Understanding Algorithmic Fairness in Health Care: A Proposed Case Study with Three Datasets

Anonymous Author(s)
Affiliation
Address
email

Abstract

Research of algorithmic fairness in machine learning (ML) has widely focused on datasets that explore criminal cases or credit loans. In this work, we present a work in progress that aims to take advantage of advances in health care and characterize ML fairness in-depth for the health domain. Our case study will focus on three datasets, including one large dataset from Brazil, and we intend to raise fairness concerns in these databases as well as identify misdiagnosis in patients with protected attributes. Finally, we aim at proposing solutions (in partnership with health care professionals) to mitigate problems related to ML unfairness.

1 Introduction

From smart assistants to cancer diagnostics, there is an increasing demand for machine learning (ML) algorithms in real-world applications. Such intelligent systems are being used in settings such as employment, credit lending and criminal justice [10, 13, 2]. With this increase in the ubiquity of intelligent systems in society, there is also a perpendicular increasing concern regarding the societal impact of ML algorithms. This rises from the the fact that decisions made with the support of such systems impact people's lives, leading to increasing concerns on how ML affects society. Such concerns have led the machine learning community to currently focus on problems related to the fairness [3, 7, 16] of intelligent systems that exploit ML algorithms.

Due to the relevance of fairness as a research topic nowadays, multiple metrics and processes to capture fairness have been recently proposed in the computer science literature [12, 17]. As a consequence, there is no clear agreement among researchers over a particular definition of fairness [5]. One popular line of thought is that fairness may be considered as the absence of algorithmic bias. However, how can we define bias? Should we use statistical or social definitions? To go from a societal notion to a metric using biases as a mediator, researchers are concerned with attributes (i.e., input to a ML algorithm) from individuals or groups that do not accurately represent a population (e.g., under-representation of race, gender or other sensitive attributes in a dataset). This leads to the notion that fairness is concerned with some kind of group or individual parity. While at a first glance this may seem sufficient, with careful considerations we make the statement that representativeness is not enough (but is still important) to capture fairness in every domain.

To make our arguments, consider two examples from the health-domain. Initially, diseases like skin cancer will naturally have a higher incidence rate towards people with lighter skin.[1] Another example is breast or prostate cancer, where some genders are naturally more affected than others. Given these characteristics, it is not yet clear how researchers may quantify fairness in health care. One point of view is that when there is a causal relationship between a sensitive attribute (e.g., gender, ethnicity) and a disease, it is expected that the lack of representativeness will not be an issue. However, when such a causal notion is not present, then we may have a representation problem. It is important to

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5403065/

understand this distinction since different outcomes by protected class may not be unfair depending on the task at hand (e.g., disease classification). Nonetheless, they would indeed be unfair if they result from suboptimal treatment decisions made in the past or the inability to follow-up treatment (e.g., due to geographical distance, monetary or transportation constraints) [6].

Overall, we can conclude that dealing with fairness is a complex matter. On one hand, the accuracy of ML systems will depend on certain sensitive attributes. On the other hand, when these systems are used in decision making, biased decisions reinforce social and historical inequalities.

Research goals: Motivated by the above discussion, our research focuses on the following research questions: RG1 - Raise fairness concerns in health datasets. As we have argued, the representativeness issue is complex in the medical domain. Yet, the use of sensitive attributes can not be ignored since they may preserve medical errors or overpass bias on predictions, which can put at risk patients' lives [11, 15]. Following the work from [9] which investigated several databases and domains, our goal is to uncover inequalities in health data and raise fairness concerns. One example of such inequalities are diseases where miss-classification rates differ across demographic variables with no justifiable cause. RG2 - Understand classifiers and propose solutions. Under the assumption that the intelligent systems are used by health professionals, the accuracy of disease classifiers (e.g., for a case such as rheumatic fever) using sensitive human attributes or non-sensitive features (e.g., x-rays) will be compared. Our contribution from this evaluation is to understand whether sensitive attributes are essential for diagnosing certain diseases and propose solutions to mitigate problems related to the unfairness of a classifier following the approach of [18].

Methodology: As our research employs data science and machine learning techniques, our initial methodology will be an exploratory data study of medical attributes considered sensitive. In light of recent laws that forbid the use of sensitive attributes on some applications, it's critical to understand to what extent do sensitive features impact ML algorithms. For this exploratory phase, we will analyze the MIMIC-III database [8], the eICU Collaborative Research Database [14] and the Telehealth Network of Minas Gerais database [1]. As an outcome of this research step, an in-depth characterization of relationships between sensitive attributes and cardiac diseases (e.g, Coronary Artery disease) is expected. With this exploratory analysis we will follow up with an evaluation of ML algorithms (e.g., from classical choices such as Naive Bayes/Decision Trees to more novel Deep Learning approaches) focusing on misdiagnosis towards sensitive attributes. Next, we aim to work with health care professionals to understand if proposed mitigations [18] are sufficient.

2 Related Work

To this date, more than 30 definitions exist and each one has details and differences that make them difficult to co-exist in the same situation. For this reason, we investigated several definitions of fairness [5, 7, 12, 17]. In parallel with research focused on ML fairness, a digital transformation in terms of health care is also happening [4, 6, 15]. While healthcare, judicial systems, and credit loans may all benefit from ML, not every domain will share the same set of fairness concerns. Nevertheless, some common trends exist as the legal definition of protected groups. This refers to groups that suffered from biased experiences in the past and yet remain vulnerable to harm by incorrect predictions or withholding of resources [15].

As stated, this work aim to understand ML fairness in-depth for the health domain. Our work will focus on two open datasets [14, 8]. Our study will also explore a dataset from the Telehealth Network of Minas Gerais [1], the first large dataset from a different demographic than the USA.

3 Contributions

As in previous work [7, 15, 18], the results from our research will be useful to comprehend if is possible to achieve and establish relevant fairness, both for health professionals and users. Overall, our research results may serve as a basis for a novel set of methodologies that need to be put in place when analyzing sensitive features from the health domain. This is particularly important in light of the recent laws (e.g., EU's GDPR and Brazil's LGPD) which regulate how such attributes may be stored and used by practitioners of the health domain. Our research also contributes by bridging findings from other spheres [9] to different application and cultural domains, something that our research also focuses on as we study data from Brazil.

References

[1] Maria Beatriz Alkmim, Renato Minelli Figueira, Milena Soriano Marcolino, Clareci Silva Cardoso, Monica Pena de Abreu, Lemuel Rodrigues Cunha, Daniel Ferreira da Cunha, Andre Pires Antunes, Adélson Geraldo de A Resende, Elmiro Santos Resende, et al. Improving patient access to specialized health care: the telehealth network of minas gerais, brazil. *Bulletin of the World Health Organization*, 90:373–378, 2012.

[2] Anna Maria Barry-Jester, Ben Casselman, and Dana Goldstein. The new science of sentencing. *The Marshall Project*, 4:2015, 2015.

[3] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H Chi. Putting fairness principles into practice: Challenges, metrics, and improvements. *arXiv preprint arXiv:1901.04562*, 2019.

[4] 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(11):981, 2018.

[5] Pratik Gajane and Mykola Pechenizkiy. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184*, 2017.

[6] Steven N Goodman, Sharad Goel, and Mark R Cullen. Machine learning, health disparities, and causal reasoning. *Annals of internal medicine*, 2018.

[7] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS Symposium on Machine Learning and the Law*, volume 1, page 2, 2016.

[8] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[9] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.

[10] Claire Cain Miller. Can an algorithm hire better than a human. *The New York Times*, 25, 2015.

[11] Sendhil Mullainathan and Ziad Obermeyer. Does machine learning automate moral hazard and error? *American Economic Review*, 107(5):476–80, 2017.

[12] Arvind Narayanan. Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp., New York, USA*, 2018.

[13] Kevin Petrasic, Benjamin Saul, James Greig, Matthew Bornfreund, and Katherine Lamberth. Algorithms and bias: What lenders need to know. *White & Case*, 2017.

[14] Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5, 2018.

[15] Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872, 2018.

[16] Nripsuta Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David Parkes, and Yang Liu. How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness. *arXiv preprint arXiv:1811.03654*, 2018.

[17] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pages 1–7. IEEE, 2018.

[18] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019.