Fair and Equitable Machine Learning Algorithms in Healthcare: A Systematic Mapping

Marcelo S. Mattos[a], Sean W. M. Siqueira[b] and Ana Cristina B. Garcia[c]

Graduate Program of Informatics - PPGI, Federal University of the State of Rio de Janeiro, Rio de Janeiro, Brazil

Keywords: Fairness, Equity, Bias, Machine Learning, Healthcare.

Abstract: Artificial intelligence (AI) is being employed in many fields, including healthcare. While AI has the potential to improve people's lives, it also raises ethical questions about fairness and bias. This article reviews the challenges and proposed solutions for promoting fairness in medical decisions aided by AI algorithms. A systematic mapping study was conducted, analyzing 37 articles on fairness in machine learning in healthcare from five sources: ACM Digital Library, IEEE Xplore, PubMed, ScienceDirect, and Scopus. The analysis reveals a growing interest in the field, with many recent publications. The study offers an up-to-date and comprehensive overview of approaches and limitations for evaluating and mitigating biases, unfairness, and discrimination in healthcare-focused machine learning algorithms. This study's findings provide valuable insights for developing fairer, equitable, and more ethical AI systems for healthcare.

1 INTRODUCTION

Ensuring fairness in healthcare is crucial for equitable access and quality care. Artificial Intelligence (AI) promises advancements in healthcare decision-making, but raises critical ethical concerns around fairness and bias.

Unfair algorithms can lead to disparities in access, quality, and health outcomes. For example, Obermeyer et al. (2019) found a widely used hospital readmission algorithm biased against black patients due to historical healthcare disparities.

AI ethics addresses the ethical implications of developing and deploying AI systems, drawing on fields like engineering ethics, philosophy of technology, and science and technology studies (Kazim and Koshiyama, 2021). Fairness is a key ethical principle, meaning AI systems should treat everyone equally, regardless of personal characteristics (Ashok et al., 2022). However, achieving fairness in AI can be challenging, as AI systems are often trained on data that reflects historical biases.

This systematic mapping of the literature delves into the intricate landscape of AI fairness in the context of medical decision-making. By conducting a comprehensive analysis, we aim to shed light on the challenges posed by algorithmic biases and the proposed solutions that can pave the way for a more equitable healthcare system. We analyzed 37 scholarly articles sourced from reputable databases, including the ACM Digital Library, IEEE Xplore, PubMed, ScienceDirect, and Scopus.

Our analysis reveals a growing interest in AI fairness in healthcare, with a recent surge in publications. This study offers a comprehensive overview of approaches and limitations for assessing and mitigating biases, unfairness, and discrimination in healthcare machine learning algorithms.

Research questions (RQs):

RQ1: What are the main statistical, technical, and ethical approaches used to assess and mitigate biases, unfairness, inequalities, and discriminations in machine learning algorithms applied to healthcare?

RQ2: What are the technical, ethical, and social limitations and challenges in the design, development, and implementation of fair and equitable machine learning algorithms in healthcare?

RQ3: What are the research gaps pointed out in the articles on fairness and equity in machine learning in healthcare?

This systematic mapping literature review aims to contribute to the study of AI ethics, specifically focusing on fairness in machine learning within the healthcare field.

[a] https://orcid.org/0009-0006-9830-6391
[b] https://orcid.org/0000-0002-0864-2396
[c] https://orcid.org/0000-0002-3797-5157

Mattos, M., Siqueira, S. and Garcia, A.
Fair and Equitable Machine Learning Algorithms in Healthcare: A Systematic Mapping.
DOI: 10.5220/0012394700003636

In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024) - Volume 3*, pages 815-822
ISBN: 978-989-758-680-4; ISSN: 2184-433X

2 RELATED WORKS

Literature review publications have been providing insights into the state of the art about fairness in the field of ethical AI. Pessach and Shmueli (2022) conducted a review on fairness in machine learning algorithms, emphasizing the importance of developing accurate, objective, and fair machine learning (ML) algorithms. They discussed causes of algorithmic bias, definitions, and measures of fairness, and mechanisms for enhancing fairness.

Garcia et al. (2023) conducted a systematic review on algorithmic discrimination in the credit domain, making important contributions by covering fundamentals of discrimination theory, the legal framework, concepts of algorithmic fairness and fairness metrics applied in machine learning.

Other reviews have focused on fairness in ethical AI in the medical field. Bear Don't Walk et al. (2022) conducted a scoping review on ethical considerations in clinical natural language processing (NLP). The review highlights ethical considerations in metric selection, identification of sensitive patient attributes, and best practices for reconciling individual autonomy and leveraging patient data. Morley et al. (2020) mapped the ethical issues surrounding the incorporation of AI technologies in healthcare delivery and public health systems.

Our systematic mapping study differs from published works due to its approach of searching the literature for studies that have investigated approaches, limitations, and methods to evaluate and mitigate biases, injustices, inequalities, and discriminations in machine learning algorithms applied in the healthcare field.

3 METHODOLOGY

We followed the guidelines of Kitchenham and Charters (2007) for conducting systematic mappings.

The systematic mapping process was conducted using three software tools: 1. **Parsifal**: Assisted in creating the search query, applying inclusion and exclusion criteria, and retrieving relevant papers; 2. **Zotero**: Used for organizing papers in a taxonomy, creating classification terms, and conducting textual analysis; 3. **Iramuteq**: Employed to visualize the results of the mapping process.

The search string design is the basis for any systematic study that will guarantee reproducibility.

The online tool Parsifal was used to identify primary studies. We entered the PICOC (population, intervention, comparison, outcomes, and context) terms and research questions into Parsifal, which automatically generated keywords and a search string. We then made the necessary adjustments.

The terms entered in Parsifal were: **Population**: healthcare field; **Intervention**: machine learning, deep learning, neural networks; **Outcomes**: evaluation of the fairness of decisions made from the use of machine learning algorithms applied in the healthcare field; **Context**: use of machine learning techniques in the healthcare field with the aim of improving decision-making processes and, at the same time, guaranteeing justice and equity in the results.

Based on the PICOC definition, the search string was: healthcare AND ("machine learning" OR "deep learning" OR "neural networks") AND (fairness OR bias OR discrimination OR equity OR justice).

We applied this search string to the ACM Digital Library, IEEE Xplore, PubMed, ScienceDirect, and Scopus databases, searching titles and abstracts without a specific time period. The searches of ACM DL, IEEE Xplore, PubMed, and ScienceDirect were conducted on April 13, 2023, and the Scopus search was conducted on May 9, 2023.

We retrieved a total of 1,013 records from the following databases: 59 from ACM DL, 70 from IEEE Xplore, 227 from PubMed, 123 from ScienceDirect, and 484 from Scopus (Figure 1).

To ensure that all duplicate records were identified, we used Zotero software in conjunction with Parsifal to check for duplicates: 349 duplicate records were removed (Figure 1).

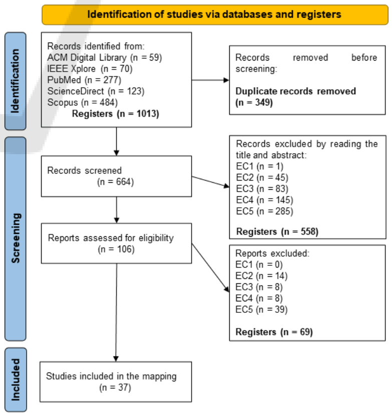

Figure 1: Prism diagram detailing the process of identification, screening, and inclusion of studies.

We applied the inclusion and exclusion criteria (Table 1). Initially, we assessed the title and abstract of the articles, leading to the elimination of 558 records out of the initial 664 considered, resulting in 106 records eligible for further assessment. Subsequently, we thoroughly read the full texts and re-

Table 1: Inclusion and exclusion criteria.

Inclusion criteria	
Criterion	**Description**
IC1	Articles must describe the application of machine learning techniques in healthcare and explore issues of fairness and equity.
IC2	Articles must report on outcomes related to fairness or equity in decision-making, with a focus on measures of fairness in machine learning.
Exclusion criteria	
Criterion	**Description**
EC1	Articles that are not written in English.
EC2	Studies that are not related to healthcare or the application of machine learning techniques in healthcare.
EC3	Studies that are preliminary reports, books, editorials, abstracts, posters, panels, lectures, roundtables, workshops, tutorials, or demonstrations.
EC4	Studies that are systematic reviews, meta-analyses, scoping studies, policy reports, guidelines, theoretical analysis, or consensus statements.
EC5	Studies that do not report on outcomes related to fairness or equity in decision-making.

applied the inclusion and exclusion criteria, which led to the removal of an additional 69 records. As a result, 37 studies were included in the mapping (Figure 1).

4 AN OVERVIEW OF THE ARTICLES

Our analysis of the 37 articles included in the systematic mapping revealed key themes and trends. In this overview, we used a word cloud (Figure 2) and a similarity graph (Figure 3).

The selected articles were published from 2019 to 2023. We observed a total of 1, 6, 8, 17, and 5 publications in the years 2019, 2020, 2021, 2022, and 2023 (5 publications by May 2023), respectively. These data suggest a growing interest in our research topic in 2022.

The word cloud (Figure 2) was generated based on article titles and abstracts to identify the most frequent words in the texts. We excluded the words that were present from our search string to identify other relevant terms.

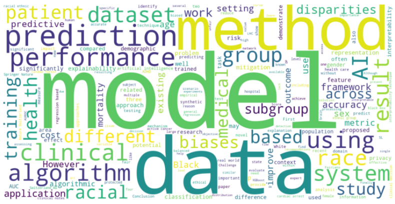

Figure 2: Word cloud.

The word "model" emerges as the most prominent word, highlighting its importance in the field of study. "Data" and "method" also highlight the importance of data and training methods in machine learning algorithms. Other noteworthy words related to application development and machine learning include dataset, prediction, performance, system, application, algorithm, training, classification, and accuracy. Words like biases, group, racial, sex, subgroup, disparities, and race are prominent in the context of fairness. Regarding healthcare, relevant words include clinical, medical, and patient. The prominent words in Figure 2 align with the findings of the similarity graph presented below.

Figure 3 illustrates a similarity analysis based on the texts of article titles and abstracts performed using the Iramuteq software. The analysis highlights the central term "model" and its connections to related concepts like healthcare, data, bias, fairness, and machine learning.

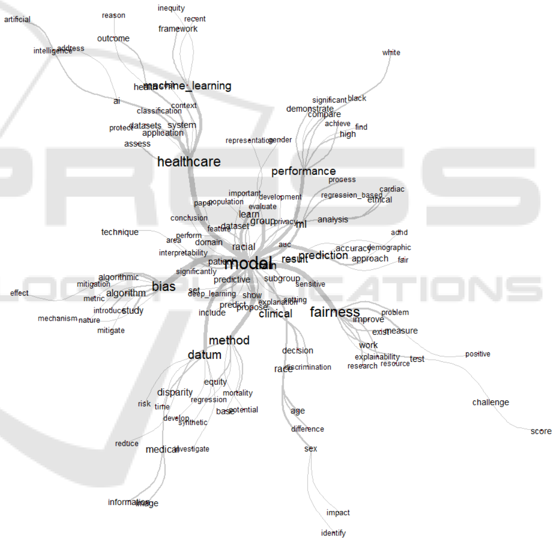

Figure 3: Similarity analysis performed in Iramuteq.

The similarity analysis reveals several relevant insights:

Healthcare Focus: Near the word "healthcare," we find terms like system, application, datasets, and AI (artificial intelligence), highlighting the focus on software development and the application of machine learning in healthcare in the selected articles.

Combating Bias: Within the context of "fairness," studies focus on combating disparities that may arise from biased machine learning algorithms used in healthcare. Words associated with "fairness" include improve, problem, research, resource, explainability, measure, test, challenge, and work. Notably, "sensitive" and "subgroup" appear between "model" and "fairness," indicating specific areas of concern.

Data and Methods: Connections to "datum" reveal important words like medical, image, information, synthetic, disparity, and investigate. These highlight the need for research in fairness, data challenges, and synthetic data generation approaches to reduce disparities and promote equity in medical AI development. Next to the word "datum," we have "method," which connects with reduction, equity, regression, and base, further emphasizing these themes.

Machine Learning in this Context: Within the "machine learning" (ML) category, relevant words include structure, inequality, and recent. Additionally, the "ml" branch (referring to machine learning) connects with words like analysis, ethical, cardiac, process, and regression based. These terms highlight current research directions and ethical considerations in this domain.

Mitigating Bias: Words related to mitigating algorithmic biases and understanding their negative impacts connect to the term "bias" through terms like algorithm, mitigation, metric, introduce, effect, mechanism, mitigate, and study. This underscores the importance of addressing bias issues in machine learning models.

This analysis provided valuable insights into the key topics and concerns surrounding fairness in machine learning for healthcare.

5 RESULTS

This study contributes by providing answers to the questions raised in the Introduction section, which are presented below:

RQ1: *What are the main statistical, technical, and ethical approaches used to assess and mitigate biases unfairness, inequalities, and discriminations in machine learning algorithms applied to healthcare?*

For the purpose of organization, we classified the approaches into three categories: statistical, technical, and ethical.

1. Statistical approaches
   There are several techniques to deal with imbalanced data, such as oversampling, undersampling, and resampling. Among these techniques, we found the following in the selected articles: 1. Undersampling (Zhang et al., 2021); 2. Resampling (Reeves et al., 2022); and 3. Stratified batch sampling (Puyol-Antón et al., 2021).

   Other articles were returned by the search string that described the application of data sampling techniques to balance imbalanced data, but they were not included in this study because they did not address fairness issues throughout the article.
2. Technical approaches
   The following are the main technical approaches described in the selected articles:
   - Adversarial training framework: Yang et al. (2023).
   - A new machine learning algorithm, called pseudo bias-balanced learning: Luo et al. (2022).
   - Assessing the impact of Swarm Learning (SL) on justice: Fan et al. (2021).
   - Differentially Private (DP): Suriyakumar et al. (2021).
   - EXplainable Artificial Intelligence (XAI) as a contribution to fairness in machine learning in healthcare: Rueda et al. (2022).
   - Connections between interpretability methods and fairness: El Shawi et al. (2019), Sahoo et al. (2022), and Meng et al. (2022).
   - Technique "de-bias" based on AIF360: Paviglianiti and Pasero (2020).
3. Ethical approaches
   The table 2 below summarizes the main studies and approaches described in the articles, aimed at ensuring that machine learning models are developed and used fairly and ethically, reducing the risk of bias related to factors such as gender, age, race, and other sociodemographic factors.

RQ2: *What are the technical, ethical, and social limitations and challenges in the design, development, and implementation of fair and equitable machine learning algorithms in healthcare?*

- Complex interactions between clinical entities: Predicting risk profiles accurately becomes difficult (Pham et al., 2023).
- Lack of interpretability: Understanding underlying mechanisms and model decisions is hindered (Chang et al., 2022; Meng et al., 2022).
- Access to healthcare data: Strict privacy laws protecting patient data in EHRs limit research reproducibility and hinder new discoveries (Bhanot et al., 2021).
- Tailoring methods for healthcare applications: Developing effective and specific models remains a challenge, as for example in medical image analysis (Stanley et al., 2022).

RQ3: *What are the research gaps pointed out in the articles on fairness and equity in machine learning in healthcare?*

Table 2: Main ethical approaches described in the articles.

Approach	Description	Article
A multi-view multi-task neural network architecture	Researchers designed a multi-view multi-task neural network architecture (MuViTaNet) and an equity variant (F-MuViTaNet) to accurately predict the onset of multiple complications and efficiently interpret its predictions, mitigating unfairness across different patient groups.	(Pham et al., 2023)
A framework for Representational Ethical Model Calibration	Framework developed to detect and quantify inequities in model performance across subpopulations defined by multiple and interacting characteristics.	(Carruthers et al., 2022)
Analysis of the effects of sociodemographic confounding factors	Shows that unfair models can produce different outcomes between subgroups, and that these outcomes can explain biased performance.	(Stanley et al., 2022)
Development of fairness metrics for synthetic data	Two metrics were developed by the researchers: 1. a disparity metric for synthetic data using the concept of disparate impact, and 2. a time-series metric to assess disparate impact overtime.	(Bhanot et al., 2021)
Experiments on race prediction from confounding factors	Suggest that race prediction models can be biased due to the presence of confounding factors.	(Duffy et al., 2022)
Fairness metric and testing for regression ML systems	Proposing a fairness metric and a fairness testing algorithm for regression machine learning systems	(Perera et al., 2022)
Four-step analytical process for identifying and mitigating biases	Provides a set of steps for identifying and mitigating biases in AI/ML algorithms and solutions.	(Agarwal et al., 2023)
Machine learning (ML) and optimization decoupling framework	Allows the ML and optimization components of the algorithm to be developed independently, which can help to reduce bias.	(Shanklin et al., 2022)
Machine learning (ML) model to reduce biases related to sex, age, and race	Uses a combination of techniques to reduce bias in machine learning models.	(Perez Alday et al., 2022)
Metrics to measure the fairness of explanation models or fidelity gaps between subgroups	The researchers introduce two new metrics: Maximum Fidelity Gap from Average, and Mean Fidelity Gap Amongst Subgroups.	(Balagopalan et al., 2022)
Quantitative evaluation of interpretability methods	Uses metrics to evaluate bias in deep learning models.	(Meng et al., 2022)
Study on predictive risks of minimal racial bias mitigation	Demonstrates that minimal racial bias mitigation can lead to worse predictive performance for minority groups.	(Barton et al., 2023)

A major gap is the need to address structural barriers and individual interactions in the health context to achieve health equity (Monlezun et al., 2022). Simply optimizing AI/ML algorithms to remove bias is insufficient. It is crucial to understand the broader social determinants of health and find subtler patterns that advocate for patients, rather than relying solely on group-level minority subgroup corrections (Li et al., 2022).

Another identified research gap emphasizes the importance of continually evaluating and auditing ML models for racial bias in clinical decision-making, even when explicit sensitive identifiers are removed from clinical notes (Adam et al., 2022).

Finally, there is a need for further research and testing of domain generalization methods in machine learning in clinical settings, exploring their impact on fairness and their performance when in the presence of bias (Zhang et al., 2021).

6 DISCUSSION

In this section, we discuss the results of our mapping.

6.1 Approaches for Machine Learning Models to Mitigate the Imbalanced Data Problem in Healthcare

This section discusses approaches for mitigating imbalanced data in healthcare machine learning models, acknowledging that such data is inherently imbalanced, uncertain, and prone to missing values (Wang et al., 2022). We specifically focus on three data sampling techniques: undersampling, resampling, and stratified batch, discussed in detail below.

6.1.1 Undersampling Use

Undersampling balances datasets by removing majority class samples. However, this risks losing valuable information (Alani et al., 2020; Reeves et al., 2022).

Zhang et al. (2021) suggested a framework for stress-test domain generalization methods in healthcare. They found that these methods can sometimes outperform traditional approaches, but may also lead to worse fairness and performance under certain conditions. They observed that directly providing the subsampled feature significantly reduces fairness and performance for both domain generalization and empirical risk minimization (ERM) algorithms. This suggests increased reliance on spurious correlations when the subsampled feature is directly provided, resulting in poor performance and fairness significantly worse under distribution shift.

6.1.2 Resampling for Fairness

Resampling methods can be classified into two main categories: oversampling and undersampling, both aiming to achieve a balanced class distribution (Alani et al., 2020).

In Reeves et al. (2022), the strategy used was resampling techniques (Blind, Separate, Equity) to balance the racial distribution in the data sample. Detailing the approach:

- Blind resampling: This is a baseline method that randomly samples a subset of the majority class (patients who do not die of suicide) to balance the training set. It does not consider racial/ethnic group membership.
- Separate resampling: This method separates the training data by racial/ethnic group and undersamples the majority class in each group to balance the data. It trains disjoint models for each racial/ethnic group.
- Equity resampling: This method divides the training data by both racial/ethnic group and class label.

6.1.3 Stratified Batch Sampling

In this approach, the data is stratified by the protected attribute(s) for each training batch, and samples are selected to ensure that each protected group is equally represented (Puyol-Antón et al., 2021).

6.2 The Importance of Explainability for Justice in Machine Learning in Healthcare

Rueda et al. (2022) highlight the inherent trade-off between precision and explainability in AI models. While high-performing models like deep learning often lack transparency, easily interpretable models typically exhibit lower precision (Holzinger et al., 2019 apud Rueda et al., 2022). This tension has significant implications for distributive justice, which concerns the fair allocation of resources. In healthcare, this means ensuring everyone has access to quality care, even when resources are limited.

Outcome-oriented justice theories prioritize precision to maximize benefits for the most people. However, Rueda et al. (2022) also bring procedural justice into the argument, which emphasizes that the process on which decisions are based is a fundamental aspect of judgments about justice.

Rueda et al. (2022) argue for procedural justice, emphasizing the importance of fair and transparent decision-making processes. Explainability plays a crucial role in achieving this, enabling verification of unbiased decisions and attributing moral responsibility.

Balagopalan et al. (2022) introduced metrics to measure the fairness of explanation models or fidelity gaps between subgroups. They argue that an explanation model can be faithful to the overall black box, but still be unfair to certain subgroups.

6.3 Mitigating Biases and Injustices Related to Sensitive Attributes

Mitigating biases related to age, race, gender, and other sensitive attributes is critical in healthcare AI. We identified various methods used by researchers, including resampling techniques (Reeves et al., 2022).

Adam et al. (2022) highlighted the importance of clinical notes in machine learning models, but also warned that these models can perpetuate biases against minorities. Notably, they demonstrated the ability to infer patient race from clinical notes even without explicit access to the attribute.

This underscores the importance of combating biases in machine learning models used in healthcare, as disparities in healthcare outcomes for minorities have been well documented, such as the finding by Lee et al. (2019) that “physicians are less likely to provide Black patients with analgesia for acute pain in the emergency room” (Lee et al., 2019 apud Adam et al., 2022).

Puyol-Antón et al. (2021) conducted an analyzing the impartiality of deep learning-based cardiac magnetic resonance imaging (MRI) segmentation models. Their work focused on the impact of gender and racial imbalance in training data and proposed three strategies to mitigate bias: stratified batch sampling, fair meta-learning for segmentation, and protected group models.

7 CONCLUSION

Our findings suggest a growing interest in the research topic, with a significant number of publications in recent years. Analyzing article titles and abstracts revealed key themes like justice, bias mitigation, interpretability, and the impact of imbalanced datasets. We identified various methods employed by researchers, including hybrid approaches, subsampling, and protected group models, to address imbalanced data and promote fairness. Additionally, explainability approaches were highlighted as crucial for achieving transparency and understanding in ML models.

Our study also emphasizes the importance of discussing and mitigating biases related to sensitive attributes like age, race, and gender. The Studies described approaches such as balancing racial proportions in datasets, examining implicit bias in clinical notes, and improving model interpretability to identify disparities across demographic groups.

This comprehensive overview of the research landscape provides valuable insights into addressing biases and injustices in healthcare ML algorithms.

REFERENCES

Adam, H., Yang, M. Y., Cato, K., Baldini, I., Senteio, C., Celi, L. A., Zeng, J., Singh, M., and Ghassemi, M. (2022). Write It Like You See It: Detectable Differences in Clinical Notes by Race Lead to Differential Model Recommendations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, pages 7–21, New York, NY, USA. Association for Computing Machinery. event-place: Oxford, United Kingdom.

Agarwal, R., Bjarnadottir, M., Rhue, L., Dugas, M., Crowley, K., Clark, J., and Gao, G. (2023). Addressing algorithmic bias and the perpetuation of health inequities: An AI bias aware framework. *Health Policy and Technology*, 12(1):100702.

Alani, A. A., Cosma, G., and Taherkhani, A. (2020). Classifying Imbalanced Multi-modal Sensor Data for Human Activity Recognition in a Smart Home using Deep Learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. ISSN: 2161-4407.

Ashok, M., Madan, R., Joha, A., and Sivarajah, U. (2022). Ethical framework for artificial intelligence and digital technologies. *International Journal of Information Management*, 62:102433.

Balagopalan, A., Zhang, H., Hamidieh, K., Hartvigsen, T., Rudzicz, F., and Ghassemi, M. (2022). The Road to Explainability is Paved with Bias: Measuring the Fairness of Explanations. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 1194–1206, New York, NY, USA. Association for Computing Machinery. event-place: Seoul, Republic of Korea.

Barton, M., Hamza, M., and Guevel, B. (2023). Racial Equity in Healthcare Machine Learning: Illustrating Bias in Models With Minimal Bias Mitigation. *Cureus*, 15(2):e35037. Place: United States.

Bear Don't Walk, O. J. t., Reyes Nieva, H., Lee, S. S.-J., and Elhadad, N. (2022). A scoping review of ethics considerations in clinical natural language processing. *JAMIA open*, 5(2):ooac039. Place: United States.

Bhanot, K., Qi, M., Erickson, J. S., Guyon, I., and Bennett, K. P. (2021). The Problem of Fairness in Synthetic Healthcare Data. *Entropy (Basel, Switzerland)*, 23(9). Place: Switzerland.

Carruthers, R., Straw, I., Ruffle, J. K., Herron, D., Nelson, A., Bzdok, D., Fernandez-Reyes, D., Rees, G., and Nachev, P. (2022). Representational ethical model calibration. *NPJ digital medicine*, 5(1):170. Place: England.

Chang, C.-H., Caruana, R., and Goldenberg, A. (2022). NODE-GAM: NEURAL GENERALIZED ADDITIVE MODEL FOR INTERPRETABLE DEEP LEARNING. In *ICLR 2022 - 10th International Conference on Learning Representations*. International Conference on Learning Representations, ICLR. Type: Conference paper.

Duffy, G., Clarke, S. L., Christensen, M., He, B., Yuan, N., Cheng, S., and Ouyang, D. (2022). Confounders mediate AI prediction of demographics in medical imaging. *NPJ digital medicine*, 5(1):188. Place: England.

El Shawi, R., Sherif, Y., Al-Mallah, M., and Sakr, S. (2019). Interpretability in HealthCare A Comparative Study of Local Machine Learning Interpretability Techniques. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 275–280. ISSN: 2372-9198.

Fan, D., Wu, Y., and Li, X. (2021). On the Fairness of Swarm Learning in Skin Lesion Classification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12969 LNCS:120 – 129. ISBN: 978-303090873-7 Publisher: Springer Science and Business Media Deutschland GmbH Type: Conference paper.

Garcia, A. C. B., Garcia, M. G. P., and Rigobon, R. (2023). Algorithmic discrimination in the credit domain: what do we know about it? *AI & SOCIETY*.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., and Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *WIREs Data Mining and Knowledge Discovery*, 9(4):e1312.

Kazim, E. and Koshiyama, A. S. (2021). A high-level overview of ai ethics. *Patterns*, 2(9):100314.

Kitchenham, B. A. and Charters, S. (2007). Guidelines for performing Systematic Literature Reviews in Software Engineering. Technical Report EBSE 2007-001, Keele University. Backup Publisher: Keele University and Durham University Joint Report.

Lee, P., Le Saux, M., Siegel, R., Goyal, M., Chen, C., Ma, Y., and Meltzer, A. C. (2019). Racial and ethnic disparities in the management of acute pain in US emergency departments: Meta-analysis and systematic review. *The American journal of emergency medicine*, 37(9):1770–1777. Place: United States.

Li, Y., Wang, H., and Luo, Y. (2022). Improving Fairness in the Prediction of Heart Failure Length of Stay and Mortality by Integrating Social Determinants of Health. *Circulation. Heart failure*, 15(11):e009473. Place: United States.

Luo, L., Xu, D., Chen, H., Wong, T.-T., and Heng, P.-A. (2022). Pseudo Bias-Balanced Learning for Debiased Chest X-Ray Classification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13438 LNCS:621 – 631. ISBN: 978-303116451-4 Publisher: Springer Science and Business Media Deutschland GmbH Type: Conference paper.

Meng, C., Trinh, L., Xu, N., Enouen, J., and Liu, Y. (2022). Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset. *Scientific reports*, 12(1):7166. Place: England.

Monlezun, D. J., Sinyavskiy, O., Peters, N., Steigner, L., Aksamit, T., Girault, M. I., Garcia, A., Gallagher, C., and Iliescu, C. (2022). Artificial Intelligence-Augmented Propensity Score, Cost Effectiveness and Computational Ethical Analysis of Cardiac Arrest and Active Cancer with Novel Mortality Predictive Score. *Medicina (Kaunas, Lithuania)*, 58(8). Place: Switzerland.

Morley, J., Machado, C. C., Burr, C., Cowls, J., Joshi, I., Taddeo, M., and Floridi, L. (2020). The ethics of ai in health care: A mapping review. *Social Science & Medicine*, 260:113172.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

Paviglianiti, A. and Pasero, E. (2020). VITAL-ECG: a debias algorithm embedded in a gender-immune device. In *2020 IEEE International Workshop on Metrology for Industry 4.0 & IoT*, pages 314–318.

Perera, A., Aleti, A., Tantithamthavorn, C., Jiarpakdee, J., Turhan, B., Kuhn, L., and Walker, K. (2022). Search-based fairness testing for regression-based machine learning systems. *Empirical Software Engineering*, 27(3). Publisher: Springer Type: Article.

Perez Alday, E. A., Rad, A. B., Reyna, M. A., Sadr, N., Gu, A., Li, Q., Dumitru, M., Xue, J., Albert, D., Sameni, R., and Clifford, G. D. (2022). Age, sex and race bias in automated arrhythmia detectors. *Journal of electrocardiology*, 74:5–9. Place: United States.

Pessach, D. and Shmueli, E. (2022). A Review on Fairness in Machine Learning. *ACM Comput. Surv.*, 55(3). Place: New York, NY, USA Publisher: Association for Computing Machinery.

Pham, T.-H., Yin, C., Mehta, L., Zhang, X., and Zhang, P. (2023). A fair and interpretable network for clinical risk prediction: a regularized multi-view multi-task learning approach. *Knowledge and information systems*, 65(4):1487–1521. Place: England.

Puyol-Antón, E., Ruijsink, B., Piechnik, S. K., Neubauer, S., Petersen, S. E., Razavi, R., and King, A. P. (2021). Fairness in Cardiac MR Image Analysis: An Investigation of Bias Due to Data Imbalance in Deep Learning Based Segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12903 LNCS:413 – 423. ISBN: 978-303087198-7 Publisher: Springer Science and Business Media Deutschland GmbH Type: Conference paper.

Reeves, M., Bhat, H. S., and Goldman-Mellor, S. (2022). Resampling to address inequities in predictive modeling of suicide deaths. *BMJ health & care informatics*, 29(1). Place: England.

Rueda, J., Rodríguez, J. D., Jounou, I. P., Hortal-Carmona, J., Ausín, T., and Rodríguez-Arias, D. (2022). ”Just” accuracy? Procedural fairness demands explainability in AI-based medical resource allocations. *AI & society*, pages 1–12. Place: Germany.

Sahoo, H. S., Ingraham, N. E., Silverman, G. M., and Sartori, J. M. (2022). Towards Fairness and Interpretability: Clinical Decision Support for Acute Coronary Syndrome. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 882–886.

Shanklin, R., Samorani, M., Harris, S., and Santoro, M. A. (2022). Ethical Redress of Racial Inequities in AI: Lessons from Decoupling Machine Learning from Optimization in Medical Appointment Scheduling. *Philosophy & technology*, 35(4):96. Place: Netherlands.

Stanley, E. A. M., Wilms, M., and Forkert, N. D. (2022). Disproportionate Subgroup Impacts and Other Challenges of Fairness in Artificial Intelligence for Medical Image Analysis. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13755 LNCS:14 – 25. ISBN: 978-303123222-0 Publisher: Springer Science and Business Media Deutschland GmbH Type: Conference paper.

Suriyakumar, V. M., Papernot, N., Goldenberg, A., and Ghassemi, M. (2021). Chasing Your Long Tails: Differentially Private Prediction in Health Care Settings. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’21, pages 723–734, New York, NY, USA. Association for Computing Machinery. event-place: Virtual Event, Canada.

Wang, Z., Liu, C., and Yao, B. (2022). Multi-Branching Neural Network for Myocardial Infarction Prediction. In *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*, pages 2118–2123. ISSN: 2161-8089.

Yang, J., Soltan, A. A. S., Eyre, D. W., Yang, Y., and Clifton, D. A. (2023). An adversarial training framework for mitigating algorithmic biases in clinical machine learning. *NPJ digital medicine*, 6(1):55. Place: England.

Zhang, H., Dullerud, N., Seyyed-Kalantari, L., Morris, Q., Joshi, S., and Ghassemi, M. (2021). An Empirical Framework for Domain Generalization in Clinical Settings. In *Proceedings of the Conference on Health, Inference, and Learning*, CHIL ’21, pages 279–290, New York, NY, USA. Association for Computing Machinery. event-place: Virtual Event, USA.