as fairness, democracy and equal opportunity. As a result, research into fair algorithms and mitigation of bias in data and algorithms, has surged in recent years. However, while it might seem clear what fairness entails, and how to achieve it, in some applications, established concepts do not translate directly to other domains. In this work, we consider healthcare specifically, illustrating limitations and challenges of fair models within medical applications and give recommendations for the development of AI in healthcare.

1. Introduction

Including measures of fairness and bias in one's algorithm assessment is a rising trend in the field of machine learning and artificial intelligence. While the need for a closer look at the process of developing specifically medical AI is widely acknowledged (7), the discussions take place on a stakeholder level and are often removed from the experience and working level of the machine learning developer. To mitigate this and to highlight the issues on a different level, we investigate the origin of human- and machine bias in healthcare via two cases and analyze the degree to which we might realistically be able to remove this bias. Based on this analysis, we discuss ethical consequences of these limitations, leading to a set of recommendations for minimizing bias and promoting fairness in medical AI.

2. Case discussions

In the two cases discussed below, bias has different sources, and we first discuss the technical implications of different types of bias and the corresponding limitations in the potential for fair medical AI. Following the technical discussion, we revisit the issue on a higher level from an ethical and societal point of view.

Case A: Gender bias in diagnostic AI Training algorithms on different subgroups can reveal classification imbalances. A recent paper (4) studied the effect of training set imbalance in image-based computer aided diagnosis. The authors studied diagnoses of 12 different thoracic diseases based on chest X-ray using a state-of-the-art classifier, and using training sets with a gender balance of 0/100%, 25/75%, 50/50%, 75/25% and 100/0% women/men, respectively. As expected, diagnostic AI performed better on women when it was specialized to diagnose women, and vice versa. However, for some diseases – pneumothorax being an example – the diagnostic AI specialized to diagnose women was actually better at diagnosing men, than at diagnosing women. Replacing some of the training set females with males emphasized this difference, but the fact remained: The best-performing algorithm for women was better at diagnosing men than women. And at the same time this was the worst-performing algorithm for men.

From the machine learner's point of view, this is a frustrating result: In many applications, it is fully feasible to ensure balance between sensitive groups in a training set, but here balancing the training set was insufficient to obtain equal performance – the classification problem appears to be more challenging for one group than for another. In the case of chest imaging, this has a plausible biological explanation: In x-ray imaging of the upper thorax women's breasts occlude the imaged organs, resulting in poorer image contrast for the relevant anatomy.

Case B: Gender differences in diagnosis of depression A correct diagnosis can only be made if the patient actually seeks treatment, and existing societal biases in access to and trust in healthcare can lead to biases in medical AI

[*]Equal contribution [1]Neurobiology Research Unit, Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark [2]Department of Computer Science, University of Copenhagen, Copenhagen, Denmark [3]Department of Food and Resource Economics, University of Copenhagen, Copenhagen, Denmark [4]Compute, Denmarks Technical University, Kgs. Lyngby, Denmark. Correspondence to: Melanie Ganz <ganz@di.ku.dk>.

Workshop on Interpretable ML in Healthcare at International Conference on Machine Learning (ICML). Copyright 2021 by the author(s).

trained on previous diagnoses. One study (6) reports that men's answers to diagnostic questionnaires for depression differ depending on whether or not they expect having to endure significant treatment following a potential diagnosis. A similar pattern was not found among women. This indicates that men with depression might be less prone to seek treatment than women, causing false negative labels in a potential training set. Another study (5) shows that the prevalence of depression among men changes when the diagnostic criteria also include aggressive symptoms. This indicates that the features used for diagnosing depression might be more informative for women than for men. Both of these problems would likely contribute to under-diagnosis of men in a diagnostic AI trained on existing diagnoses and seem to have their origin in socio-cultural biases.

3. Extrapolating from cases to the underlying machine learning challenges

The standard workflow of a machine learning expert who gets involved in a medical AI problem is described in Fig. 1. Usually the problem is defined by clinical practitioners and the data to build the medical AI algorithm is also collected by clinical stakeholders. The machine learner's playground then consists of the data, e.g. features and labels, on whose basis a predictive algorithm is developed. The algorithm as well as its predictions are then reported back to the clinical stakeholders and often the clinical interpretation as well as use of the knowledge derived from the predictor, such as relevant features, are also again interpreted by clinicians. Keeping our workflow shown in Fig. 1 as well as the two case examples in mind, we observe different potential issues in the use of medical AI.

Imbalanced training data Imbalanced training sets are an obvious source of bias in predictive models. While this can be alleviated by providing balanced training data sets, this requires access to sufficiently large and diverse numbers of samples from every group, which can be unfeasible in medicine. In cases where we lack data from given groups, transfer learning or data augmentation may help even out the bias, although these methods assume a similar variation across groups.

Different feature distributions However, as illustrated above, there are also causes of bias that are much harder to alleviate. In the image-based diagnosis example, we see that the diagnostic features (the image) are simply more informative for one group (men) than for the other (women). The same holds in the case of depression if the diagnostic criteria should be broader than presently in order to be effective for the under-represented group (men). In machine learning terms, this translates to having different feature distributions for the two groups, where the diagnostic task is simply easier for one group than for the other given the same features.

Different levels of label noise Another challenging cause of bias can again be seen in the depression example, where the self-reporting of disease is likely to be lower for group II (men) than for group I (women). In machine learning terms, this means that group II will have higher label noise than group I.

Appropriate proxy label to detect bias Adding to these problems comes the challenging task of checking whether one or more of these problems is present. Having different predictive distributions for two groups is not a problem per se – for instance, one should be making far more breast cancer diagnoses among women than among men. Bias consists of systematic predictive errors that are made at different rates for one group than for another. But in order to detect predictive error, we need access to ground truth labels, which is impossible in cases where the diagnostic criteria or features themselves are biased. In order to detect such errors, we rely entirely on finding a good proxy label for diagnosis – but this requires rethinking the whole basis for medical diagnostics. Is the goal to produce an optimal diagnostic accuracy? To keep the patient alive as long as possible? Or to ensure the best possible perceived quality of life?

4. A broader view: Ethical implications

The above technical discussion assumes that fairness requires a form of classification parity across groups with respect to a given performance measure (2) such as true positive rate (3) or positive predictive value/precision (1). Classification parity definitions suggest that inequality in some aspect of diagnostic performance across groups is always unfair. The cases presented above give us reason to consider more carefully, why inequalities are unfair.

What does it take for inequality to be unfair? How should we respond to the fact that in some contexts the best-performing algorithm for group I (women in Case A) works better for group II (men in Case A), while at the same time being the worst-performing algorithm for group II? Should we level down and make the performance for group II worse to ensure equality across groups? We may find it relevant that in this scenario, the inequality in performance across the two groups does not appear to be a result of discriminatory practices, sample bias, or measurement bias. Rather the performance inequality is due to biological differences between the groups. How should this fact influence our ethical assessment of the ensuing inequality? One might argue that if the intentions are good, and measures have been taken to ensure equality, then the resulting model is

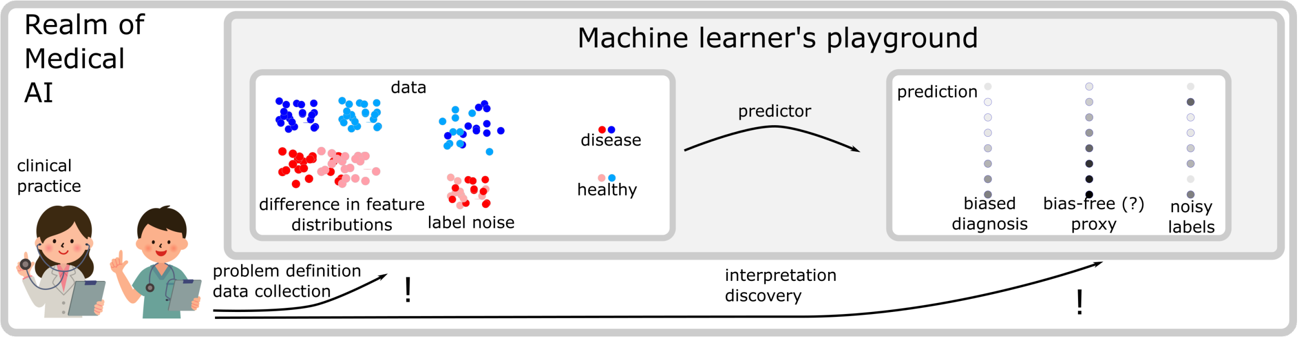

Figure 1. The standard workflow of a machine learning expert who gets involved in a medical AI problem.

fair.

Conversely, one might also argue that the discrepancy is (still) unfair because it means that group II gets better treatment, more resources, and a longer and healthier life than group I. According to this line of argument, the inequality might require that society redirects resources towards developing better diagnostic procedures, e.g. novel imaging techniques, for the under-performing group.

In the context of Case B, we would instead expect that when a trained algorithm performs better for group I (women in Case B) than for group II (men in Case B), this may be a result of the way in which the true label has been established for the cases in the training data. Thus in the case of depression the diagnostic criteria may be biased in two directions: group I may be over-diagnosed and group II may be under-diagnosed. This may not be due to natural differences, but to prejudices among doctors and patients, as well as the way in which diagnostic tests are formulated. Again, classification parity definitions of fairness (2) will seem to require equality even when this results in levelling down the performance for one group without improving the performance for any group. This seems problematic in the medical context.

These two cases leave us with a general question: Is there a difference, ethically, between biases that have their origin in biology as opposed to socio-cultural conditions? One could argue that since we can actually change the socio-cultural dimensions, their nature is fundamentally different from that of anatomy. On the other hand, it is unlikely within the powers of the data scientist to single-handedly change the socio-cultural dimensions affecting a training set.

One might argue that it is just bad luck for women that they are harder to diagnose with pneumothorax based on x-rays using an algorithm than men (Case A). The inequality does not amount to a fairness problem. Still, there are many cases in which we think that we have an obligation to compensate or try to equalize natural inequalities. Following this line of argument, we should devote extra resources to developing diagnostic tools for female patients that perform on par with those for men, e.g. via improved noise modelling or alternative imaging modalities. Still, equality may be hard to achieve if such research leads to methods that improve performance for women, but improve it equally or even more for men.

In the context of the likelihood of under-diagnosis of men with depression, the need for a change in diagnostic criteria and socio-cultural stereotypes cannot be affected by the machine learning developer. Therefore, in the next section we outline our recommendations not only for the developers, but also for the users of medical AI as well as give societal recommendations.

5. Recommendations

Our discussion above highlights that there are no one-size-fits-all solutions in deriving fair algorithms for medical AI. In general, care needs to be taken when assessing the performance of medical AI. It is important that not only the machine learning practitioners and data analysts are aware of the caveats of the algorithm, but that they also communicate the caveats clearly to the clinicians and other clinical stakeholders. In the end, it is the clinicians, the end users of medical AI, who are left with the responsibility of justifying the decisions taken with the help of the algorithms. Moreover, most of the ethical considerations and questions are of a societal nature, addressing distribution of resources and societal implications of bias. Decisions on this level should not be made by AI developers alone and we advocate for a broader discussion such as presented in (7). Nevertheless, based on our case examples we make the following recommendations on the level of developers, users, and society, respectively.

Recommendations for developers of healthcare AI

- Report classifier performance not only for the data set as a whole, but also for relevant vulnerable subgroups such as the population stratified by e.g. gender, age and ethnicity

- Report classifier performance when trained on one and applied on another subgroup in order to realistically assess stability in a clinical setting

- Report, to the greatest extent possible, which features were used by the algorithm and how (feature stability, etc.)

- Define and describe what a false positive and false negative means in the given clinical context

- In medical AI, the ground truth output value, such as diagnosis, can be impossible to obtain 100% correct, and should be considered a proxy for the wanted variable. This proxy should be chosen with care and its influence assessed.

Recommendations for users of healthcare AI

- Consider whether fairness as equality entails that an algorithm is morally wrong all things considered, including its utility.

- Consider whether the bias-causing features are biological or socio-cultural in origin and whether it makes a difference to the unfairness of the ensuing bias.

Societal recommendations

- Rethink healthcare, and medical AI in particular, in the context of resources. Are resources distributed fairly, and to those who benefit most from them?

6. Acknowledgements

The authors thankfully acknowledge support by the Independent Research Foundation Denmark (grant number 9131-00097B, Bias and Fairness in medicine).

References

[1] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

[2] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.

[3] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[4] Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23):12592–12594, 2020.

[5] Lisa A Martin, Harold W Neighbors, and Derek M Griffith. The experience of symptoms of depression in men vs women: analysis of the national comorbidity survey replication. *JAMA psychiatry*, 70(10):1100–1106, 2013.

[6] Sandra T Sigmon, Jennifer J Pells, Nina E Boulard, Stacy Whitcomb-Smith, Teresa M Edenfield, Barbara A Hermann, Stephanie M LaMattina, Janell G Schartel, and Elizabeth Kubik. Gender differences in self-reports of depression: The response bias hypothesis revisited. *Sex Roles*, 53(5-6):401–411, 2005.

[7] Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine*, 25(9):1337–1340, 2019.