BIRMINGHAM CITY
University

Faculty of Computing, Engineering and the Built Environment

The Generalisation of Fairness: A Study of Bias Mitigated AI Models on Clinical Prediction Data Under Temporal Data Drift

Project Proposal for CMP7200

Submitted by:
Ashionye Aninze

In partial fulfilment of the requirements for the degree
of
MSc Artificial Intelligence

June 23, 2025

Abstract

With the global AI in healthcare market growing, these predictive models are increasingly trusted to inform patient care decisions; however, this integration risks amplifying existing health inequalities if models perpetuate biases found in historical data. Furthermore, model performance and fairness can degrade over time due to temporal data drift, posing a significant challenge for the development of equitable and reliable clinical systems. This research aims to investigate whether applying bias-mitigation techniques to a clinical prediction model enhances its fairness and robustness when evaluated on a chronologically distinct and more diverse patient subset.

Using the MIMIC-IV dataset, this research will split the data into historical (2008-2014) and contemporary (2015-2019) data subsets. A baseline deep learning model and a fairness-mitigated model will be trained on the historical data to perform a clinical prediction task. The performance (accuracy) and fairness (Equalised Odds) of both models will then be comparatively evaluated on the contemporary subset to assess the generalisability of the fairness intervention.

The project aims to provide empirical evidence on the relationship between fairness and robustness in the context of data drift. The findings will contribute to best practices for developing and maintaining equitable AI models suitable for deployment within large-scale healthcare systems.

Contents

List of Tables

List of Figures

1 Introduction

This chapter introduces the research project by establishing the context and motivation for the work. It will outline the problem of algorithmic bias and temporal data drift within modern clinical AI systems. Finally, it will present the specific aim and objectives that have been developed to address this challenge.

1.1 Background and Rationale

With regulators like the U.S. FDA having already approved 130 AI-enabled medical devices for clinical use by 2020, predictive models are increasingly trusted to inform patient care decisions (Benjamens et al. 2020). However, this integration risks amplifying existing health inequalities if models perpetuate biases found in historical data. Furthermore, research conducted by Almuzaini et al. (2022) demonstrates that a model's performance and fairness can degrade over time due to fairness drift, where shifts in the patient data cause a once fair model to become biased after deployment. This poses a significant challenge for the development of equitable and reliable clinical systems suitable for a dynamic healthcare environment.

A failure to address these biases has significant real-world consequences for patients. As highlighted in a paper by Aninze & Bhogal (2024), unchecked algorithmic bias can lead to systemically unfair or inaccurate predictions for specific demographic groups, potentially worsening the inequalities that modern medicine seeks to resolve. This project will provide the empirical evidence needed to inform more ethical and effective AI implementation in a real-world healthcare context.

This research is personally motivated by my upcoming role as a data scientist within the UK healthcare industry. Developing a deep, practical understanding of how to build and maintain fair and robust AI systems is, therefore, not only an academic goal but a critical element of my future professional responsibilities.

1.2 Aims and Objectives

The primary aim of this research is to investigate whether applying bias-mitigation techniques to a clinical prediction model enhances its fairness and robustness when evaluated on a chronologically distinct and more diverse patient subset.

This aim will be achieved through the completion of the following key objectives:

1. To quantify the performance and fairness degradation of a deep learning model when trained on a historical clinical dataset (MIMIC-IV A) and evaluated on a contemporary, diverse subset (MIMIC-IV B).

2. To implement a bias-mitigation technique (data re-weighting) to create a fairness-aware model and train it on the same historical data.
3. To critically evaluate whether the fairness-mitigated model demonstrates improved generalisation and robustness compared to the baseline model when both are tested on the contemporary data subset.
4. To analyse the ethical implications of the findings and provide recommendations for the validation and maintenance of equitable AI systems within real-world clinical settings.

1.3 Research Questions

To provide a clear focus for this investigation, the research will be guided by the following two primary questions:

1. **Establishing the Problem:** To what degree does a deep learning model, trained on a historical clinical dataset, exhibit performance degradation and fairness drift when evaluated on a contemporary, chronologically distinct dataset?
2. **Testing the Solution:** Does the application of a bias-mitigation technique lead to an improvement in model robustness and fairness on the contemporary dataset compared to a non-mitigated baseline?

2 Literature Review

This chapter provides a critical review of the relevant academic literature, structured into three sections to build a case for the proposed research. This review will examine the current literature across three key areas: algorithmic bias in healthcare, the challenge of temporal data drift, and established fairness mitigation techniques. By analysing foundational and contemporary work, this review will identify a critical research gap and thereby establish the necessity for this project's investigation.

2.1 Algorithmic Bias

Sufian et al. (2024) states that Algorithmic bias refers to systematic and repeatable errors that result in unfair outcomes. It displays as a change in results or performance of predictive models that may lead to unequal allocation or outcomes between subgroups. These biases arise when models learn inappropriate correlations from training data, leading to poor performance on minority

or underrepresented groups (Mazumder & Singh 2022). Algorithmic bias often stems from deficiencies in the data used to train the models and the inherent structure of the algorithms themselves.

Algorithmic bias in healthcare is a critical concern as Artificial Intelligence (AI) and machine learning (ML) models are increasingly integrated into clinical settings to improve diagnosis, treatment, and patient outcomes (Davis et al. 2025). While ML offers many benefits, these systems are susceptible to learning pre-existing biases from training data, which can lead to unfair or inaccurate predictions, heightening existing health inequalities.

In a paper by Robik et al. (2021), they state that DL models are designed to minimise loss on training datasets, but real-world data often contains invalid correlations and hidden biases. For instance, a model might incorrectly predict non-blond hair colour for blond males if non-blond males are a significantly larger group in the training data. Models trained on such biased data learn incorrect correlations and produce biased predictions, degrading their performance, especially on minority groups.

Data Imbalances

Training data frequently contains unwanted correlations or biases due to imbalances or skew involving sensitive attributes like age, gender, race, or socioeconomic status (Aninze & Bhogal 2024),(Sahiner et al. 2023), (Chen et al. 2021). For example, models trained predominantly on data from white males may underperform for women or individuals from other ethnic backgrounds. Small sample sizes for specific outcomes or subgroups can also lead to bias Meerwijk et al. (2024).

Historical Inequities and Social Determinants of Health

Healthcare disparities are often a reflection of historical and current socioeconomic inequities. Research has shown that vulnerable groups such as individuals with lower socioeconomic status, psychosocial challenges, or those belonging to immigrant communities are often under-represented in health data (Sufian et al. 2024). Models can learn and perpetuate these biases if trained on data reflecting past discriminatory practices, such as minority groups receiving less frequent or lower quality healthcare. The potential gaps in representation result in datasets that fail to capture the nuances of the desired populations, leading to delayed diagnoses, lower quality treatment, and the perpetuation of health inequities.

2.2 Temporal Data Drift

Temporal data drift in healthcare refers to the phenomenon where the statistical properties of the data used to train ML models change over time. The tendency of model accuracy to drift over time is a well-documented consequence of temporal changes in clinical practice, patient populations, and information systems Davis et al. (2025). This can be due to changes in patient populations over time (COVID-19), like immigration, an ageing population, or deployment in different clinical settings (Kore et al. 2024).

Data drift describes differences between the data used for training an ML model and the data encountered during its real-world function (Sahiner et al. 2023). In a temporal context, this means that source and target data samples originate from the same context but at different points in time (Kore et al. 2024). As the real-world data evolves, it no longer matches the stationary data the model was trained on, causing its predictions to become unreliable.

Data drift can be categorised into the following:

Covariate Shift/Input Data Drift A change in the characteristics of the input data (Sahiner et al. 2023).

Concept Drift A change in the relationship between the input data and the target variables that the ML model is trained to predict.

Label Shift A change in the distribution of the target variables themselves (Chen et al. 2021).

Fairness Drift The development of algorithmic biases post-deployment due to temporal changes in data or the model updating process (Davis et al. 2025).

Recurring Concept Drift A specific type of temporal drift where a previously observed data distribution reappears after being absent for a period Li et al. (2021).

Within a medical study, risk prediction models trained at one hospital and deployed at another might perform worse due to differences in patient populations, as seen with the Epic Sepsis Model (ESM) Sahoo et al. (2022). Patient populations may differ along both observable and unobservable attributes. Such biases can have dire consequences in healthcare applications like medical diagnosis, potentially discriminating against minority groups based on age, gender, religion, sexual orientation, ethnicity, or race.

2.3 Fairness Mitigation Techniques

Fairness mitigation techniques generally aim to reduce the impact of unwanted correlations arising from biases in training data. In a paper Wang & Singh

(2025) states that fairness methods can be categorised based on when they interlope in the machine learning pipeline: pre-processing, in-processing, and post-processing.

Pre-processing Techniques

These methods modify the input data before training to reduce existing biases, such as adjusting sample weights or balancing distributions (Almuzaini et al. 2022).

In-processing Techniques

These approaches adjust the learning algorithm during the training process itself, usually by adding fairness constraints or regularisation terms for the objective (Chen et al. 2021).

Post-processing Techniques

These methods adjust the model's predictions after training to ensure fair outcomes based on sensitive attributes (Chen et al. 2021).

There has been many techniques proposed to address bias, with some being specifically adapted for healthcare. Research by Mazumder et al. (2022) describes self-supervision and self-distillation techniques. Typically used for improving representation learning and generalisation, they have been effective in bias mitigation, particularly in limited data settings. They significantly reduce bias levels for models and can be integrated with other mitigation approaches to boost their efficacy. Adversarial debiasing is a machine learning technique used to mitigate bias in models by training two neural networks in opposition to each other. This involves an adversarial classifier that tries to predict the protected attribute, while the main model learns representations that minimise the adversary's ability to do so (Mazumder & Singh 2022).

While a wide collection of bias mitigation techniques exists, their application is often limited by the complex trade-off between achieving fairness, maintaining accuracy, and ensuring the model remains useful over time.

3 Research Methodology

This chapter details the approach taken to answer the research questions outlined in the introduction. It begins by establishing the project's foundation, justifying the research philosophy, approach, and strategy using the Research Onion framework. Following this, the chapter will describe the specific, practical phases of the experimental design, from dataset preparation to model evaluation.

3.1 Research Design and Philosophy

To provide a clear and justifiable structure for this study, the research design follows the Research Onion framework (Saunders et al. 2007). This ensures that the chosen methods are aligned with the project's underlying philosophical assumptions and research goals.

Research Philosophy: Positivism

The philosophical stance of this project is Positivism (Park et al. 2019). Positivism argues that knowledge must be derived from objective, practical observation and the testing of formal hypotheses. This philosophy is highly appropriate here because the project's goal is not to interpret subjective experiences, but to generate observable, quantifiable data (model accuracy, fairness metrics) to test a specific hypothesis about the relationship between bias mitigation and model robustness. The research will be conducted from the detached perspective of an objective observer, analysing numerical results to uncover generalisable patterns.

Research Approach: Deductive

Continuing from the positivist philosophy, this project will adopt a deductive approach (Karen 2025). A deductive approach starts with an existing theory or a specific hypothesis and then designs an experiment to test its validity. This project begins with the clear hypothesis that fairness-mitigated models will demonstrate greater robustness on new data. The entire experiment, which involves training models and evaluating them on a separate temporal subset, is

designed specifically to confirm or deny this hypothesis.

Research Strategy: Experimental

The core research strategy is a quantitative Experiment (Bolinger et al. 2021). By creating two separate conditions and applying them to two different data subsets (historical and contemporary), this experimental design serves to measure the cause-and-effect relationship between the fairness intervention (the cause) and model performance (the effect).

3.2 Research Phases

The experimental strategy will be developed in four phases, ensuring a methodical progression from data preparation to final analysis.

1. **Dataset Preparation :** The initial phase will involve acquiring and preparing the MIMIC-IV dataset. Using Python with the Pandas library, the dataset will be partitioned into two temporally distinct subsets based on hospital admission dates. A historical training set (2008-2014) and a contemporary evaluation set (2015-2019). EDA will be conducted on both subsets to understand their statistical properties, identify potential data quality issues, and document the demographic and clinical differences between them.

2. **Baseline Model Development and Bias Quantification:** In this phase, a baseline deep learning model (GRU architecture) will be developed in Python using the Keras/TensorFlow library. The unmitigated model will be trained on the historical data subset to perform a clinical prediction task, such as 30-day hospital readmission. The trained model will then be evaluated on the contemporary subset to establish its performance and to quantify the extent of fairness drift using metrics such as Equalised Odds and Demographic Parity.

3. **Bias Mitigated Model Development:** A second, bias mitigated model will be developed using the same architecture as the baseline. Fairness techniques, primarily data re-weighting, will be implemented during the

training process on the same historical data subset. This technique accommodates the importance of training samples to reduce the model's reliance on sensitive demographic features and promote fairness.

4. **Comparative Analysis and Evaluation:** The final phase will consist of a direct comparison of the baseline and bias-mitigated models. Both models will be evaluated against the contemporary data subset. A quantitative analysis will compare their performance and fairness metrics to determine if the mitigation technique resulted in a more robust and equitable model. This analysis will end in a critical discussion of the findings, the trade-offs observed, and the practical implications for developing and validating fair AI systems in real-world clinical environments.

3.3 Ethical Considerations

This project will be conducted with the highest regard for ethical principles and in full compliance with university policy.

The dataset, MIMIC-IV, contains sensitive patient health information. Although the data is de-identified, access is strictly controlled to ensure it is used responsibly. Adherence to the following rigorous data access protocol is a central component of this project's ethical strategy:

Formal Ethics Training: Before accessing the data, the researcher is required to complete the Data or Specimens Only Research course. This mandatory training ensures a thorough understanding of data privacy, security, and the ethical principles of conducting research with human-centred data.

Researcher Credentialing: The researcher must be successfully credentialed by the PhysioNet platform, a process which verifies academic affiliation and legitimacy.

Legally Binding Data Use Agreement (DUA): A DUA must be signed with PhysioNet. This agreement contractually obligates the researcher to protect the confidentiality of the data, not attempt to re-identify any individuals, and use the data solely for the academic purposes outlined in this proposal.

All data will be stored and handled per the DUA and university guidelines to ensure security and confidentiality throughout the project's lifecycle. Similarly,

the research itself is rooted in an ethical motivation: to investigate and find ways to mitigate algorithmic bias, with the ultimate goal of contributing to more equitable and fair healthcare outcomes.

4 Project Management

This section details the key planning components that will ensure the project is completed on schedule and to a high standard. It proposes a detailed project schedule in the form of a Gantt chart and an assessment of potential risks with their affiliated mitigation strategies. Following this plan is important for meeting all objectives and demonstrates the thorough planning required by the module.

4.1 Project Schedule

A detailed Gantt chart has been developed to ensure all objectives are met within the required time frame. The schedule, visualised in the Gantt chart below (Figure 1), is broken down into five distinct phases.

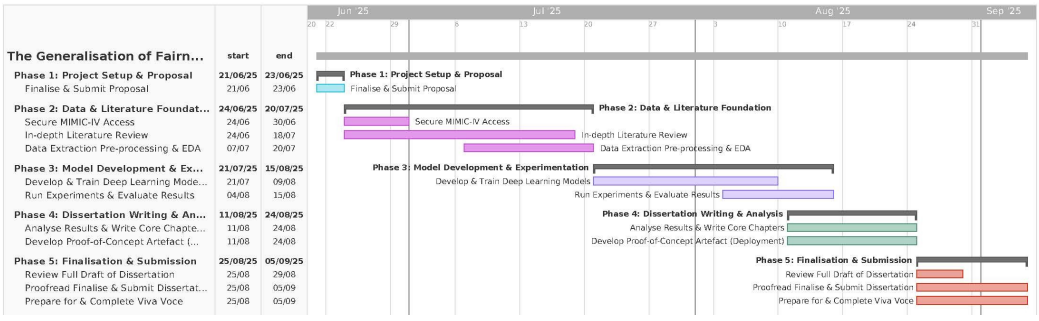

Figure 1: Project Schedule Gantt Chart

4.2 Risk and Contingency Plan

The following table outlines the risks identified for this research project, assesses their potential impact and likelihood, and details the specific contingency plans that will be implemented should these issues arise.

Table 1: Risk Assessment and Mitigation Strategies

Risk Identified	**Impact (1-5)**	**Likelihood (1-5)**	**Mitigation Strategy**
Data Quality Issues in MIMIC-IV	4	2	An extensive Exploratory Data Analysis (EDA) will be performed immediately after data extraction. If key features are unusable, imputation techniques will be tested.
Model Underperformance	5	2	The project will begin with a simpler, Gated Recurrent Unit (GRU) model to establish a benchmark. If the deep learning model fails, hyperparameters (learning rate) and architecture (Transformer models instead of GRU) will be systematically adjusted.
Insignificant or Null Results	3	3	A null result is a consequential finding. The analysis would redirect to a deep investigation into why the intervention failed, analysing feature importance and error patterns. The project's contribution would then become an evaluation of the chosen mitigation technique's limitations.

5 Conclusion

In conclusion, this proposal outlines a research project developed to address the important challenge of fairness drift in clinical AI. The primary aim of this research is to investigate whether applying bias mitigation techniques to a clinical prediction model can enhance its fairness and robustness when evaluated on a chronologically distinct and more diverse patient subset. To achieve this, the research will adopt a positivist philosophy with a deductive, experimental approach. The methodology involves preparing the MIMIC-IV dataset, which will be split into historical and contemporary data subsets. A baseline deep learning model and a bias-mitigated model will be developed. The mitigated model will incorporate bias interventions, such as adversarial debiasing. Both models will be trained on the historical data subset and evaluated on both historical and contemporary datasets to quantify performance and fairness degradation and to assess the generalisability and effectiveness of the fairness intervention. Ultimately, this project will provide crucial empirical evidence to inform the development of more robust and equitable AI systems for real-world clinical deployment.

References

Almuzaini, A. A., Bhatt, C., Pennock, D. M. & Singh, V. K. (2022), 'Abcinml: Anticipatory bias correction in machine learning applications', *2022 ACM Conference on Fairness, Accountability, and Transparency* .

Aninze, A. & Bhogal, J. (2024), 'Artificial intelligence life cycle: The detection and mitigation of bias', *International Conference on AI Research* **4**, 40–49.
URL: *https://papers.academic-conferences.org/index.php/icair/article/view/3131*

Benjamens, S., Dhunnoo, P. & Meskó, B. (2020), 'The state of artificial intelligence-based fda-approved medical devices and algorithms: an online database', *npj Digital Medicine* **3**.
URL: *https://www.nature.com/articles/s41746-020-00324-0*

Bolinger, M. T., Josefy, M. A., Stevenson, R. & Hitt, M. A. (2021), 'Experiments in strategy research: A critical review and future research opportunities', *Journal*

of Management **48**, 77–113.
URL: *https://journals.sagepub.com/doi/abs/10.1177/01492063211044416*

Chen, R. J., Chen, T. Y., Lipkova, J., Wang, J. J., , D., Lu, M. Y., Sahai, S. & Mahmood, F. (2021), 'Algorithm fairness in ai for medicine and healthcare'.
URL: *https://arxiv.org/abs/2110.00603*

Davis, S. E., Dorn, C., Park, D. J. & Matheny, M. E. (2025), 'Emerging algorithmic bias: fairness drift as the next dimension of model maintenance and sustainability', *Journal of the American Medical Informatics Association* **32**(5), 845–854.
URL: *http://dx.doi.org/10.1093/jamia/ocaf039*

Karen, S. L. (2025), 'Compare and contrast inductive and deductive research approaches.'.
URL: *https://eric.ed.gov/?id=ED542066*

Kore, A., Bavil, E. A., Subasri, V., Abdalla, M., Fine, B., Dolatabadi, E. & Abdalla, M. (2024), 'Empirical data drift detection experiments on real-world medical imaging data', *Nature Communications* **15**.
URL: *https://link.springer.com/article/10.1038/s41467-024-46142-w?fromPaywallRec=false*

Li, P., Wu, M., He, J. & Hu, X. (2021), 'Recurring drift detection and model selection-based ensemble classification for data streams with unlabeled data', *New Generation Computing* **39**, 341–376.
URL: *https://link.springer.com/article/10.1007/s00354-021-00126-2*

Mazumder, P. & Singh, P. (2022), 'Protected attribute guided representation learning for bias mitigation in limited data', *Knowledge-Based Systems* **244**, 108449.
URL: *http://dx.doi.org/10.1016/j.knosys.2022.108449*

Mazumder, P., Singh, P. & Namboodiri, V. P. (2022), 'Fair visual recognition in limited data regime using self-supervision and self-distillation', *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* pp. 3889–3897.
URL: *https://ieeexplore.ieee.org/document/9707026*

Meerwijk, E. L., McElfresh, D. C., Martins, S. & Tamang, S. R. (2024), 'Evaluating accuracy and fairness of clinical decision support algorithms when health care resources are limited', *Journal of Biomedical Informatics* **156**, 104664–104664.
URL: *https://pubmed.ncbi.nlm.nih.gov/38851413/*

Park, Y. S., Konge, L. & Artino, A. R. (2019), 'The positivism paradigm of research', *Academic Medicine* **95**, 690–694.
URL: *https://journals.lww.com/academicmedicine/fulltext/2020/05000/the*$_p ositivism_p aradigm_o f_r$

Robik, S., Kushal, K. & Christopher, K. (2021), 'Are bias mitigation techniques for deep learning effective?'.

Sahiner, B., Chen, W., Samala, R. K. & Petrick, N. (2023), 'Data drift in medical machine learning: implications and potential remedies', *British Journal of Radiology* **96**.

Sahoo, R., Lei, L. & Wager, S. (2022), 'Learning from a biased sample', *arXiv.org* .
URL: *https://arxiv.org/abs/2209.01754*

Saunders, M., Lewis, P. & Thornhill, A. (2007), *Research Methods for Business Students*, 4th edn, Pearson Education, Harlow.

Sufian, M. A., Alsadder, L., Hamzi, W., Zaman, S., Sagar, S. & Hamzi, B. (2024), 'Mitigating algorithmic bias in ai-driven cardiovascular imaging for fairer diagnostics', *Diagnostics* **14**, 2675–2675.
URL: *https://pubmed.ncbi.nlm.nih.gov/39682584/*

Wang, Y. & Singh, L. (2025), 'Impact on bias mitigation algorithms to variations in inferred sensitive attribute uncertainty', *Frontiers in Artificial Intelligence* **8**.
URL: *http://dx.doi.org/10.3389/frai.2025.1520330*