Fairness in Machine Learning for Healthcare

Muhammad Aurangzeb Ahmad
Department of Computer Science
University of Washington Tacoma
KenSci Inc
Seattle, WA, USA
maahmad@uw.edu

Dr. Arpit Patel
Department of Bioinformatics and Medical Education
University of Washington
Seattle, WA, USA
arpitp@uw.edu

Dr. Carly Eckert
Department of Epidemiology
University of Washington
KenSci Inc
Seattle, WA, USA
millerc7@uw.edu

Vikas Kumar
KenSci Inc
Seattle, WA, USA
vikas@kensci.com

Ankur Teredesai
Department of Computer Science
University of Washington Tacoma
KenSci Inc
Seattle, WA, USA
ankur@kensci.com

ABSTRACT

The issue of bias and fairness in healthcare has been around for centuries. With the integration of AI in healthcare the potential to discriminate and perpetuate unfair and biased practices in healthcare increases many folds. The tutorial focuses on the challenges, requirements and opportunities in the area of fairness in healthcare AI and the various nuances associated with it. The problem healthcare as a multi-faceted systems level problem that necessitates careful consideration of different notions of fairness in healthcare to corresponding concepts in machine learning is elucidated via different real world examples.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Machine learning**; **Machine learning algorithms**; • **Applied computing** → **Health care information systems**; **Health informatics**.

KEYWORDS

healthcare ai, machine learning in healthcare, fairness, fatml, fate ml

ACM Reference Format:
Muhammad Aurangzeb Ahmad, Dr. Arpit Patel, Dr. Carly Eckert, Vikas Kumar, and Ankur Teredesai. 2020. Fairness in Machine Learning for Healthcare. In *26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '20), August 23–27, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3394486.3406461

KDD '20, August 23–27, 2020, Virtual Event, USA

ACM ISBN 978-1-4503-7998-4/20/08.
https://doi.org/10.1145/3394486.3406461

1 EXTENDED ABSTRACT

Responsible machine learning is central to driving adoption of machine learning in healthcare. While the focus of deployment of responsible machine learning system has largely been on robustness and interpretable machine learning, fairness is now becoming a pivotal issue in healthcare AI/ML. Even though there is already a large and growing body of literature on fairness in machine learning in general, a focused emphasis on requirements for fair and unbiased systems deployed in healthcare settings is lacking. This tutorial is motivated by the need to comprehensively study fairness in the context of applied machine learning in healthcare. Although the issue of fairness in healthcare AI may seem recent, one of the earliest examples of algorithmic discrimination from the 1970s in fact comes from healthcare where an algorithm employed by St. George's Hospital Medical School in the UK was discriminating on the basis of race and gender in making initial screening decisions for applicants to medical school.

Even outside of AI/ML, healthcare and medicine have a long history of implicit or explicit bias which has been well documented: There is a large body of research showing that minority patients receive poorer quality of care despite similar disease severity, clinical presentation and medical insurance. Healthcare has been rife with examples of algorithmic discrimination. A study started in 1958 on normal human aging did not include any women for 20 years till 1978. Another study found that older women were less likely to get lifesaving interventions as compared to older men. Bierman [4] found that older women were less likely to be given lifesaving interventions as compared to men, Chen et al [6] observed women are less likely to be given analgesia, Tamayo-Sarver et al [12] noted that Blacks were less likely to be given opioids as compared to Whites and Latinos etc. In a scathing incitement of the field of medical research published in Hastings Center Report in 1992, Rebecca Dresser [8] highlighted that most published medical studies used White Male as the norm with exclusion of women and minority populations, and thus with questionable generalizability. What these examples demonstrate is that applied AI/ML in healthcare not only has to navigate algorithmic bias but other human biases that may creep in during the healthcare delivery process.

In recent years there have been increasingly vocal calls for fairness in machine learning in a vast number of areas; however, what constitutes fairness is healthcare is quite different from most other domains [7]. In this tutorial we will extensively cover the definitions, nuances, challenges, and requirements for the design of fair and unbiased machine learning models and accompanying systems in healthcare. We note that the problem of fairness in healthcare does not directly map onto machine learning since the goal of the standard objective functions in machine learning predictive models is to create accurate models for the majority class which may be at the expense of the protected class [5]. Fair ML can reduce the bias in AI/ML systems by addressing bias in the data (Selection/sample bias, Response bias, Publication bias, Prejudicial bias, Measurement bias, Hawthorne effect, Social desirability bias, Self-reporting bias, Linking bias, Temporal bias), bias in Algorithms (Pre-existing, Technical, Emergent) and bias in the delivery [10]. We not only discuss what notions of fairness are applicable in what scenarios in healthcare but also describe how one would select the right interpretable machine learning algorithm for a given problem in healthcare [3]. In this tutorial we will highlight the scope, the limitations and the pitfalls associated with applied AI/ML in healthcare with the acknowledgement that machine-based decision making has the potential to be much more transparent as compared to human decision making.

The generalizability of AI algorithms across subgroups is critically dependent on factors like representativeness of included populations, missing data, and outliers. Generalizability and representativeness are also important considerations when interpreting randomized clinical trials (RCT) [9] and many of the data related issues that are present in RCT are also applicable to AI and machine learning models. Consider Electronic Health Records (EHRs) which are basically observational databases, the data in EHRs reflects not just the health of the patients but also their interactions with the healthcare system e.g., the date of a code for a diseases is when the physician made the diagnosis, not when the patient first developed the disease [2]. Similarly, the billing code used for an office visit might be influenced more by reimbursement policies of the organization or the government than the original reason for the visit.

When AI models are deployed in real world settings in healthcare, predictions start affecting the outcomes. This is because actions are taken on the basis of machine learning predictions which in turn invalidate the predictions unless retraining and re-tuning of the models is not done. Any discussions of fairness in AI should take the feedback loop and the delayed effects of action and its consequences into account [11]. This is further complicated by the fact that most machine learning models treat the world as relatively simple closed systems, in reality delivering healthcare is a complex phenomenon which has disparate impacts on individuals over time. Need for fairness in healthcare is not limited to algorithm design but also to other aspects of the software engineering process including input data, model parameters, and visualization and embedding of model results in the (healthcare) workflow. Additionally, the type of fairness needed (Unawareness, Demographic Parity, Equalized Odds, Predictive Rate Parity, Individual Fairness, Counterfactual fairness) is highly dependent upon the use case, the protected population, the risk associated with the outcome etc. We map requirements for fairness in machine learning across the spectrum of healthcare problems and machine learning solutions e.g., a machine learning system for emergency department utilization in healthcare has different requirements for fairness as compared to a system that predicts a patient's mortality i.e., vastly different notions of fairness may be applied in the two settings.

An oft neglected topic in fairness is healthcare is delivery; Adelman [1] noted that socioeconomic status, gender and ethnicity have implicit and explicit effect on how healthcare is delivered e.g., studies have shown that clinicians are less likely to believe black women when they complain about pain, which translates into less care given to them and which ultimately translates to significantly worse outcomes for black women. AI/ML based systems also offer the possibility to help identify and potentially reduce such biases. We employ a number of examples in healthcare that are based off of our own experience of deploying machine learning models in a commercial setting healthcare system in the United States, Europe, Asia and Australia. Based on a comprehensive survey of literature on fairness in machine learning in the context of healthcare we describe a framework which can be used to evaluate AI/ML systems across the axes of healthcare. We then map this framework and various machine learning algorithms to multiple scenarios such as risk prediction for readmissions, mortality prediction, disease progression, and diagnosis detection. Lastly, it is important to emphasize that purely algorithmic answers to the question of fairness in healthcare AI/ML is not the correct answer to problems in healthcare since the AI/ML models constitute only a small part of the complex network of delivery of healthcare so a system level view of healthcare is often needed to create a system that is relatively unbiased. We conclude the tutorial with a discussion on constraints and pitfalls for interpretable machine learning within healthcare and solicit audience perspective by conducting a use-case selection and exploration exercise.

REFERENCES

[1] Larry Adelman. 2007. Unnatural causes: Is inequality making us sick? *Preventing Chronic Disease* 4, 4 (2007).

[2] Denis Agniel, Isaac S Kohane, and Griffin M Weber. 2018. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *Bmj* 361 (2018).

[3] Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. 2018. Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*. 559–560.

[4] Arlene S Bierman. 2007. Sex matters: gender disparities in quality and outcomes of care. *Cmaj* 177, 12 (2007), 1520–1521.

[5] Reuben Binns. 2018. Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*. 149–159.

[6] Esther H Chen, Frances S Shofer, Anthony J Dean, Judd E Hollander, William G Baxt, Jennifer L Robey, Keara L Sease, and Angela M Mills. 2008. Gender disparity in analgesic treatment of emergency department patients with acute abdominal pain. *Academic Emergency Medicine* 15, 5 (2008), 414–418.

[7] Robyn M Dawes, David Faust, and Paul E Meehl. 1989. Clinical versus actuarial judgment. *Science* 243, 4899 (1989), 1668–1674.

[8] Rebecca Dresser. 1992. Wanted single, white male for medical research. *The Hastings Center Report* 22, 1 (1992), 24–29.

[9] Thomas R Fleming and David L DeMets. 1993. Monitoring of clinical trials: issues and recommendations. *Controlled clinical trials* 14, 3 (1993), 183–197.

[10] Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236* (2016).

[11] Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed impact of fair machine learning. *arXiv preprint arXiv:1803.04383* (2018).

[12] Joshua H Tamayo-Sarver, Susan W Hinze, Rita K Cydulka, and David W Baker. 2003. Racial and ethnic disparities in emergency department analgesic prescription. *American journal of public health* 93, 12 (2003), 2067–2073.