Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230184

Fairness in Artificial Intelligence: Regulatory Sanbox Evaluation of Bias Prevention for ECG Classification

Arian RANJBAR[a,1], Kristin SKOLT[b], Kathinka Theodore AAKENES VIK[b], Beate SLETVOLD ØISTAD[b], Eilin WERMUNDSEN MORK[b] and Jesper RAVN[b]
[a] *Medical Technology and E-health, Akershus University Hospital, Norway*
[b] *Norwegian Data Protection Authority, Oslo, Norway*
[c] *Equality and Anti-Discrimination Ombud (LDO), Oslo, Norway*

Abstract. As the use of artificial intelligence within healthcare is on the rise, an increased attention has been directed towards ethical considerations. Defining fairness in machine learning is a well explored topic with an extensive literature. However, such definitions often rely on the existence of metrics on the input data and well-defined outcome measurements, while regulatory definitions use general terminology. This work aims to study fairness within AI, particularly bringing regulation and theoretical knowledge closer. The study is done via a regulatory sandbox implemented on a healthcare case, specifically ECG classification.

Keywords. Artificial Intelligence, Fairness, Bias, Ethics, Regulation, GDPR

1. Introduction

As Artificial intelligence (AI)-enabled solutions within healthcare are getting closer to production, a raised attention has been directed towards potential ethical issues. This development includes regulation of AI which increasingly use terminology or requirements of fairness and fair use.

Defining fairness has a long history within ethics. Although extensive research has been conducted on the topic, there is no general consensus on a definition. On the other hand, to provide fair models within machine learning, fairness needs to be quantifiable. Such algorithmic fairness definitions have previously been widely explored, [1]. Analysis often relies on either individual fairness, i.e. "mapping similar people similarly", or statistical parity. Common among them is the need to set the measurement with respect to a metric, protected attributes and desired outcomes.

This work aims to investigate fair use of AI within healthcare from a regulatory perspective. How is fairness defined, and what can be done to ensure fairness? More importantly, can algorithmic methodology and regulatory frameworks be aligned? The investigation is carried out within a regulatory sandbox on a case study of an AI-enabled ECG classifier.

[1] Corresponding Author: Arian Ranjbar, Akershus Universitetssykehus HF, 1478 Lørenskog, Norway; E-mail: arian.ranjbar@ahus.no. This research was partly funded by Nasjonalforeningen for folkehelsen.

2. Method

To evaluate the fairness and bias definitions from a regulatory perspective, a regulatory sandbox is implemented together with the corresponding authorities. The sandbox as a method was developed to enable testing of new technology or methodology, currently not (or recently) covered by or compliant with existing regulatory frameworks, [2].

3. Case Study

The case study regards a machine learning based ECG decision support system, developed internally at Akershus University Hospital using regular production systems, [3]. Relevant regulation is found in GDPR and Norwegian Anti-Discrimination Act.

A very brief summary of the outcomes follows. Fairness is central to several regulations such as GDPR. However, it has no clear definition, but rather follows a dynamic principle, i.e. adjusts over time in accordance to general societal perception. Corresponding guidelines also keep advice at a high level, not generally compatible with explicit quantification. The Norwegian equality and Anti-Discrimination Act, elaborate on fairness a a prohibition of discrimination with respect to twelve specified protected grounds, e.g. gender, ethnicity, age etc.; useful from a quantitative perspective. Bias similarly has no legal definition, however is often referred to in relation to fairness. In this regard four sources of bias are specifically pointed out: bias arising from already existing biases within healthcare, data collection, design- and deployment principles, and application injustice or misuse; all extensively researched within machine learning, [1].

Although the particular risk of bias in the ECG algorithm is judged as low, countermeasures may be necessary from a regulatory perspective such as: data control, monitoring systems with re-training capabilities, and establishment of training routines for clinicians. Notably, from a regulatory perspective, the full treatment chain may be considered in a fairness evaluation. Thus, even if a specific algorithm produces biases, it might be considered fair if patients receive the same treatment in the end.

4. Conclusion

This study took the first step in aligning fairness from a regulatory and algorithmic perspective, using the sandbox methodology. Particularly finding explicit traits to protect for and necessary bias reduction procedures. However, evaluation may still be challenging since information regarding protective grounds is generally not available, and may be in conflict with other regulation. Further research is also needed regarding metrics and outcomes to optimize for, with respect to compliance.

References

[1] Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A survey on bias and fairness in machine learning. ACM Computing Surveys (CSUR). 2021;54(6):1-35.

[2] Leckenby E, Dawoud D, Bouvy J, Jonsson P. The sandbox approach and its potential for use in health technology assessment: a literature review. Applied health economics and health policy. 2021.

[3] Ranjbar A, Ravn J, Ronningen E, Hanseth O. Enabling Clinical Trials of Artificial Intelligence: Infrastructure for Heart Failure Predictions. Proceedings of Medical Informatics Europe. 2023.