

Fairness-Aware and Explainable Machine Learning for Healthcare Prediction: A Systematic Review

Ananthi B¹, Kaviya R²

¹Assistant Professor, Department of Computer Science and Engineering,
Vivekanandha College of Engineering, for Women, Tamilnadu, India

²PG Scholar, Department of Computer Science and Engineering,
Vivekanandha College of Engineering, for Women, Tamilnadu, India

Abstract—Machine learning models trained on structured health data, such as electronic health records, insurance claims, registries and population surveys, are increasingly used for risk prediction and resource allocation. When the data or the problem formulation reflects existing inequities, these models can reproduce or amplify them, so that some patient groups receive less accurate predictions or less access to care. This paper reviews methods for mitigating bias in such models. We organise the literature by the stage at which mitigation is applied (pre-processing, in-processing and post-processing), summarise the fairness metrics used to judge success, and compare prior reviews and empirical studies. The evidence shows that pre-processing methods such as reweighing dominate current practice and can reduce measured disparities, that removing the protected attribute is often insufficient, that fairness constraints can reduce performance, and that no single model or metric is best. Gaps include inconsistent reporting, limited external validation and intersectional analysis, and weak links between fairness metrics and health outcomes. We conclude with practical recommendations for developing and reporting fairer clinical prediction models.

Index Terms—Algorithmic Bias, Fairness, Bias Mitigation, Structured Health Data, Electronic Health Records, Machine Learning, Health Equity.

I. INTRODUCTION

Prediction models built from routinely collected health data now guide decisions about who receives preventive care, extra follow-up or scarce resources. A widely used commercial algorithm for population health management demonstrated racial bias because it relied on healthcare expenditure as an indicator of patients' health needs. As a result, Black patients were

assigned lower risk scores than other patients with similar health conditions, reducing their chances of receiving additional medical support. Researchers estimated that addressing this bias could increase the proportion of Black patients identified for extra assistance from 17.7% to 46.5% [1]. This example highlights the importance of identifying bias in healthcare datasets and developing effective mitigation strategies to improve fairness in machine learning systems [4].

. It can arise from the way a prediction target is defined, and it can have large consequences for care.

Health data are particularly prone to such problems. Access to care, diagnostic practices and documentation all differ across groups, so observed records are an imperfect picture of true health needs. Researchers have argued that machine learning must be designed deliberately if it is to advance health equity and not undermine it [2], and that the use of race as an input in clinical algorithms deserves careful reconsideration [3]. A growing body of work proposes technical methods to detect and reduce bias in models built on structured, tabular data, which is the most common form of data in clinical risk prediction.

Several reviews exist. A systematic review of bias in models built from electronic health records (EHRs) catalogued This example highlights the importance of identifying bias in healthcare datasets and developing effective mitigation strategies to improve fairness in machine learning systems [4].and a scoping review examined fair machine learning with real-world health data [5]. This paper complements them by focusing on bias mitigation for structured health data, organising methods by the stage at which they act, and comparing what the empirical evidence says about their

effectiveness and costs. Section II gives background on bias, fairness metrics and mitigation methods, Section III explains the scope of the review, Section IV reviews the literature, Section V compares reported findings, Section VI analyses the evidence and open problems, and Section VII concludes.

II. BACKGROUND

A. Sources of Bias in Structured Health Data

Bias can enter at every stage of the machine learning life cycle, from data collection to deployment [6], and surveys of the field give broad taxonomies of the types of bias involved [7]. For structured health data, five sources are most often discussed. Label or proxy bias occurs when the recorded outcome measures something other than the intended concept, as in the cost example above [1]. Representation bias arises when some groups are under-represented, so that models fit them poorly. Measurement bias arises when variables are recorded or valued differently across groups. Missingness bias arises when the pattern of missing data depends on group membership or on care received. Historical bias reflects past inequities in care that are present in the data even when it is collected accurately.

B. Fairness Metrics

Fairness is judged with group metrics computed over a protected attribute such as race, sex or insurance status. Demographic parity compares the rate of positive predictions across groups, with disparate impact as a ratio form. Equal opportunity and equalised odds require similar true positive rates, and for equalised odds also false positive rates, across groups [9]. Calibration and predictive parity require that a given score means the same risk in each group [10]. Individual fairness asks that similar individuals be treated similarly [8]. These definitions cannot in general be satisfied together when base rates differ between groups [10], [11], and enforcing calibration and equalised error rates at the same time is generally not possible [12]. Choosing a metric is therefore a clinical and ethical decision, not only a technical one.

C. Mitigation Methods and Tools

Mitigation methods are grouped by the stage at which they act. Pre-processing changes the data before training, for example through resampling, reweighing

of instances [13], suppression of sensitive or proxy features, relabelling, or better imputation. In-processing changes the learning algorithm, for example by adding a fairness penalty to the loss or by adversarial debiasing, in which a second network tries to predict the protected attribute from the model output and the main model is trained to defeat it [14]. Post-processing adjusts the outputs of a trained model, for example by setting group-specific decision thresholds or by equalising error rates [9], [12]. Open-source toolkits such as AI Fairness 360 [15] and Fairlearn [16] implement many of these methods and metrics, which has made comparison on health data easier.

III. REVIEW SCOPE AND METHOD

The review addresses four questions. RQ1: which sources of bias are reported in models trained on structured health data? RQ2: which mitigation methods are applied, and at which stage? RQ3: how are fairness and predictive performance evaluated together? RQ4: what gaps remain? Eligible work concerns machine learning models trained on structured health data (EHRs, claims, registries or surveys) that measure or reduce bias. Imaging-only, text-only and non-health studies are excluded.

The evidence base of this draft consists of four prior systematic, scoping or mapping reviews [4], [5], [17], [18], whose own searches covered PubMed, Web of Science, IEEE Xplore, ACM Digital Library, Embase and Scopus, and four primary studies [1], [19], [20], [21] identified by targeted searches. It is therefore a structured synthesis and not yet a full protocol-driven review. To report it as a systematic review in the PRISMA sense, the authors should run and document their own database search and insert the counts of records identified, screened and included here: [records identified = __; screened = __; included = __].

IV. LITERATURE REVIEW

A. Prior Reviews

[4] Chen and colleagues, in Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic health record-based models, followed PRISMA guidelines and searched PubMed, Web of Science and IEEE for articles published between January 2010 and December 2023. They identified key biases, strategies

for detecting and mitigating them throughout model development, and metrics for assessing bias. The mitigation strategies they found predominantly involved data collection and pre-processing, such as resampling and reweighting, and they call for standardised, detailed reporting and for systematic real-world testing.

[5] Huang and colleagues, in A scoping review of fair machine learning techniques when using real-world data, examined metrics for quantifying fairness, public datasets for fairness research and bias mitigation approaches. They identified 35 review articles and then individual studies of fair machine learning with real-world data. Existing studies often report positive outcomes from pre-processing techniques. Open questions include the root causes of bias in models, extending fairness research to more settings, and bias in multi-modal data.

[17] Yang and colleagues, in A survey of recent methods for addressing AI fairness and bias in biomedicine, searched PubMed, the ACM Digital Library and IEEE Xplore for articles from 2018 to 2023, filtered 10,041 results to 55 articles, and described distributional methods (data augmentation, perturbation, reweighting and federated learning) and algorithmic methods (representation learning, adversarial learning, loss-based and causality-based methods).

[18] Mattos, Siqueira and Garcia, in Fair and Equitable Machine Learning Algorithms in Healthcare: A Systematic Mapping, analysed 37 articles from five sources. They report growing interest and give an overview of approaches and limitations for evaluating and mitigating bias and discrimination in healthcare machine learning.

B. Empirical Studies on Structured Health Data

[1] Obermeyer, Powers, Vogeli and Mullainathan analysed a widely used algorithm for identifying patients with complex health needs. The bias came from using cost as a proxy for need, and reformulating the algorithm so that it no longer uses costs as a proxy for needs eliminates the racial bias in predicting who needs extra care. This is a mitigation through a change of the label, and it illustrates that the most effective remedy can lie in problem formulation, not in a fairness algorithm applied afterwards.

[19] Park and colleagues, in Comparison of methods to reduce bias from clinical prediction models of postpartum depression, used Medicaid claims data for White and Black pregnant women. Logistic regression, random forest and gradient boosting models were trained to predict postpartum depression and postpartum mental health service use. Reweighting, Prejudice Remover and removing race from the models were compared using disparate impact and equal opportunity difference. For the logistic regression model for postpartum depression, reweighting improved disparate impact from 0.31 to 0.79 and equal opportunity difference from -0.19 to 0.02, while removing race gave inferior results (disparate impact 0.61, equal opportunity difference -0.05).

[20] Pfohl, Foryciarz and Shah, in An Empirical Characterization of Fair Machine Learning for Clinical Risk Prediction, studied the effect of penalising group fairness violations during training across several observational healthcare databases, clinical outcomes and sensitive attributes. Procedures that penalise differences between the distributions of predictions across groups caused nearly universal degradation of several performance metrics within groups, and the authors urge researchers to engage with the broader sociotechnical context and not rely on the algorithmic fairness frame alone.

[21] Dang and colleagues, in Fairness and bias correction in machine learning for depression prediction across four study populations, studied bias in models for depression across four public datasets (LONGSCAN, FUUS, NHANES and UK Biobank). Standard machine learning approaches regularly showed biased behaviour. Standard mitigation techniques and a new post-hoc method were effective in reducing unfair bias, but no single model provided equality of outcomes. The authors stress analysing fairness during model selection and reporting transparently on debiasing.

V. PERFORMANCE METRICS

Table I summarises the reviewed studies. Findings are those reported by the authors and are not directly comparable, because populations, outcomes, fairness metrics and mitigation methods differ.

Table I. Comparison of Reviewed Studies On Bias Mitigation for Structured Health Data

Study	Data / task	Mitigation approach	Fairness metrics	Key reported finding
Obermeyer et al. [1]	Care-management risk algorithm used at scale; cost as the prediction target	Change of label (stop using cost as a proxy for need)	Risk score versus health status by race	Remedy would raise Black patients identified for extra help from 17.7% to 46.5%; bias removed by reformulating the label
Park et al. [19]	Medicaid claims; postpartum depression and mental health service use (LR, RF, XGBoost)	Reweighting; Prejudice Remover; removing race	Disparate impact; equal opportunity difference	Reweighting (LR, PPD): DI 0.31 to 0.79; EOD -0.19 to 0.02; removing race inferior (DI 0.61; EOD -0.05)
Pfohl et al. [20]	Multiple observational healthcare databases; clinical risk prediction	In-processing: penalising group fairness violations	Conditional prediction parity; calibration; cross-group ranking	Penalties caused near-universal degradation of performance within groups
Dang et al. [21]	Depression prediction; LONGSCAN, FUUS, NHANES, UK Biobank	Standard methods plus a new post-hoc threshold method	Group fairness metrics across sex, ethnicity and other attributes	Standard models biased; mitigation effective; no single best model
Chen et al. [4]	Systematic review (PRISMA); EHR-based AI; 2010 to 2023	Mostly resampling and reweighting (pre-processing)	Performance and fairness metrics	Call for standardised reporting and real-world testing
Huang et al. [5]	Scoping review; fair ML with real-world data	Pre-processing most often reported positive	Fairness quantification metrics	Gaps: root causes of bias, wider settings, multi-modal data
Yang et al. [17]	Survey; 55 articles from 10,041 results; biomedicine	Distributional and algorithmic methods	Not itemised here	Broad catalogue of methods and their strengths and weaknesses
Mattos et al. [18]	Systematic mapping; 37 articles; healthcare ML	Approaches for evaluating and mitigating bias	Not itemised here	Growing interest; overview of approaches and limitations

VI. ANALYSIS

Several patterns emerge from the reviewed work.

Pre-processing dominates. Prior reviews find that reported mitigation in EHR-based work is mostly pre-processing, especially resampling and reweighting [4], [5]. This is understandable, since these methods are model-agnostic and fit existing workflows. The evidence in [19] shows that reweighing can substantially improve disparate impact and equal opportunity on a clinical prediction task, though for one dataset and two groups.

Removing the protected attribute is often not enough. In [19], excluding race was less effective than reweighing, which is consistent with the view that other variables act as proxies for the removed attribute. At the same time, the clinical meaning of using or omitting race is itself debated [3], so the choice needs clinical input and not only a fairness score.

Some bias cannot be fixed after the fact. The cost-as-need problem in [1] was solved by changing the target, not by adjusting a trained model. Pfohl and colleagues argue that algorithmic fairness methods cannot correct causal quantities that observational criteria do not capture, and encourage a broader sociotechnical view [20].

Fairness and performance interact, and metrics conflict. Fairness penalties degraded within-group performance in [20]. Theory shows that common fairness criteria cannot all hold when base rates differ [10], [11], [12]. Studies should therefore state which criterion is targeted, why, and what is traded away.

No single method or model is best. The four-population study found no model that provided equality of outcomes, and recommends examining fairness during model selection [21]. Results on one dataset or condition should not be assumed to transfer. Reporting and validation are weak. Reviewers call for standardised reporting of methods and for real-world

testing [4]. The empirical studies above are retrospective and mostly assess one or two protected attributes at a time.

Open challenges. (i) Intersectional groups and small subgroup sizes, which make fairness estimates uncertain. (ii) External validation and robustness under data shift. (iii) Linking fairness metrics to clinical benefit and health equity outcomes, not only to statistical parity. (iv) Bias in missing data and in outcome labels. (v) Bias across modalities beyond tabular data [5]. (vi) Consistent definitions and reporting standards.

Recommendations for practice. (1) Audit the prediction target and key variables for proxy and measurement bias before modelling. (2) Report performance and fairness for each subgroup, with confidence intervals. (3) Select fairness metrics with clinical stakeholders and justify them. (4) Compare several mitigation methods, at different stages, rather than a single one, and report the performance cost. (5) Validate externally and monitor after deployment. (6) Document data, methods and limitations in a standard format.

VII. CONCLUSION

Bias in models trained on structured health data is a documented problem with real consequences for patients. The literature offers a toolbox of pre-processing, in-processing and post-processing methods, and the available evidence suggests that simple pre-processing such as reweighing can reduce measured disparities, while removing a protected attribute alone is often insufficient and fairness penalties can lower performance. The strongest lesson is that mitigation starts before modelling, with the choice of target, variables and metrics. Progress now depends on better reporting, external validation, intersectional analysis and evidence that reduced disparities in metrics translate into more equitable care. A full protocol-driven systematic search, reported according to PRISMA, would strengthen and extend the synthesis presented here.

REFERENCES

[1] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. Upper Saddle River, NJ, USA: Prentice Hall, 2009.

[2] Lighthill, "Artificial Intelligence: A General Survey," in *Artificial Intelligence: A Paper Symposium*. London, U.K.: Science Research Council, 1973.

[3] M. Minsky and S. Papert, *Perceptrons*. Cambridge, MA, USA: MIT Press, 1969.

[4] J. J. Hopfield, "Neural Networks and Physical Systems with Emergent Collective Computational Abilities," *Proceedings of the National Academy of Sciences*, vol. 79, pp. 2554–2558, 1982.

[5] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning Internal Representations by Error Propagation," DTIC Document, 1985.

[6] M. Wooldridge and N. R. Jennings, "Intelligent Agents: Theory and Practice," *The Knowledge Engineering Review*, vol. 10, pp. 115–152, 1995.

[7] L. A. Zadeh, "Fuzzy Logic—A Personal Perspective," *Fuzzy Sets and Systems*, vol. 281, pp. 4–20, 2015.

[8] N. Spinrad, "Mr Singularity," *Nature*, vol. 543, p. 582, 2017.

[9] E. Horvitz, *One-Hundred-Year Study on Artificial Intelligence: Reflections and Framing*. Stanford University, 2014.

[10] S. Lohr, "The Age of Big Data," *The New York Times*, 2012.

[11] P. Alston, "Lethal Robotic Technologies: The Implications for Human Rights and International Humanitarian Law," *Journal of Law, Information and Science*, vol. 21, p. 35, 2011.

[12] Young and M. Yung, "Deni..." [Reference incomplete in the supplied text.]