Quantitative Ethics in Healthcare Artificial Intelligence

Dr Isabel Straw
University College London
Center for Doctoral Training in AI-Enabled Healthcare

Submitted to University College London (UCL) in partial fulfilment of the requirements for the degree of Doctor of Philosophy.

Declaration

I, Isabel Straw, confirm that the work presented in my thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

The deployment of Artificial Intelligence (AI) in medicine has brought issues of health equity to the forefront. Inequitable performance of medical AI algorithms affecting different demographic groups may widen health inequalities, negatively affecting historically marginalised populations. In this research, I identify and characterise bias in healthcare algorithms. My research provides three key contributions to the domain of Machine Learning (ML) fairness and Healthcare AI. First, I provide a conceptual analysis, evaluating the roots of AI bias in healthcare, adopting an anthropological and sociological perspective. Second, I establish a quantitative framework for evaluating and addressing demographic inequities in algorithmic performance. Third, I introduce a novel application of causal modelling for evaluating bias in AI models, taking into account the nuanced challenges associated with achieving ML fairness in medicine.

This research significantly contributes to our understanding of AI bias in healthcare, by differentiating between inequities arising due to (1) unintentional harms (e.g. from a lack of representation in datasets), and (2) intentional harms (e.g. from politically shaped medical scoring systems). In taking such an approach, I demonstrate that resolving AI bias in healthcare depends on identifying and targeting the origin of the inequity. For AI Bias that stems from under-representation and the misuse of statistical averages, I evaluate the (in)applicability of traditional fairness methods and explore the role of high-dimensional representation learning for improving model individuation. Secondly, for biases stemming from harmful medical tools, I demonstrate that causal modelling can be an effective approach for uncovering and counteracting these inequities.

This study has limitations, including small datasets, missing demographic data, and a narrow focus on two medical domains, which together limits the generalisability of the results. Despite these constraints, my work highlights the need for context-specific solutions to create equitable AI systems in healthcare and the need for socio-technical methodologies that integrate an anthropological understanding of the roots of AI bias.

Impact Statement

In keeping with UCL's Research Strategy that aims to deliver research for public benefit, the outputs of my PhD are pertinent to all members of the public whose lives are increasingly shaped by algorithmic systems. Across all social sectors we are witnessing an advancing presence of Artificial Intelligence (AI) systems in decision-making processes, affecting welfare allocation, job hiring, and healthcare treatment [1–3]. In 2023 the UK launched the first AI safety summit and founded a new centre for examining and mitigating evolving AI risks. These considerations form a central component of the UK National Health Service's (NHS) long term plan, which highlights the role of AI technologies in meeting the growing demands on healthcare services and enhancing quality of patient care. Central to these initiatives is a focus on algorithmic equity and "AI Fairness", which places an emphasis on ensuring AI models benefit the whole population and do not disadvantage any group on the basis of a protected characteristic.

Furthermore, in January 2024, the UK Parliament released governmental documentation describing the policy implications of Artificial Intelligence (AI) [1]. Highlighted in this report were both the risks and benefits that AI tools pose to democracy and society more widely [1]. Given a number of high-profile cases that have demonstrated algorithmic discrimination, model fairness has become a central tenet of advancing AI in the UK. Policy documents focused on the ethics of AI are increasingly appearing on the international stage, from organisations including the United Nations, the Organisation for Economic Co-operation and Development (OECD), and the World Economic Forum [4, 5]. These vital instruments require a bed of empirical research on which to base their recommendations. The work in this thesis speaks to all of these themes, providing essential evidence on the state of AI equity in healthcare, advancing the scientific discourse in this domain.

Throughout my doctoral research I have produced research articles exposing discriminatory biases in healthcare AI models, publishing these works in leading international journals [6–8]. To further impact the field, I have acted as a peer-reviewer for diverse journals spanning AI ethics and Machine Learning (ML) in healthcare. In addition to publishing in academic journals, I have contributed to policy reports from the United Nations on the topic of AI Ethics and "AI and Gender", in which my own research has

helped shape international guidance [9, 10]. The full list of peer-reviewed publications produced in relation to this work are provided on the UCL declaration form that follows this impact statement. My efforts to improve AI equity in healthcare and the public sector more widely has also been acknowledged by several national awards, including the UK "We are Tech Women" Tech100 awards, the "Outstanding contribution to the Public sector" award from the Women in IT UK Summit, and in being featured as a "Rising Stars in AI Ethics" on the 2021 List of 100 Women in AI.

In addition, some of my most impactful activities have evolved through the creation of my non-profit company during my pHD - bleepDigital. I founded this organisation to provide educational material to the public and health professionals on the risks emerging at the intersection of advanced technologies and healthcare. After obtaining grant funding and forming a team of ten, I have delivered teaching on healthcare AI ethics at UK medical schools, developed research and policy material in collaboration with global advocacy groups, and arranged public engagement events focus on addressing digital harms. One highlight was leading our *"Tech back your bits"* event in collaboration with the London Vagina Museum, launched in London in July 2024. Our team ran an open workshop for the public with demonstrations of AI Bias, medjacking, and biotech harms, followed by an *"Ask the Experts"* evening panel with representation from industry (Google Deepmind), academia, and healthcare (the NHS). The event resulted in several articles and a public engagement award from the UCL Institute of Healthcare Engineering.

Beyond the non-profit space, I co-led an internationally-attended academic workshop during my PhD, developed in collaboration with colleagues at UCL Department of Science, Technology, Engineering and Public Policy (STEaPP). The workshop saw over fifty attendees spanning government, regulatory bodies, security and intelligence, healthcare, and academia, exploring the challenges of emerging digital technologies in patient care, resulting in a comprehensive workshop report and a first-authored publication in the Journal of Medical Internet Research (JMIR) [11]. During this time, I also co-led the development of a new educational module for UCL medical students focused on AI in healthcare, writing the curriculum content for specific sections focused on AI bias.

Lastly, I have delivered over thirty talks during my doctoral training, to national and international audiences including UK Homeland Security, the United Nations, Refuge UK, and the Royal Colleges of Emergency Medicine (RCEM) and Paediatrics and Child Health (RCPCH). In these talks, I combined my research on algorithmic bias and AI Ethics, with my additional research into tech-abuse and the cybersecurity of healthcare technologies, examining diverse means by which evolving technologies may cause harm, and the complex solutions required to safeguard patients. A few highlights include:

- **Algorithmic discrimination and AI Bias in Healthcare.** Invited Guest Speaker to the Contemporary Debates in Bioethics Series, Institute of Bioethics, University of Basel, Switzerland. November 2023.
- **Examining the risks of AI-Enhanced Harms, AI Jailbreaks & Tech-Abuse in Clinical Settings**. UK Refuge Tech Summit: Leading the Change Against Technology-Facilitated Abuse. London, UK. September 2024.
- **Evaluating bias in healthcare Artificial Intelligence (AI).** Algorithms For Her (2) Conference. Millennium Gallery, Sheffield, UK. March 2023.
- **Addressing Bias in Cardiological Artificial Intelligence: An Evaluation of Performance Disparities in Medical Machine Learning for Heart Failure.** Joint Center for Doctoral Training (CDT) in Artificial Intelligence in Healthcare, Milton Keynes. May 2022. Awarded Best Research Poster Prize
- **Digital threats to life: the post mortem evaluation of deaths mediated by technology**. Opening keynote speaker for Digital Forensics ICDDF Conference 2023, invited by UK Homeland Security. London, UK.
- **Safeguarding patients from technology-facilitated violence and abuse: International and humanitarian challenges**. One hour presentation for the Global Webinar Series, United Nations (UNFPA) Technical Division, May 2023.
- **Dont believe it when you see it: Gender, tech-abuse and deepfakes**. Opening talk at the Science Fiction Cinema. London, UK. August 2022.
- **All information should be free (except for the brain data you wanted to keep inside your head) - The Cybersecurity of Deep Brain Stimulators**. DEFCON Homecoming, Biohacking Village. USA. August 2022.
- **Artificial Intelligence and Public Health in the 21st Century.** Keynote speaker, United Against Inequities in Disease Conference, USA, April 2022.
- **Hacked devices, faulty implants and cyberattacks: Examining medical Artificial Intelligence (AI) and cybersecurity through the lens of patient care**. Whittington Hospital, London, Senior Registrar Emergency Medicine Training Day. December 2022.
- **"Bias in the blood: Investigating the equitability of machine learning algorithms built from biochemical datasets**. UCL. London UK. March 2022.

Acknowledgements

I dedicate this thesis to my Grandad, Allan, who would have read it cover to cover and loved seeing it in its final form. This work is also dedicated to my friend Tolu, who was an unwavering support throughout this period and an inspirational scholar - *it is not how long you live, but how well.* I also dedicate this work to my partner, Adrian, whose reassurance, endless encouragement, patience, and humour have made this journey not only possible, but enjoyable. I would like to thank my Dad, who nurtured my love of science from a young age. His example as an academic, alongside my Grandad, instilled in me an appreciation for learning, a constant curiosity, and a dedication to the field I love. To all my family, friends, colleagues, and mentors - thank you for your support, encouragement, and patience over these past years.

I would also like to extend my gratitude to my supervisory team, Professor Parashkev Nachev and Professor Geraint Rees, the team at the UCL Centre for Doctoral Training in AI-enabled healthcare, and the mentors who have supported me throughout my journey including Dr Leonie Tanczer and the "Gender and Tech" Team. All of your insights and support have helped shape this thesis and my growth as a researcher.

A special thanks to the patients who volunteer for the UK Biobank, and the patients who donated their data to the Indian Liver Patient Dataset and UCI Machine Learning Repository. The generous provision of this medical data, with the open-access support of these respective institutions, facilitated the development of my research. Finally, this research would not have been possible without the tireless efforts of the academics, non-profits, advocacy groups, policymakers, and members of the public working to tackle the historic inequities in medicine and the evolving biases of digital systems. Their dedication to this cause continues to inspire and drive meaningful change, laying the groundwork for research like mine.

UCL Research Paper Declaration Form: Referencing the doctoral candidate's own published works.

The following details are provided for each manuscript and referenced in the permissions below:

a) What is the title of the manuscript?
b) Please include a link to or doi for the work:
c) Where was the work published?
d) Who published the work?
e) When was the work published?
f) List the manuscript's authors in the order they appear on the publication:
g) Was the work peer reviewd?
h) Have you retained the copyright?
i) Was an earlier form of the manuscript uploaded to a preprint server (e.g. medRxiv)? If 'Yes', please give a link or doi, if 'No', please seek permission from the relevant publisher and check the box next to the below statement:

Straw, I., Rees, G. and Nachev, P., 2024. Sex-Based Performance Disparities in Machine Learning Algorithms for Cardiac Disease Prediction: Exploratory Study. ***Journal of Medical Internet Research*****, *26*, p.e46936. https://doi.org/10.2196/46936**
I published this peer-reviewed manuscript in the Journal of Medical Internet Research as the lead author during my PhD, which describes some of my key PhD outputs from Chapters 3 and 4. I developed the idea for this paper, the methods, carried out the empirical work and wrote the manuscript. An earlier form of the manuscript as uploaded to this preprint server: https://preprints.jmir.org/preprint/46936
X I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.

Straw, I. and Wu, H., 2022. Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction. ***BMJ health & care informatics*****, *29*(1). doi: 10.1136/bmjhci-2021-100457**
I published this peer-reviewed manuscript in the British Medical Journal Health and Care Informatics as I transition from my Masters, onto the PhD programme that pertains to this thesis. I developed the idea for this paper, the methods, carried out the empirical work and wrote the manuscript, of which some relates to my later work in the PhD. This work is therefore reference in Chapters 1 and 3.
X I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.

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), p.170. https://doi.org/10.1038/s41746-022-00716-4**
I co-authored this peer-reviewed manuscript in Nature Digital Medicine, with colleagues in my PhD lab, where the lead author was R Carruthers. I contributed to the ideas and discussion, which helped form the final manuscript. This work is referenced as part of the background literature of Chapter 1, 3 and 4.
X I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.

Straw, I. and Callison-Burch, C., 2020. Artificial Intelligence in mental health and the biases of language based models. PloS one, 15(12), p.e0240376. https://doi.org/10.1371/journal.pone.0240376
I published this peer-reviewed manuscript in PLOS One during my Masters programme, and reference it in my PhD as part of the background literature that is relevant in Chapter 5. For this paper, I developed the idea, designed the methods, performed the empirical experiments, and wrote the thesis.
X I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.

Straw, I., Brass, I., Mkwashi, A., Charles, I., Soares, A. and Steer, C., 2024. Insights From a Clinically Orientated Workshop on Health Care Cybersecurity and Medical Technology: Observational Study and Thematic Analysis. *Journal of Medical Internet Research*, *26*, p.e50505. https://doi.org/10.2196/50505

I published this peer-reviewed manuscript in the Journal of Medical Internet Research during my PhD, and it is referenced in the impact statement as part of my wider policy impact relating to my doctoral research. For this paper, I conceived the idea with the second author, developed the methods, and wrote the final manuscript.

X I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.

Van Niekerk D, Pérez-Ortiz M, Shawe-Taylor J, Orlič D, Siminyu K, Deisenroth M, Fasli M, Adams R, Drobnjak I, Moorosi N, Holmes W, Oliver N, Mladenic D, Eliassi-Rad T, Firth-Butterfield K, Straw I, Chair C, Aneja U, Kay J, and Siegel N. *"I don't have a gender, consciousness, or emotions. I'm just a machine learning model."* International Research Centre on Artificial Intelligence (IRCAI) under the auspices of UNESCO, United Nations. 2023. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000387189

I am a co-author on this published policy report from the United Nations, where I acted as an expert author writing on the topic of AI and Gender. The copyright belongs to the United Nations (UNESCO). For this report I was one of a team of global experts and contributed to sections of the report and final edits for the overall manuscript.

X I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.

Van Niekerk, D., Peréz-Ortiz, M., Shawe-Taylor, J., Orlic, D., Kay, J., Siegel, N., Evans, K., Moorosi, N., Eliassi-Rad, T., Tanczer, L.M. and Holmes, W., 2024. Challenging Systematic Prejudices: An Investigation into Bias Against Women and Girls. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000388971?posInSet=1&queryId=78ffd99d-3775-4c93-8971-b89292739e74

I am a co-author on this published policy report from the United Nations, where I acted as an expert author writing on the topic of AI and Gender. The copyright belongs to the United Nations (UNESCO). For this report I was one of a team of global experts and contributed to sections of the report and final edits for the overall manuscript.

X I acknowledge permission of the publisher named under 1d to include in this thesis portions of the publication named as included in 1c.

Candidate: [redacted]

Date: 01/11/2024

Supervisor/Senior Author signature (where appropriate): [redacted]

Date: 10/11/2024

Abbreviations

ACS, Acute Coronary Syndrome p.57
AI, Artificial Intelligence ..... p.1 - 191
AUC, Area Under the Curve p.24, 70, 123, 129
BCS, Breast Cancer Survivability p.32
BNP, B-type Natriuretic Peptide p.57
BP, Blood Pressure p.63, 71, 73, 95
CAD, Coronary Artery Disease p.60, 71, 85, 90
CFA, Causal-Fairness Adjusted p.170
CPK, Creatinine Phosphokinase p.62, 94
CVD, Cardiovascular Disease p.55
DSM, Diagnostic and Statistical Manual of Mental Disorders p.51
EBM, Evidence-Based Medicine p.45, 46
ECG, Electrocardiogram p.63, 71, 95
EHR, Electronic Health Record p.123
EF, Ejection Fraction p.56, 71
EO, Equal Opportunity p.36
FAGTB, Fair Adversarial Gradient Tree Boosting p.90, 96, 105-106
FNR, False Negative Rate p.37, 66, 79, 83, 97, 104, 136
FN, False Negative p.67
FPR, False Positive Rate p.37, 66, 79, 83, 91, 97, 104, 136
GAN, Generative Adversarial Network p.28
HF, Heart Failure p.56-57, 60-61, 70, 109
HFpEF, Heart Failure with Preserved Ejection Fraction p.56
HFrEF, Heart Failure with Reduced Ejection Fraction p.56
ICU, Intensive Care Unit p.33
IMD, Index of Multiple Deprivation p.21
LCS, Lung Cancer Survivability p.32
LLM, Large Language Model p.27, 124
LGBTQ+, Lesbian, Gay, Bisexual, Transgender, Queer+ p.108, 170
LR, Logistic Regression p.27, 129

Contents

List of Figures

List of Tables

Chapter 1

Background

> The so-called holy grail of medicine has always been to provide the right treatment to the right patient at the right time.
>
> *Kravitz (2014) [12]*

1.1 Introduction

Health equity has been defined as the ability for everyone to attain his or her full health potential regardless of socially-determined circumstances [13]. The focus of this thesis is dedicated to understanding the impact of evolving systems of Artificial Intelligence (AI) on realising this aim. In the adoption of AI in healthcare, ethics has become the foremost concern, in particular the question of equity [2, 14–17]. With the accelerating deployment of AI in society, those versed in issues of equity have queried how these new digital tools will affect the health disparities that already exist within our population [2, 15, 16, 18, 19]. Through this work I examine the impact of AI on the existing landscape of health equity, focusing on issues of AI bias in healthcare, and explore methodologies for evaluating and addressing demographic inequities in healthcare AI systems. To set the scene, the thesis will begin by providing an overview of health disparities, including a brief history of the health equity domain. Following on, a definition of AI will be provided, accompanied by an overview of the AI algorithms relevant to this manuscript. The methodology, aims and structure of the thesis will then be set out to frame the subsequent empirical chapters. In the closing chapter I summarise the key elements of this empirical work, and discuss both the potential positive and negative implications of AI for achieving equity in healthcare.

1.1.1 History of health equity

Anand argues that the historic attention given to the specific egalitarianism in health rests on the premise that health is a special good [20]. By this, Anand means that health

has both intrinsic and instrumental value, directly affecting a persons wellbeing and being closely tied to inequalities in the most basic freedoms and opportunities a person can enjoy [20, 21]. Hippocrates stated that "health is the greatest of human blessings" and Descartes declared health to be "the first good and the foundation of all the other goods of this life" [20, 22]. The truth that health equity underpins equity to achieving ones life potential and desires gives reason as to why it has received such consistent attention throughout time and has become a pertinent question in the adoption of AI within the healthcare domain.

In "Nicomachean Ethics", Aristotle first introduced the concept of "epikeia", which is often translated to equity [23]. In the legal context, Aristotle viewed equity as a necessary supplement to legal justice, in cases where the law falls short due to its generality. "Epikeia" ensures that justice is served in individual cases where strict adherence to the law would result in unfairness [24]. Here, Aristotle identified the key notion that extends to our understanding of health equity today. Aristotle argues that applying the same legal consequences to both a starving child and a wealthy man who steal food does not account for their differing circumstances, and thus, is not equitable [23]. His principle relates closely to health equity, where medical treatment must often be personalised to an individual's circumstances to emsure optimal care. In medicine, what may be considered "just" - the application of dogma for all - may be inequitable when applied to heterogeneous patient populations. Patients differ biologically (e.g. on the basis of sex), physiologically (across the lifespan) and medically (with co-morbidities). Thus, to treat all according to one medical template will result in varied treatment effects and inequities in healthcare outcomes.

In jurisprudence, Aristotle considered cases where the universality of dogma should not be applied as exceptions [23]. Foucault differed from him on this point [25]. Foucault viewed inequities in society and healthcare as manifestations of power relations and consequences of "biopower" [25]. Foucault's concept of biopower explored how health inequities are rooted in broader social, political, and economic structures [25]. These structures determine who has access to healthcare, whose health is prioritised, and who benefits from the implementation of healthcare policies [25]. Foucault saw biopower as an inevitable result of the "episteme" that we live in, defined as the interweaving network of assumptions about the world that condition the beliefs and propositions that are accepted as true [25, 26].

Through a foucauldian lens, health inequities are seen as inevitable outcomes of a society in which power is unevenly distributed, leading to the recurring marginalisation of specific population groups [25, 26]. In addition, Foucault's notion of power also encom-

passes authority, which is the power viewed as legitimate by both the governing, and the governed group. The power associated with authority plays a particularly pertinent role in medicine, where health professionals are seen as legitimate authorities whose decisions are trusted by patients and the population at large. In medicine, power operates through these dual channels - as both oppressive and legitimate - creating complex power dynamics that influence healthcare inequities.

Aristotle and Foucault differed in the weight they assigned to the influence of the social system on our lives, with Foucault taking a structuralist stance and Aristotle maintaining an individualistic lens. Throughout this chapter, I will explore both the structuralist and individualistic lenses that can be applied when examining disparities in healthcare outcomes.

1.1.2 Measures of health equity

Moving forward in time to the 21st century, research into health equity is an active and dynamic domain, with scholars increasingly attempting to quantify the healthcare disparities present within the population. In Micheal Marmots seminal Whitehall study, that observed civil servants for a period of 25 years follow up, the authors demonstrated a steep social gradient of all-cause mortality [27, 28]. While the extremes were not entirely surprising (richest living for longer, the poorest dying earlier), the defined step-wise change in mortality with social position indicated a more direct link between social circumstance and health than previously thought [27]. With each step up the ladder of wealth and power in the civil service, the health of the civil servant improved, and stress and morbidity fell [27]. Marmot identified the combination of high demand, high stress, and a lack of control over ones life as the formative factors contributing to deteriorating health at the lower end of the professional ladder [27, 28].

Marmot's work built on landmark documents from the 20th Century that examined health inequalities through the lens of class structure, in particular, the Black Report. The Black report examined four possible explanations of class difference in health: (1) measurement artefact; (2) natural or social selection; (3) materialist/structuralist and (4) cultural/behavioural. The structuralist framework came out on top, highlighting the impact of social factors on health outcomes. This stance emerged from empirical evidence that demonstrated a strong relationship between adverse material conditions and poorer health, which predominantly affected the lower social classes. Subsequent research has continued to support this view, reinforcing the key mediating role that social determinants play in the evolution of health inequities [27, 28].

Since these studies, the question of causation in health inequalities has been explored from a variety of angles, with scholars looking at the deleterious health effects of environmental exposures, genetic factors, health-related behaviours and psychological stress [20, 21, 27–29]. Since the original Whitehall study, the observed gradient of health inequalities in the UK appears to have worsened, with Marmots recent international comparisons demonstrating the steepness of inequality in the UK [28]. These international comparisons have raised the question of whether governmental policy is a key driver of health inequalities, for if the gradient can vary presumably as an unintended consequence of government policies, then it should be possible to vary it as an intended consequence.

In order to attain a deeper understanding of healthcare disparities, economists and public health specialists have attempted to develop metrics for quantifying these trends. In "The Health Gap" Marmot provides a range of methods for evaluating health equity across different medical applications [29]. Global comparisons of child mortality and poverty rates provide international rankings for paediatric health; measures such as the "Index of Multiple Deprivation (IMD)" can be used to explore relationships between deprivation and attainment of development milestones amongst children; and examining the Gini Coefficient in the context of intergenerational mobility can provide a measure for present inequality within a society [20, 29]. These metrics are similar in their focus on a specific outcome or indicator of disadvantage (e.g., measures of poverty, life expectancy). The launch of the Health Inequality Monitor in 2022 by the World Health Organization offers a comprehensive overview of health metrics considered to be indicative of inequality in different geographic regions [30, 31].

We can consider these methods as an outcome-orientated approach to equity, as they focus on the end-point manifestations of inequities during life that result in untimely death (e.g. inequities in life expectancy or child mortality) [30, 31]. While useful in quantifying inequalities within society, such approaches cannot ascertain the underlying causal pathways driving these inequalities. If one wishes to address inequities to inform social change, a deeper exploration of the underlying causal structures and contributing factors is necessary.

Anand et al attempted to move further up the pipeline of factors that contribute to health inequities, exploring the manner in which resource allocation results in health disparities [20]. Resource allocation may refer to the access to health services in a region, the availability of healthcare practitioners, therapeutic interventions or health-determining resources such as nutrition, education and housing [20]. From the WHO indicator list we can see that resource metrics extend to the availability of health investigations (diagnostics, testing), health promotion resources (vaccination, postnatal care coverage), and different

health treatments [30, 31]. Examining disparities in the availability of such provisions can, in part, explain contributing factors to inequities in health outcomes.

These upstream factors that influence health inequalities, are widely referred to as the **"determinants of health"** [20, 27, 29]. Comprehensively described by Gareth Williams, these key elements of our social structure have garnered attention over the past century for their impact on individual health outcomes [32]. The World Health Organisation's (WHO) "commission on the social determinants of health" provides a deep dive into this topic, and attributes the marked health inequities seen globally to the *"unequal distribution of power, income, goods and services, globally and nationally"*. The report details the impact of inequalities in access to *"health care, schools, and education, condition of work and leisure, their homes, communities, towns, or cities"* [33]. The commission focuses not only on the absence of supportive health structures, but also the unequal distribution of health-damaging experiences (e.g. exposure to crime) that may be a result of policy choices, economic arrangements, and political decision-making. As a result, the WHO definition of health equity is broad and encompasses determinants relating to different social sectors and disciplinary domains. In this thesis, I focus on the following definition of health equity taken from the World Health Organisation (WHO) [34],

> Equity is the absence of unfair, avoidable or remediable differences among groups of people, whether those groups are defined socially, economically, demographically, or geographically or by other dimensions of inequality (e.g. sex, gender, ethnicity, disability, or sexual orientation). Health equity is achieved when everyone can attain their full potential for health and well-being.
>
> *World Health Organisation (2010) [34]*

1.2 Medical Intelligence & Artificial Intelligence (AI)

The World Health Organisation (WHO) relate the health inequalities we see around the world to inequalities in the distribution of power, income, goods and services [33]. In the early 21st century the new modalities through which power is operating increasingly encompasses digital systems, equipped with varying levels of Artificial Intelligence (AI) capability [35]. Over the past decade, AI systems have permeated society, with applications across the domains of policing, healthcare, education and banking [2, 36]. The AI models deployed in these contexts now play a central role in key decision making activities, such as job hiring or individual risk profiling in the criminal justice system [2, 36]. As these algorithms become embedded in our digital infrastructure they become a conduit for power, acting as a mediator of the known determinants of health (e.g. AI

models that determine the allocation of social housing), and as a determinant in their own right as they become decision-makers in the provision of healthcare resources (e.g. AI algorithms that diagnose disease and indicate the need for care [6, 17, 18, 37]). To now bring together the topics of health equity and Artificial Intelligence (AI), I will first define what I mean by AI and the systems that this term encompasses.

Artificial intelligence (AI) refers to a constellation of technologies, that enable machines to sense, comprehend, act, and learn with human-like levels of intelligence [38]. The term AI first appeared in 1989, coined by computer scientist John McCarthy who defined it as "the science and engineering of making intelligence machines" [39]. The concept of intelligent machines had been proposed before, with pioneers such as Alan Turing publishing papers on "Computing Machinery and Intelligence" in the 1950s [40, 41]. A major shift in the AI domain occured with the development of Machine Learning (ML), which evolved as a subfield of AI in the latter half of the 20th century. Machine Learning (ML) involves the development of self-learning algorithms that derive knowledge from data in order to make predictions (explored in depth below) [42]. These models were distinct from classic algorithms that were traditionally designed by humans, who would derive rules from large volumes of data and integrate these rules into building predictive models. Machine learning developed as an alternative to human-directed learning, and this approach has increasingly outperformed traditional methods over the past fifty years [42, 43].

Before evaluating the ethical risks associated the integration of AI in healthcare, it is important to note that the challenges associated with the deployment of computational models in medicine are not new. The mainstay of modern medical practice is "Evidence Based Medicine", which refers to the process of using the latest research to inform clinical decision making [44]. When clinicians are tasked with treating a patient, their clinical management is determined by the existing knowledge base and guidelines drawn from research on other patient groups [44]. Herein lies an inherent challenge within medicine that potentiates issues of bias and inequity. Treatment choices for an *individual patient* are based on intelligence derived from studies and statistical insights derived *from groups of patients.* This approach takes findings from a research sample based on a population, and applies the derived intelligence to the context of individuals. Inferring from groups to individuals in this manner will always fail to descend to the level of the individual, precluding personalised decision-making and potentiating the perpetuation of health inequities that are marked by how well individuals resemble the original reference group [7].

In the development of Evidence Based Medicine, researchers have built and deployed a plethora of statistical models for different datasets, conditions, and contexts, in order to derive intelligence that can be applied to new patients in the future [45]. Unlike AI,

these traditional models tend to be low dimensional, simple and linear. It is this approach that informs the thresholds used for blood tests, the levels for medication prescriptions and the parameters for diagnosing disease [46]. Yet, inferring individual level treatment from group-level statistics will always fail to meet the optimal treatment for an individual patient, especially for those who were poorly represented in the original group. These issues have been outlined by researchers applying traditional statistical models to healthcare, highlighting that the rigidity of biostatistical models and their lack of individuation contribute to health inequalities [46, 47].

Therefore, the challenge of ensuring AI models do not exacerbate demographic inequalities, and are applicable across heterogeneous patient populations, is not a new issue in medical modelling but one that reflects historical issues in the field [46, 48, 49]. Researchers have proposed that the capabilities of complex AI models (particularly representational learning) may be the best step forwards for addressing these historical issues, as only by these means can we develop methods that are personalised for each patient and descend to the level of the individual [7, 47, 50]. As of yet, this promise has not been realised and we have seen AI biases emerging that predominantly affect historically marginalised groups [6, 7, 37, 51]. Before diving into these case examples, I will first review some key terminology and definitions in the field of healthcare AI.

AI is now understood to refer to the integration of statistical approaches with ML methods, to facilitate learning by the software about the data [52]. For the purpose of my research, I adopt the definition of AI that has been proposed by the United Nations, in the UNESCO Recommendation of AI ethics which defines Artificial Intelligence as:

> Systems which have the capacity to process data and information in a way that resembles intelligent behaviour, and typically includes aspects of reasoning, learning, perception, prediction, planning or control
>
> *UNESCO*
> *(United Nations, 2021 [53]*

-

There are three key domains within Machine Learning (ML) which I will briefly review, to then be explored in greater depth in the empirical chapters:

1. Supervised Machine Learning
2. Reinforcement Learning
3. Unsupervised Machine Learning

1.2.1 Supervised Machine Learning (ML) models

In the case of supervised ML models, the model is given labelled data in the training period, in order to fit a predictive model that can then make new predictions on unlabelled data inputs [42]. Classification tasks refer to those where the class labels are discrete, whereas regression tasks involve labels where the outcome signal is continuous [42]. In either case, the model is given a number of predictor variables for predicting the outcome, and attempts to find a relationship between those variables that allows for consistent future predictions. These predictor variables are commonly called the **"features"** of the model, and the outcome is referred to as the **"target variable"**. I use these terms throughout this thesis.

1.2.2 Reinforcement Machine Learning (ML) models

In Reinforcement ML models, the algorithms learn through rewards received during training [52]. The goal is to develop a system that improves its performance based on interactions with the environment [42]. Unlike supervised learning, feedback is not provided in the form of labelled data, but instead through a reward signal that measures the action of the model [42].

1.2.3 Unsupervised Machine Learning (ML) models

Unsupervised ML models differ from supervised and reinforcement learning, in their use of data of unknown structure [42]. Unsupervised ML models are not given labelled data for prediction, instead the structure of the data is explored to extract meaningful information without knowing the target outcome or reward function [42, 52]. A common example is clustering algorithms, through which a dataset may be separated into meaningful subgroups without prior knowledge of the nature of the data [7, 42, 52]. Another key area of unsupervised ML is in dimensionality reduction, which involves the compression of data from a high dimensional subspace to a small dimensional subspace while retaining the most pertinent information [42]. Representation learning emphasises that compactness is just one desirable property of a representation, and also aims to capture the essential structure of the data for future tasks [42]. Furthermore, self-supervised learning has emerged as a key approach within unsupervised learning, in which a model generates its own supervisory signal from the data to improve learning efficiency and effectiveness [42].

1.2.4 The development of ML algorithms

In order to understand how demographic disparities may be propagated through ML systems it is necessary to understand the pipeline of ML model development [43]. Throughout

this thesis I will build a range of different ML models, and thus here I summarise the existing literature regarding these models and the important steps in model development. For any ML model, there are several foundational steps that must be undertaken:

1. **Data pre-processing**: The raw data must first be processed to ensure it can be inputted into the model for training. Datasets may need to be scaled to ensure optimal performance, dimensional reduction techniques may be required for high-dimensional data and the division between the training and test data must be determined. A full review of data-preprocessing techniques are provided Raschka and Mirjalili [42].
2. **Feature Engineering and Selection:** Often datasets include a wide array of features, of which not all may be relevant to the target outcome of interest. Computational techniques that evaluate feature relevance, such as correlation scores and information gain metrics, allow the tailored selection of features for model prediction [42].
3. **Training and Model Selection:** When it comes to ML model selection, there is no "one size fits all" approach, as different models are better suited to different tasks [42]. Choosing a model, or training a series of different models for comparison, is an important step in the development pipeline. In this section it is essential to also tune the parameters of the model, as the default settings of a model may not be best for the chosen problem, referred to as "hyper-parameter optimisation" [42]. In this stage the "learning" occurs, in which the ML model summarises patterns in the training data and makes generalisations that will be applied to future instances [43].
4. **Model Evaluation:** The next step is to evaluate the model performance, for which a range of performance metrics exist including Accuracy, F SCore, ROC AUC Score and error rates (described in greater detail in Chapter 3).

In the sequence of events listed above we reference the "learning process" that occurs in the third stage of "Training and Model Selection". This is the stage that has been identified as carrying serious risks for model bias [2, 43]. The process of learning involves generalising from previous examples, in order to form predictions about future unseen data. This is a process of induction: *"drawing general rules from specific examples - rules that effectively account for past cases, but also apply to future, as yet unseen cases, too"* [43]. For the learning process to be effective the algorithm must be provided with *good* examples, with a sufficiently large and diverse number of examples that represent the heterogeneity of the relevant population [43]. Furthermore, models require careful evaluation to assess which rules are being learnt from the training data [43].

Barocas and colleagues detail this issue extensively in their review of ML fairness methods, explaining that there will always be some patterns in the training data that we want the

model to learn (e.g. smoking is associated with cancer), while other patterns may reflect stereotypes that we might wish to avoid learning (e.g. girls like pink, boys like blue) [43]. The model itself has no way of distinguishing between these two types of patterns during learning, and establishing which are a result of social norms or judgements [43]. Without intervention, harmful associations may be learnt, embedded and amplified by a model, resulting in the downstream harms of AI bias that are reviewed below.

In stage 3 the developer must also select an ML model. When selecting a ML model for a specific task (e.g. predicting breast cancer) a range of possible algorithms exist. The key models I explore in this thesis include:

1. **Logistic Regression (LR) Models** Logistic regression is a classification model, particularly suited to linearly separable classes, and one of the most widely used in the classification industry [42].
2. **Decision Tree Classifiers**: Decision trees classifiers represent decisions, and decision-making processes, through tree-like graphs, where nodes represent tests on the features, and branches represent the outcome of these tests [54].
3. **Random Forest (RF) Models**: Random forest models are a supervised ML algorithm that are formed from the construction of several decision trees [52]. Each tree within the ensemble casts a vote on the final predictive decision, the final decision is made based on the majority vote of the trees [42].
4. **Support Vector Machines (SVMs)**: Support Vector Machines (SVMs) are another supervised ML model used for both classification and regression, which finds a hyperplane that best divides a dataset into classes with the maximum margin. The model maximises the margin for separating data points to determine classes [42, 55].
5. **Neural Networks**: Neural networks are a type of ML model designed to simulate human brains, and are composed of layers of interconnected nodes (neurons), where the weight of each connection is adjusted during learning. Neural networks are a foundational component of "deep learning", due to the use of multiple layers to learn high-level features from data [42, 56].
6. **Generative Models**: Generative models are one of the rapidly evolving forms of AI, which learn the underlying distribution of a dataset in order to generate new samples that are similar to the training data. They've receive extensive media attention due to the evolution of content such as "Deepfakes" [7, 42, 57]. Key types of generative models include Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs). VAEs learn to encode data into a hidden/laten space, and then decode it back, allowing developers to generate new similar data samples. GANs make use of both a generator, and a discriminator, which work together to produce high-quality generated data [7, 42, 57].

1.2.5 Artificial Intelligence in healthcare

Now that we have reviewed the foundational concepts underpinning the AI domain, I turn to the field of healthcare and explore how these models are being implemented in the medical context. The healthcare AI field has seen significant advances in recent years, due to the enhanced computing power of modern technologies and the vast amounts of digital data available for model development. The application of AI in healthcare can be examined across the medical specialities or through the lens of the patients journey.

In their comprehensive review of the role of AI in healthcare, Secinaro and colleagues categorise AI applications by their use in population screening, patient diagnostics and clinical-decision making [58]. In patient management, AI has been used to uncover new drugs and develop personalised patient treatment plans [59]. When we consider the range of medical specialties that exist, we can see AI models emerging across these varied domains. Jiaang et al review uses of AI in the early detection, diagnosis, and treatment of stroke [60], Hogarty and colleagues explore widespread applications in dermatology [61] and Ray and colleagues provide and overview of AI models in psychiatry [62]. The superior performance of AI models for predicting disease occurrence and progression has been demonstrated in the case of heart disease, liver disease, Alzheimer's disease, Autism, Post Traumatic Stress Disorder (PTSD), psychosis, breast cancer, and lung cancer, with new applications constantly emerging [6, 18, 63–67]. A recent 2024 review from Chafari and colleagues details the latest advances in medical AI over the past five years, with new innovations appearing in genomic medicine and drug discovery [68].

While deployment of AI in healthcare is a rapidly evolving space, the notion of algorithms and predictive systems in medicine has existed for a long time, stemming from the foundations of Evidence Based Medicine reviewed in Section 1.2. Brabrand and colleagues provide an overview of the commonly used clinical scoring frameworks and "traditional algorithms" that classify patients in terms of their risk of disease, deterioration or mortality [69]. For example, the CHA2DS2-VASc score is one example of the many scoring metrics used in medicine to rank and determine care for a patient, in the case of CHA2DS2-VASc this is a risk stratification score for patients experiencing cardiac emergencies [70].

In recent years these traditional scoring metrics have come under fire due to the presence of biases within their frameworks and the resulting negative impact on marginalised groups [69, 71]. As a solution, researchers have proposed that the advanced modelling of AI, and its potential for personalising disease predictions to one individual's data, may vastly improve clinical processes that have historically relied on this inflexible scoring scales [69, 71]. In contrast however, other researchers in the field of computational fair-

ness have highlighted that key elements of AI modelling may mean these algorithms are at greater risk of generating, perpetuating and exacerbating the biases that already pervade the medical domain [2, 15]. Central to these concerns of the inequitable performance of algorithms is the issue of "AI Bias". To understand this issue in greater depth I will now explore the topic of AI bias across multiple domains, explaining how this relates to our concern of health equity.

1.3 AI Bias

Since the advancement of ML methods over the past decade, AI systems have been adopted across various disciplinary domains [2]. Yet, whilst ML methods have increased the power of computational systems, researchers have demonstrated that these improvements have not been distributed equally throughout society, with many authors exposing intrinsic biases within AI algorithms that disadvantage marginalised groups [2, 3, 15, 51, 72, 73]. This form of inequity has been described as *"epistemic inequity"*, defined as *"the inequitable distribution of the knowledge"*. The training data that feeds these AI systems are effectively knowledge about the population, that can be used to inform decisions. Thus, when they underperform for specific groups for which there was little "knowledge" in the original training data, this can be conceptualised as a form of "epistemic inequity", where by the ignorance of the model towards this population manifests in poorer algorithmic performance for these individuals [7, 50]. These concepts of epistemic equity and AI bias will be explored in greater depth in Chapter 2.

Across domains, AI bias is understood as the *differential performance of AI algorithms on the basis of a particular feature, usually a protected characteristic (e.g. race, gender)* [74]. The problem of AI bias has been reviewed extensively by Cathy O'Neil in her seminal book "Weapons of Math Destruction" [2]. Here, O'Neil describes issues of AI bias and discrimination in the criminal justice system, predictive policing models, job hiring and college ranking systems [2]. Eubanks built on this foundational work, detailing the role of AI systems in the USA in social services and child protection processes for informing decisions on family intervention and support [3]. These cases have highlighted the critical impact that AI may have on people's lives, with further research unveiling discrimination in AI algorithms used within employment processes, university admissions, loan granting and criminal risk assessment [75].

In the UK, AI applications are rapidly emerging across a wide range of public sector domains [1]. Recent reports from the UK Government have described uses of AI to assist with managing workers (e.g. allocating work and determining pay) and aiding with local decisions in social housing allocation and benefit claims [1]. If these AI models exhibit

discriminatory biases against specific demographic groups, members of these groups may see their access to fundamental resources restricted [2, 3].

The impact of AI bias on marginalised groups can compound when multiple biases affect intersecting elements of an individual's identity [15]. A central work that highlighted this issue came from Buolamwini and Gebru, who exposed intersectional biases in computer vision systems that disadvantaged women, particularly those with darker skin [15]. Their project - "Gender Shades" - demonstrated the inferior performance of AI-based facial recognition systems in identifying the faces of Black women [15]. The authors utilised the commonly used six-point Fitzpatrick classification system for determining skin type, in order to classify the individual's skin tone shade in their analysis. Interestingly, even at this early stage in the algorithmic analysis, the authors identified this traditional scoring system as a source of potential bias. This widely used classification system affords three categories to people perceived as White, reducing the heterogeneity of the rest of the world and their skin colours to three coarse categorisations - poorly accounting for the heterogeneity in tones [15]. Across the datasets used for these facial recognition systems, light males were the most represented unique subjects in all datasets, with darker females receiving the least representation [15]. The researchers' evaluation of three classifiers tasked with identifying gender from the dataset of faces, found that the algorithms had a lower performance on females, with the lowest performance on darker females and a maximum difference in error rate between the best (White males) and worst classified groups (Black females) at 34% [15].

A recent example of AI bias that caught headlines in the UK and resulted in the first mass student protest against AI, was that of the "Ofqual A Level Results Algorithm" (Figure 1.1). In August 2020, Ofqual (The Office of Qualifications and Examinations Regulation) used a decision-making algorithm to replace standardised A Level examinations for secondary school students, which had been cancelled that year due to the Covid-19 pandemic [76–79]. However, on the allocation of grades by the algorithm, 36.5% of students received a lower grade than that submitted in their teachers' predictions, and this appeared to disproportionately affect students from lower economic backgrounds [76–79]. It was found that the algorithm incorporated a school's historic results and classroom size into the grade prediction, which resulted in students at less wealthy schools being penalised with lower predicted grades by the algorithm [76–79]. As a result, the government abandoned the algorithm and returned to the grades that had previously been submitted by secondary school teachers [76–79].

Figure 1.1: Student protesters in the UK, marching in response to the A-level results algorithms that was accused of being biased against students from lower socioeconomic groups.

1.3.1 AI Bias & health inequity

The case studies described above shine a light on the risks that AI poses when integrated into a disciplinary domain that has a history of inequitable practice. Historic inequities are often represented within the data that an AI model relies on, and if not carefully teased out these inequities can be learnt, perpetuated, solidified and/or amplified by AI models [2, 3, 75].

In healthcare, there are many paths by which AI systems may affect health inequities. Firstly, the examples discussed so far illustrate how issues of AI bias may affect an individuals access to employment, education and mediate an individual's relationship with the criminal justice system [2]. Each of these elements - employment, education and exposure to the criminal justice system - are a determinant of health in themselves [20, 27, 28]. Thus, inequities in the performance of AI models within each of these domains may potentiate downstream effects on healthcare disparities. In this thesis, I will not dive into each subdomain that relates to each social determinant of health, as each could be a thesis in itself. Instead, the remainder of this thesis will focus on AI models that are built specifically for healthcare purposes. In doing so, I narrow my scope to focus on how the differential performance of AI systems used in healthcare may contribute to health inequalities, while acknowledging that this is only one area through which AI may mediate health inequalities in the population. As described by Chen and colleagues, the deployment of insufficiently fair AI systems in medicine may undermine the delivery of equitable care [80].

1.3.2 AI bias: related work

To begin, I will first review the existing research in AI Bias applied to healthcare and evaluate the gaps that exist in our understanding of how AI bias may impact health equity. Firstly, it should be noted that the term AI "bias" itself has been challenged, with researchers highlighting that "bias" may not always be a bad thing in medicine. As detailed by Cirillo and Colleagues, some biases (termed "desirable bias") may be beneficial, when they involve taking into account demographic differences in order to recommend tailored and more effective treatments for patients [81]. For example, some medical conditions are strongly associated with a demographic feature (e.g. Alzheimer's disease and age), thus here, integrating an "age bias" when considering a diagnoses for a patient may be appropriate [81]. In my research I am focusing specifically on the "undesirable bias" that Cirrillo and Colleagues identify, which refers to biases discriminating on the basis of a protected characteristic, that has no place in informing the AI's decision [81]. I now turn to a series of examples of AI discrimination within different medical domains.

AI Bias in medical specialties

A wide range of AI algorithms have been built for different purposes across the range of medical specialties, from radiology, to dermatology, to cardiology [17, 52, 63, 82]. Within these studies, a small number have focused specifically on the issue of bias in model performance. Seyyed-Kalantari and colleagues exposed bias in radiology algorithms [17]. The team examined three large radiology datasets containing chest X-rays of healthy and unhealthy individuals, and demonstrated that models were more likely to falsely predict that patients were healthy if they were members of underserved populations [17]. The authors demonstrated disparities in the model error rates across the protected attributes of patient sex, age, race, and insurance-type (as a proxy for socioeconomic status) [17]. Daneshjou and colleagues adopted a similar approach examining models trained on visual data, exposing performance disparities in state-of-the art dermatology algorithms that performed worse on darker skin tones [82]. Afrose and colleagues examined algorithmic performance disparities across a range of ML tasks developed for hospital inpatients including (i) 5-year breast cancer survivability (BCS) prediction, (ii) in-hospital mortality prediction, (iii) 5-year lung cancer survivability (LCS) prediction and (iv) decompensation prediction from the clinical benchmark [18]. The authors describe significant differences in prediction across racial groups, with the lowest algorithmic performance occurring for Black patients (compared to White and Hispanic patients) [18].

These papers largely show issues of AI Bias in diagnostic and predictive processes, where model errors lead to missed disease or missed opportunities for treatment. Further research has showcased the means by which AI bias may impact inequities in organisational

processes within healthcare [37]. A key paper from Obermeyer in the USA highlighted performance inequities in a model used to refer patients for healthcare attention in the hospital setting [37]. The study found that an algorithm used to allocate healthcare to patients was less likely to refer black people than white people who were equally sick, to programs that improved care for patients with complex medical needs [37]. The cause in this instance appeared to be the use of proxies, an issue previously well described by Cathy ONeil in Weapons of Math Destruction [2]. The model assigned risk scores on the basis of total health-care costs accrued over the course of a year, however due to many barriers and other factors, the lower access to healthcare for Black patients meant that their reduced use of resources was misinterpreted as lower need. As a result, black patients received lower risk scores even while experiencing higher levels of morbidity, and were not referred to specialist care programs [37].

In a recent systematic review of medical AI ethics, Tang and colleagues identified 36 empirical studies that focused on ethical issues of AI in healthcare, highlighting several cases of demographic bias in algorithmic performance [83]. The methodological approaches of these papers focused predominantly on differences in model errors across demographic groups. Borgese and colleagues focused on Natural Langauge Processing (NLP) models used to predict unhealthy alcohol use disorder, illustrating a model bias that result in the under-prediction of unhealthy alcohol use for Hispanic patients, compared to Non-hispanic White patients admitted for trauma [84]. Estiri et al examined the performance of a model predicting intensive care unit (ICU) mortality and 30-day psychiatric readmission with respect to race, gender, and socioeconomic status, and demonstrated higher error rates for older patients [85]. Furthermore, the models performed marginally better for female and Latinx patients, compared to male patients [85]. A study from Larrazabal and colleagues demonstrated gender bias in models built to predict respiratory disease from datasets of chest X Ray images [86]. The authors demonstrate that the existing gender imbalance in medical imaging datasets manifests in the under-performance of computer-aided diagnosic systems, particularly affecting female patients [86].

AI bias in psychiatry

In this paper I not only focus on AI bias in the medical specialities, but also on algorithms deployed in psychiatry. Thus, here I look to research focused on AI models developed for psychiatric conditions. Thompson and colleagues describe fairness issues in an opioid misuse classifier that incorporates natural language processing (NLP) techniques [87]. The authors identify the range of sources from which bias can emerge when building an ML model, from sample bias, to measurement bias to representation bias, to historical bias [87]. The team examine type II errors (i.e. false negative classifications) across groups, looking at age range, sex, and race/ethnicity, choosing to focus on False Negatives (FNs)

due to the harm that could be incurred from missing treatment [87]. In their results, the team identified a higher false negative rate (FNR) amongst the Black subgroup, compared to the White Subgroup. In my own research I have previously identified biases in Large Language Models (LLMs) with regards to psychiatric diagnosis, identifying stereotypes within language models that mirror historic biases in the psychiatry - reviewed in more detail in Chapter 5 [51].

Evaluating model bias & fairness metrics

To determine whether an AI system exhibits undesirable bias, it is necessary to have a series of tools and metrics capable of describing the performance of an AI system [42]. In the Machine Learning (ML) domain these are referred to as evaluation metrics, and include a series of well-established indicators that have been described for comparing the relative performance of different systems [42]. These metrics can be used to evaluate the performance of an AI model overall, but can also be used to distinguish the differential performance of a model for demographic subgroups within the dataset [15, 17, 18]. These metrics will be described in full in the technical background of Chapter 3, however here I provide a brief overview of these measures in Table 1.1

Table 1.1: Commonly used metrics for evaluating the performance of an AI model, defined in words (full technical details and equations provided in Chapter 3)

Evaluation Metric	**Summary**
Accuracy	The proportion of correct predictions, determined by dividing the number of correct predictions by all observations [41, 42].
Receiver operating characteristic (ROC) Score	The ROC curve plots the true-positive rate (TPR) of a model on the y-axis, and the false-positive rate (FPR) on the x-axis. The area under this curve is referred to as the AUC. The ROC score, commonly referred to as the Area Under the ROC Curve (AUC-ROC) Score, measures the area underneath the ROC curve. An ROC score of 1.0 represents a perfect classifier, while a value of 0.5 suggests a performance no better than random guessing. [41, 42]. In the case of AUC ROC=0, all predictions are incorrect.
Precision and Recall	Recall is synonymous with the True Positive Rate, thus optimising for recall minimises the chances of missing positive cases (e.g. missing disease) [42]. Precision focuses on the correctness of positives, and is important when false positives may be negatively consequential (e.g police profiling algorithms).

In the studies of medical AI bias reviewed above, a range of evaluative metrics are used in an attempt to quantify bias, many derived from the generic evaluation metrics provided in Table 1.1. These metrics have been referred to as "Fairness Notions", with fairness being closely tied to issues of AI Bias and one of the major subdomains of AI Ethics. In the ML literature, one widespread definition of fairness was proposed by Mehrabi and colleagues in their review of fairness metrics:

> The absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics.
>
> *Mehrabi (2021) [88]*

In Table 1.2 I have provided details on the range of fairness concepts and metrics that have been proposed in the computational fairness domain. A comprehensive review of these fairness notions was first described in detail in a 2018 tutorial at the Conference on Fairness, Accountability and Transparency, titled *"21 fairness definitions and their politics"* [89]. In this seminar Arvind Narayanan detailed the diversity of possible metrics that exist for evaluating the fairness of a model, exploring how each comes with it's own assumptions and challenges [89].

Firstly, Narayanan discusses the concept of "Group Fairness", which involves evaluating whether a model's outcomes systematically differ between demographic groups (as opposed to between individuals) [89]. When applied to a binary classifier, this approach examines whether the sensitive attribute is statistically independent from the prediction/-model outcome, satisfying Equation 1.1 below [43]. The foundational idea of independence between the sensitive attribute and model outcome has been explored through many related fairness notions, including demographic parity, statistical parity, group fairness and disparate impact (see Table 1.2).

Equation for statistical independence between model outcome (Y) and the sensitive attribute (A) [43]

$$\mathbb{P}\{\hat{Y} = 1 \mid A = a\} = \mathbb{P}\{\hat{Y} = 1 \mid A = b\} \tag{1.1}$$

Table 1.2: Summary of existing fairness approaches, notions and metrics used in machine learning research

Fairness Notion or Metric	**Aim of fairness metric**
Fairness through unawareness	To exclude the sensitive attribute when training the model.
Demographic or Statistical Parity	To ensure a models prediction is statistically independent from the sensitive attributes and that the decision rate (e.g. allocation of financial loan) is equal across all groups.
Conditional Statistical Parity	Requires the decision rate to be equal across groups, when conditioned on specific factors/ legitimate attributes. Thus allowing for disparities if they're explained by legitimate reasons.
Equalised Odds	Considers a models prediction in the context of the true outcome, requiring that subpopulations have equal error rates.
Conditional Use Accuracy	All groups have equal positive predictive value (PPV) and negative predictive value (NPV). The focus is on the predicted outcome, not the actual outcome (unlike EO).
Counterfactual fairness	Fairness is achieved for every individual, if the probability of being predicted positively (e.g. hired) is the same had the individual been in the other demographic group.
Equal Opportunity (EO)	Requires only True Positive Rate (TPR) to be equal across groups.
Balance for Negative Class	Ensures that the average predicted probability for negative cases is the same across groups, meaning that the likelihood of predicting negative outcomes should be consistent across groups.
Balance for Positive Class	Ensures the average predicted probability for positive cases is the same across groups.
Predictive Equality (PE)	Requires only False Positive Rate (FPR) to be equal amongst groups
Predictive Parity (PE)	Only positive predictive value (PPV) needs to be equal amongst group.
Calibration and Well Calibration	Calibration ensures that predicted probabilities corresponds to actual probabilities of the positive class, with Well Calibration being a stricter condition that holds true across all probability thresholds and groups.
No unresolved discrimination	On a causal graph there must be no directed path from the sensitive attribute A to the predictor Y, except via a resolving variable.
No proxy discrimination	Ensures that on a causal graph, there is no indirect path from the protected attribute to the predicted outcome through proxies. .
Causal Discrimination	The model should produce the same prediction for individuals who differ only in the sensitive attribute, while possessing identical other attributes.
Fairness through awareness	Ensures that predictions are fair by accounting for individual similarities based on relevant non-sensitive features, hence implying that similar individuals should have similar predictions

The concept of *Independence* is the first of the three main broad notions of observational group fairness, of which the other two are *Separation* and *Sufficiency* [90]. *Independence* is connected to Demographic parity, *Separation* is related to Equalised Odds, and *Sufficiency* connects to the concept of Predictive Parity (Table 1.2). Separation and sufficiency differ from Independence in that they focus on error rate disparities [43, 90].

Separation was proposed to overcome the limitation of independence, which may fail to address scenarios were a particular demographic group (A = a) may be more or less well represented in the strata defined by the target variable [43]. For example, a model diagnosing breast cancer would likely have a lower rate of positive predictions for males, which would be expected given the difference between the sexes in breast cancer prevalence. To account for such cases, researchers proposed the *conditional independence statement*, also referred to as *conditional demographic parity* [90]. Here, independence of the model decision from the sensitive attribute is required where individuals of differing attribute (e.g. sex), otherwise have the same rating, thus requiring Y to be independent of A given R (Equation 1.2a to 1.2b). Given that here we are looking at $\mathbb{P}\{\hat{Y} = 1 \mid Y = 1\}$, we can see that this is referring to the True Positive Rate, which also implicates the False Negative Rate (as TPR = 1 - FNR) [43]. Thus, the notion of Separation requires that all demographic groups experience the same TPR and FNR [43].

Equation for conditional independence between model outcome (Y) and the sensitive attribute (A), given score (R) [43]

$$\mathbb{P}\{\hat{Y} = 1 \mid Y = 1, A = a\} = \mathbb{P}\{\hat{Y} = 1 \mid Y = 1, A = r\}, \tag{1.2a}$$

$$\mathbb{P}\{\hat{Y} = 1 \mid Y = 0, A = a\} = \mathbb{P}\{\hat{Y} = 1 \mid Y = 0, A = r\}. \tag{1.2b}$$

Through this approach to group fairness developers have often sought to ensure the error rates of a model are consistent across demographic groups, however if the prevalence of an outcome (e.g. disease) differs significantly between these groups, establishing equivalent false positive or false negative rates may be impossible [89]. As stated by Barocas et al, *"when the propensity of positive outcomes differs between groups, an optimal predictor will generally have different error rates"* [43]. This is particularly relevant in medicine, where differences in prevalence of disease across demographic groups may render it impossible to equalise error rates across sensitive attributes.

Chouldechova's impossibility theorem explains the issue of conflicting fairness notions in depth, utilising the case of bias in recidivism prediction instruments. Chouldechova focuses on two population fairness notions from Table 1.2 - (i) Equalised odds and (ii) Predictive Parity, which are defined as follows:

1. **Equalised Odds**: For this notion to be satisfied, both the false positive rate and the true positive rate must be equal across different groups, given the same actual outcome.
2. **Predictive Parity:** To meet this criteria the positive predictive value (proportion of true positives among all positive predictions) must be equal across groups. In such, the likelihood of a positive prediction being correct should be the same for all demographic groups.

Chouldechova's impossibility theorem states that it is generally impossible for a model to both Equalised Odds and Predictive Parity simultaneously, unless either:

1. (1) The model has 100% accuracy, or
2. (2) The prevalence of the outcome is the same across all groups.

Given that these conditions are unlikely to be met for the majority of models, one must instead face the challenge of choosing one metric over the other. In choosing a fairness metric it is therefore essential to consider the domain in which it is being applied, and the limitations that a specific notion may have in certain disciplinary contexts.

Narayanan goes on to explain how specific fairness metrics differ in their importance to difference stakeholders. For example, in the case of a model predicting criminal recidivism, the defendant will be more concerned regarding false positives, where as the decision-maker may be concerned regarding false negatives [89]. Here we see that when choosing a fairness approach, we cannot rely purely on mathematical means but also require a philosophical understanding of the implications of different metrics.

The third metric focused on observational group fairness is that of *Sufficiency*. Sufficiency is based on the concept that the probability of a predicted outcome should be the same across different groups, given the actual outcome. Hence, in this approach one looks at error rates in the context of the actual outcome. The predictor $\hat{Y}$ satisfies sufficiency with respect to a sensitive attribute A and outcome Y if for all groups a in A and predictions y in $\hat{Y}$:

$$P(Y = 1 \mid \hat{Y} = y, A = a) = P(Y = 1 \mid \hat{Y} = y)$$

In the Zoo of fairness metrics Castelnovo and colleagues dive into the issue of domain-specific fairness metrics, highlighting how each of the notions presented in Table 1.2 may be suited to different real world scenarios [90]. The authors also distinguish between the two foundational challenges in fairness within machine learning [90]:

1. How to measure and assess fairness (and bias)
2. How to mitigate bias in models when necessary

The landscape of methods for addressing issues of bias and fairness is as complex as the

fairness notions themselves. In Chapter 4 I review these different methods extensively and discuss their applicability to our identified issues of bias in healthcare.

AI bias & causal fairness notions

The most commonly used fairness notions are observational and rely on examining correlations between demographic variables and a target outcome [75]. The papers discussed so far that look to identify differences in algorithm performance between groups adopt this approach, and thus despite identifying inequities, they are limited in their ability to identify causes of the underlying inequity and therefore provide solutions [75]. As described in depth by Makhlouf and colleagues, over-reliance on such metrics may fail to identify cases of bias in cases of statistical anomalies such as Simpson's or Berkson's Paradoxes, and fail to identify causal processes [75]. Increasingly, scholars are attempting to improve on these methods, moving beyond observational approaches for evaluating fairness and integrating methods from causal modelling. To understand these approaches, we will take a brief detour into the domain of causal machine learning, which is underpinned by the foundational works of Judea Pearl [91].

Judea Pearl's contributions to the field of causal modelling have provided a framework that transcends traditional statistics, in order to address issues of confounding and unearth causal effects in observational data [91]. The key components of his approach that involve the development of causal graphs and do-calculus are reviewed in depth in Chapter 5, and have provided advanced methods for researchers to disentangle issues of correlation and causation. In the domain of personalised healthcare, the advance of ML models has been leveraged to examine complex causal pathways to disease, that were previously too opaque to traditional statistical models [92, 93]. Understanding the causal factors connecting disease agents, resulting in a pathological phenotype is one of the greatest advances of medical AI [92, 93].

Through the deployment of causal models, researchers have demonstrated how these methods can unpack true relationships between variables in complex healthcare datasets [92–94]. As detailed by Jones and colleagues, if you have a set of random variables A, B, C, D that correspond to age, bladder cancer, cigarette consumption and deafness, respectively, these variables may be associated with one another through correlations [94]. However, if we want to know whether C (smoking) causes B (bladder cancer) we must establish whether intervening on the value of C changes the distribution of B. If such a change occurs, we may conclude that smoking more cigarettes may increase the risks of bladder cancer, however the reverse is not true - people who develop bladder cancer do not become smokers. Distinguishing between association and intervention is the central component of Pearl's causal hierarchy [91, 95].

Adopting a similar approach in ML fairness is a powerful means for understanding why inequities arise in model performance, which is essential for appropriately targeting interventions. For example, in the UK the A-Level results algorithm was criticised for bias after being demonstrated to under-predict academic attainment in students from state/non-fee paying schools. In this example, engaging with causal machinery is a useful mechanism for addressing the confusion that confounding variables may introduce to the algorithmic bias assessment. Following an evaluation of the models mechanism, it became apparent that the model had used classroom size as a predictor of academic attainment, which tends to correlate with socioeconomic status (i.e. students in private schools tend to have smaller classroom sizes). Thus, despite the model not incorporating socioeconomic status into its predictive process, the use of classroom size as a proxy emerged as a socioeconomic disparity in performance. In this example, understanding the causal mechanism is essential to ensure the same mistake does not occur again.

Questions of causality are of particular interest to those investigating AI accountability, for the application of causal inference methods fit well with the legal frameworks of anti-discrimination laws [96]. In the USA it is necessary in a case examining discrimination for the plaintiff to demonstrate a causal connection between the alleged discriminatory practice and the observed statistical disparity [96]. An AI algorithm deployed within healthcare may be approached in a similar manner, such that if a disparity in performance is observed, demonstrating a causal connection may be useful to support claims that the AI is engaging in biased or discriminatory practice. Plecko and Bareinboim develop a framework for causal fairness analysis grounded in legal frameworks and translated into mathematical language [96]. Within the legal domain their focus on disparate treatment and disparate impact, defined as follows:

- **Disparate treatment** = enforces the equality of treatment of different groups, prohibiting the use of the protected attributed (e.g. race) in the decision process. Within these legal formulations it is expected that a similarly situated person who is not a member of the protected class would not have suffered the same fate [43].
- **Disparate impact** = focuses on equality of outcomes between protected groups, such that discrimination is identified if a supposed neutral practice has an adverse impact on members of the protected group [43, 96]

Challenges of causal fairness in medicine

Perhaps unlike algorithms used in credit scoring or the prison system, the deployment of causal fairness approaches in medicine faces the additional challenge of unpicking the causal role that demographic features may play in the manifestation of disease. The well-established biopsychosocial model of medicine proposed by George Engel in 1977 described

the interplay between biological, psychological and socio-economic ad socio-environmental factors contributing to a patients disease [97]. Advances in many medical fields have come through understanding disease in this integrated manner, whereby the impact of social factors on brain processes have been demonstrated to influence the trajectory of disease [97]. It is well understood that factors such as income, social relationships, experiences of adversity and geography affect the likelihood of disease [97]. Yet, as detailed by McCradden and colleagues, the current methods of algorithmic fairness have not accounted for the complex causal relationships within disease pathogenesis which may involve demographic factors [14].

Ethical complications arise when we consider the fact that difference does not always entail inequity, and sometimes there may be good reason to expect differences on the basis of protected characteristics, when that characteristic impacts the occurrence of disease [14]. As stated by the authors, it may be difficult to distinguish between a computational system that acknowledges difference, and one that is propagating discrimination [14]. The aim of causal modelling is to identify which variables have a direct effect on the target variable. One of the key challenges in causal fairness research in medicine therefore, is understanding the role that a demographic feature plays on the pathway to disease. This issue will be explored in depth in Chapter 5.

1.3.3 Research areas not covered

In this chapter we have reviewed the landscape of healthcare AI, examined the constituents of AI models and the stages at which bias may emerge, paying particular attention to the period of learning in ML development, in which a model may struggle to differentiate between learning desirable patterns in training data (e.g.. smoking is associated with lung cancer) and unwanted patterns that reflect social constructed norms. The nuances of differentiating between desirable and undesirable patterns of learning are a foundational issues for addressing AI bias, and requires both an anthropological and computational lens for forming comprehensive socio-technical solutions. In Chapter 2 I dive into this challenge in greater depth, providing a framework for conceptualising the different types of AI bias and their potential respective solutions.

Furthermore, this thesis concentrates on the application of AI in healthcare and the specific issue of bias in model performance. Yet, model bias represents only one facet of the broader field of AI ethics. This study will not delve into other significant areas of AI ethics such as Explainable AI and Trustworthy AI, which has been reviewed extensively in the wider literature [98]. While these are crucial ethical issues, they are not central to our issue of model performance bias.

In addition, this thesis will not touch on the topic of AI policy and regulation, for which there is a vast and growing body of research on ethical guidelines and standard-setting instruments for AI systems [98–102]. Notable comprehensive reviews by Schiff and colleagues, and Kluge and colleagues, have documented over 200 different ethical frameworks and standards developed by governments, NGOs, industry, and academic bodies [4, 5]. Across these documents, certain ethical pillars tend to recur, including the themes of accountability, fairness, transparency, justice, non-maleficence, responsibility and privacy [4, 5, 103]. The purpose of this thesis is to focus on the question of fairness, focusing specifically on inequalities related to differential performance in medical AI algorithms. I will be examining technical solutions in model design and development, as opposed to wider questions that relate to policy decisions on model adoption, deployment and monitoring in different domains.

Lastly, the AI algorithms I will focus on will largely include those built from electronic health records, therefore comprising of predominantly structured data such as blood tests results, diagnostic scores, and clinical measurements. As a result, I neglect the domain of research examining bias in machine vision systems and linguistic models (e.g. natural language processing), that predominantly focus on unstructured data [17, 18]. While these model types will not be our primary concern, our exploration of fairness notions still applies across these diverse subdomains.

1.3.4 PhD summary and contribution

In conclusion, the research of this thesis examines the performance of healthcare AI algorithms across different demographic groups, identifying issues of algorithmic bias and their impact on health equity. The research contribution is unique in several ways.

Firstly, I combine both qualitative and quantitative methods to produce a rigorous evaluation of AI bias, beginning with a conceptual analysis evaluating the roots of inequity in healthcare. The complex issues tackled in this thesis require an interdisciplinary approach, drawing on themes from computer science, public health, and medicine. Building on the conceptual foundations laid out in Chapter 2, I apply existing methods of AI fairness to healthcare contexts that have not yet been assessed. I explore a series of remediation techniques which have not be previously deployed on healthcare AI algorithms. Finally, I examine the role of causal machinery in addressing questions of bias in medical AI systems and evaluate the unique challenges of modelling causality in cases of apparent demographic bias in healthcare AI.

PhD Objectives

- Summarise the existing research literature on health equity and AI bias, critically appraising the current knowledge base (Chapters 1 & 2)
- Develop a conceptual analysis examining the origins of bias in medicine and present a new approach for how to approach historic inequities in medical care (Chapter 2).
- Evaluate existing machine learning models used within healthcare, with the aim of exposing any biases and performance inequities present (Chapters 3, 4 & 5).
- Apply novel methods, including causal fairness approaches, to evaluate and mitigate biases present within healthcare AI models (Chapters 3, 4 & 5).

Chapter 2
The conceptual and historical landscape of health equity

> Typically, the paradigm patient or research model has been the 70 kilogram man. Traditional studies on diseases which affect both sexes have characteristically used male subjects exclusively, with the results extrapolated or generalized, as if to suggest that males are the generic humans. The problem with this male model is that information is extrapolated to women with effects ranging from incorrect to lethal.
>
> *Bess (2019) [104]*

2.1 Introduction

The primary concern of this work is AI bias that occurs at the point of model prediction, whether it is a model that predicts the diagnosis of a condition, the most effective treatment for a patient, or whether a patient will need a referral to further care services. Our focus is therefore on AI medical decision-making, which at it's core closely mirrors the existing state of play in human medical decision-making. To understand the importance of AI model fidelity, and the means by which inequity may arise from this focal point, we can draw parallels to a typical clinical encounter between a patient and a doctor.

In clinical decision-making, knowledge precedes action, as every decision in medicine relies on prior knowledge to guide treatment strategies. Clinicians consult the literature, referred to as the "evidence base", in order to make informed decisions about patient care. As we will see throughout this thesis, it is this knowledge, along with its derived intelligence, that stands as a primary resource shaping the landscape of healthcare inequities - with our without the application of AI. Ensuring equity for all patients begins with providing actionable knowledge equitably at the point of care. For every patient who enters a medical clinic, the quality of the care they receive will depend on the closeness of that clinician's knowledge, and their respective reference evidence base, to the patient's individual context.

Deciding treatment for a patient requires individualised actionable knowledge; yet, in unfamiliar clinical scenarios doctors must infer information from elsewhere to assess and manage the patient. To do this, one infers from historic group-level data and applies this to the individual, however the availability of this individual actionable knowledge varies across demographic groups. This concept of tailored care highlights the difference between "Epistemic Equality", which applies a uniform medical template to all, and "Epistemic Equity", which adapts knowledge to meet individual needs [7, 50]. The main stay of current medical practice and medical modelling focuses on Epistemic Equality, which we argue is a root source of health inequity and central to the evolving issues of AI bias [7, 50].

- **Epistemic equality** = Equal knowledge about the optimal management of an individual patient
- **Epistemic equity** = Such knowledge equally close to the possible maximum for a given individual

The nuance of the discrepancy between epistemic equality and epistemic equity underpins a central challenge in modern Evidence Based Medicine (EBM) and the health equity domain. For truly individualised care, we require group-to-individual level inferences of knowledge, whereas EBM has long focused on group-group inferences of knowledge [44]. Traditionally, EBM relies on Randomised Control Trials (RCTs) as a gold standard, which use a research sample to define (i) an average result, (ii) parameters for the group, and (iii) thresholds for wellness and disease. These parameters are then applied to the population as a whole, and patients are treated respectively to this individual mean. The utility of this model for the individual however, will be influenced by that individual's resemblance to the original reference group.

To understand issues of epistemic inequity in greater depth, it is beneficial to turn to the philosophical literature on this topic. In "L'Archéologie du savoir" (1969) Foucault discusses the episteme, defined as the means by which thought processes arise in society, with resulting patterns of knowledge, that define particular historical periods [25]. According to his works, each historical period is characterised by an interweaving network of assumptions and beliefs about the world. In Western Medicine, the modern episteme that has shaped medical knowledge was inherited from the scientific rationalism of the enlightenment period, the power structures in place at the time, and the later developments of evidence based medicine in the 21st Century. Understanding the origins of inequity in medicine requires an understanding of the roots of the modern medical episteme.

2.2 Epistemic equity & the biomedical model

The premise of EBM and predictive ML models is to look to the past, in order to make guesses about the future. The models require us to represent a population as a probability distribution, yet this is an approach that has not always been widely accepted. At its initiation, the application of statistical methods to human populations, so widespread in the modern day, was scientifically and politically contentious. One pioneer of the movement was Adolphe Quetelet, a 19th century astronomer who built a scientific programme of social physics in which he examined the presence of statistical laws in human populations [43, 46]. Here, the concept of the average man emerged, characterised by mean values of various characteristics that followed a normal distribution (e.g. height) [43]. Quetelet viewed averages as an ideal to be pursued and his work became highly influential. Unfortunately his work was also taken up by Eugencists, including Francis Galton whose theories of Eugenics were heavily influenced by Quetelet.

The creation of an average in medicine required the construction of a prototypical individual, to which others could be compared. Yet, the definition of a standard, so central to the parameterisation of normality that occurred with scientific rationalism, was initially at odds with the heterogeneous view of human existence that presided before. Idealisation of the normal has not always been a historical fact - in contrast, Vesalius in his studies of anatomy proposed that exceedingly rare variants represented the ideal we should strive towards, for example, believing that six sacral vertebrae were preferable over five [105].

The standardised model that superseded previous thought systems has been described as the natural outcome of the Enlightenment, a period concerned with order, reason and reproducibility; a technical solution to the complexity of life. Having established and classified the laws of nature, scientists gradually achieved the skills of judging what is most typical and what deviates from the commonly observed pattern; scientists developed the habit of defining normality and with it, anomaly [46]. The scientific rationalism of Descartes and the Enlightenment lay the foundations for the demographic statistical methods that followed, and the definition of a standardised body that each one of us are fitted into every time we seek out biomedical assistance [46]. In medicine, the integration of ideas of normalisation into clinical practice led to the mistaken assumption that what is statistically abnormal is always pathological, and that no pathology lies within what is statistically normal [46].

2.3 Concepts of normality & health equity

> The normal itself is an abnormality
>
> *G.K.Chesterton*
> *[106]*

To discuss an average requires a definition of what this means. The average is typically considered to be the normal of the Normal, or Gaussian, distribution [46]. Normality is an idealised theoretical model, of which height serves a good example, where the heights of most people are clustered around an average value, while particularly tall or particularly short people have growth values at opposite ends of the curve [46, 106]. In "Normality as a Biological Concept", Wachbroit further defines different meanings of Normal, one being the statistical concept and the other being an evaluative concept [107]. Normality understood as a statistical concept can be defined as an average expressed by measures of central tendency such as the mean, median or mode [46, 107]. The evaluative concept of normality considers conventional, cultural, institutional and ethical norms [107]. Yet increasingly, both the statistical and evaluative notions of Normality in healthcare have been criticised by researchers highlighting the disservice this approach does to heterogenous patient populations.

The biological absurdity of normality

The fact that anatomy is not consistent was an early conclusion drawn by the anatomists of the Renaissance [46]. Depicted clearly in anatomical illustrations of Bartolomeo Eustachi, the morphology of the vital human organs (e.g. the kidneys), glands (e.g. adrenal) and gonadal vessels vary considerably in structure [46, 105]. In a comprehensive review of anatomical variation and diversity, Zytkowski and colleagues explored the limitations of concepts of bodily normality. Beyond these examples, medical researchers have argued against normality as a biological function [46, 106, 107]. As laid out by Chadwick, the extent of difference among individual members of a species, in both the animal and human population, is integral to species survival [106]. Thus, in biology the scientific concept of interest is variation, rather than normality, as variation is the force that allows for species adaptation, evolution and survival [106]. With diversity being a matter of our species survival, to force ourselves into theoretical conformity is to move against our very nature [46, 106]. Amundson goes so far to suggest that the concept of normality is a biological error, since "diversity of function is a fact of biology" [105].

Furthermore, as stated by Wachbroit, the limitation of statistical normality is that *"What is statistically norm may vary with changes in the population"*. Variation can be advan-

tageous; however it may be seen as a deficit when framed through the lens of normality. In a classic study of adolescent populations living on different diets, Ryle et al found considerable variation in the size of the individuals thyroid glands. Rye argues that this symptom may represent a normal adaptation to a specific environment and should not be interpreted as a meaningful clinical sign [46].

In "When Normal Does not Exist", Lock and Nguyen highlight that there is no way to define a biological norm or deviations from it without reference to specific populations and their sociocultural characteristics [46]. Chadwick stated that "in the medical literature an ideal human model is a 70 kg male with 32 teeth, no mental disorders, and a clean genetic slate" [46, 106]. Thus, the concept of Normal can become a powerful position, elevating a specific population above others. These conclusions echo the arguments made by Foucault in 19th Century, who argued that medicine was regulated more in accordance with normality than with health [25, 46].

The statistical limitations of normality

The approaches of biomedicine are anchored in an idea of a universal somatic body where health and illness are conceived as opposite poles along a biological continuum [46, 106]. In biostatistics, the "normal distribution" refers to a common pattern of variation around an average, otherwise known as "the Bell Curve" [46]. Since the Bell Curve was first proposed it has been the subject of contentious debate. Eugenia Cheng criticises its simplistic approach to populations in her appraisal of statistical measures used to evaluate gender-based differences in educational outcomes [108]. To quote Cheng, *"Averages are a way of condensing a collection of numbers to just one number, a process which sacrifices a huge amount of information and nuance in the process"* [108].

Our reliance on the Normal Distribution means that we take a single average and extrapolate it to a population without taking into account intersecting relationships and the complexity of high dimensional interactions [108]. One of the key limitations of traditional statistical models is their lack of flexibility and inability to account for intersecting factors that contribute to outcomes such as health and disease, or as relevant to our focus, experiences of marginalisation. In the sociological literature this is referred to as "intersectionality" [109, 110]. Introduced by Crenshaw in 1989, the term "intersectionality" was coined to describe how different forms of oppression such as racism, classism and sexism interact to create unique experiences for individuals who belong to multiple marginalised groups [109, 110]. These ideas were further advanced by the writings of Audre Lorde, in key works such as "The Master's Tools Will Never Dismantle the Master's House", which emphasised the importance of understanding the multiplicity of oppressions faced by Black and Queer women [109, 110].

Our experiences of health and disease, and the factors that contribute to marginalisation and health inequity, are ultimately multifaceted and require an intersectionalist approach. Yet, the reductionism present in traditional statistical models, and some AI models, are limited in the degree of intersectionality that they can capture. Here, the role of AI and representational learning is key for understanding and addressing issues of health (in)equity, and will be explored in greater detail in the following quantitative chapters.

The historical construction of normality

Silvia Federici offers an additional perspective on the roots of bodily normalisation, arguing that the standardisation of the body was not driven by philosophical thought related to the Enlightenment, but instead was driven by an economic intention to mechanise the body starting in the 1300s - from which the philosophy of scientific rationalism later emerged [111]. Federici contends that the mechanisation of the body was a response to the economic crises following the Black Death, with the need to replenish the labour force acting as a primary motivator, rather than abstract scientific rationalism [111]. Federici argues that after the Black Death (1347-1350 AD), the aristocratic and landlord class became concerned with their loss of feudal workers to disease and their diminishing labour force [111]. The following century saw peasant revolts and wars across Europe, with a power struggle between the feudal and landlord class, in which the feudal class resisted news laws of land enclosure, private capital, the eradication of the commons and the money-wage [111].

According to Federici, the capitalist economic system's foundations were laid during this time, transforming the peasant into a worker whose body was increasingly viewed as a machine for labor and procreation [112]. Figure 2.1 from Federicis research illustrates this shift in thought as the population came to view humans as a machine, and a tool for production and procreation over the following centuries.

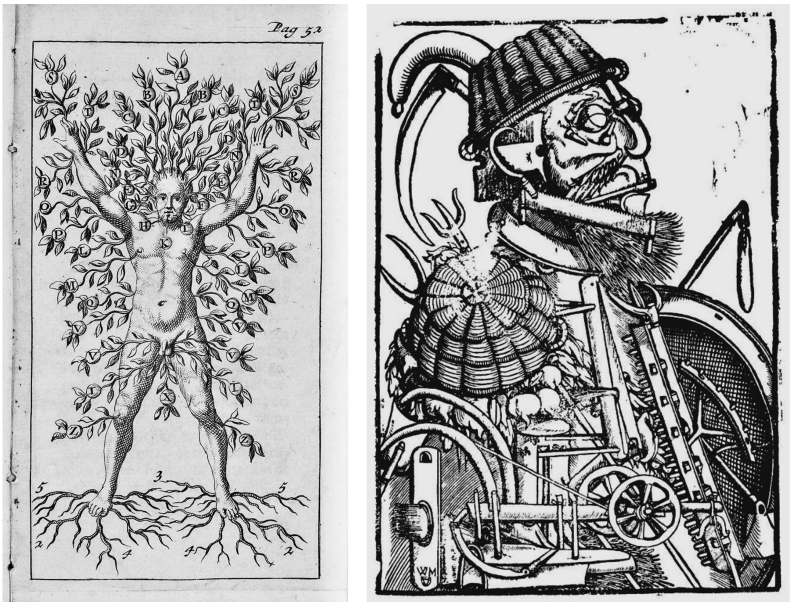


Figure 2.1: **Right**: An illustration of a nineteenth century worker by Henry Kamen (1972), taken from Silvia Federici's book "Caliban and the Witch: Women, the Body and Primitive Accumulation" to illustrate the new mechanical conception of the body where the peasant is represented as nothing more than means of production. **Left**: J. Case. *Compendium Anatomicum* (1696). In contrast to the "mechanical man" is this image of the "vegetable man" in which the blood vessels are seen as twigs growing out of the human body, capturing the holistic view of personhood that was superseeded by the mechanical approach.

Federici discusses the mechanisation and standardisation of the human body as a crucial element in the transition to a capitalist society [111]. The mechanistic view of the body, supported the emerging capitalist need for a disciplined and controlled workforce, transformed individual bodies into labour power that could be commodified and managed [111]. Federici's work provides further insights as to why lines of inequity emerged from the production of the standardised body, that go beyond statements of representation. Commonly, issues of inequity related to normality and averages are attributed to the neglect of marginalised patient groups in the original sample, yet Federici provides an additional perspective on how these harms emerged [46].

Federici gives the history of the societal members who were considered "unhelpful" in the reproduction of a labour force following the Black Death, namely women who chose not to reproduce (often in senior social positions), those in same-sex relationships not raising children, and elderly women [111]. These individuals were at odds with the capitalist imperative of increased procreation after the Black Death, and as a result were targeted and oppressed through the outlawing of homosexuality; the genocidal "witch trials", and the banning of female reproductive choice in the 1300s-1700s AD [111]. The capitalistic society that evolved during the era following The Black Death required both (i) the mechanisation of the human body as a force for labour, and (ii) the eradication of societal members who did not contribute to the reproduction of the labour force [111]. Thus, Federici draws together the mechanisation of the body, and the standardisation and parameterisation of normality, with the exclusion of specific demographic groups from power [111].

In the sections above we have reviewed both the philosophical history of Normality and the economic undercurrents of this domain. One on hand, researchers argue that the development of the "Normal body" related to the intellectual breakthroughs of the enlightenment and scientific rationalism, while on the other, authors argue that these were intentionally elevated philosophies that served the wider economic imperative of mass labour and capitalistic growth at the time [46, 105, 111]. Either way, the mechanisation of both society and the individual resulted in a rigid mechanism that attempted to view all individuals as an extrapolation from a defined ideal.

The political applications of normality

We are now 700 years from where Federici based her exploration and the power shifts that have occurred across the last seven centuries have further reshaped the defined "Normal body" within society, including the rise, expansion, and escape from the slave trade, Western colonialism, and indigenous genocide in the global colonies. Throughout these historical periods, Western medicine has been used as a vehicle for political ideologies

and has perpetuated discriminatory ideas of scientific racism, European eugenics, female inferiority and racial purity [46, 48, 49, 113]. The combination of politicised ideologies that seek to create hierarchies between demographic groups, combined with statistical techniques in which one group or sample is defined as the ideal or "Normal", laid a context in which a medical practice emerged that better served those in positions of social power at the time [46, 48, 49, 113].

Furthermore, science has long been a conduit for asserting political power with the use of "expert opinions" and "evidence" being exploited for political means. For example, homosexuality was only removed from the diagnostic manual of psychiatric diagnoses in 1973, with the history of medical mistreatment of the LGTBQ+ community being well documented [114]. Similarly, feminist researchers have pulled apart diagnostics frameworks in the Diagnostic and Statistical Manual of Mental Disorders (DSM) which have been demonstrated to target and oppress women [115]. In particular, the definition and application of the diagnostic framework for "hysteria" was often weaponized throughout the 1900s to silence women who challenged sexist structures in society [116, 117].

In addition, the medical sciences has a history of mistreating different racial groups to advance the interests of those in power, through the construction of bogus scientific theories that maintain structures of oppression and white supremacy [46, 49, 118]. For instance, "Protest psychosis", as defined by Bromberg and Simon, was a constructed condition that viewed Black male participation in the civil rights movement, as a contributing factor for aggressive and volatile schizophrenic symptoms in the Black community [118]. In addition, doctors have historically advocated for the use of flawed techniques such as phrenology in to classify the "likely criminality" of predominantly Black men during the 1900s [119, 120]. The experimentation and exploitation of racial groups in the medical arena has received increased attention in the bioethics literature following a series of exposing reports including (i) The Tuskegee Trials, (ii) The harm of Puerto Rican women in contraception development, and (iii) The Nazi experimentation on predominantly Jewish people and other oppressed minorities [46, 121].

The legacy of racist medical science and the intentional targetting of specific societal groups still impacts clinical guidelines today. In "Breathing Race into the Machine", Braun exposes how the incorrect application of physiological and anatomical knowledge in respiratory medicine has affected racial disparities in asthma care [122]. In this review of the roots of lung function tests, Braun reports how racist science stemming from slavery plantations manifests in modern management guidelines for paediatric asthmatic patients [122]. Further, the use of race as a clinical marker in medical modelling is under review across a wide range of domains, with additional criticisms being made of the use

of race in establishing renal function. The race adjustment to eGFR (a clinical measure of kidney function), has been reported to lead to falsely high reports of eGFR amongst black patients, resulting in delayed referral and treatment [49]. Angela Saini's comprehensive critique of racism in medicine identifies the current NHS guidelines that harmfully allocate different blood pressure medications based on confused notions of race, disadvantaging patients of colour [49]. In the later chapters I explore this challenge in greater depth, comparing the works of researchers who advocate for the inclusion of demographic features such as Race (to ensure appropriate care), with other researchers who highlight the harm that emerges when we use outdated ideas of features such as Race, that have been incorrectly represented in the scientific literature [49, 122].

Normality & demographic harms

Aside from the harmful impact of these intentional abuses of medical power, extensive research has also demonstrated the perhaps more unintentional negative impact that the "Normal Body" has on marginalised groups, due to issues of representation. Elinor Cleghorn sheds light on this in "Unwell Women", systematically reviewing the neglect of female physiology which has continued from ancient Greece to modern day [48]. The author highlights that frequently in medicine, female patients are treated according to male parameters, and it is therefore unsurprising that female illness often falls within the defined normal range and remains unacknowledged [46, 48, 123]. In fact, women are up to 75 per cent more likely to experience adverse reactions to prescription drugs compared to men, largely attributed to their exclusion from medical trials and the inattention paid to their physiological responses [124, 125]. The issue extends to all clinical trials, not just human ones, with Karp and Reavey describing the history of sex bias in pre-clinical trials in which only male animals (e.g. male mice) have been used [125].

In *"Eve: How the Female Body Drove 200 Million Years of Human Evolution"*, Cat Bohannon has examined this phenomena through an anthropological lens, expanding her scope to encompass the neglect of female bodies across all animal species and mainstream biological research. Bohannon details in depth how the scientific focus on male bodies and behaviour has limited the fields of evolutionary biology and human anthropology, neglecting significant pathways that allowed the development of modern civilisation [126]. Lucy Cooke in "Bitch" goes further to examine species beyond the mammalian domain, detailing how a great deal of the foundations of biological determinism and our "understanding" of biological sex differences, stem from science developed through the male gaze, shaped by Victorian ideologies [127]. Cooke explains how Darwin himself with subject to advancing his ideas within a context defined by fixed Victorian ideals, thus a feminist approach that centered species in which females were the aggressor were unlikely to be adopted [127].

Susan Whites illustration of the *"Sex Change of the Vitruvian Man"* brings this issue back to medicine (Figure 2.2). Da Vincis world-famous Virtuvian Man that has shaped the development of anatomical texts in medicine, an image that was built purely based on male physiology and body geometry. Susan White drew the Virtuvian Woman to shine a light on the neglect of female physiology and anatomy within society (Figure 2.2). The androcentric nature of anatomical teachings manifests in deleterious health outcomes for women, such as the higher rates of hip transplant failures amongst female patients [128]. Upstream inequities in anatomical knowledge translate to downstream inequities in health outcomes. The results of treating male anatomy as the normal extends beyond the field of medicine, for example, Caroline Perez has demonstrated that women suffer from a higher risk of death in motor vehicle accidents due to the use of male models in crash testing [113].

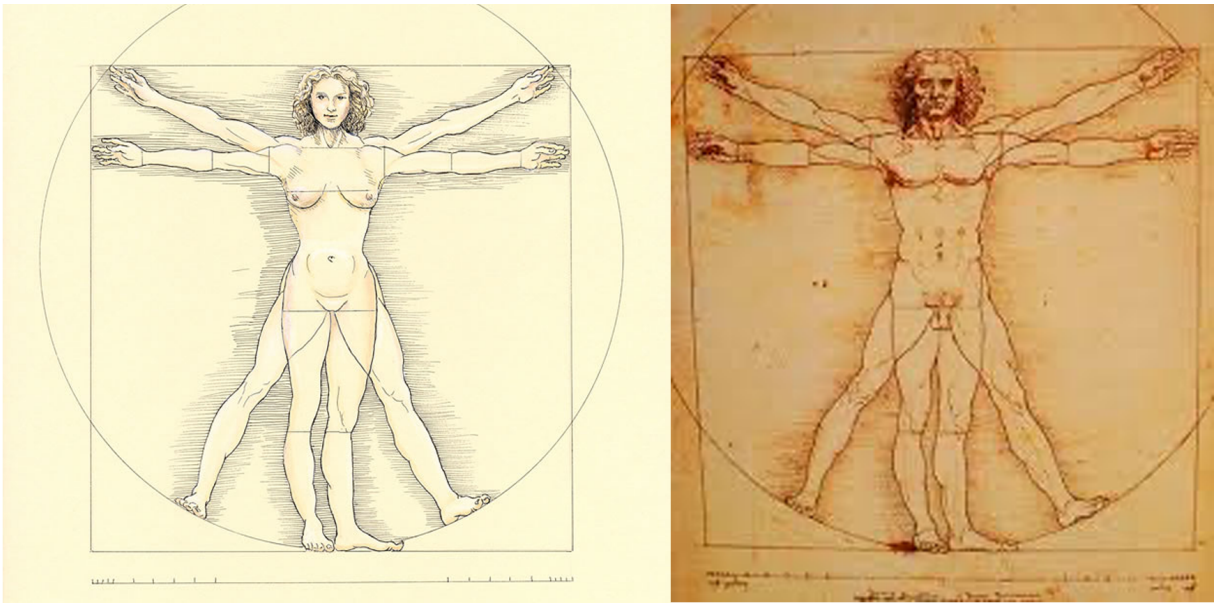


Figure 2.2: Left: Susan Whites illustration of the Sex Change of the Vitruvian Man, Right: A copy of The Vitruvian Man drawn by Leonardo Da Vinci in 1490

2.4 Conclusion

In this Chapter I have explored how the concept of "normality" in the medical sciences has historically been constructed, its relationship to epistemic inequity, and how this acts a source of health inequity in the population. I have examined how the creation of a "normal" body may result in unintentional harms, due to the exclusion of certain groups from the reference sample, and also how concepts of normality and pathology have been historically weaponised to elevate certain demographic groups, resulting in a legacy of intentional harms.

In the history of modern medicine, the emphasis on normalisation and parametrisation, with respect to an average, created a domain foundation that has failed to serve specific patient groups leading to persistent health inequalities in the population. If healthcare AI is built without an awareness of these fundamental flaws, there is a risk that these historic inequities may be perpetuated in new digital models. In contrast, the advanced computational functions of AI models may instead offer an opportunity for creating more complex representations of patient populations, addressing these issues of epistemic uncertainty that are common to traditional low-dimensional models. Taking this approach could be advantageous for minoritised group, accounting for errors of the past, and is a key theme explored in the chapters below.

To ensure medical AI models do not perpetrate historic discrimination that has been enacted by the medical domain, researchers must be acutely aware of the anthropological history of bogus scientific theories and diagnostic frameworks that were built to maintain power structures. Both these legacies of unintentional and intentional harms may manifest in emerging healthcare AI models if they are not tackled at source during model development. For the purpose of this research, I use these two arms that underpin the history of healthcare biases as a framework for designing my machine learning experiments: firstly, evaluating for issues related to lack of representation, and secondly examining harms that stem from biased medical tools and frameworks. In the following chapters, I explore both phenomena, evaluating how these two anthropological roots of health inequity may manifest in emerging AI systems.

Chapter 3

Exposing bias in cardiac algorithms

> Cardiovascular disease (CVD) has historically been perceived as a male disease; however, it is the leading cause of mortality and morbidity worldwide for both men and women. Despite this, CVD is understudied, underdiagnosed, and undertreated in women.
>
> *Gauci et al (2022) [129]*

3.1 Introduction

The focus of this chapter is on the identification of AI bias in healthcare algorithms, looking specifically at algorithms built in to predict heart disease. Heart disease is of particular interest from a health equity standpoint, due to the history of demonstrated demographic differences in disease presentation and the treatment biases that exist within the domain [123, 130–137]. In 2022 the British Heart Foundation released a report titled "*Bias and biology: The heart attack gender gap*", in which they detailed the intersecting factors that result in female patients receiving worse care then their male counterparts [133]. The findings of the report revealed that for cardiac diseases (i) women are more likely to be misdiagnosed or to experience delays to diagnosis, (ii) they are less likely to receive optimal treatment, and (iii) they are more likely to experience difficulties accessing cardiac rehabilitation [133]. When it comes to heart attacks, women are more often misdiagnosed as having panic attacks, leading to their clinical presentations being dismissed or incorrectly managed [133]. Women from Black, Asian and Ethnic Minority backgrounds, plus those who live in socioeconomic deprivation, are at greater risk of suffering from these biases and receiving sub-optimal care [133]. African-American patients appear to be at a particularly heightened risk of unequal care, with research demonstrating that they are less likely than White patients to receive necessary diagnostic tests and revascularisation treatments, even after controlling for other clinical and social factors [137].

The report from the British Heart Foundation adds to the volume of academic research that describes overlooked sex-based differences in cardiac disease and cardiac care, whereby persistent male-centric conceptions of heart disease disadvantage female patients. Gauci and colleagues highlight how cardiovascular disease has historically been perceived as a

"male disease", despite the condition also being a leading cause of death for women worldwide [129]. Such perceptions have been linked to the under-treatment of women in cardiac care, exemplified by greater delays in emergency response times to cardiac crises, delays in diagnosis and longer wait times for accessing key treatments [129, 138–140]. One form of heart diseases that has garnered increasing attention for it's disproportionate impact on women is Heart Failure.

Heart failure is a clinical syndrome in which the heart is unable to pump enough blood to the body to adequately to meet metabolic demands [132]. Heart failure affects more than 64 million people globally and the prevalence is increasing due to the ageing population [132]. In the UK cardiovascular fatality figures are on the rise for the first time in 50 years [63].

The medical community groups Heart Failure (HF) into two types depending on the patients ejection fraction (EF) value, which is the proportion of blood pumped out of the heart during a single contraction, given as a percentage [63]. "Heart failure due to reduced ejection fraction (HFrEF)", previously known as "Heart failure due to Left Ventricular (LV) Systolic Dysfunction" or "Systolic Heart Failure", is the first sub-type which is characterised by an ejection fraction smaller than 40% [63]. In the second HF subtype - "Heart Failure with Preserved Ejection Fraction" (HFpEF), formerly called "Diastolic Heart Failure" or "Heart Failure with Normal Ejection Fraction" - the heart's ability to contract is maintained, however the advancing stiffness of the heart wall prevents heart relaxation in diastole, precluding filling [63].

Historically, the severity of HF disease was measured using the New York Heart Association (NYHA) functional classification, which comprises four classes ranging from Class 1 (with the patient experiencing minimal symptoms) to Class 4 (with the patient experience symptoms at rest) [63, 141]. In recent years the NYHA classification system has been criticised, with researchers and clinicians highlighting that the system fails to predict basic markers of disease progression [63, 141]. Furthermore, scholars in the field of health equity have exposed the differences that exist between a patients subjective experience of their HF disease, and the grade assigned by the clinician, highlighting risks of clinician bias in allocating these scores [142]. In response, there has been a renewed interest in computational methods for predicting HF disease progression and death, with researchers exploring the use of ML for cardiac modelling [63, 141, 143].

In their review of sex differences in heart hailure, Sobhani and colleagues highlight that since 1984, each year, more women than men die of heart failure despite more men being diagnosed [134]. The reasons appear to be multifaceted. Heart Failure is now known

to present differently in male patients compared to female patients, with each group following a different trajectory [130]. Females often experience increased symptoms of fluid overload, often presenting with a greater overall burden of symptoms and lower quality of life [131]. Further, when females present they tend to be older on average and sustain a higher Ejection Fraction (EF) towards the end of their life [131]. In addition, when women develop cardiac complications, such as acute coronary syndrome (ACS), they more often present with "atypical" symptoms, such as jaw pain and nausea compared to men [134, 136]. One could argue however that these symptoms are only considered "atypical" due to the male-centric model of cardiology that dominates medical education [104, 123, 129]. Furthermore, amongst females, Black patients appear to experience a higher rate of "atypical symptoms", resulting in a higher risk of misdiagnosis and delayed care [135].

The symptomatic differences between male and female patients have been attributed to differences in the underlying pathological changes in the heart tissue, with females having higher systolic and diastolic left ventricular (LV) stiffness compared to males [130]. Unsurprisingly therefore, the efficacy of various cardiac treatments also appears to differ between the two groups, with therapeutic options of RAAS (renin-angiotensin-aldosterone system) inhibitors and beta-blockers offering differential effects to the sexes.

The sex differences observed in the response rates to available treatments have been attributed to historic inequities in cardiovascular research [104, 123, 129]. Sullivan and colleagues describe the impact of the under-representation of females as participants in HF randomised control trials (RCTs) on evauating sex-specific efficacy and safety of treatments [130]. The current "sex-neutral" HF guidelines that are used in the NHS are largely based on data reflecting experiments on males [130]. The problem extends to the biochemical testing that is used to evaluate heart disease, such that cardiac blood tests have been demonstrated to underperform for female patients [134]. In particular, the use of "unisex" troponin reference intervals for detecting death of the cardiac tissue, have been criticised for resulting in an under-diagnosis of heart attacks in women [134]. Troponin is one example of the biomarkers drawn from patient blood tests that are used to predict disease occurence and disease outcome. Previous research has reported that standard troponin criteria fail to detect one out of five acute myocardial infarcts occurring in females [134].

Biomarkers such as troponin provide useful information for predicting disease progression, and have have received attention from the ML community as a useful form of structured data that may be easily interpreted by a ML model. Further cardiac biomarkers being explored within this research include Creatinine Kinase, C-reactive protein and B-type natriuretic peptide (BNP) [134]. ML models built from feature sets composed

mainly of biomarkers have been developed for evaluating other clinical conditions such as Alzheimer's disease, kidney injury, and predicting long term outcomes for diabetes patients. [144–146]. Researchers are increasingly applying these methods to the field of cardiology, demonstrating that ML models built from datasets of cardiac biomarkers can outperform traditional statistical models and clinical risk scores for predicting disease progression in cardiac care [63, 147–149].

While datasets of cardiac biomarkers may provide useful clinical information, our research so far has demonstrated that these data sources may under-represent marginalised patient groups, and hence any reliance on these datasets in model development may result in algorithmic biases that perpetuate historic inequities in care. Thus, for my initial exploration into AI bias in healthcare, I chose to focus on algorithms built from medical data, particularly those incorporating biochemical markers, for predicting heart disease. I have extrapolated computational fairness techniques from other domains, and applied these to the cardiology context to assess for similar issues of AI bias [15]. In doing so, I draw on seminal studies, such as that from Buolamwini and Gebru, in which their evaluation of performance disparities in facial recognition algorithms demonstrated a significantly higher error rate for Black Women [15]. My methods replicate these key works, however I adjust the approach to ensure it's appropriateness to the medical context.

Chapter Research Aim

In this Chapter, I have begun with an exploration into the field of AI bias in healthcare, focusing on the field of cardiology due to the known inequities that exist in current medical practice. I have examined the existing research in the cardiac ML domain, summarising the challenges that relate to issues of health equity. Moving forwards, this chapter will investigate existing ML studies for the potential of AI bias, from which I develop a quantitative approach for evaluating algorithmic performance inequities in medical ML models.

3.2 Methods

My analysis consists of two stages, (1) a literature review of papers describing ML models used to predict heart failure, and (2) a quantitative analysis of identified models, evaluating inequities in algorithm performance. The flowchart in Figure 3.1 provides an overview of this approach. I began by scoping the published literature for studies that utilised ML methods for predicting Heart Failure, and collected these articles for my own review (see Stage 1). In Stage 2, I utilised the open-source datasets uncovered during the systematic

review to rebuild the models reported in the cardiac ML domain and interrogate these models for performance bias.

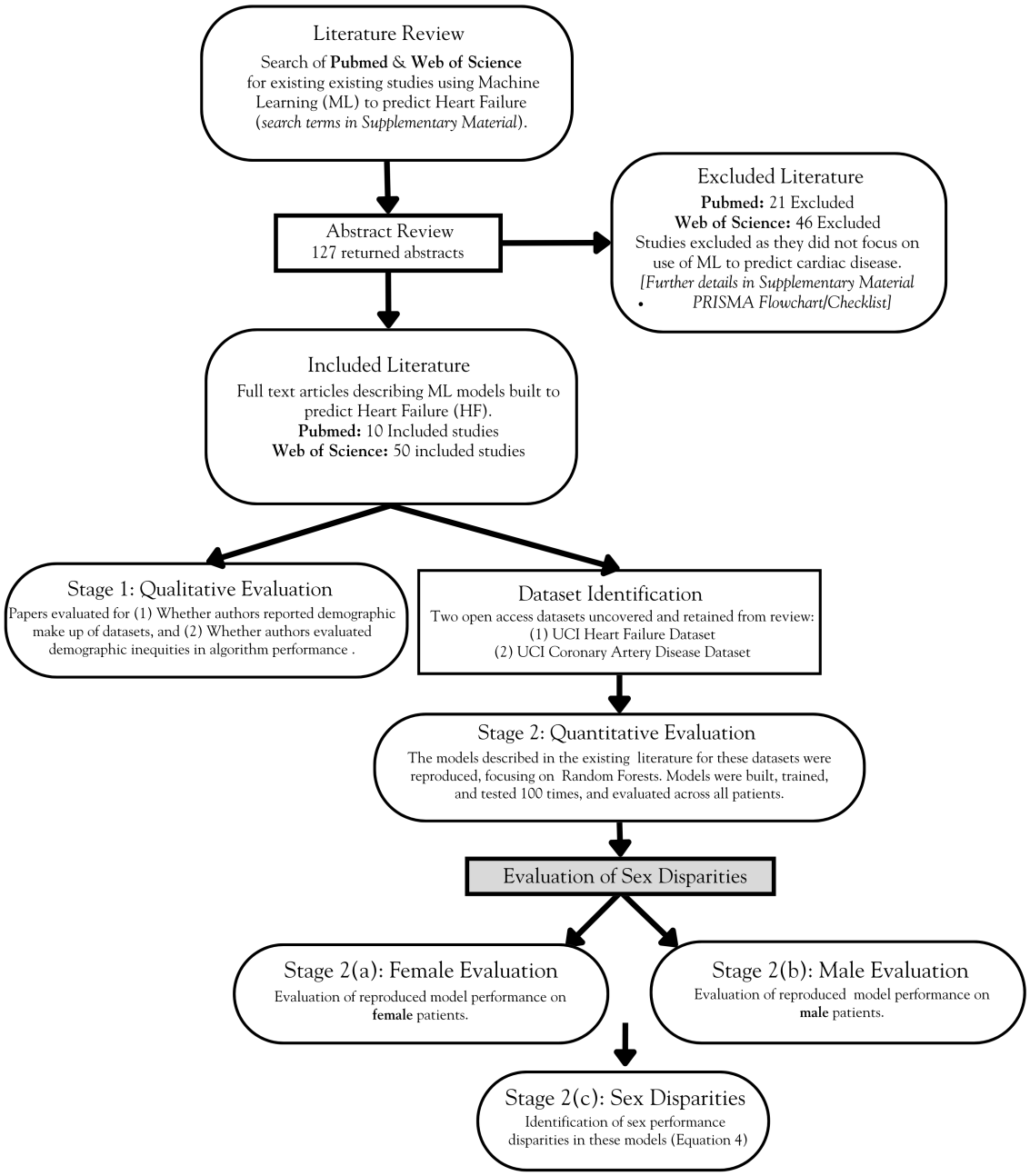


Figure 3.1: A flowchart detailing the steps of methodological steps of Chapter 3, including (1) the initial literature search and qualitative evaluation of identified studies, plus (2) the identification of datasets and interrogation of algorithms for demographic performance biases.

3.2.1 Methods: Stage 1

Stage 1 involved a literature review. We searched PUBMED and Web of Science between 1st April 2022 and 22nd May 2022 to identify ML algorithms used to predict cardiac disease adhering to PRISMA Guidelines for systematic reviews (search terms and PRISMA documentation is provided in the Supplementary material). All abstracts were reviewed, and articles were included for full text review if they met the following criteria:

- (1) The target diagnosis was Heart Failure (HF)
- (2) The model utilised biochemical markers to predict disease
- (3) The computational methods involved a Machine Learning (ML) approach (including supervised, unsupervised, and deep learning)

The full texts of the papers that met these criteria were reviewed, to evaluate questions specific to our research aim. In assessing these texts, I evaluated whether authors:

- (1) Reported demographic make-up of datasets
- (2) Evaluated demographic inequities in algorithm performance, meaning that the authors specifically examined differences in model fidelity by demographic groups defined by protected characteristics.

Throughout the literature review, any identified open-source datasets were maintained for use in Stage 2.

3.2.2 Methods: Stage 2

The results of the literature review are detailed briefly here as the information regarding the returned datasets is required for the description of Stage 2. The literature review returned two open source datasets, which were then used for further analysis. In this stage, we used these datasets to (i) rebuild the models described in existing publicatoins, and (ii) identify inequities in the performance of algorithms on the basis of Sex. The uncovered datasets were:

- **Dataset 1**: The UC Irvine (UCI) Machine Learning Repository Dataset for Heart Failure Prediction [150]
- **Dataset 2**: The UC Irvine (UCI) Machine Learning Cleveland Heart Disease dataset for identifying Coronary Artery Disease (CAD) [151]

The datasets were obtained from the UC Irvine (UCI) Machine Learning Repository, which is an open-access database provided datasets across a range of disciplinary domains [150, 151].

Datasets and model environment

The first dataset imported was the UCI Heart Failure (HF) Clinical Records Dataset, which was released by Ahmad and colleagues in 2017 and has been used extensively for

cardiac ML modelling [63, 143]. The dataset contains the records of 299 Heart Failure patients at the Allied Hospital in Faisalabad (Punjab, Pakistan). The complete dataset consists of 299 rows (patients) and the 13 feature columns.

The initial study published with the release of the dataset was a survival analysis, in which Cox regression models was used to model mortality from a range of demographic, biochemical and clinical variables [143]. As described in depth by Chicco and colleagues, the initial papers focused on this dataset utilised traditional biostatistical methods, including one study that looked at sex-specific models for cardiac disease modelling. Zahid and colleagues adopted a similar approach to the original dataset authors using Cox models to build sex-specific predictions and identifying significant differences in the predictive power of variables in the dataset [152]. Since these original papers, data scientists have demonstrated that ML models outperform traditional statistical models with greater predictive accuracies, however these authors have not examined the sex-differences in model development that were identified in the early statistical studies [63, 143, 152].

The ML models that have been built for this dataset outperform the original statistical studies, which achieved concordance indexes of 72 - 77% for predicting HR deaths. Chicco and colleagues produced one of the first papers utilising ML methods on the dataset, matching and outperforming these original models with Random Forest algorithms achieving accuracies of 74% and ROC Scores of 80% [63, 143]. Bashir and colleagues went further, applying a range of classification algorithms to the HF Dataset and achieving accuracies of 84.17% with the Random Forest Model [153]. Senan and colleagues outperform both these earlier works, again demonstrating Random Forest models to be the most effective and achieving accuracies of 97.68%. Of note, none of these Machine learning papers examined sex differences in algorithmic performance or looked at the sex-differences in features that were referenced in earlier statistical works by Zahid and colleagues [152].

Secondly, we imported the UCI Coronary Artery Disease (CAD) Dataset [154]. The Coronary Artery Disease database, available from IEEE Dataport, is one of the largest available online datasets describing cardiovascular diseases in individuals. Previous researchers have applied a range of algorithms to this dataset for the purpose or predicting cardiac disease, spanning supervised and unsupervised ML techniques. Latha and colleagues built Random Forest Models, Multilayer perceptrons and Naive Bayes models, achieving accuracies of 84.5% [148]. Miao et al chose to use adaptive boosting models, resulting in accuracies of 80.14% [149]. Atallah et al developed K-Nearest Neighbour models, Logistic Regressions, and Random Forests, achieving accuracies of 90% [155]. We extract the original full database which integrates hospital data across multiple clinical

sites in Cleveland, Hungary, Switzerland and Long Beach, published by Siddartha et al and used widely in the research literature [154]. In this dataset the target variable is the diagnosis of Coronary Artery Disease.

For all experiments in this thesis, models were built on Jupyter notebook within the Anaconda distribution and written in Python code. To initialise my working environment, I began by importing a range of python libraries commonly deployed in Machine Learning, including: Math, ScikitLearn, NumPy, Pandas, Seaborn and Matplotlib.

Data Exploration and Descriptive Statistics

Data exploration is the primary stage of the ML process and involves file importation, formatting, descriptive statistics and configuring datatypes. Table 3.1 and Table 3.2 provide the variables included in our datasets and their initial datatypes. The features in Table 3.1 and 3.2 are a mix of demographic and physiological risk factors, and biochemical markers, that relate to cardiac disease. Clinical measurements include blood pressure and ejection fraction, biochemical markers include serum creatinine, sodium, platelets, and creatinine phosphokinase (CPK).

To begin our evaluation of the data, descriptive statistics were performed on both datasets. Firstly, I assessed the balance of the sexes within the dataset, and broke this down by the target variable (death in Dataset 1, and disease in Dataset 2). As outlined in the introduction, there is an extent of research describing sex differences in disease presentation and progression, thus the next step was a sex-stratified descriptive evaluation of each dataset variable - paying particular attention to differences in predictive biomarkers. For the variables of each dataset (detailed Table 3.1 and Table 3.2), I calculated the mean and variance for sexes separately, further stratifying by those affected by death (Dataset 1) and disease (Dataset 2).

Feature Evaluation

Feature evaluation is an essential step in any ML model development, but may be even more pertinent in the case of ML bias evaluations. Previous research has demonstrated the value of examining the shifts in feature ranking that occur with training data stemming from different demographic subgroups [18]. In their article examining performance disparities affecting racial minorities and younger patients, Afrose and colleagues demonstrated that the analysis of feature ranking via SHAP values highlighted the importance of demographic-specific feature rankings that would go undetected if not separated out through different subgroup datasets [18]. The authors illustrate that when predicting

Table 3.1: Description of Features for Dataset 1 (Heart Failure)

Feature	**Description**	**Datatype**
Age	Age of the patient (years)	Integer, continuous
Anaemia	Reduced count of red blood cells or haemoglobin	Integer, binary
High blood pressure	Whether the patient has high blood pressure	Integer, binary
Creatinine phosphokinase (CPK)	The level of CPK enzyme in the blood	Float(64), continuous
Diabetes	Previous diagnosis of diabetes	Integer, binary
Ejection Fraction	The percentage of blood that is ejected from the heart with each contraction	Numerical, continuous
Sex	Sex of patient	Numerical, binary
Platelets	Count of platelets in the blood	Numerical, continuous
Serum creatinine	The level of creatinine in the blood	Numerical, continuous
Serum sodium	The level of sodium in the blood	Numerical, continuous
Smoking	Whether the patient smokes	Integer, binary
Time	Follow up period of patient	Integer, continuous
Death event (Target variable)	If a patient died during the follow-up period	Integer, binary

Table 3.2: Description of Features for Dataset 2 (Coronary Artery Disease)

Feature	**Description**	**Datatype**
Age	Age of the patient (years)	Numerical, continuous
Sex	Sex of patient	Numerical, binary
Chest pain type	Category of chest pain	Numerical, categorical
Resting BP	Level of blood pressure at rest	Numerical, continuous
Cholesterol	Serum cholesterol	Numerical, continuous
Fasting blood sugar	Blood sugar levels on fasting of >120mg represented as 1 if true, 0 if false	Boolean
Resting ECG	Result of electrocardiogram while at rest	Numerical, categorical
Max Heart Rate	Maximum heart rate achieved	Numerical, continuous
Exercise Angina	Whether angina is induced by exercise	Nominal
Old Peak	Whether there is exercise induced ST depression on the ECG in comparison with state of rest	Numerical, continuous
ST Slope	The slope of the peak exercise ST Segment	Nominal

breast cancer survival, models trained on samples with increased representation of Asian patients exhibit different feature preference, compared to models trained on original data that had majority White patients [18]. Our approach mirrors similar methodology, however, we examine these issues through the lens of sex.

A range of approaches have been deployed for the purpose of feature evaluation, including examining correlation metrics between features and a target variable (e.g. Pearson's Correlation Coefficient), and more novel methods based on game theory, such as Shapley Values (Shapley Additive exPlanations Values). Shapley values were proposed in 2017 and have become a widely accepted uniform measure of feature importance [18, 156, 157]. As described by Fryer and colleagues, Shapley values have become particularly popular in the "Explainable AI (XAI)" literature, due to their use in interpretable feature attribution [112]. SHAP values measure how much each feature contributes to the model's prediction, facilitating rankings of feature importance [112]. The method relies on a game theory approach, in which each feature is treated as a "player" and the mark of their importance is based on their final contribution to prediction performance. As SHAP values are model-agnostic, they can be used to interpret a range of ML models [112].

In my experiments I chose to focus on Random Forest (RF) models as these were most commonley reported as high performing in the literature (detailed below), and thus I also used RF specific methods for evaluating feature importance. Feature rankings in Random Forest can be determined using a metric called Gini importance, which measures the reduction of the Gini impurity of the dataset when a specific feature is used for splitting. The higher the Gini importance, the more important the feature is for the model. For both datasets, I evaluated feature importance using Pearson's Correlation Coefficients, RF Gini Importance and SHAP Values for the dataset overall (all patients) and for sex-specific subsets (males and females separately), and compare the feature rankings between these groups.

Reproduction of original models

The datasets described above were found within studies that had used the data to build ML models for predicting cardiac disease and death. My next step was to then rebuild these same models, and go a step further by examining whether sex-based disparities existed in the model performance. As we were rebuilding the models of existing published studies, our model selection was guided by the choices of these previous papers. Our focus was therefore on Random Forest (RF) algorithms, which were reported to be the most effective predictive models for both datasets in the existing literature [52, 63].

We built our Random Forest models using the Science Kit (SciKit) Learn package, originally developed by David Cournapeau in 2007 and one of the most commonly used libraries in the ML community [158]. The following steps were followed for both datasets:

1. **Data Splitting:** The dataset was split into training and test sets, allocating 70% to training and 30% for testing.

2. **Feature Selection:** In this chapter, I use the full feature set in building models to replicate the methods of the uncovered papers from the literature search. In the following chapter, I dive into the impact of selecting specific feature subsets on inequities in algorithm performance.
3. **Hyperparameter Tuning:** The depth of the trees ('max depth') and the number of trees ('n estimators') was tuned using GridSearchCV - a tool from the scikit learn library in Python that facilitates hyperparameter optimisation. GridSearchCV takes a parameter grid as input (a defined dictionary), and creates combinitations of the grid values, evaluating the impact on model performance. For each combination, cross-validation was performed (n=3), meaning that 3 subsets were formed from training data, with each acting as a validation set in turn. The best hyperparameters were then used to configure the final model.
4. **Model Training:** The RF classifiers were then trained with optimal hyperparameters using the 70% training data.
5. **Model Prediction and Evaluation:** The trained classifier made predictions on the test dataset, predicting either death from heart failure (Dataset 1) or diagnosis of coronary artery disease (Dataset 2). The results were collected and then stratified by Sex, to analyse the performance for the males and females separately. Model performance was assessed using common ML metrics (Accuracy, F1 Score, ROC AUC score, Precision and Recall) and error rates (e.g. False Negative Rate), detailed below.
6. **Repeated Experiments:** Steps 1-5 were integrated into a loop, such that these steps were repeated across 100 experiments, in order to assess for the model's stability and reliability across different iterations.
7. **Final Evaluation and Statistical Significance**: After the 100 experiments concluded, the full results were aggregated for all patients and for the sexes separately. The mean and standard deviation for each performance metric was calculated across the 100 runs, and the mean difference was then calculated between the males and females. To evaluate for statistical significance, independent t-tests were performed where the data was normally distributed, and Mann-Whitney U tests were performed where the data was not normally distributed. Kolmogorov-Smirnov Tests were used to assess for normality [159].

Performance bias & evaluation metrics used

In my evaluation of AI bias I look specifically at algorithmic "performance bias", such that I am examining demographic differences in model fidelity. I do this to focus the research of this thesis on one area of AI ethics, specifically examining computational solutions in ML fairness. My approach facilitates an in-depth analysis of the specific ethical issues and computational challenges relating to model fidelity, which is the focus of the machine

learning experiments of this research. By adopting this approach, I neglect other areas of the AI pipeline where bias may arise, and this issue is covered in greater depth in the discussion.

Global performance metrics

First, I made use of the global evaluation metrics which are commonly used in the evaluation of ML models (e.g. Accuracy and ROC Scores). A summary and the respective equation fo each of these metrics is provided below [Equations 3.1 to 3.2]. Utilising these metrics in my analysis ensured I was using the latest ML methods, and allowed me to compare the results of my experiments to the algorithmic performance referenced in the published literature.

1. **Accuracy**: The proportion of correct predictions, determined by dividing the number of correct predictions by all observations [41, 42]. The metric is known to perform less well when there is significant uneven class distribution.
2. **Precision and Recall**: The metrics of Precision and Recall depend on the underlying error rates of the model (see below) and provide a more detailed assessment of the type of errors occuring in model performance. Recall is synonymous with the True Positive Rate, thus optimising for recall minimises the chances of missing positive cases (e.g. missing disease) [42]. Alternatively, Precision focuses on the correctness of positives, and is important when false positives may be negatively consequential (e.g. police profiling algorithms).
3. **Receiver operating characteristic (ROC) curve**: The ROC curve plots the true-positive rate (TPR) of a model on the y-axis, and the false-positive rate (FPR) on the x-axis. The area under this curve is referred to as the AUC. An AUC of 1 indicates optimal performance, in which the model can perfectly distinguish between classes, whereas an AUC of 0.5 demonstrates no ability to distinguish [41, 42]. In the case of AUC=0, all predictions are incorrect.

Equations 3.1 - 3.2: Global Performance Metrics, calculated from False Positives (FP), False Negatives (FN), True Positives (TP) and True Negatives (TN) [42]

$$\text{Error Rate (ERR)} = \frac{\text{FP} + \text{FN}}{\text{FP} + \text{FN} + \text{TP} + \text{TN}} \tag{3.1}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{FP} + \text{FN} + \text{TP} + \text{TN}} = 1 - \text{ERR} \tag{3.2}$$

Clinically tailored performance metrics

Moving beyond global performance metrics, I also examined differences in specific algorithmic error rates, drawing from the foundational work from Buolamwini and Gebru, who demonstrated that a range of ML algorithms for facial recognition performed poorly on darker skinned females [15]. In their paper the authors examined the differential performance of ML algorithms used for facial recognition across different gender and racial groups, widely cited as one of the first papers exposing AI bias in the field [15]. The authors identified a difference in error rate between the best and worst classified groups as 34.4%. The calculations are derived from the difference in error rates between the different subgroups (e.g. lighter skinned males and darker skinned females) [15]. Of note, the authors here focus on the True Positive Rate (TPR), as this is potentially a more consequential harm where facial recognition systems are being used by policing for criminal identification. We have used the same techniques but focused on the False Negative Rate (FNR), as the impact of missed diagnosis and treatment is arguably more consequential in our clinical context.

Examining error rate disparities have been a strongly proposed measure for evaluating bias in the ML fairness literature, with several key papers adopting this approach for identifying model under-performance that disadvantages specific subgroups [17, 87, 160, 161]. Allen and colleagues describe the importance of identifying disparities in False Negative Rates in mortality prediction tools, where a failure may lead to a lack of timely care and increased risk of death [160]. As detailed by Afrose and colleagues, previous research that focuses only on global metrics such as ROC Scores and Accuracy may neglect subtle disparities manifesting within different error rates [18]. Similar works from Thompson et al and Rajkomar and colleagues have described the impact of disparities in false negatives and false positives on patients from racial and ethnic minority groups [17, 87, 161].

For bias in AI to be understood in clinical terms, the evaluative performance metrics must be placed in their clinical context. For example, in the case of a algorithm being used to prescribe a potentially toxic medication, the impact of false positives (unnecessarily prescribed medications) may have more adverse effects. In our context where we are examining the prediction of disease and death, the False Negative Rate (FNR) may be more consequential as neglect of disease may result in delayed treatment and worse disease outcomes. The existing research has highlighted the importance of examining these specific metrics within the medical domain [18, 81, 87]. I therefore chose to evaluate models using the metrics presented in Table 3.3. For every model that is built within the repeated experiments, these evaluative metrics were calculated for the population of patients overall, and for the male and female patients separately.

Table 3.3: Algorithm evaluation metrics defined by the number of True Positives (TPs), False Positives (FPs), True Negatives (TNs) and False Negatives (FNs), presented with their associated clinical implications

Evaluation Metric	**Equation**	**Clinical Implications**
True Positive Rate (TPR) (Recall)	$\frac{TP}{TP+FN}$	Correct diagnosis that patient has disease
False Positive Rate (FPR)	$\frac{FP}{FP+TN}$	Misdiagnosis of disease when patient is healthy
True Negative Rate (TNR)	$\frac{TN}{TN+FP}$	Correct diagnosis that patient is healthy
False Negative Rate (FNR)	$\frac{FN}{FN+TP}$	Misdiagnosis that patient is healthy when patient has disease
Precision	$\frac{TP}{TP+FP}$	Accurate identification of actual positive cases
F1 Score	$2 \cdot \frac{Prec \times Rec}{Prec+Rec}$	Harmonic mean of Precision and Recall, providing a balanced measure for uneven class distributions

3.2.3 Evaluating model disparities

Once the results for the 100 model experiments were obtained, sex differences in each algorithm evaluation metric were calculated (Equations 3.2 to 3.1 and Table 3.3). First, the sex performance disparity was calculated using Equation 3.3 for each of the evaluation metrics (e.g. mean male ROC score - mean female ROC score). The difference is calculated across the 100 runs of our experiment, and the overall average sex disparity is calculated (along with standard deviation) (Figure 3.1).

Equation for Sex Performance Disparity

$$\textbf{Sex performance disparity} = \text{Score for males (mean)} - \text{Score for females (mean)} \tag{3.3}$$

When calculating the average sex difference for each metric, accompanying statistical tests were also performed to evaluate for the significance of any identified disparities. To choose our statistical test for evaluating the significance of differences between the males and females, the distribution of the model performance data was first evaluated for normality. There are various methods for testing the normality of continuous data, of which the Shapiro Wilk test and Kolmogorov Smirnov test are two popular options [159]. Given that we were dealing with a larger sample size ($n > 50$), we opted for Kolmogorov Smirnov tests to ascertain normality [159]. In cases whether the data was normally distributed, independent t-tests were performed to assess for the significance of differences between males and females. Where the data was not normally distributed, Mann-Whitney U tests were performed.

In summary, our methodological approach was split into two stages, starting with a qualitative evaluation of the existing cardiac ML literature, followed by a quantitative analysis of performance inequities in ML algorithms built from two open source cardiology datasets:

1. (1) **Stage 1:** A qualitative evaluation of the existing literature on cardiology ML algorithms, evaluating whether authors reported demographic make-up of datasets and/or inequities in algorithm performance.
2. (2) **Stage 2:** A quantitative evaluation of two cardiac datasets focused on elucidating any under-reported biases in algorithm performance, through (i) Rebuilding models described in published literature, (ii) Examining these models for statistically significant disparities in algorithm performance, across a range of performance metrics.

3.3 Results

In Stage 1 our initial literature review of papers discussing ML models built to predict cardiac disease returned 127 articles, of which 60 met the criteria for full review, and three highlighted sex differences in model performance [63, 162–168]. Throughout the returned articles there was a consistent under-representation of female patients in the datasets, and none of the returned papers investigated racial or ethnic differences on dataset representation or algorithmic performance. The majority of articles reported global performance metrics for all patients, but failed to examine specific error rates or stratify by sex [63, 162–167].

Many papers relied on proprietary or private datasets, which were not openly available due to the confidential nature of patient records, precluding further secondary analysis of the algorithms described. Amongst these articles, those that reported the demographic make up of their datasets consistently demonstrated an over-representation of male patients [162, 164, 166, 167]. Nakajima and colleagues built ML models for predicting life-threatening arrhythmia's and cardiac death amongst HF patients, for which their dataset consisted of 72% male patients [166]. Panahiazar and colleagues demonstrate the superior performance of ML methods for predicting HF Mortality, compared to traditional scoring metrics (e.g. The Seattle Heart Failure Model (SHFM)), however these improvements are not examined the the sexes separately [164]. The team built a series of ML classification Models (Random Forests, Support Vector Machines, and Logistic Regression) on a dataset of 5044 patients (52% male, 94% white), achieving AUC scores of 80-81% for all patients, with no breakdown by sex [164]. The neglect of sex differences in disease and algorithm performance was consistent across studies from differing countries [63, 162–167]. Joon-myoung and colleagues present a study of 2165 patients from 10 university hospitals of the Korean Acute Heart Failure (KorAHF) registry, in which their deep-learning algorithm for predicting HF mortality achieves AUC scores of 0.88, however potential sex differences of bias is not mentioned [165].

Adler and colleagues build tree-based models for predicting mortality from a hospital dataset of 822 hospitalized and ambulatory patients with HF. The author's description of the 3 datasets used demonstrate an under-representation of females, varying from 28% to 41% inclusion of females [167]. These authors do examine differences in performance, but not specific error rates [167]. One article from Tison and colleagues was an exception, which focused specifically on females with heart failure and highlighted that heart failure was more common in people who were older, Caucasian, with a higher mean number of pregnancies, a higher BMI and were less likely to have Medicare [168].

3.3.1 Descriptive Statistics

In Stage 2 we examined the UCI Heart Failure (HF) Clinical Records Dataset (Dataset 1) and UCI Coronary Artery Disease (CAD) Dataset (Dataset 2) that were uncovered in our literature review. Initial descriptive statistics of both datasets revealed a greater number of male patients than female patients (194 Males vs 105 Females Dataset 1; 564 Males vs 182 Females Dataset 2; Table 3.4 - Table 3.5). In Dataset 1 there was almost a 70:30 split with regards to the target variable for both sexes (67.7% healthy vs 32.4% HF Deaths for females; 68.0% healthy vs 32% HR Deaths for males). In Dataset 2 there was a smaller proportion of females with the target variable (CAD disease) compared to the males (78.0% Healthy vs. 22.0% diseased for females; 44.0% Healthy vs 56% Diseased for males) (Table 3.4 - Table 3.5)). In both datasets we therefore saw an under-representation of female disease, firstly due to the overall low number of females in Dataset 1, and in Dataset 2 this is compounded by a reduced proportion of diseased females amongst the female patient group.

The sex-stratified descriptive statistics of the two datasets revealed subtle sex differences in the presentation of both Heart Failure (HF) and Coronary Artery Disease (Table 3.4 to Table 3.5). For Dataset 1 focused on HF mortality, Table 3.4 presents the mean value of the biochemical markers and clinical measurements for the sexes separately and stratified by the target outcome (death from heart failure). Here we see that in the case of deaths from Heart Failure (HF Deaths), males tend to be older than their female counterparts, with a higher Creatinine Phosphokinase, lower likelihood of diabetes, lower Ejection Fraction (EF) and lower blood pressure (BP).

Table 3.5 provides the mean and standard deviation for each of the features in Dataset 2 (CAD Dataset). Of note, the original dataset reports a total participant count of 1190, however this fell after duplicate and null values were removed, giving the total count of n=746 presented below. The outcome variable of CAD diagnosis assigns a value 1-5 of disease severity based on the narrowing of heart vessels, this is changed to a binary outcome where values above 2 are considered diseased (diseased = 1). The dataset contains 76 attributes, but all published experiments refer to a subset of just 10 of them which are presented in Table 3.5. Here, we see that female patients with CAD have a higher resting BP than their male counterparts. The variable for resting ECG is also higher for females which appears to relate to higher incidence of Left Ventricular Hypertrophy. Furthermore, we see in Table 3.5 specific difference in the biochemical features of males and females, with sick females demonstrating a far higher cholesterol level than sick males (mean values; 279.18 Female Sick vs. 247.50 Male Sick). In fact, the cholesterol level of a healthy female (mean 249.2, SD 62.2) almost mirrors that of a diseased male (mean

247.5, SD 61.9). Thus, when the data is sex-mixed, as opposed to sex-stratified, models may struggle to extract predictive features that offer conflicting information depending on whether a patient is male or female.

Table 3.4: Descriptive statistics of variables in Dataset 1 (Heart Failure) for 299 patients, by Target (Death) and Sex

	Female (n=105)				**Male (n=194)**			
	Survived		**Death**		**Survived**		**Death**	
Metric	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Attack Rate (%)	71 (67.6%)		34 (32.4%)		132 (68.0%)		62 (32.0%)	
Age (yr)	58.6	10.6	62.2	12.3	58.8	10.7	66.9	13.5
Anaemia	0.5	0.5	0.6	0.5	0.4	0.5	0.4	0.5
CPK (mcg/L)	462	518	508	780	583	853	759	1532
DM	0.5	0.5	0.6	0.5	0.4	0.5	0.3	0.5
EF (%)	41.9	11.6	37.5	14.6	39.4	10.4	31.2	10.7
High BP	0.4	0.5	0.5	0.5	0.3	0.5	0.4	0.5
Platelets (k/mL)	289	98.7	259	107.6	254	94.9	254	94.1
Creatinine (mg/dL)	1.1	0.6	1.9	1.6	1.2	0.7	1.8	1.4
Sodium (mEq/L)	137.4	3.6	135.5	6.7	137.1	4.2	135.3	3.8
Smoking	0.0	0.1	0.1	0.3	0.5	0.5	0.4	0.5

*Values denote mean (Mean) and standard deviation (SD) unless indicated. For Death, 1 = mortality. Abbreviations: CPK = Creatinine Phosphokinase, DM = Diabetes Mellitus, EF = Ejection Fraction, BP = Blood Pressure. Full dataset details at `https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records`

,

Table 3.5: Descriptive statistics of the variables in Dataset 2 (Coronary Artery Disease) (n=746), stratified by Target (CAD Diagnosis) and Sex

	Fem. (n=182)				**Male (n=564)**			
	Healthy		**Diseased**		**Healthy**		**Diseased**	
Rate (%)	142 (78%)		40 (22%)		248 (44%)		316 (56%)	
	M	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
Age (yr)	51.1	9.6	56.0	7.2	49.6	9.1	55.8	9.0
Chest pain	2.7	0.9	3.7	0.7	2.8	0.9	3.6	0.8
BP (mmHg)	128.8	16.7	143.4	20.7	131.0	15.8	135.2	17.4
Chol (mg/dL)	249.2	62.2	279.2	60.1	232.8	50.2	247.5	61.9
Fasting Sugar	0.1	0.3	0.2	0.4	0.1	0.3	0.2	0.4
Resting ECG	0.6	0.8	0.8	0.9	0.5	0.8	0.7	0.8
Max HR (bpm)	149.2	21.6	139.2	21.7	149.0	24.0	129.4	22.2
Angina	0.1	0.3	0.6	0.5	0.1	0.3	0.7	0.5
Old Peak (mm)	0.4	0.6	1.5	1.4	0.4	0.7	1.5	1.1
ST Slope	1.3	0.5	2.0	0.4	1.2	0.5	2.0	0.5

*CAD = Coronary Artery Disease. Details and full dataset available at https://dx.doi.org/10.21227/dz4t-cm365 and feature descriptions provided in Table 3.2

3.3.2 Feature Evaluation

In our examination of feature importance we identified further sex differences. The features available in each dataset detailed in Table 3.1 and Table 3.2 were ranked in terms of feature importance with respect to the target variable for that dataset (disease or death), for all patients and then for the sexes separately.

Table 3.6 and Table 3.7 provided feature rankings measured by Pearson's Correlation Coefficient for Dataset 1 and Dataset 2 respectively. Figure 3.3 to 3.4 presents the results when ranking features using the Random Forest measure of Gini Importance, and Figures 3.6 to Figure 3.7 presents the findings when using SHAP Values. Across all of these results, we see significant differences in feature rankings between the sexes.

For Dataset 1, we see consistent sex differences across the different methods for feature ranking. Firstly, Ejection Fraction consistently serves a different level of importance for females vs males, which is unsurprising given the previous research we have outlined that details sex differences in this clinical metric (Table 3.6). The differences are more pronounced when examining feature importance with the RF specific methods of Gini importance. For Dataset 1, we see from Fig 3.2 and 3.3 that while Ejection Fraction is the highest ranked feature for males, this falls to 5th place for the females. Further, with the RF method, Platelets and Creatinine Phosphokinase are selected as the 2nd and 3rd features for females in Dataset 1, which were not previously identified from the Pearson

Correlation methods. For Dataset 2 we see differences again for the RF method compared to using the Pearson approach, with Cholesterol and Resting Blood Pressure being ranked higher at positions 3 and 4 for female patients (Figure 3.4). For males these features are ranked 7th and 8th respectively (Figure 3.5). On examining SHAP values we again see Ejection Fraction being ranked higher for males compared to females in Dataset 1 (Figure 3.6 to Figure 3.7.

Table 3.6: Dataset 1: Features with greatest correlation with target outcome (death) measured by Pearson correlation coefficient for full dataset and sex-stratified subsets

Rank	All Patients	Female Patients	Male Patients
1	Serum Creatinine	Serum Creatinine	Ejection fraction
2	Ejection fraction	Serum sodium	Age
3	Age	Smoking	Serum Creatinine
4	Serum sodium	Ejection Fraction	Serum Sodium
5	High blood pressure	Age	Creatinine phosphokinase

Table 3.7: Dataset 2: Features with greatest correlation with target outcome (CAD diagnosis) measured by Pearson correlation coefficient for full dataset and sex-stratified subsets

Rank	All Patients	Female Patients	Male Patients
1	ST Slope	ST Slope	ST Slope
2	Exercise Angina	Exercise Angina	Exercise Angina
3	Old Peak	Old Peak	Old Peak
4	Chest pain type	Chest pain type	Chest pain type
5	Maximum Heart Rate	Resting Blood Pressure	Maximum Heart Rate

Figure 3.2: Dataset 1 (HF): Ranking of features for **Female Patients**, measured by Gini Importance for Random Forest Models

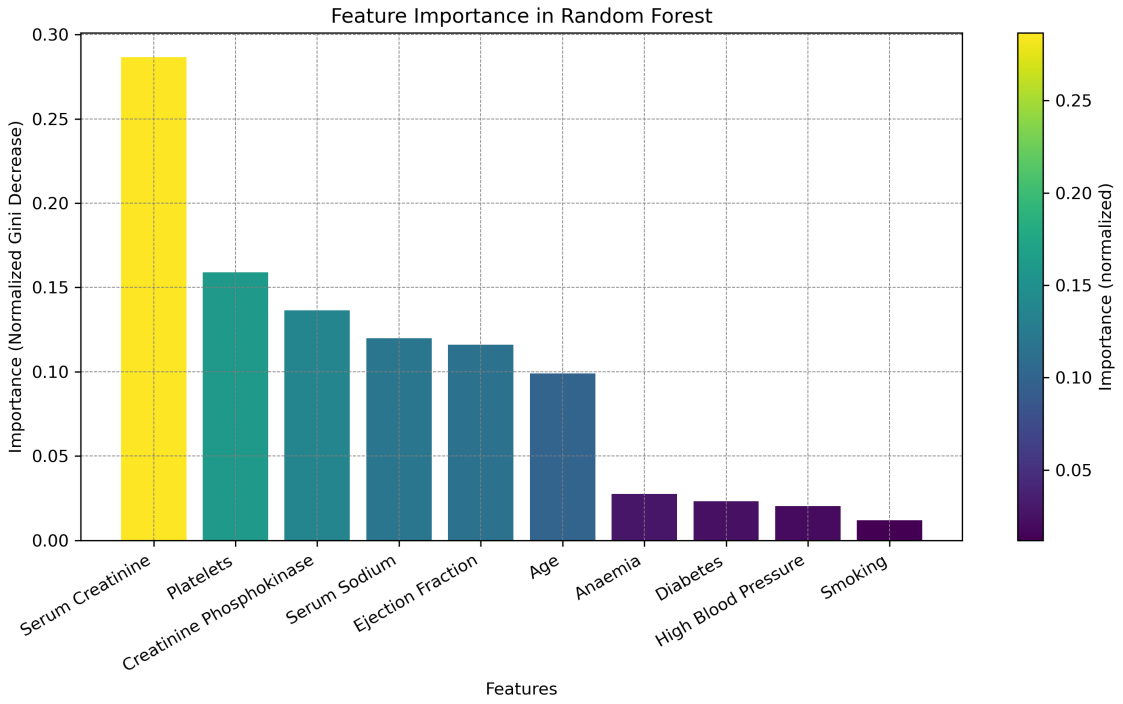


Figure 3.3: Dataset 1 (HF): Ranking of features for **Male Patients**, measured by Gini Importance for Random Forest Models

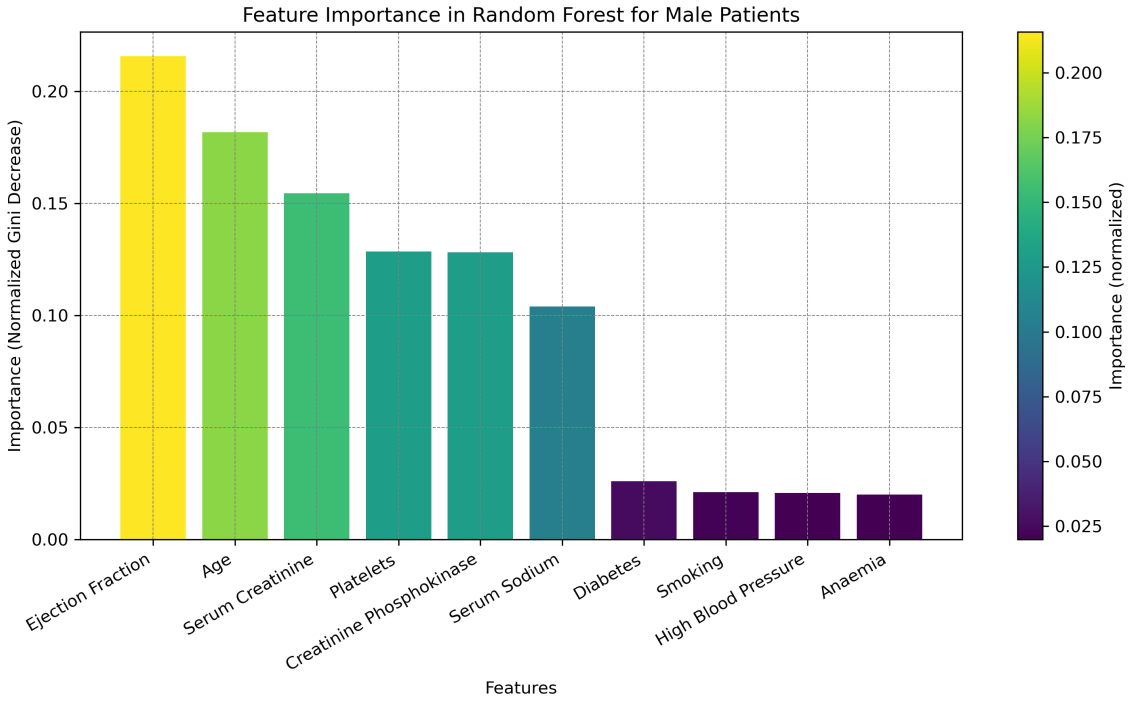


Figure 3.4: Dataset 2 (CAD): Ranking of features for **Female Patients**, measured by Gini Importance for Random Forest Models

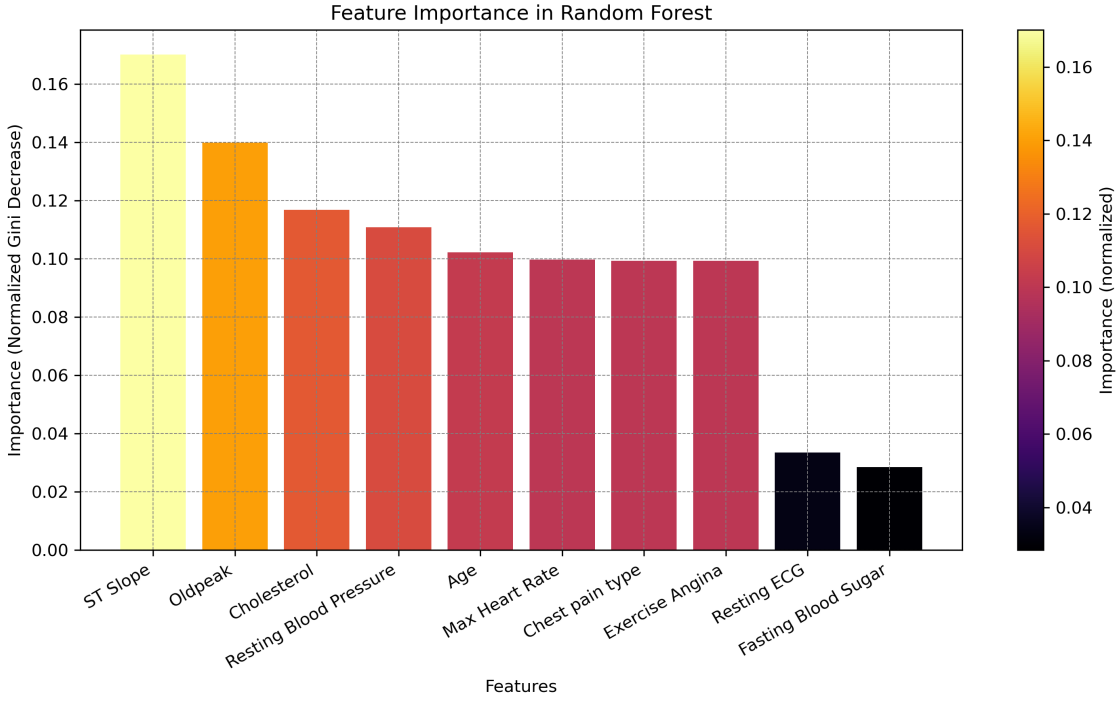


Figure 3.5: Dataset 2 (CAD): Ranking of features for **Male Patients**, measured by Gini Importance for Random Forest Models

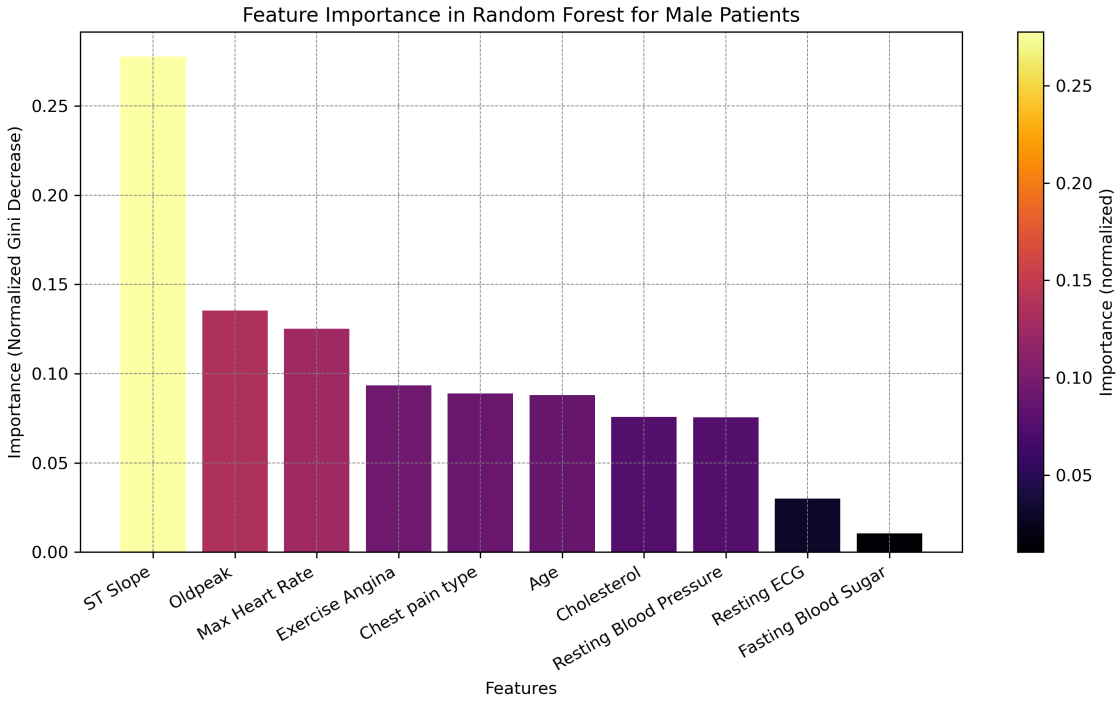


Figure 3.6: Dataset 1 (HF): Comparison of feature rankings for all patients and the male and female subsets, using SHAP Values. Features are listed in descending order of their impact on model prediction.

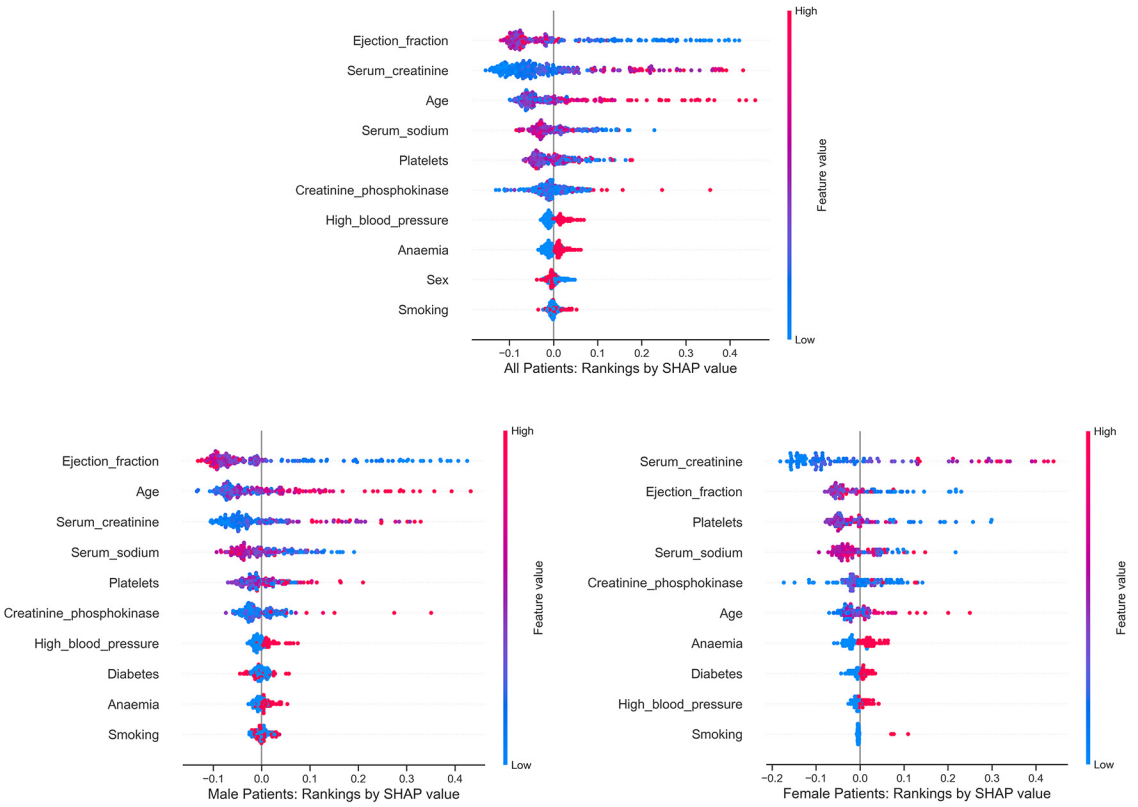
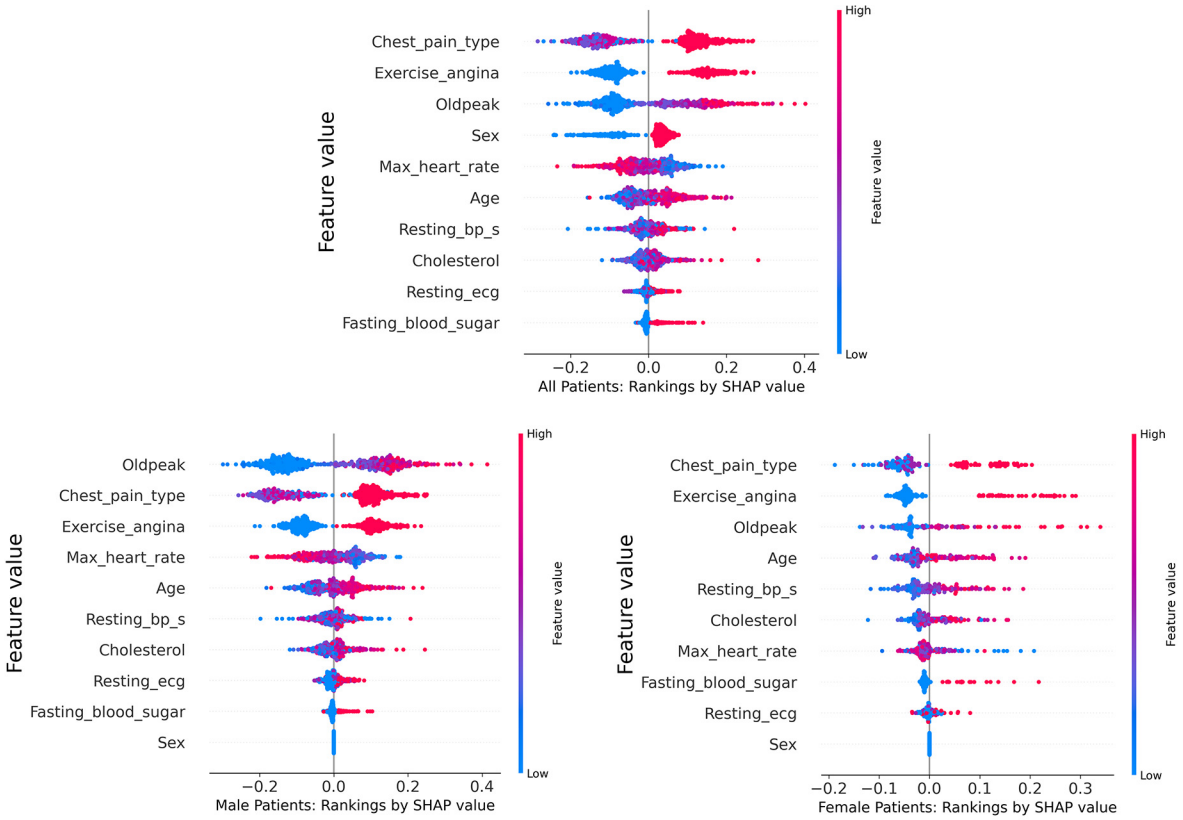


Figure 3.7: Dataset 2 (CAD): Comparison of feature rankings for all patients and the male and female subsets, using SHAP Values. Features are listed in descending order of their impact on model prediction.



3.3.3 Model Results and Performance Disparities

Our next step in Stage 2 was to rebuild the models described in existing publications uncovered in our literature review. Focusing on Random Forest models, we rebuilt these algorithms and achieved the same predictive accuracies of current studies: 84.24% (3.51 SD) for Dataset 1, and 85.72% (1.75 SD) for Dataset 2 [63]. The model performance was then evaluated separately for the sexes, and sex disparities in performance were identified for each evaluative metric (Equation 3.3 and Table 1.2). Table 3.8 provides the sex disparities in model performance for the global metrics and the model error rates, calculated from Equation 3.3. Here we see a significant difference in overall performance that benefits male patients, with a significantly higher False Negative Rate for female patients (Sex disparity of -7.53%, $p<0.01$) for Dataset 1. For Dataset 2, we see similar results, with a higher overall performance for males (ROC disparity of 3.86, $p<0.01$) and significantly higher False Negative Rate for females (-11.66%, $p<0.01$).

These findings are further illustrated in Figure 3.9, where the average point estimates for each model performance score are provided for the sexes separately on the violin plot. The visualisation of the differences in these point estimates, and the disparities identified in Table 3.8, demonstrate that we see a pattern of models under-predicting disease in females (with a higher female FNR and TNR) and over-predicting disease in males (with a higher FPR and TPR) (Figure 3.9).

Table 3.8: Sex disparities in the performance of Random Forest (RF) Models for Dataset 1 and Dataset 2. The disparity is the mean difference in algorithmic performance between the males and female subsets across 100 experiments, with a positive number indicating a higher value for males (See Equation 3.3). The asterix (*) indicates that this difference was statistically significant ($p<0.05$)

Disparity in Model Performance	**Dataset 1 (Heart Failure)**	**Dataset 2 (Coronary Artery Disease)**
Accuracy Disparity (%)	*1.63 (0.03)	0.32 (0.50)
ROC AUC Disparity (%)	*3.14 (<0.01)	*3.86 (<0.01)
FNR Disparity (%)	*-7.53 (<0.01)	*-11.66 (<0.01)
FPR Disparity (%)	1.26 (0.07)	*3.94 (<0.01)

Figure 3.8: Dataset 1: Performance of reproduced cardiac ML models for Dataset 1 (Heart Failure), for the female patients (left of violin plot) and male patients (right of violin plot), measured by global performance metrics and error rates.

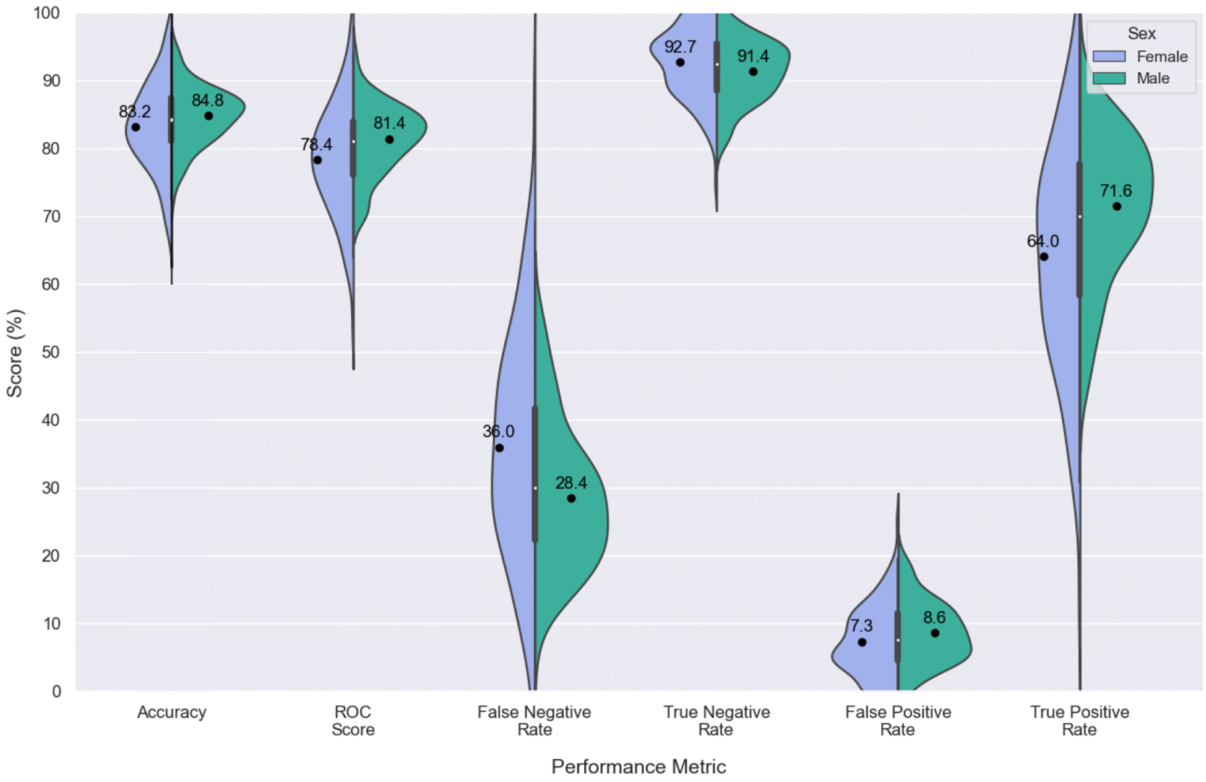
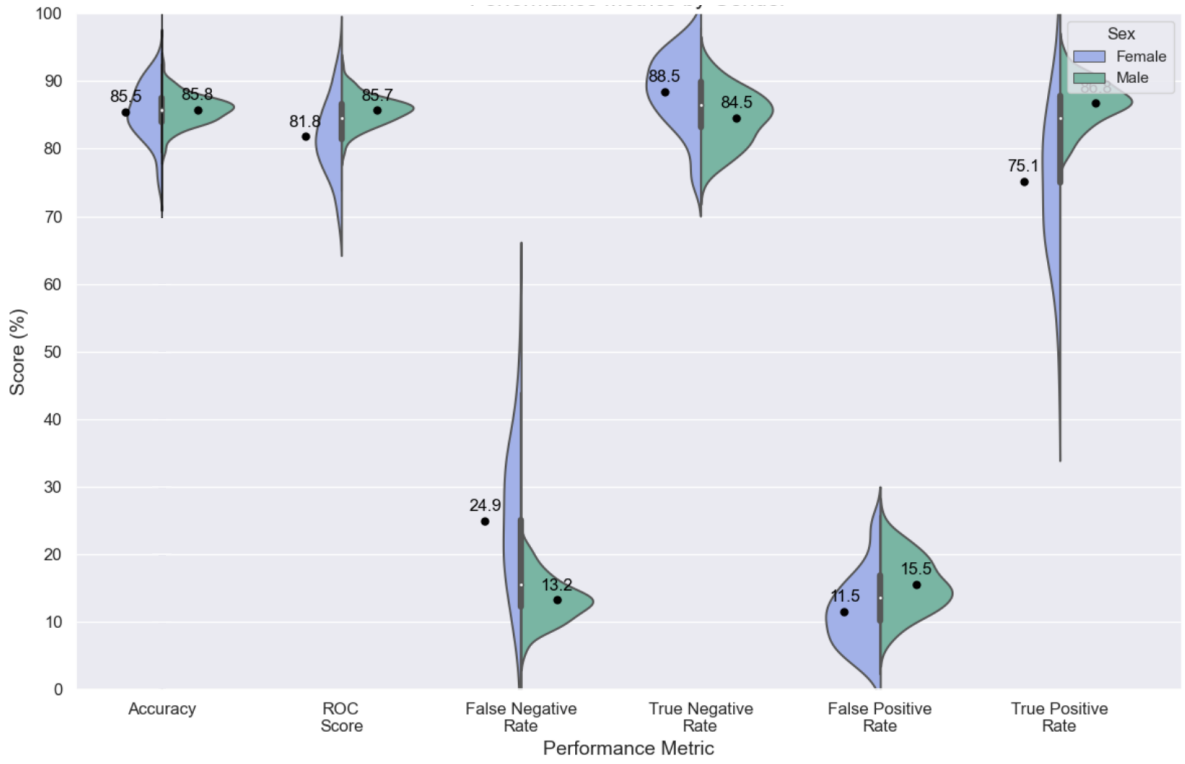


Figure 3.9: Dataset 2: Performance of reproduced cardiac ML models for Dataset 2 (CAD), for the female patients (left of violin plot) and male patients (right of violin plot), measured by global performance metrics and error rates.



3.4 Discussion

In this Chapter I have exposed an important gap in existing cardiac ML research, with significant implications for digital health equity. I find that the majority of published ML studies predicting heart failure fail to acknowledge the under-representation of female patients in their datasets, and do not perform stratified model evaluations, thus failing to assess sex disparities in algorithmic performance. The secondary evaluation of two cardiac datasets exposed a neglected sex disparity in model performance, highlighting the importance of integrating these methods into future studies that use ML methods for cardiac modelling. In this approach I identified several potential sources of algorithmic bias.

3.4.1 Sex representation & AI bias

First, I detected under-representation of females in training datasets that may produce inequalities in model fidelity. These findings are similar to studies of AI bias in other healthcare domains, where the lack of demographic representation in training data manifests in algorithmic performance inequities. In particular, this has been reported in dermatology algorithms and medical vision models that fail to use samples with diverse skin colours [82]. Unfortunately I could not assess for representation regarding other demographic characteristics (e.g. race or socio-demographic class) due to the omission of these labels in the original datasets. These findings support the research of authors such as Hung and colleagues, who have identified the inconsistent reporting of race and ethnicity in medical datasets as a barriers to effective algorithmic equity evaluations [169].

3.4.2 Feature rankings and demographic evaluation

In my evaluation of model feature selection I demonstrated significant differences in the feature rankings of models, that varied depending on the ranking methodology chosen. The findings unearthed that using Gini Importance, as a method specific to the RF model, demonstrated different insights to generic methods (e.g. Pearson correlation coefficients), highlighting that feature evaluations must be tailored to the specific model context. Furthermore, the feature rankings observed for "all patients", more closely reflected those rankings produced for the male sub-sample, likely relating to the over-representation of males in both datasets. The identified sex-differences in feature importance's mirrored that of the historic domain, indicating that these agents may act along sex specific pathways in the manifestation of disease. Given that these features appear to hold differing degrees of information for the separate sexes, it may be that sex-mixed training data contains conflicting information on the relationship between feature variables and the target outcome.

3.4.3 AI Bias in Cardiac Models

In the introduction I defined AI fairness as *"the absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics"* [88]. In this chapter I have examined ML fairness through the lens of Sex and identified an under-performance of the ML models for female patients, thus violating this initial definition of AI fairness.

My experiments demonstrated a higher prediction of disease amongst males and a higher overall performance for male patients. In my initial descriptive statistics I identified a lower presence of female disease in both datasets, firstly due to the lower number of overall female participants, and secondly due to the lower incidence of disease amongst the females. Consequently, both datasets contained less information on female illness and hence the ML models may have struggled to learn these pathways to the same degree of accuracy as the males. The evaluation of feature importance highlighted that sickness presents differently between the sexes, and thus the lack of information on how female sickness manifests in the datasets, may result in algorithms less effective in detecting female disease.

My approach to evaluating bias, namely using the Sex-disparity metric and examining inequities across a range of evaluative scores, proved effective. It is important to note that had I only focused on global performance metrics such as accuracy, I would have failed to identify the larger disparities in performance that were reflected in the False Negative Rate (FNR). These findings highlight the importance of selecting a fairness notion based on the clinical context in which a model is deployed, and not limiting bias evaluations to traditional performance scores. This adds to the work of Afrose and colleagues, who have also exposed the importance of examining demographic-specific error rate disparities when considering questions of ML fairness.

3.4.4 Avenues for further research

The pattern that identified in the performance disparities included an under-prediction of disease for females, and an over prediction of disease for males, likely related to a lack of representation of females, and sex-differences in feature importance and pathways to disease. In this Chapter I did not explore methods for addressing these biases, as this preliminary research was focused on exposing neglected biases that are present in published algorithms, however this avenue is explored in the Chapters below. Further, in this research chapter I focused specifically on RF Models, as I was examining the existing models reported in the literature. A range of supervised and machine learning models exist beyond this, some of which have been proposed as better options from a bias

perspective (E.g. XG Boost). Thus, in the following chapters I explore similar questions over a wider range of models.

3.4.5 Limitations

My experiments were limited by the datasets available, restricting the complexity of methodological approaches I could deploy. The overall number of patients was small, with only 194 Males and 105 Females Dataset 1, and 564 Males and 182 Females in Dataset 2. Compounded by the fact that these datasets were from the same source (the UCI Machine Learning repository), this limited the generalisability of my results. Additionally, the small dataset size precluded my ability to run additional tests into the impact of dataset representation (i.e. examining whether models would perform better on their sex-specific subsets). I dive into this issue in greater depth in the next chapter

Furthermore, I opted to focus on Random Forest models, as these had previously been cited to be the most effective for these datasets. In doing so, I narrowed the research scope and have not considered whether specific models are more prone to bias than others. Finally, in this section I did not implement methods for addressing algorithmic bias. In the following chapter I will now explore the utility of a range of fairness notions and approaches for addressing these issues.

3.4.6 Conclusion

In summary, this chapter exposed concerning biases in cardiac ML algorithms that result in worse predictions, and potentially health outcomes, for female patients. These findings underscore a critical need for ensuring the inclusivity of data when developing healthcare AI algorithms. Furthermore, I have demonstrated that the current practice of extrapolating findings from predominantly male data to the general population is inadequate and results in algorithmic performance inequities that reflect and perpetuate historic healthcare disparities. This research sets the stage for future studies into cardiac ML modelling and exposes specific challenges to the medical domain, such as the need for models to learn different patterns between features and the target outcome on the basis of demographic features.

Chapter 4

Tackling bias in cardiac algorithms

> For millennia, medicine has functioned on the assumption that male bodies can represent humanity as a whole. As a result, we have a huge historical data gap when it comes to female bodies, and this is a data gap that is continuing to grow as researchers carry on ignoring the pressing ethical need to include female cells, animals and humans, in their research.
>
> *Caroline Criado-Pérez (2019) [113]*

4.1 Introduction

In Chapter 3 I exposed disparities in the algorithmic performance of predictive models built from two open-source cardiology datasets: (i) The UCI Heart Failure (HF) Clinical Records Dataset and (ii) The UCI Coronary Artery Disease (CAD) Dataset [150, 151]. My review of the cardiology ML literature showcased that the issue of algorithmic performance disparities was largely neglected in existing studies and the widespread use of confidential and proprietary datasets precluded further evaluation of potential disparities in published models. For the models that I was able to replicate and evaluate from an equity standpoint, I demonstrated significant sex disparities in algorithm performance that disadvantaged female patients. These findings highlighted the importance of integrating bias analyses into algorithm development. Moving forwards, this chapter will dive into the potential methods for addressing inequities in algorithm performance. Here, I review the methods of algorithmic fairness that have been deployed in other domains and re-purpose these approaches for our issue of algorithmic bias in cardiology models.

Outside of academia, industry entities have paid increasing attention to the issue of AI fairness, releasing guidance and toolkits for improving algorithm development. IBM has created a public GitHub repository called AI Fairness 360, providing popular metrics for fairness and details of various bias mitigation approaches [170]. Similarly, Microsoft have launched a specialist group for fairness, accountability, transparency, and ethics in AI building a repository of ethically-orientated computational techniques [171]. These approaches often extend beyond questions of algorithmic fairness, to encompass additional ethical issues such as data and algorithm transparency, explainability and intepretability [170, 171]. For the purpose of this thesis, I am focused solely on technical solutions (at

the level of algorithm development) and questions of fairness [172]. The common methods utilised for this subdomain of AI ethics are reviewed here.

In the ML Fairness research, technical solutions to issues of algorithmic bias can be largely grouped into three categories dependant on which stage in the model development pipeline they are used [170, 171, 173]:

1. Pre-processing techniques
2. In-Processing techniques
3. Post Processing techniques [173].

Pre-processing techniques involve making modifications to the dataset before the algorithm is trained, with the goal of reducing biases that may be learned from the training data [88, 173]. Common approaches involve dataset resampling and modifying feature selection [88, 173]. In-processing techniques involve the integration of fairness techniques directly into the model development, such as the use of fairness constraints during training (e.g. regularisation terms that penalise unfair predictions) [88, 173]. Post-processing techniques are applied after a model has been built, and involve the adjustment of a model output to achieve fairness goals (e.g. adjusting thresholds for different subgroups to inform classification outputs) [88, 173]. Hort and colleagues provide a useful review of the range of methods that exist within each category and the models to which they are most suited [173].

Once an issue of algorithmic bias has been identified, the choice between pursuing pre-processing, in-processing or post-processing techniques to address the issue often depends on the type of ML model being used and the type of bias that needs addressing [88, 173, 174]. Pre-processing techniques are the most common for supervised ML models, particularly logistic regression, decision trees, random forests and support vector machines [88, 173, 174]. Common methods in this domain involve data resampling or dataset balancing, where the minority class is over-sampled in order to improve representation in the dataset. Alternatively researchers have examined the benefit of reweighing approaches, in which different weights are assigned to different samples based on their representation in the data [173, 174]. In contrast, complex ML models, such as deep learning algorithms, often require in-processing methods that embed fairness techniques into the training process [88, 173, 174]. Post-processing techniques are useful when alterations to the initial data or the model itself are impractical [88, 173, 174].

My research on cardiac ML models so far has focused on supervised ML techniques and random forests, for which pre-processing bias mitigation techniques have been proposed as the most effective [88, 173]. In this research I have access to the two UCI datasets, thus I can explore the efficacy for dataset preprocessing techniques on algorithm perfor-

mance and model disparities. Furthermore, the focus of this thesis is predominantly on fairness in model performance, whereby we are examining how to address the differential performance of models for different groups. Thus, the methods of post-processing, which are applied after a model has made it's predictions are less relevant to our work. In the following sections I detail the existing research on bias-mitigation that is most relevant to our domain, focusing on pre-processing techniques (fairness through unawareness, fairness-aware feature selection and data representation) and in-processing techniques (adversarial training) [88, 173].

4.1.1 Fairness through unawareness

One of the first methods explored for addressing AI bias was to remove or ignore a sensitive attribute in algorithm development, in the hope that this would prevent the model discriminating on the basis of this feature [2, 15, 88, 173, 174]. Unfortunately, this approach has been demonstrated to be largely ineffective, and sometimes harmful, due to the presence of proxy features in ML training data [2, 43]. For example, Barocas and colleagues explain how models can identify the gender of an individual from their search history and website preferences, even if gender is removed from the training data [43]. The authors refer to this as *redundant encoding*, where by the model is able to predict the sensitive attribute from other variables in large feature spaces [43]. The issue of redundant encodings for sensitive attributes has been demonstrated in healthcare. For example Poplin and colleagues trained deep-learning models on large clinical datasets and were able to predict age, gender, smoking status from retinal scans [175]. In another example, Banerjee and colleagues demonstrate how deep learning models can be trained to predict race from radiological images with high accuracy [176]. Interestingly, the authors noted that if the algorithm then secretly used this knowledge of race to misclassify patients from a specific racial group, radiologists with access to the same data that the model had been trained would remain unaware of this harm [176].

In my assessment of sex bias in cardiac ML algorithms I utilise the "Fairness through unawareness" approach, to ascertain whether the criticisms described above apply in the context of classifiers that predict cardiac disease. This is particularly important in healthcare where a patient's membership to a specific group may actually inform their health trajectory (e.g. Sex is associated with certain conditions). Thus, removing sensitive attributes from the dataset may reduce the ability of models to predict disease, potentially further disadvantaging members of marginalised groups. Furthermore, previous researchers have used this technique as a baseline measure in assessing various techniques for mitigating against bias [174]. I therefore include this approach in the following methods to assess for both it's harms and benefits in addressing algorithmic bias.

4.1.2 Fairness-aware feature selection

The analysis of sex-specific feature importance in Chapter 3 demonstrated the differential role of features across patient subgroups. Recently, this issue has gained attention in the ML fairness literature, with researchers advocating for "fair feature selection" in ML development [177]. Zawad and colleagues have highlighted that much of the ML fairness literature focuses on re-calibrating models and adjusting the preprocessing of training data, yet research into fairness in feature selection has been relatively nascent [177]. The authors go on to demonstrate the value of partitioning datasets by sensitive attributes (e.g. gender) and then applying a range of feature selection techniques for each partition, to improve the fairness of models [177]. Similarly, Afrose and colleagues demonstrated the value of group-specific feature rankings for correcting racial biases in predictive healthcare algorithms [18]. The team built models for predicting breast cancer and hospital mortality, and demonstrated that customised models built with an awareness of subgroup-specific feature importance reduced biases in algorithm performance [18].

In the background section and empirical research of Chapter 3, I demonstrated the differential value that various features had for the sexes separately. These findings reflected that of historic research, including (i) a lower predictive importance of Ejection Fraction for female patients, (ii) sex differences in the predictive power of biochemical markers (e.g. cholesterol). The first example is a clinical feature, one that is based on a measurement of physiological performance (EF is the percentage blood ejected from the heart in one beat). The second example refers to differences in biochemical features, these are the pathological markers derived from the bloods and other biological tissues, as described in Chapter 3. In this chapter I explore the topic of "fair feature selection" on algorithmic bias, focusing on these differences in biochemical and clinical features.

4.1.3 Dataset representation for fairness

A commonly identified source of AI bias is that of the lack of representation of a demographic subgroup in the training data [2, 15, 43]. Referred to as "Representation Bias", this issue arises from how we sample the population, whereby inequities in the data collection can result in a lack of diversity in datasets [88]. As a result, representation bias happens when the training data under-represents some parts of the target population and consequently fails to generalise well [178]. The issue can emerge due to selection bias where the sampling method only reaches a portion of the population, or if the population of interest changes significantly from that used in model training [178]. For example, in medical datasets a commonly under-represented group is pregnant patients as they are less likely to be involved in clinical research and trials, thus consequently the models that result form such research may be less robust for this group [48, 179].

In the previous chapters I have reviewed the issue of representation in medicine, especially in cardiology where the research reviewed in Chapter 3 highlighted the lack of attention given to female physiology and biology [133, 134]. The issue has been demonstrated across a wider range of domains, with researchers highlighting the lack of diversity in medical image datasets leading to biases that affect patients with darker skin [82]. To address biases manifesting from poorly representation in datasets, fairness researchers have proposed a range of approaches [178]. Shahbhazi and colleagues detail the landscape of techniques for addressing representation bias, for datasets formed from both structured (e.g. tabular) and unstructured (e.g. image) data [178]. Amongst their proposed remedies, the authors highlight that the most beneficial approach may be to add more data through enhanced collection processes, however depending on context this may not be possible [178]. In cases where the data collection process cannot be revisited, the available dataset can be resampled to reduce the presence of the majority group, or researchers can pursue oversampling approaches creating synthetic examples of the minority group [178]. I dive into these options in greater depth in the methods section of this chapter.

4.1.4 Adversarial training for fairness

One in-processing technique is "Adversarial Debiasing", in which the predictive ML model, and an adversarial model, are trained at the same time to ensure the sensitive attribute cannot be predicted - thus, attempting to integrate fairness in model training [173]. Described in depth by Zhang and colleagues, the method involves building a model in which one maximises the predictor's ability to predict Y (the target outcome), while minimising an adversary's ability to predict Z (the sensitive feature) [180]. The authors demonstrate the effectiveness of this approach, in their models built to predict income from the UCI adult census dataset [180]. The team produce a predictive model that maintains high levels of accuracy with regards to predicting income, while also reducing inequities in error rates across demographic groups [180].

This "Fairness through adversarial training" approach stands apart from many of the other fairness methods summarised in the introduction and presented in Table 1.2. In contrast to methods that attempt to remove the sensitive attribute from the dataset (e.g. Fairness through unawareness), or resample the dataset to upregulate the presence of the minoritised group (Fairness through representation), adversarial training instead specifically targets the machine learning process and the learnt relationship between the target variable and the sensitive attribute. Such an approach is particularly interesting for the healthcare context, where sensitive attributes may play a role in disease pathways (e.g. the biological effects of sex), and thus training against the attribute could be

counter-intuitive. Furthermore, existing research into RF models used for this purpose have focused on pre-processing bias mitigation techniques - here we build on this research, integrating methods of adversarial training and examining its applicability to healthcare.

Chapter Aim

In this chapter I apply these identified ML fairness techniques to our problem of algorithmic bias in cardiac ML models, and assess their efficacy in reducing performance disparities. I have opted for the range of techniques detailed above, meaning that I do not cover the full range of fairness methods. I expand on this further in the discussion section and following chapter.

4.2 Methods

In this Chapter I build on the methods of Chapter 3, integrating a range of bias mitigation approaches into the algorithm design, and evaluating the resulting impact on algorithmic performance disparities. My focus here is still on the two datasets discussed in depth in Section 3.2.2; (i) Dataset 1: The UC Irvine (UCI) Machine Learning Repository Dataset for Heart Failure Prediction [150] and (ii) Dataset 2: The UC Irvine (UCI) Machine Learning Cleveland Heart Disease dataset for identifying Coronary Artery Disease (CAD) [151].

Figure 4.1 provides an updated image of the flowchart from Figure 3.1 in Chapter 3, now detailing the additional steps of our bias mitigation approaches that have been added to the previous methods. As demonstrated in Figure 4.1 below, these bias mitigation techniques are divided into:

1. **(i) Changes made to training data**
2. **(ii) Changes made to feature sets**
3. **(iii) Adversarial training (FAGTB Technique)**.

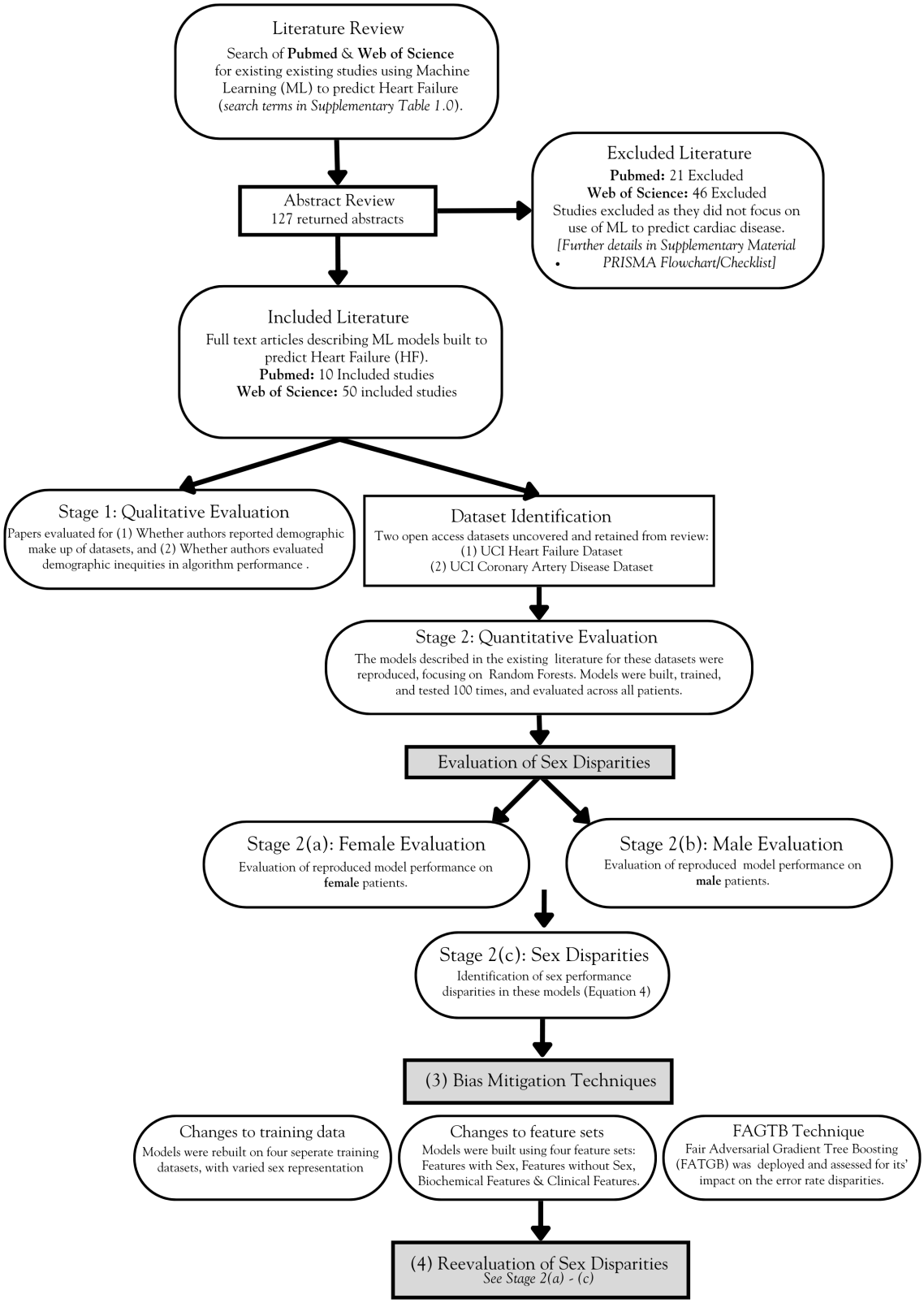


Figure 4.1: Methods of Chapter 4: An updated flowchart based on Figure 3.1 from Chapter 3, now including the added the steps to our methodology of bias mitigation

Pre-processing Techniques: Changes to Training Data

One widely proposed bias mitigation technique includes pre-processing the training data of a model to account for demographic representation [18]. Prior research has demonstrated that model fairness can be improved when training is performed on demographically balanced or demographically stratified datasets [18]. My first approach therefore was to create a range of training datasets that varied in their sex representation, and to assess the impact of this on performance disparities.

For both Dataset 1 and Dataset 2, I created four training datasets:

1. Original Sex-Imbalanced Training Data
2. Sex-Balanced Training Data
3. Female-only Training Data
4. Male-only Training Data

To form the sex-balanced dataset, I utilised the oversampling function of SMOTE(), which has been proposed as an effective method for improving representation of underserved populations in machine learning datasets [181]. The SMOTE package generates new minority data points, based on existing minority samples, through linear interpolation [18, 181]. This function was used to oversample the females in the dataset, to bring their representation up to the level of the males (Tables 4.1 to 4.2). For SMOTE, the main hyperparameter is the k-nearest neighbours (k-NN) parameter, which specifies the number of nearest neighbours to a data point to consider, when generating synthetic samples for the minority class [181]. We used the default number of 5 for this parameter, which is described in the package documents. It is worth noting that in datasets where the minority class is very sparse, or highly diverse, using only 5 neighbours can lead to a loss of local structure and a failure to capture the distribution of the minority class (potentiating overfitting). We explore these challenges further in the discussion.

To create the sex-specific datasets (e.g. Female only) the data for one sex was then extracted from the balanced dataset (Table 4.1 and Table 4.2). I maintained the original sex-imbalanced dataset for comparison. The data counts for each of the new training datasets are provided in Table 4.1 and Table 4.2, for Dataset 1 and Dataset 2 respectively. Note, for Dataset 2, I have also listed the values for the original Cleveland Dataset discussed in Chapter 3, and the combined dataset across five hospitals which we opted to use, plus the case counts after duplicates were removed to give the final training data case counts (further detailed in Section 3.2.2 of the Methods).

Table 4.1: Case counts for the training data subsets of Dataset 1 (Heart Failure). *Nb. Unlike in Table 4.2 below, no erroneous or duplicate records needed to be removed*

Sex	HF Death	Original unbal. dataset	Balanced Dataset	Female only	Male only
0	0	71	153	153	X
0	1	34	41	41	X
1	0	132	132	X	132
1	1	62	62	X	62
Total		299	388	194	194
Total Training (n*0.7)		209	272	136	136

Table 4.2: Case counts for Dataset 2 (Coronary Artery Disease) training data subsets. For sex specific training samples the data was the sex subset of the balanced training data, i.e. for females 248 well & 263 unhealthy. *Nb. Erroneous values included 172 instances Cholesterol = 0, and 1 instance Resting blood pressure = 0*

Sex	CAD Diagnosis	Original Cleveland Dataset	Dataset 2 (Combined dataset)	Dataset 2, (Duplicates removed)	Final Unbalanced Training Dataset	Final Balanced Training Dataset
0	0	72	211	143	142	248
0	1	25	70	50	40	263
1	0	92	350	267	248	248
1	1	114	559	458	316	263
Total Data		303	1190	918	746	1022
Total Training (n*0.7)		212	833	643	522	715

Pre-processing Techniques: Changes to Feature Selection

To understand why models make certain decisions, researchers in the domain of Explainable AI have demonstrated how feature evaluation may provide important information regarding model performance for different subpopulations [18, 157]. In Chapter 3 we introduced Shapley Values, which have been widely accepted as a unified measure of feature importance since their proposal in 2017 [156]. Our analysis of feature rankings by Shapley Values, Correlation Coefficients and Gini Importance demonstrated the differential value of specific clinical measurements (E.g. Ejection Fraction) and biochemical features (e.g. Cholesterol) for the males and female subsets (Figure 3.2 to Figure 3.7). Our findings here reflected the wider medical research domain that has described that biochemical markers may have different predictive power for each sex, and clinical measurements of cardiovascular status (e.g. blood pressure, Ejection Fraction) differ between the sexes over the course of cardiac disease [131, 134]. In this section I use this information to delineate four different feature subsets that vary in this feature information, to examine whether certain feature subsets perform better for different demographic groups. We compare the impact of using the clinical features, biochemical markers, and the full feature set (Table 4.3 to Table 4.4). Further, I also created a feature subset that does not include Sex, thus deploying the "Fairness through unawareness" approach discussed in the introduction. The four resulting feature subsets for each Dataset are detailed in Table 4.3 and Table 4.4 below.

Table 4.3: Feature Subsets for Dataset 1 (Heart Failure)

Features with Sex	Features without Sex	Clinical Features	Biochemical Features
Sex	Age	Anaemia	CPK
Age	Smoking	Diabetes	Serum Creatinine
Anaemia	Anaemia	Ejection Fraction	Platelets
CPK	CPK	High Blood Pressure	Serum Sodium
Diabetes	Diabetes		
Ejection Fraction	Ejection Fraction		
High blood pressure	High blood pressure		
Platelets	Platelets		
Serum Creatinine	Serum Creatinine		
Serum sodium	Serum sodium		
Smoking			

CPK = Creatinine Phosphokinase, full details of feature available in Table 3.1

Table 4.4: Feature Subsets defined for Dataset 2 (Coronary Artery Disease)

Features with Sex	Features without Sex	Clinical Features	Biochemical Features
Cholesterol	Cholesterol	Chest pain type	Cholesterol
Fasting blood sugar	Fasting blood sugar	Resting BP	Fasting blood sugar
Age	Age	Resting ECG	
Chest pain type	Chest pain type	Max Heart Rate	
Resting BP	Resting BP	Exercise Angina	
Resting ECG	Resting ECG	Old peak	
Max Heart Rate	Max Heart Rate	ST Slope	
Exercise Angina	Exercise Angina		
Old peak	Old peak		
ST Slope	ST Slope		
Sex			

Resting BP = Resting Blood Pressure, full details of feature available in Table 3.2

Variations in Model Development

Our new set of training data and feature subsets were then used to run multiple experiments, exploring performance across all permutations that combined each training subset with each feature subset. The models were rebuilt as per Section 3.2.2 of Chapter 3, represented in Figure 4.1, splitting training data randomly in each experiment into 70% training and 30% test subsets. The methodology of Chapter 3 was reused to assess the consistency of our results and to identify performance disparities, such that models were built, trained and tested over 100 runs, with the average performance metrics calculated with standard deviation. Our final series of experiments were therefore performed across the four training datasets (sex-imbalanced, sex-balanced, female only and male only), and the four feature sets giving 16 total experiments:

1. **Experiments 1 - 4:** Original Imbalanced Training Data Experiments (across four feature subsets)
2. **Experiments 5 - 8:** Balanced Training Data Experiments (across four feature subsets)
3. **Experiments 8 - 12:** Female Training Data Experiments (across four feature subsets)
4. **Experiments 12 - 16:** Male Training Data Experiments (across four feature subsets)

Re-evaluation of Sex Disparities

The sex disparities in algorithm performance were re-calculated for each performance metric, across the sixteen experiments detailed above that utilised variations in feature subset and training data. As detailed in Chapter 3, models were evaluated using global evalu-

ation metrics (e.g. Accuracy) and specific error rates (e.g. False Negative Rate [FNR]) (Table 3.3). The difference between male and female performance scores were calculated to give the models "Sex performance disparity" outlined in Chapter 3 (Equation 3.3). To evaluate for statistical significance of differences identified across the 100 experimental runs, Kolmogorov-Smirnov Tests were used to assess for normality of the data, following which independent t-tests were performed where the data was normally distributed, and Mann-Whitney U tests were performed where the data was not normally distributed [159]. These average performance disparity was calculated for each performance metric, as previously described in Chapter 3.

Fair Adversarial Gradient Tree Boosting (FAGTB)

The final approach I explored for addressing algorithmic bias was the in-processing technique of adversarial training. I implemented the fairness technique of Fair Adversarial Gradient Tree Boosting (FAGTB), proposed by Grari and Colleagues for mitigating bias in decision tree classifiers [182]. In their article the authors propose using a fairness regulariser that aims to remove correlation between the sensitive attribute and the target value [182]. The objective of the model is to predict the target with gradient tree boosting, while minimizing the ability of an adversarial neural network to predict the sensitive attribute [182]. They authors apply these methods across four datasets, spanning the domains of income prediction (the Adult UCI Dataset), criminal justice (the COMPAS Dataset) and financial services (a credit defaulting and bank marketing dataset). Their approach has not been applied in healthcare, thus here I used these methods to evaluate the effect on reducing disparities in the cardiac ML algorithms.

In their attempts at achieving fairness, the authors focused on "Demographic Parity", for which a classifier is considered fair if the prediction Y from features X is independent from the protected attribute S [182]. The authors define multiple ways to assess this objective, here I focus on the use of DispFNR and DispFPR, which mirror our approach of evaluating disparities in False Negatives and False Positives. The metrics of DispFNR and DispFPR are defined in Equations 4.4 to 4.4 below.

$$P(\hat{Y} = 1 \mid S = 0) = P(\hat{Y} = 1 \mid S = 1) \tag{4.1}$$

Figure 4.2: Equation for Demographic Parity [182]

$$D_{\mathbf{FPR}} : \left| P(\hat{Y} = 1 \mid Y = 0, S = 1) - P(\hat{Y} = 1 \mid Y = 0, S = 0) \right| \tag{4.2}$$

Figure 4.3: Equation for Disparate False Positive Rate (DispFPR) [182]

$$D_{\mathbf{FNR}} : \left| P(\hat{Y} = 0 \mid Y = 1, S = 1) - P(\hat{Y} = 0 \mid Y = 1, S = 0) \right| \tag{4.3}$$

Figure 4.4: Equation for Disparate False Negative Rate (DispFNR) [182]

To implement this fairness technique I replicated the same methods of the original paper, building both a Gradient Boosting Classifier and the FAGTB model to predict the cardiac outcomes in Dataset 1 and 2 [182]. As per the original paper, I repeat 10 experiments by randomly sampling two subsets (80 training and 20 test set) and report the accuracy and fairness metrics for the test set. The Fairness metrics include "Disparate Mistreatment", "Disparity FNR" and "Disparate FNR". In keeping with the rest of our paper, I focus on the Disparate FNR. The closer the values of DFPR and DFNR to 0, the lower the degree of disparate mistreatment of the classifier.

4.3 Results

I begin by providing an overview of the performance disparities identified across the set of 16 experiments, with all variations in training data and feature selection, assessing whether sex-disparities in performance persisted despite the changes. The results are then broken down into the order of the bias-mitigation techniques described, progressing through the pre-processing techniques (changes to training data and feature selection) to the in-processing techniques (integration of adversarial training).

4.3.1 Re-evaluation of performance disparities: Dataset 1

For Dataset 1, Table 4.5 demonstrates that in 13 out of 16 experiments the False Negative Rate (FNR) was higher for females, meeting the threshold of statistical significance (mean difference of -17.81% to -3.37%, $p<0.05$). Figure 4.5 represents this disparity in performance graphically, providing the point estimates of the FNR for the Sexes separately, highlighting that the disparity in FNR persisted across the variations in training data and selected features. On Figure 4.5, the average FNR scores for the females (left side of the violin plot) can clearly be seen to sit above the average scores for the males (right side of violin plot), across the 16 experiments which vary in training data and feature subset. Thus, here we are seeing that despite these adaptations to model development, the sex-based disparities in algorithm performance persist. A smaller disparity in the False Positive Rate (FPR) was statistically significant for males in 13 out of 16 experiments (-0.48% to +9.77%, $p<0.05$) (Table 4.5). On examining the individual error rates, we see consistencies in the sex disparities across feature sets, most notably an over-prediction of disease for males (higher FPR) and an under prediction of disease for females (higher FNR - Table 4.5).

The sex performance disparities in the global performance metrics of Accuracy and ROC varied depending on the underlying shifts in the error rates for each sex (Table 4.5 and Figure 4.6). In the original models trained on sex-imbalanced data the Accuracy was marginally higher for males (84.8% males vs 83.2% females), yet the trend reversed when trained on the sex balanced and sex specific datasets. We explore this in more depth in our review of pre-processing techniques below.

Table 4.5: Sex performance disparities for models built from Dataset 1 (Heart Failure Disease) – Disparities calculated as performance for males minus performance for females. Asterisks (*) indicate statistical significance.

Model Performance Disparity (%)	**Features With Sex**	**Features Without Sex**	**Biochemical Features**	**Clinical Features**
		Sex-Imbalanced Training Data		
Acc Disparity	*1.63 (0.03)	-0.72 (0.30)	0.10 (0.88)	-0.50 (0.49)
ROC Disparity	*3.14 (<0.01)	0.43 (0.61)	1.51 (0.09)	0.47 (0.60)
FNR Disparity	*-7.53 (<0.01)	*-3.84 (0.02)	*-5.15 (0.01)	*-3.49 (0.049)
FPR Disparity	1.26 (0.07)	*2.97 (<0.01)	*2.11 (<0.01)	*2.56 (<0.01)
		Sex-Balanced Training Data		
Acc Disparity	*-4.78 (<0.01)	*-7.25 (<0.01)	*-9.42 (<0.01)	*-3.63 (<0.01)
ROC Disparity	*7.0 (<0.01)	*4.27 (<0.01)	0.15 (0.83)	*8.32 (<0.01)
FNR Disparity	*-17.81 (<0.01)	*-13.91 (<0.01)	*-3.37 (0.04)	*-16.09 (<0.01)
FPR Disparity	*3.90 (<0.01)	*5.37 (<0.01)	*3.07 (<0.01)	-0.54 (0.24)
		Female Training Data		
Acc Disparity	*-10.95 (<0.01)	*-9.75 (<0.01)	*-12.32 (<0.01)	*-9.64 (<0.01)
ROC Disparity	0.60 (0.57)	0.57 (0.23)	*-2.92 (<0.01)	-0.53 (0.07)
FNR Disparity	*-7.42 (<0.01)	*-10.91 (<0.01)	-2.24 (0.27)	*1.55 (0.01)
FPR Disparity	*8.61 (<0.01)	*9.77 (<0.01)	*8.08 (<0.01)	*-0.48 (0.04)
		Male Training Data		
Acc Disparity	*-5.46 (<0.01)	*-5.73 (<0.01)	*-8.73 (<0.01)	*-2.46 (<0.01)
ROC Disparity	*4.98 (<0.01)	*4.54 (<0.01)	*-1.59 (0.049)	*8.32 (<0.01)
FNR Disparity	*-13.96 (<0.01)	*-13.32 (<0.01)	-1.68 (0.33)	*-16.58 (<0.01)
FPR Disparity	*4.00 (<0.01)	*4.24 (<0.01)	*4.86 (<0.01)	-0.06 (0.35)

Figure 4.5: Dataset 1 (Heart Failure): A series of violin plots showing the sex stratified performance (False Negative Rate [0-100%]) of the Random Forests trained across the four feature sets, on the different variations in training data. The plots show male (orange) and female (grey) FNR alongside each other, in groups of four (divided by a line) according to the training data used (Sex-Imbalanced, Sex-Balanced, Female & Male). The Feature Set used is indicated within each training data group (Features with Sex, Features Without Sex, Biochemical Features & Clinical Features)

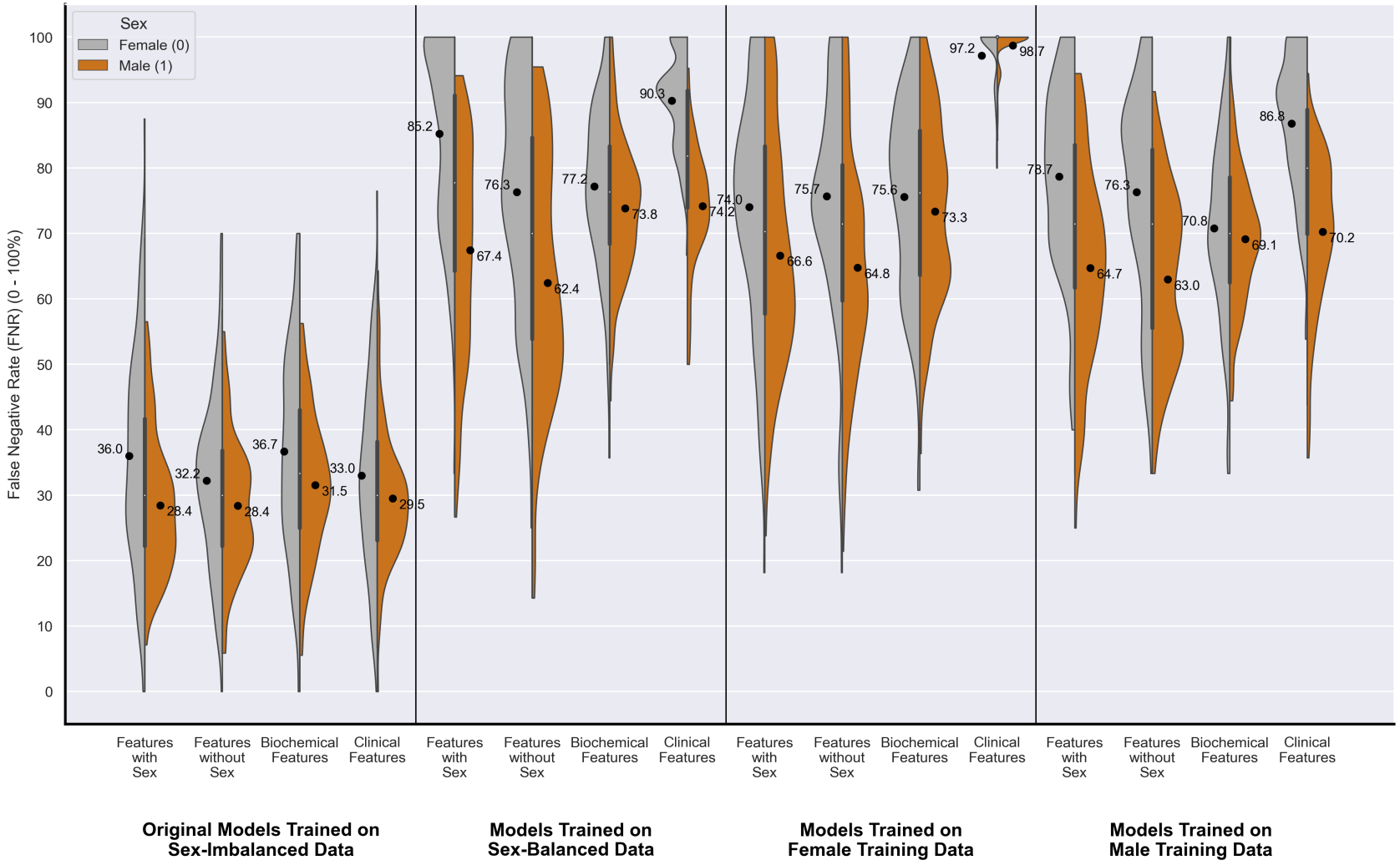
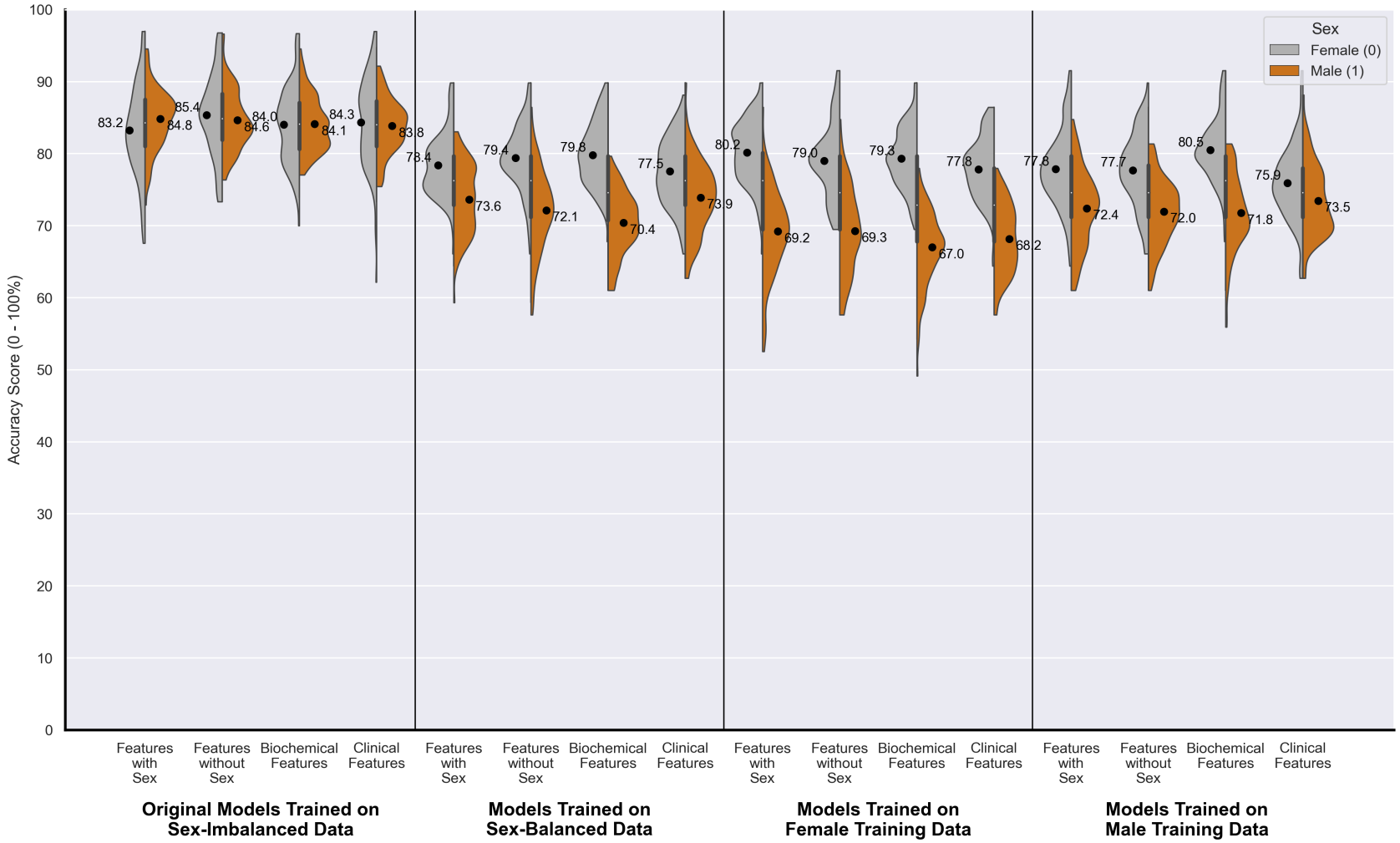


Figure 4.6: Dataset 1 (Heart Failure): A series of violin plots showing the sex stratified performance (Accuracy [0-100%]) of the Random Forests trained across the four feature sets, on the different variations in training data. The plots show male (orange) and female (grey) Accuracy alongside each other, in groups of four (divided by a line) according to the training data used (Sex-Imbalanced, Sex-Balanced, Female & Male). The Feature Set used is indicated within each training data group (Features with Sex, Features Without Sex, Biochemical Features & Clinical Features)



4.3.2 Re-evaluation of performance disparities: Dataset 2

The findings for Dataset 2 were similar to those for Dataset 1, such that models built on the original sex-imbalanced dataset demonstrated a higher FNR for females (mean difference of -10.81% to -12.52%, $p<0.05$) and a higher FPR for males (3.94% to 4.71%, $p<0.05$) (See Table 4.6). Figure 4.7 visualises the disparity graphically, and demonstrates that, unlike Dataset 1, the disparity in error rates reversed when training on sex-balanced data and female-only data (Figure 4.7). Figure 4.8 illustrates the disparity in Accuracy between the Sexes, where we see that the direction of the disparity varies dependant on the training data and feature set, explore in greater depth below (Figure 4.8).

Table 4.6: Sex performance disparities for models built from Dataset 2 (Coronary Artery Disease) – Disparities calculated as performance for males minus performance for females. Asterisks (*) indicate statistical significance.

Model Performance Disparity (%)	Features With Sex	Features Without Sex	Biochemical Features	Clinical Features
Sex-Imbalanced Training Data				
Acc Disparity	0.32 (0.50)	0.64 (0.17)	0.13 (0.8)	0.25 (0.61)
ROC Disparity	*3.86 (<0.01)	*4.24 (<0.01)	*3.05 (<0.01)	*3.91 (<0.01)
FNR Disparity	*-11.66 (<0.01)	*-12.52 (<0.01)	*-10.81 (<0.01)	*-12.38 (<0.01)
FPR Disparity	*3.94 (<0.01)	*4.04 (<0.01)	*4.71 (<0.01)	*4.57 (<0.01)
Sex-Balanced Training Data				
Acc Disparity	*-4.01 (<0.01)	*-5.12 (<0.01)	*-7.32 (<0.01)	*-2.86 (<0.01)
ROC Disparity	*-3.89 (0.01)	*-4.91 (0.01)	*-7.18 (<0.01)	*-2.75 (<0.01)
FNR Disparity	*7.69 (<0.01)	*10.54 (<0.01)	*15.59 (<0.01)	*6.61 (<0.00)
FPR Disparity	0.10 (0.87)	-0.72 (0.19)	-1.23 (0.29)	-1.11 (0.06)
Female Training Data				
Acc Disparity	*-9.25 (<0.01)	*-11.34 (<0.01)	*-11.49 (<0.01)	*-8.69 (<0.01)
ROC Disparity	*-8.97 (<0.01)	*-10.95 (<0.01)	*-11.10 (<0.01)	*-8.45 (<0.01)
FNR Disparity	*18.98 (<0.01)	*22.60 (<0.01)	*27.23 (<0.01)	*17.86 (<0.01)
FPR Disparity	-1.04 (0.07)	-0.70 (0.20)	*-5.02 (<0.01)	-0.96 (0.09)
Male Training Data				
Acc Disparity	*6.38 (<0.01)	*5.66 (<0.01)	*-1.66 (0.02)	*6.10 (<0.01)
ROC Disparity	*6.30 (<0.01)	*5.57 (<0.01)	*-1.59 (0.049)	*8.32 (<0.01)
FNR Disparity	*-13.96 (<0.01)	*-13.32 (<0.01)	-1.68 (0.33)	*-16.58 (<0.01)
FPR Disparity	*4.00 (<0.01)	*4.24 (<0.01)	*4.86 (<0.01)	-0.06 (0.35)

Figure 4.7: Dataset 2 (Coronary Artery Disease): A series of violin plots showing the sex stratified performance (False Negative Rate [0-100%]) of the Random Forests trained across the four feature sets, on the different variations in training data. The plots show male (orange) and female (grey) FNR alongside each other, in groups of four (divided by a line) according to the training data used (Sex-Imbalanced, Sex-Balanced, Female only & Male Only), and the respective feature subsets.

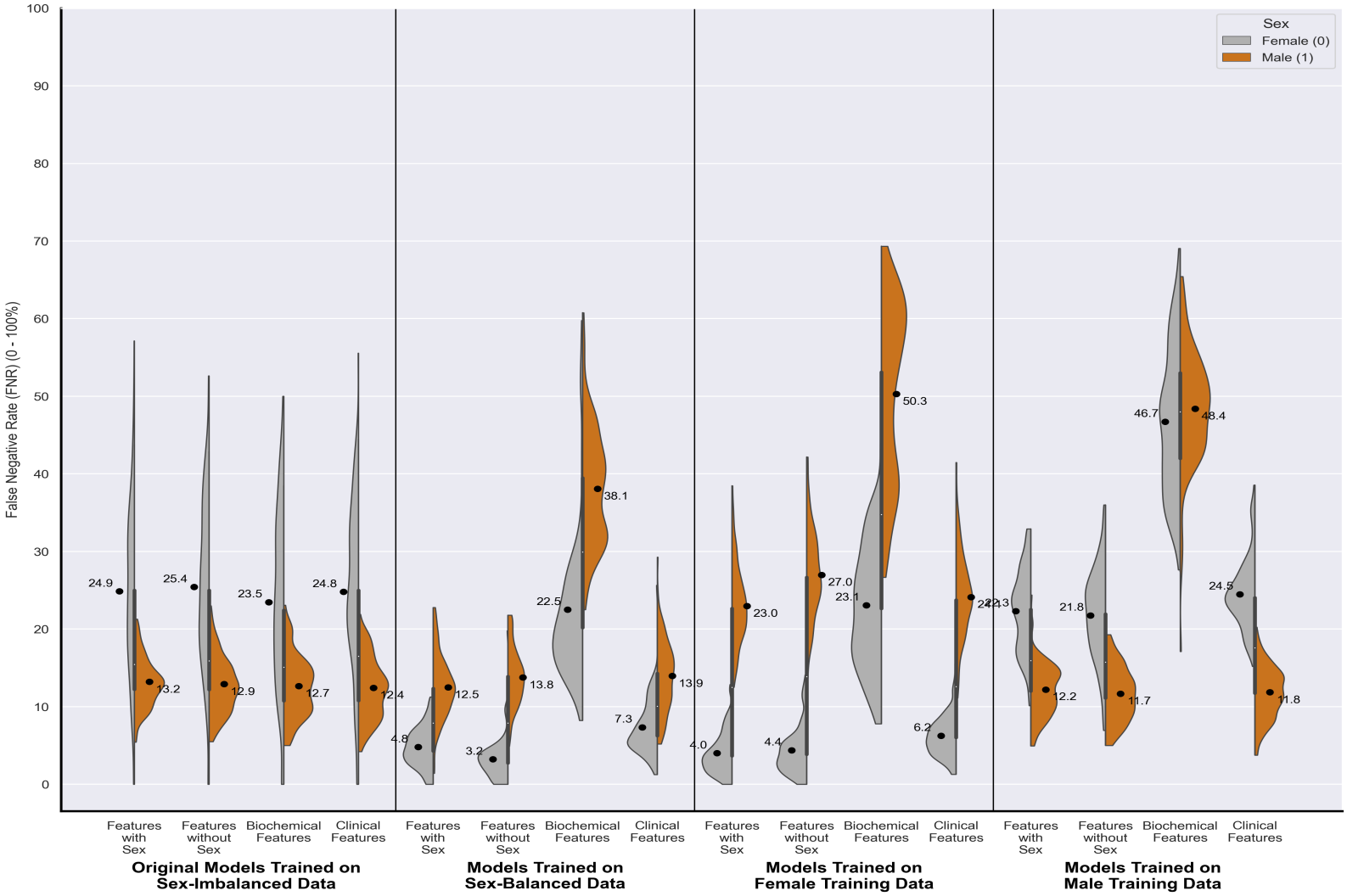
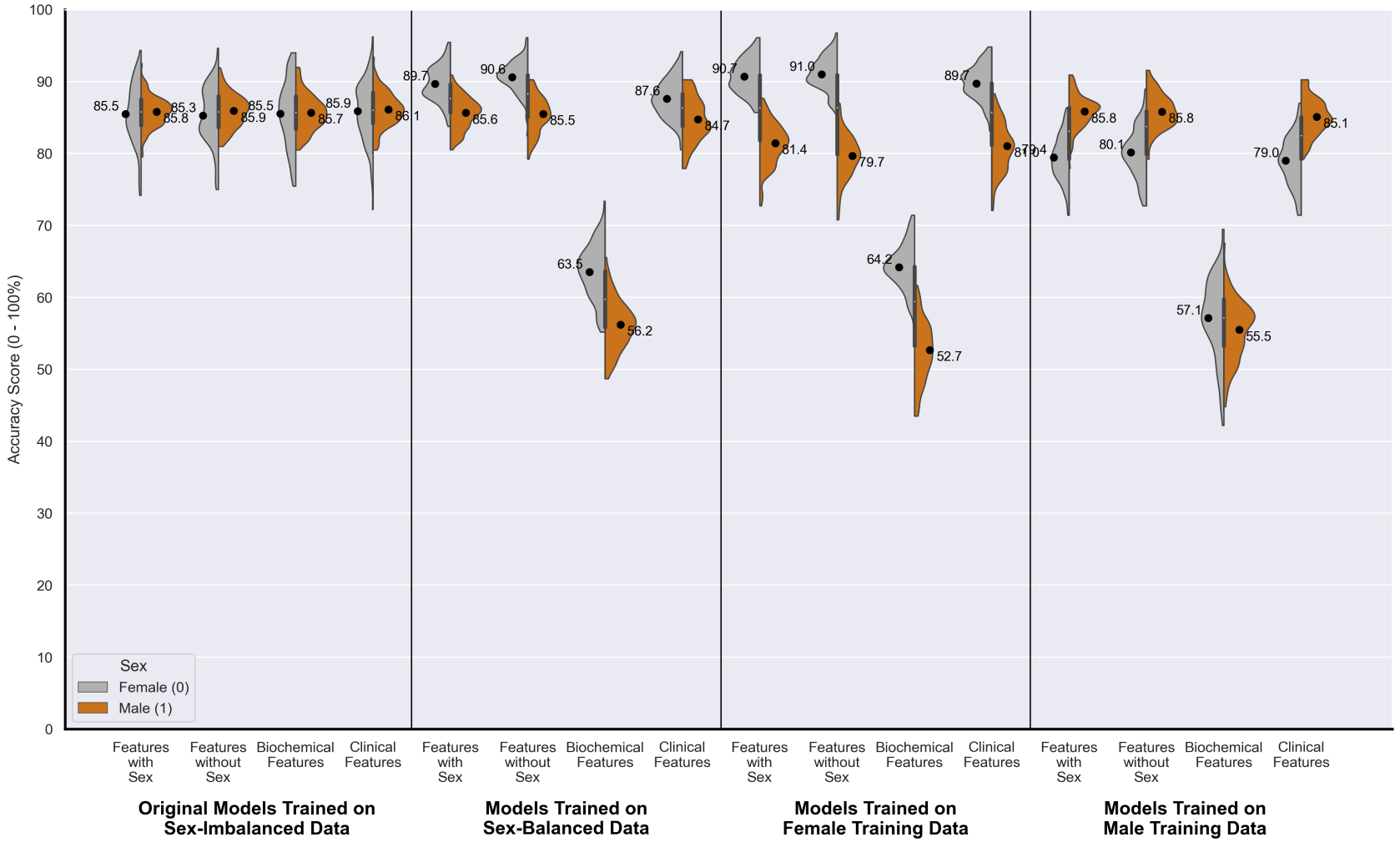


Figure 4.8: Dataset 2 (Coronary Artery Disease): A series of violin plots showing the sex stratified performance (Accuracy [0-100%]) of the Random Forests trained across the four feature sets, on the variations in training data. The plots show male (orange) and female (grey) Accuracy alongside each other, in groups of four (divided by a line) according to the training data used (Sex-Imbalanced, Sex-Balanced, Female only & Male Only), and the respective feature subsets.



4.3.3 Pre-processing techniques: Changes to training data

Training on sex balanced data led to a fall in mean accuracy for all patients in Dataset 1 (76.0% (3.46 SD) vs. 84.24% (3.51 SD)), with a more substantial drop in mean accuracy for males (73.61% (4.84 SD) vs. 84.84% (4.16 SD)) (Table 3, Figure 5). The opposite trend was seen in Dataset 2, with models trained on balanced data outperforming models trained on imbalanced data for all patients (87.65% (1.77 SD) vs. 85.72% (1.75 SD)) and for females (89.66% (2.44 SD) vs. 85.48% (4.12 SD)) (Table 3). The models trained on balanced data in Dataset 2 reduced the FNR for both sexes (Females 4.79% (2.58 SD) vs 24.86%, 11.35 SD; Males 12.48% (4.11 SD) vs. 13.19% (3.26 SD)) (Table 3). The differences between the datasets may relate to underlying differences in the two cardiac conditions. Further, the failure to improve performance with balanced training data may reflect the issues of mixing data that has conflicting indicators for disease, discussed further in Section 3.5.

Table 4.7: Model results for sex-specific subsets, looking at the Features Including Sex subset.

	Dataset 1 (Heart Failure)				Dataset 2 (Coronary Artery Disease)			
	Imbalanced Data (n=209)	Balanced Data (n=272)	Female Data (n=136)	Male Data (n=136)	Imbalanced Data (n=522)	Balanced Data (n=715)	Female Data (n=358)	Male Data (n=358)
All Patients, Acc. (SD)	84.24 (3.51)	76.0 (3.46)	74.68 (3.53)	75.12 (3.71)	85.72 (1.75)	87.65 (1.77)	86.06 (1.67)	82.63 (1.94)
Females, Acc. (SD)	83.21 (6.37)	78.39 (19.68)	80.15 (4.43)	77.85 (5.21)	85.48 (4.12)	89.66 (2.44)	90.69 (2.38)	79.44 (3.20)
Males, Acc. (SD)	84.84 (4.16)	73.61 (4.84)	69.20 (5.96)	72.39 (5.32)	85.80 (2.14)	85.65 (2.23)	81.44 (3.02)	85.82 (2.30)
Females, FNR (SD)	35.98 (16.72)	85.25 (14.58)	74.04 (17.68)	78.66 (14.0)	24.86 (11.35)	4.79 (2.58)	4.00 (2.74)	22.32 (5.25)
Males, FNR (SD)	28.45 (10.41)	67.43 (16.6)	66.62 (17.32)	64.70 (14.9)	13.19 (3.26)	12.48 (4.11)	22.97 (5.20)	12.20 (3.41)

Pre-processing techniques: Sex Specific Training Data

For Dataset 1, mean accuracy for all patients when trained on imbalanced data (84.24%, 3.51SD) falls when training both on female specific data (74.68%, 3.53SD) and male specific training data (75.12%, 3.71SD), likely related to the smaller training data. For Dataset 2, mean accuracy for all patients when trained on imbalanced data (85.72%, 1.75SD) improves when training on female specific data (86.06%, 1.67SD) and falls when training on male specific training data (82.62%, 1.94SD) (Table 4.7. The overall improve-

ment seen in the Dataset 2 models when trained on female data, relates to the increase in accuracy for females (91.36% 2.32 SD, vs. 85.48%, 4.12SD) co-occurring with a smaller decrease in accuracy for males (81.44%, 3.02, vs. 85.80%, 2.14SD) (Table 3). Unsurprisingly, performance for each sex was lowest when trained on the opposing sex (Table 4.7). In Dataset 1, same-sex training was preferable to opposite-sex training, however, did not improve results compared to the models built from imbalanced and balanced training data, likely relating to the smaller sample size (Table 4). In contrast, Dataset 2 had greater training data available and demonstrated that sex specific training is beneficial to both sexes above the imbalanced models (Table 4.7).

Pre-processing techniques: Changes to feature sets

Models built on the biochemical features subset gave the worst performance in terms of Accuracy and Error Rates (Figures 4.5 and 4.7). For Dataset 2 biochemical features included just Cholesterol and Fasting Blood Sugar, and so the fall in performance may relate to information loss. Additionally, Table 3.2 highlights the different biochemical profiles for sick males and females, with sick females demonstrating a far higher Cholesterol level (mean values; 274.54 Female Sick vs. 248.54 Male Sick).

In-processing techniques: Adversarial training

Table 4.8 compares the performance of the Fair Adversarial Gradient Tree Boosting (FAGTB) algorithm to the standard Gradient Boosting Classifier for both datasets, highlighting the global performance metrics (e.g. Accuracy) and the disparity metrics reported for this model: the Disparate False Positive Rates (DispFPR) and Disparate False Negative Rates (DispFNR). For Dataset 1, both models achieve a similar accuracy (Gradient Boosting 71.3% vs FAGTB 71.2%), suggesting that the fairness constraints present within the FAGTB Model had little effect on overall performance. The DispFPR is consistent across both models at 0.08, and the DispFNR only shows a slight improvement with the FAGTB Model (0.20 FAGTB Model vs 0.21 Gradient Boosting Algorithm) (Table 4.8).

For Dataset 2, there was as notable drop in the overall performance with accuracy falling from 86.3% with the Gradient Boosting Model, to 82.9% for the FAGTB Model. Similarly to Dataset 1, we see that the DispFPR persists across both models (0.06 Gradient Boosting Classifier vs. 0.06 FAGTB Model). In the case of the FNR, we see that for Dataset 2 the FAGTB shows a substantial improvement (0.19 FAGTB Model vs 0.28 Gradient Boosting Classifier).

In summary, despite the integration of the adversarial training with the FAGTB Model, the DispFNR remainined consistently higher than the DispFPR, affecting female patients.

Compared to the Gradient Boosting Classifier, the FAGTB reduced the DispFNR for both datasets (0.20 vs 0.21, Dataset 1; 0.19 vs 0.28, Dataset 2), however the DispFNR that disadvantaged female patients still persisted. The fall in DispFNR and DispFPR that occurred with FATGB was associated with a fall in overall accuracy for both datasets.

Table 4.8: Results of Bias Mitigation with Fair Adversarial Gradient Tree Boosting (FAGTB)

Evaluation Metric	**Gradient Boosting Classifier**	**FAGTB Model**
Dataset 1: Heart Failure		
Accuracy	71.3%	71.2%
DispFPR	0.08	0.08
DispFNR	0.21	0.20
Dataset 2: Coronary Artery Disease		
Accuracy	86.3%	82.9%
DispFPR	0.06	0.06
DispFNR	0.28	0.19

4.4 Discussion

In this Chapter I have explored the utility of a range of fairness notions that have been proposed in the ML fairness literature for addressing the inequities in algorithmic performance identified in Chapter 3. In doing so, I have identified challenges specific to the space of healthcare AI, and identified issues in these fairness approaches which are particularly applicable to researchers examining ML equity in medical models. We now explore these concepts in the order of fairness notions applied to our problem of inequity in cardiac ML.

4.4.1 Fairness through representation

Our analysis in Chapter 3 exposed an under-representation of females in training datasets. Despite introducing oversampling techniques to address this omission, the disparities in performance persisted suggesting that addressing dataset representation alone is not a sufficient measure for mitigating bias. Further, our experiments demonstrated that oversampling could reduce overall performance, which may result from the mixing of conflicting data (i.e., male vs female feature rankings). In addition, oversampling with synthetic instances solely from the dataset at hand does not provide the machine with more information, it simply redirects attention and therefore cannot easily compensate for demographic under-representation [47]. We also found that in some cases resampling the dataset to focus on just one sex led to a drop in performance, even for the minority group, which we

relate to a loss in informational power. Re-balancing the data by sub-sampling inevitably changes the number of available training points. This may result in a seemingly paradoxical drop in performance for the minority class and explains why simple stratification of demographic groups may be an inadequate solution to model bias

4.4.2 Fairness aware feature selection

Our evaluation of performance disparities across the range of feature subsets illustrated that the gap in performance persisted throughout these variations in feature information (Figures 4.5 to 4.8). Furthermore, restricting the available features in an attempt to account for sex-differences in biochemical and clinical information resulted in a drop in performance due to the loss of information, particularly where only the biochemical features were used. Moreover, it is possible that these biochemical markers may be less effective predictors for female patients in general, which may stem from the historical issues detailed in Chapter 2 regarding the neglect of female bodies in medicine [48, 123]. The existing biochemical markers used for identifying cardiac disease were drawn from majority male samples, and hence these blood panels that have been selected for quantifying and measuring cardiac disease may not be tailored to female physiology [48, 123, 134]. Thus, it is possible that the overall predictive power of these features may be less for female patients, compared to male patients, resulting in a disparity in performance that is hard resolve. There is a growing body of research that critiques the use of unisex thresholds in medicine for biochemical tests, our sex-stratified analysis of the cardiac datasets and the identified sex differences in feature rankings supports these proposals [134].

4.4.3 Fairness through unawareness & adversarial training

We consider these two fairness approaches together because they are underpinned by the same idea of removing knowledge regarding the sensitive attribute in model development - firstly by removing the feature itself ("Fairness through unawareness") and secondly by training against the sensitive feature with adversarial modelling. Both approaches are trying to remove the sensitive attribute from the model development process, yet this notion in itself may be flawed in the context of healthcare.

Firstly, our findings support the existing reports in the ML community stating that "Fairness through unawareness" is an ineffective technique. The disparity in model performance when training on features without sex, was as persistent as when training on features that included sex. As a biological factor, sex affects the other biomarkers and clinical metrics within the dataset, due to the complex interplay between the biological effects of sex and these features (e.g. the effect of sex-specific hormonal changes on other biochemical

markers). Consequently, the remaining features within the dataset co-vary with Sex and the notion of blinding a model to the effects of sex may not be possible.

Secondly, we saw that the implementation of the FATGB model and adversarial training also failed to resolve the disparity in the global performance of the algorithm and the specific error rates (Table 4.8). Given the explanation provided above, this result is perhaps unsurprising as the foundational concept of the FAGTB model is to try and separate the model prediction from the sensitive attribute. However, with sex playing a role in disease progression, it's effects will ultimately be embedded within other features and thus removing its effects may not be possible. In fact, when the biological effects of sex plays a role in the pathway by which disease emerges, training against knowledge of sex may reduce the ability of the model to predict the outcome overall, potentially negatively impacting the minoritised group. Attempting to blind models to the sensitive feature such as sex, either through removal or adversarial training, may end up being counter-intuitive, as these attributes are closely intertwined with other predictors in the dataset that are vital to improving model performance. Unlike other areas of algorithmic bias, e.g. credit card lending or criminal justice predictions, in healthcare, an attribute such as sex may form an integral part of the pathway an algorithm is attempting to model and predict.

4.4.4 Conclusion & Limitations

In this chapter I have reviewed the inequities in the performance of cardiac ML algorithms found in Chapter 3 and evaluated the applicability of a range of fairness techniques for addressing these disparities. In this chapter we have examined the (in)applicability of traditional fairness metrics in healthcare, where the nuanced relationship between the sensitive attribute and target variable, mean that standard ML fairness practices may fail to resolve the issue of performance inequity. Specifically, we see that removing the sensitive feature is inappropriate when that feature plays an important role in the biological pathway of disease, and training against a models ability to predict that feature forces the model to unlearn potential pertinent information on disease manifestation

In terms of limitations, these findings are limited by the small size of the uncovered datasets, reducing their potential generalisability. I propose that larger studies focused on this issue are required to fully investigate the problem. These datasets also came from the same source, as I found a limited number of open-access databases due to the confidential nature of patient data and issues of proprietary ownership. In addition, I focused on Random Forest (RF) models to replicate the papers uncovered in the literature search, however ML models may differ in their degrees of performance disparity, and an evaluation across the range of ML model options is an important next step.

My research was further limited by the available information in the datasets. The absence of race/ethnicity data precluded the evaluation of their effects. Furthermore, the absence of other demographic data in the studies we identified prevented the investigation of health inequities that might impact the LGBTQ+ community, disadvantaged socioeconomic groups, or other subgroups. Previous research has described historic and institutional biases that contribute to worse health outcomes for these groups, and evolving AI systems require the same scrutiny to ensure these harms do not become embedded within digital systems [183–185].

Lastly, throughout these chapters I have used the terms male and female to reference biological sex, so as not to conflate sex and gender. With the on-going problematic conflation of sex and gender in medicine, stratification of model performance by either sex or gender is often impossible, which was noted in our own work [30, 183–185]. Beyond the features discussed above, there are a wide range of additional factors that we cannot account for. For example, CK was a key feature in HF modelling yet existing studies have demonstrated the variation in these levels for manual labourers and athletes, illustrating how occupation may impact a patients physiology [186].

4.4.5 Avenues for further research

To account for the complex interactions that potentiate disease, and the heterogeneous nature of patient cohorts, we require more complex modelling capable of capturing the full range of intersecting factors influencing patient health (e.g., sex differences may be mediated by income). Unsupervised high-dimensional representation learning may be the path forward for this purpose [7]. In addition to improving representation, unsupervised techniques enable us to detect neglected sub-populations without predetermining a characteristic of interest, facilitating the identification of previously overlooked disadvantaged. In this sense, AI may provide a route forward to uncovering and addressing bias, by deploying more complex modelling that can improve patient representation and by revealing previously neglected disparities in the provision of care. This is an avenue being explored in greater depth within our research lab, described in detail by the work of Carruthers (2022) [7].

Finally, there are further sources of inequitable performance that the evaluative methods of this chapter cannot distinguish between. It may be that the sex-differences in physiological expression of disease means that the prediction is harder to extract from one population. As a result, one sex may require more complex models than another,

with differing architecture and degrees of flexibility. It may also simply be that there are differences in the predictability of one group compared with another, such that if the physiology of one group is more opaque, it may ultimately not be possible to resolve the observed disparities. McCradden and colleagues detail this challenge further in their review, highlighting that differences across groups may not always indicate inequity [14]. There are complex causal relationships between biological, environmental, and social factors that underpin the differences in disease rates seen across population subgroups [14]. While it is imperative that models should not promote different standards of care according to protected characteristics, differences between groups may not necessarily reflect discriminatory practice [14]. In the next chapter I will dive into this issue in greater depth, exploring how ML methods can be deployed to untangle complex causal relationships and ascertain the reasoning for the disparities observed in algorithmic performance.

Chapter 5

Causal fairness applied to psychiatry algorithms

> We should not allow models to promote different standards of care according to protected identities that do not have a causative association with the outcome.
>
> *McCradden (2020) [14]*

5.1 Introduction

In the last Chapter we saw that achieving ML fairness in healthcare faces the particular challenge of distinguishing between (1) when a sensitive attribute (SA) contains important information for predicting the target outcome, and (2) when its presence is inappropriate and may potentiate discrimination. Unlike in other domains (e.g. predictive policing algorithms), in healthcare the sensitive attribute may be a key agent in the manifestation of a target outcome. In the case of sex, the biological effects of sex (e.g. hormonal factors) influence other bodily biomarkers (e.g. cholesterol) which act as key indicators of disease (e.g. in cardiovascular disease). Thus, techniques such as "Fairness through unawareness" that remove the SA, may actually remove essential information regarding the pathway to the target outcome for that group, unintentionally disadvantaging that group by removing important information on the pathogenesis of the condition.

Central to achieving fairness in healthcare ML therefore, is understanding the causal pathways to disease, as only then can sensitive attributes be treated appropriately. As detailed in the introduction, Anand argues that inequalities related social arrangements (e.g. discrimination relating to a protected characteristic) cause greater aversion than inequalities resulting from a chosen behaviour (e.g. smoking) [20]. In healthcare, it is necessary to understand when an inequality on the basis of a sensitive attribute is due to (i) resolvable social arrangements (e.g. discriminatory policies on the basis of sex) or (ii) differences stemming from biological pathways associated with the sensitive attribute (e.g. higher rates of breast cancer in female patients, or melanoma in lighter skinned patients). To achieve fairness in healthcare ML, we must therefore go beyond the techniques reviewed in Chapter 1, and implemented in Chapters 3 and 4, and instead implement methods capable of teasing apart the complex pathways related to sensitive attributes. For this, I turn to causal modelling.

In our review of healthcare AI fairness so far, I have considered a plethora of case studies demonstrating the negative impact of AI bias on historically marginalised patient groups. Seyyed-Kalantari and colleagues have demonstrated that machine vision models built to diagnose chest X-Rays may miss pathological signs in marginalised populations [17]; Oberymeyer and colleagues exposed an AI model that under-referred black patients for hospital care [37]; and further research has demonstrated that AI may fail to detect disease from female blood tests or exhibit damaging stereotypes that influence psychiatric diagnoses [6, 51]. These papers play an essential role in exposing the issue of AI bias in medicine, however in this chapter I argue that these research methods do not go far enough. In this chapter we will look beyond the existing research that identifies and quantities disparities in algorithmic performance, to explore the underlying causal mechanisms that lead to these disparities. This approach is an essential next step for ensuring computational interventions that aim to address AI bias are appropriately targeted and work effectively for all demographic groups. Thus, this chapter will progress through the following steps:

Step 1: Causal modelling and causal fairness: I shall examine the topic of causal modelling and causal fairness, highlighting domains in which these computational methods have been applied for evaluating issues of algorithmic inequity. This section will provide a thorough overview of the domain, to provide the reader with a foundational understanding of the methods that follow.

Step 2: Causal modelling and fairness in medicine: I will examine the existing uses of causal modelling in medicine and how these methods may translate to fairness applications. I will discuss both causal models, and the use of causal methods for understanding issues of causality across a range of ML models. Further, will examine the specific challenges of using causal fairness methods in medical modelling.

Step 3: Empirical research - Causal fairness in psychiatric algorithms: In this chapter I then take the novel approach of applying causal fairness frameworks to issues of bias in psychiatry algorithms - exploring a new area of medical AI bias and fairness resolution. In addition to using methods of feature evaluation to understand causal pathways, I evaluate the utility of causal methods for medical AI fairness, and propose a framework for improving the methods for addressing AI bias in healthcare.

5.1.1 Causal Modelling and causal fairness

In the field of computational fairness, an evolving area of research is focusing on how the causal frameworks first proposed by Judea Pearl can be applied to questions of al-

gorithmic bias [91, 95]. These methods of causal modelling are an essential approach for understanding the relationships between the variables in a dataset that are used for model development. Through the causal lens, researchers consider a model unfair, if there is an unfair *causal effect* between a sensitive attribute in the dataset and the model's decision [187]. The methods of causal fairness have been applied to evaluate these pathways and examine AI bias in the finance sector and in job hiring, yet these approaches have not been used to assess AI bias in healthcare models [75].

To review these methods, we can return to the example first provided in Chapter 1. In this example, Jones and colleagues provide a useful clinical analogy for causal modelling, describing a set of random variables A, B, C, D that correspond to age, bladder cancer, cigarette consumption and deafness, respectively, which may be associated with one another through correlations [94]. The authors highlight that if we want to know whether C (smoking) causes B (bladder cancer), then we must establish whether intervening on the value of C changes the distribution of B. If such a change occurs, we may conclude that smoking more cigarettes may increase the risks of bladder cancer, however the reverse is not true - people who develop bladder cancer do not become smokers. Distinguishing between association and intervention is the central component of Pearl's causal hierarchy, that was explored in greater depth in the introduction [91, 95].

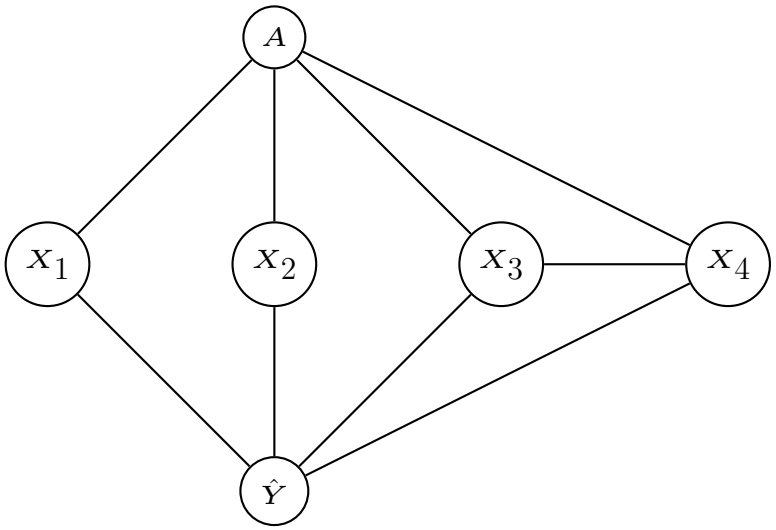
In 2024, Plecko and Bareinboim released a "Causal Toolkit for Fair Machine Learning" in which they detail the value of causal methods for evaluating AI bias, demonstrating the applicability of their methods across algorithms used in credit-scoring, college admissions processes, and criminal justice decisions [96]. The authors relate the importance of causal methods, to the fairness metrics of disparate treatment and disparate impact [96].

In the case of disparate treatment, it is expected that there is no direct effect of a sensitive attribute on the outcome, whereas the disparate impact doctrine ensures the sensitive attribute has no effect on the outcome at all [96]. As a parallel, Plecko and Bareinboim highlight that anti-discrimination laws often require the demonstration of a strong causal connection between an alleged discriminatory practice and an observed statistical disparity [96]. Yet, as stated by the authors - "statistical measures alone cannot distinguish between different causal mechanisms that transmit change and generate disparity in the real world". Thus, if we intend to hold AI models to the same standards as humans when it comes to questions of discrimination, there is a clear need for methods that can determine how much of an observed disparity can be attributed to a causal path from the sensitive attribute to the AI decision [96]. To examine the causal connections between a sensitive attribute and a model outcome we can turn to the standard causal frameworks first proposed by Pearl [91, 95].

Causal graphs

Causal relationships can be visually expressed using *graphical causal models (GCMs)* [187]. Of these, there are two common types. Firstly, *directed aycyclic graphs* (in which a node cannot be an ancestor of itself), and secondly partially directed acyclic graphs (PDAGs) [91, 95]. The graphs allow for the evaluation of causes and causal effects, such that if there is a *directed path* from A to Y, then A is a potential *cause of Y* (see Figure 5.1) [187].

Figure 5.1: An example of a causal graph, where the prediction $\hat{Y}$ is obtained by a function f which takes $X_1, \ldots, X_4$ as input features, which may be influenced by Sensitive Attribute (A). Our graph is adapted from the previous work from Pan and colleagues [188]



Pan and colleagues developed an approach for examining path-specific effects, that accounts for both directed and undirected relationships in causal pathways [188]. In designing their approach, the authors adapt causal graphs described in other domains, to the issue of protected characteristics and model disparity [188]. Figure 5.1, adapted from Pan and colleagues paper on causal fairness, takes a traditional causal graph and focuses on the relationship between a sensitive attribute (A) (e.g. Race) and the predicted outcome (Y) (e.g. diagnosis of disease). These graphs can be drawn either on the basis of domain knowledge, or through using causal discovery algorithms [188]. Active paths from A to Y, are potential sources of model disparity, however if there are no active paths between A and Y, the authors argue that this indicates zero model disparity on this basis of the sensitive attribute (A). It is important to note that it may be difficult to establish whether 5.1 is a Directed Acyclic Graph (DAG) (in which relationships are directed), or a Partially Directed Acyclic Graph (PDAG) in which the causal direction between factors cannot be determined.

Causal fairness in medicine

There is an extent of research on causal modelling in healthcare that could be transferred to the domain of healthcare AI fairness. Feuerriegel and his team provided an update on the latest causal machine learning efforts in healthcare in their 2024 article *"Causal machine learning for predicting treatment outcomes"*. They summarise the domain of medical causal machine learning (ML), in which researchers estimate individualised treatment effects for patients under different treatment scenarios [189]. These effects are considered "causal quantities". Unlike traditional ML approaches, causal ML quantifies changes in outcomes due to a specific treatment, so that treatment effects can be estimated which may significantly advance patient care through the personalisation of their therapies [189]. In their comprehensive review of causal ML, Feuerriegel et al cite numerous applications by which these methods may advance healthcare, yet the authors do not mention the relevance of these methods to questions of fairness [189]. To transfer these methods to the fairness domain, one may take the approach of considering the sensitive attribute (e.g. sex) as the "treatment", thus evaluating the impact of one's membership to a particular group on their treatment outcomes.

It is worth noting that causal modelling has previously been used to untangle other complex issues in healthcare modelling, such as the statistical challenge of Simpsons Paradox [188, 190]. Described in depth by Von Kügelgen and colleagues, causal modelling provides a route for unpacking Simpson's paradox, whereby a trend appears in several groups of data, but disappears or reverses when the groups are combined [190]. Most recently, this was observed in the modelling of COVID19 mortality rates [190]. These efforts are particularly relevant, as the use of causal modelling to untangle demographic differences in COVID-19 mortality across groups, is directly applicable to the question of algorithmic bias. We therefore look at this example in greater depth below, exploring how the methods may be extrapolated to questions of algorithmic bias.

Causal modelling in population health

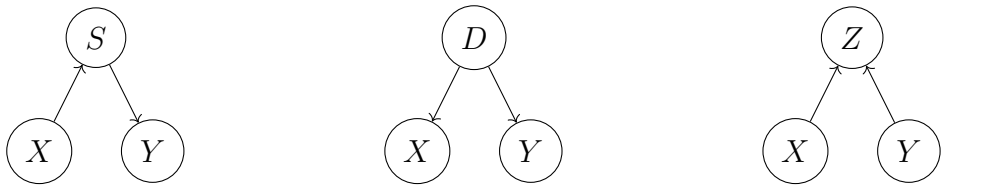
In the field of health equity, causal modelling has previously been used to better understand inequalities observed in epidemiological data. In recent years we saw this in the COVID-19 pandemic, where researchers applied causal frameworks to better understand the mortality disparities emerging between male and female patients. In their paper examining gender and sex bias in COVID-19 epidemiological data, Diaz and colleagues consider a range of possible explanations for the observed disparity in disease severity between the sexes. To begin, the authors acknowledge the disparity in COVID19 mortality that exists between male and female patients, with males experiencing a higher case fatality ratio

than females. The authors then detailed a wide range of hypothesis that emerged in the research literature attempting to account for this disparity, including:

1. The impact of Sex on vaccine acceptance, response and outcomes [191]
2. Biological differences in the immune system of males and females, affecting the patient's capacity for fighting the infection [192].
3. Male patients appeared to be a greater risk of cardiac complications of COVID19 [193].
4. The relationship between gender and having a responsible attitude to COVID19 mitigation efforts [194]
5. The relationship between gender, likelihood of smoking, and the impact of smoking on COVID19 mortality [195]

In order to untangle the causal questions underpinning the observed disparities, the authors utilised causal graphs, inference methods and causal mediation analysis [195]. Firstly, the authors built causal graphs to examine the relationships between Sex, Gender, COVID-19 mortality and the remaining variables in the dataset. Díaz-Rodríguez and colleagues provide a deep dive into the role of causal graphs for unpacking healthcare disparities, highlighting the importance of distinguishing between the three key causal structures in questions of health equity. These three structures presented in Figure 5.2 illustrate the relationship between a feature (X), the prediction (Y) and the remaining variables in the dataset (S, D, Z), which may be a: (a) Mediator, (b) Confounder or (c) Collider (Figure 5.2).

Figure 5.2: Causal modelling: Basic structures of causal graphs.



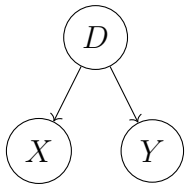
(a) Mediator (b) Confounder (c) Collider

Figure 5.2 provides a visual illustration of the different causal pathways between X and Y, dependant on whether the third variable is a Mediator, Confounder or Collider. All of which have different implications for fairness. In (a) we can see that the Mediator variable (S) influences the causal effect of X on Y [95, 195]. In (b) the confounder can be understood as a common cause of the other two variables (X and Y). In both these instances X will correlate with Y, yet only in the Mediator pathway is X a cause of Y. Finally in the case of the collider, this is a variable caused by the two other features, and X and Y won't be correlated unless we condition on Z [91, 95, 195]. The pathways determine the type of causal effect present, such that there are direct effects (X $\rightarrow$Y), indirect effects (X $\rightarrow$A $\rightarrow$Y and X $\rightarrow$B $\rightarrow$Y), and path specific effects (e.g. only X $\rightarrow$B $\rightarrow$Y).

When considering fairness and protected characteristics, if X is a protected characteristic, then a direct effect between X and Y would be considered unfair (as the sensitive attribute should not be use to predict the outcome). Alternatively is there is path from the protected characteristic to the outcome via an explaining variable (i.e. a mediator), this may be considered acceptable or fair. To appreciate why understanding these separate paths is important, we can consider the following hypothetical scenarios.

Hypothetical 1: Confounder Scenario

Let's first consider a hypothetical algorithm that is predicting the cost of health insurance premiums (Y), from a set of features that includes Income (X) and Disability Status (D), where the causal structure is underpinned by Figure 5.3. When the researchers evaluate this model's outputs it appears that it's discriminating against low-income groups, predicting higher insurance premiums for those from a lower income background. Yet, in this hypothetical scenario, disability (D) is acting as a confounder. The presence of a disability has a causal effect on insurance, leading to higher costs. The presence of disability also causes lower income, due to occupational barriers and institutional discrimination against those with disabilities. Thus here, identifying the causal pathway is essential in understanding the true cause of the observed income-based disparity and targeting interventions in an appropriate manner. Is it important to note, that when examining disparities related to Sex or Race, the confounding pathway may be less relevant as another variable will rarely be able to cause Sex or Race.



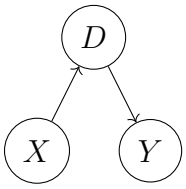
(b) Confounder

Figure 5.3: Structure of the Confounder Scenario

Hypothetical 2: Mediator Scenario

More relevant to the case of sensitive attributes such as Race of Sex are mediator pathways. Identifying a responsible mediator in mediated-causal pathways can be an essential way to unpack whether a disparity is underpinned by a fair or unfair causal path. The case study of sex bias in Berkeley College's admission process provides a useful example [187]. In this case, a statistical evaluation of the admissions data demonstrated that female applicants were rejected more often than male applicants, however it was later unearthed that this was due to females more often applying to departments with lower

admission rates. In Figure 5.4, X represents Sex, D represents department, and Y represents admission decision. In this case, an unfair causal path would have involved Sex [X] directly informing rejection, however instead a "fair" path was uncovered where by Sex [X] informed departmental choice (D), which affected rejection rate. The latter path was considered fair, as the impact of Sex on departmental choice is not under the remit of the college's control.



(a) Mediator

Figure 5.4: Structure of a mediator causal graphs

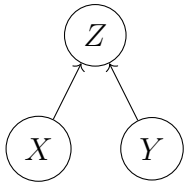
Hypothetical 3: Collider

In our final example, we consider the causal path that involves a collider. The collider is a variable caused by two other variables. In the presence of the collider, X and Y are not correlated, however if we condition on the collider they become correlated [91, 195]. Described in depth by Digitale and colleagues, collider bias occurs when statistical analysis conditions on a variable (collider) that is influenced by at least two other variables, and this conditioning inadvertently introduces associations between these influencing variables that are not causally related [196].

The authors provide an example from the paediatric context, looking at patients having their HbA1C measured (a marker of blood sugar control) (Figure 5.5). In this example X represents obesity, Y represents diabetes symptoms, and Z represents the measurement of HbA1C. A clinician may measure HbA1C (Z) either because a child is obese (X), or because they have symptoms of diabetes (Y). If we then just analyse group Z (children who had their HbA1C tested) it may appear that obsesity protects against diabetes, because children *who are not obese*, show symptoms of diabetes. One could then incorrectly infer that the kids who are obese are less likely to have diabetes, compared to those who are not obese.

5.1.2 Methods for causal fairness

In order to apply these causal frameworks to questions of algorithmic fairness in healthcare one must understand the causal pathways between sensitive attributes and the target variable that a model is predicting. Herein lies a particular challenge that is unique to



(c) Collider

Figure 5.5: Structure of a collider causal graphs

healthcare modelling. The research so far on algorithmic fairness in healthcare rarely accounts for the complex causal relationships between biological, environmental and social factors that influence the different medical conditions affecting various protected identities [14]. McCradden and colleagues have provided a detailed overview of this issue, highlighting the fact that methods of algorithmic fairness have not historically accounted for the complex causal effects of sociodemographic features that contribute to the differing prevalence of medical conditions across protected identities [14]. The task of building causal models in medicine is particularly challenging as factors such as the social determinants of health are understood to play an important role in disease, however the mechanism is poorly described [14]. As stated by the authors - *Sometimes, it is appropriate to incorporate differences between identities because there is a reasonable presumption of causation.* They go on to state that *"difference does not always entail inequity"*, giving the example of biologically differences between sexes that can affect the efficacy of pharmacological compounds, whereby incorporating these differences into prescribing practices would not be considered not unjust [14].

In other domains (e.g. financial loan decisions), it is determined that a fair decision should not be based on any knowledge of the sensitive attribute such as gender, race, or sexual orientation [195]. Yet in medicine sensitive attributes may play an essential role in causal pathways, especially in cases where certain diseases and conditions are informed by this characteristic. For example, some conditions are sex-specific (e.g. ovarian cancer) or may involve pathways whereby biological sex influences the progression of disease (e.g. Parkinson's disease) [197]. The challenge emerges in differentiating between times where it may be appropriate to differentiate on the basis of a demographic feature when building a model, and when it is irrelevant.

In their review of causal methods for medical imaging models, Jones and colleagues discuss this issue in the context of age. [94]. Age is a sensitive attribute and age-related biases may need to be removed from models, yet for some medical conditions age is a clinically meaningful risk factor e.g. in Alzheimer's Disease [94]. If a developer deployed bias-mitigation methods such as adversarial training on the age attribute, this would

inappropriately worsen performance by forcing the model to neglect important clinical information [94]. In contrast, when predicting a mental health diagnosis such as substance misuse, researchers have described the damaging impact that age bias from clinicians can have, leading to an under-diagnosis of cases in the elderly [198]. Here we see that in some medical conditions age may be an important feature to consider, while in other domains it may be a spurious factor that contributed to widening healthcare disparities. One of the key challenges in medical causal research is determining when a sensitive attribute is, and is not, relevant to the pathway to disease [14].

Causal methods (i) : Mediation analysis with causal effects

In the above section we see that understanding the causal path in a situation of algorithmic disparity is essential to understanding whether there is true discrimination occurring and how a solution may be found [199]. Causal graphs provide a useful avenue for displaying these mechanisms, however these methods must be paired a quantification approach that exposes the effect of different paths on the outcome. **Causal mediation analysis** is useful for this purpose, as it allows one to distinguish between different potential pathways and uncover the causal structure of a process. The papers we have discussed so far have utilised causal analysis to understand disparities existing in the real world, however we will be applying these methods to unpacking the decision-making processes of algorithms in simulated environments. Thus, the following distinction is important:

1. The application of causal frameworks to understand why a disparity exists in reality (e.g. sex differences in COVID19 mortality rates) [195, 200]
2. The application of causal methods to understand why a disparity exists in an algorithm's performance (e.g. a racial bias in an healthcare AI model)

In the literature there have been a range of studies seeking to explain algorithmic fairness through either (i) feature-based, or (ii) path-specific explanations [188]. In the last chapter we introduced Shapley values, which can provide a useful means for better understanding the features involved in a causal pathway. Shapley value based methods allow modellers to examine the individual contributions of input features to the final outcome, and outcome disparity [188]. These feature based explanations facilitate the examination of feature contributions to model disparity, yet this approach ignores the causal structure of features themselves [188]. Newer methods go beyond this, integrating diagrammatic methods for visualising causal pathways with causal metrics for quantifying causal effects [188].

The most common non-causal fairness metric is statistical or demographic parity, which uses the total variation (TV) to evaluate the relationship between X and Y. TV measures the difference between the conditional distribution of Y when we observe X changing [195]. The causal version of TV is **Total effect (TE)** [195]. Total effect (TE) is defined

in terms of experimental probabilities, measuring the effect of the change of X from X1 to X0 on Y = y along all causal paths from X to Y [195].

$$TE_{X_1,X_0}(y) = P(Y = y|\text{do}(X = x_1)) - P(Y = y|\text{do}(X = x_0)) \tag{5.1}$$

: Equation for Total Effect

In the process of mediation analysis, it is essential to differentiate between paths of causal effect between two variables, examining direct and indirect paths. To do this we utilise Pearl's definitions of Natural Direct and Natural Indirect Effects (NDE and NIE respectively).

1. **Natural Direct Effect (NDA)**: Measures the direct causal effect between two variables. As demonstrated by Equation 5.2, where the mediator variables are represented by **Z**) [195].

$$NDE_{X_1,X_0}(y) = P(y_{X_1,Z_{X_0}}) - P(y_{X_0}) \tag{5.2}$$

: Natural Direct Effect (NDE)

2. **Natural Indirect Effect (NIE)**: The NIE measures the indirect effect of X on Y, defined in Equation 5.3 [195]. The approach is limited in that it cannot distinguish between the fair (explainable) and unfair (indirect discrimination) effects [195].

$$NIE_{X_1,X_0}(y) = P(y_{X_0,Z_{X_1}}) - P(y_{X_0}) \tag{5.3}$$

: Natural Indirect Effect (NIE)

3. **Path Specific Effect (PSE)**: To address the limitations of the NIE, that cannot account for path-specific effects, further methods have been introduced to examine path specific effects **(PSE)** [195]. Given a path set T, the T-specific effect is defined by the relationships in Equation 5.4.

$$PSE^{T}_{X_1,X_0}(y) = P(y_{X_1|x_0 \to X_1}) - P(y_{X_0}) \tag{5.4}$$

: Path-Specific Effect (PSE)

In causal mediation analysis one examines **NDA, NIE and PSE** to quantify the contribution of various paths in a causal graph on the outcome, in order to untangle the underlying causal contributions. These methods will form the first part of our causal approach, detailed in the methods section below.

Causal Methods (ii): Counterfactual fairness

In addition to quantifying the effect of various causal paths, researchers have introduced further methods for examining causal fairness. Another approach that we will explore is *Counterfactual fairness*, which requires equality between the "observed" outcome and the "counterfactual" outcome, for every individual in a dataset. Kusner and colleagues described a causal framework for examining individual-level fairness, through the application of these counterfactual fairness methods. Through this lens, the authors state that a decision is fair toward an individual, if it *"coincides with the one that would have been taken in a counterfactual world in which the sensitive attribute were different"* [187, 201]. Thus, evaluating the counterfactual fairness of a model relies on the construction of counterfactual worlds where the sensitive attribute is flipped (e.g. males are treated as females). Described in depth by Kusner and colleagues, outcome Y is considered counterfactually fair if under any assignment of the values A = a and for any individual in the population, the follow equation is satisfied:

$$P(y_{x_1} \mid A = a, X = x_0) = P(y_{x_0} \mid A = a, X = x_0) \tag{5.5}$$

: Counterfactual fairness

Hence the notion of counterfactual fairness is satisfied if the probability distribution of Y, is the same in the actual and counterfactual worlds, for every possible individual [195, 201]. It should be noted that Chiappa and colleagues challenged the approach of Kusner, stating that counterfactual fairness assumes the entire effect of the sensitive attribute on the decision is problematic, neglecting to consider the nuances of causal relationships whereby the sensitive attribute might affect the decision along both fair (e.g. via explaining variables) and unfair pathways (e.g. via proxies) [2, 187]. To account for these challenges of path-specific effects in fairness, Chiappa proposed a technique that went beyond brute counterfactual fairness for evaluating algorithmic equity [187]. The authors propose that **path-specific counterfactual fairness**, determines a decision to be fair if *"it coincides with the one that would have been taken in a counterfactual world in which the sensitive attribute along the unfair pathways were different"*.

In my approach I combine both of the methods described above, exposing the relationships between features through causal graphs, quantifying specific paths using causal effects, and implementing counterfactual experiments to evaluate the overall influence of the sensitive attribute on the target variable.

5.1.3 Empirical research: Causal fairness in psychiatry algorithms

So far we have discussed algorithmic bias in general terms and the role of causal fairness in modelling issues of algorithmic inequity across a range of domains. I will now turn to the specific medical domain of this chapter, in which we will be exploring bias in algorithms used to predict psychiatric outcomes amongst patients. In the next set of experiments I chose to focus on psychiatric care as the target variable, due to the history of demonstrated inequities in mental health care, detailed in greater depth in the background chapter [48, 49, 51, 115]. Furthermore, the burden of mental illness is known to be increasing globally, at a time when resources are low, and AI has been proposed as a mechanism for addressing the gap in care. The advance in psychiatric AI modelling, and the unaddressed history of biases within the domain, make this an important area for bias investigation, in order to understand the potential impact of AI on health equity in psychiatry [41, 51, 115].

Graham and colleagues have provided a comprehensive overview of AI applications in mental health care, summarising twenty-eight studies that vary in use of structured and unstructured data for predicting the psychiatric outcomes of depression, schizophrenia, suicidal ideation and attempts, plus further psychiatric diagnoses [41]. Of the 28 studies reviewed, the authors find that 23 rely on supervised ML techniques, deriving data from a range of sources including electronic health records (EHRs), mood rating scales, and brain imaging data [41].

Empirical work within the domain has demonstrated high success rates in building predictive psychiatric models, yet evaluations of demographic bias within model performance is notably scarce. Lacy and colleagues focus on the adolescent population, creating models capable of predicting cases of anxiety, depression, attention deficit, disruptive behaviors and post-traumatic stress with an AUC of 0.94 [202]. Yet in their conclusive remarks the authors acknowledging that the lack of demographic stratification may limit the generalisibility of their results.

Psychiatric AI is a particularly challenging area for studying AI bias, as the underlying causal structures of psychiatric diagnoses, the influence of social determinants, and even the diagnoses themselves are undergoing increased scrutiny for their validity [115]. Taylor draws attention to the historic discrimination perpetuated by the psychiatry discipline, exposing a history of biased diagnostic practices along gender lines, that has led to the pathologisation of women, non-binary and trans-patients who have experienced trauma [115]. Her work builds on a growing body of research that has

exposed psychiatric diagnostic biases and treatment inequities along demographic lines, stemming from stereotyped clinical frameworks, clinician bias and the lack of diverse voices present in shaping of epistemological psychiatric discourse [51, 123, 203–205].

Unlike other medical disciplines, the diagnostic criteria for psychiatric conditions relies largely on scores derived from self-reported experiences or subjective evaluations from professionals (as opposed to blood tests or clinical investigations). Thus, mental health clinical data is often in the form of qualitative statements and written clinical notes [41]. These diagnostic frameworks themselves have been called into question for their impact on bias and health inequalities [115, 203]. For example, the traditional metrics used to rate severity of Autism Spectrum Disorder has been identified as focusing predominantly on male expressions of the condition, leading to widespread neglect of the female experience [206, 207]. Furthermore, in my own previous work, I have highlighted biases that exist within Large Language Models (LLMs) - an evolving subdomain of AI - that perpetuate harmful stereotypes relating to the history of discrimination in psychiatry [51]. This research was predominantly exploratory, examining the association between terms in word embedding structures [51]. Taken from this paper, Table 5.1 highlights of these previous key findings, including the different mental health diagnoses that we found to be associated with varied demographic subgroups with the LLM's dataset [51].

Table 5.1: Results table from previous work on bias in Large Language Models, illustrating psychiatric stereotypes associated with different subgroups [51]

Race Label	**Most Closely Related Mental Health Diagnosis**	**Vector Similarity**
Latino	substance_abuse	0.22431692
African_american	schizoaffective_disorder	0.1818381
Native_american	substance_abuse	0.2724196
Asian	compulsive_hoarding	0.0947723
Hispanic	ADHD	0.17809318
White	alcoholism	0.11180493
Black	bipolar_disorder	0.12816364

Chapter Research Aim

The experiments of this chapter take the novel approach of applying causal fairness methods to question of algorithmic inequity in healthcare. As I have discussed, deploying causal methods for computational fairness have been explored in other domains but not in healthcare, which may be due to the unique challenge in medicine of untangling the pathophysiological contributions of sensitive attributes to disease. I have chosen to focus

on psychiatric diagnoses as this is also an unexplored area, and one known to be prone to bias. In selecting the causal quantities to focus on, I draw on state-of-the-art research and examine **counterfactual outcomes** and **average causal effects** for evaluating causal pathways to model decisions [189].

5.2 Methods

Our focus is on models that predict which patients in a population will require psychiatric care, and to do this I used a dataset from the UK Biobank which reports whether patients have utilised secondary psychiatry care services [208]. The UK Biobank is a major prospective study with significant involvement from the UK Medical Research Council and the Wellcome Trust, and has become an important open-access resource for medical researchers across the UK and worldwide [208]. I utilised a subset of the UK Biobank dataset, previously described by Ruffle et al to predict high blood pressure amongst a patient cohort, and which contains rich additional clinical information including metrics that measure patients mental health and uptake of psychiatric services [209]. From this extensive dataset, they key target variable selected was "Psychiatric Care", alongside a range of demographic and clinical factors detailed below. We selected key demographic variables for comparing algorithmic performance across difference subgroups, and specific clinical variables related to the target variable of "Psychiatric care".

Demographic Variables

The demographic variables examined included Age, BMI, Handedness and Sex. For demographic subgroup performance, we reformatted Age and BMI into categorical variables for the evaluation of group fairness. Originally a continuous variable, age was divided into three groups to form variable "Age Group" with young (40-50 years), middle (50-60 years) and older age (60 - 70 years). BMI was binarised to into "Healthy" (BMI <25) and "Overweight" (BMI >25) (Table 5.3). We chose retain handedness as a demographic measure as a form of control. Handedness could be considered a spurious difference between patients to which there are little reports of medical bias, and hence we wouldn't expect to see discrimination on this basis. By including all these demographic variables, in this chapter we are able to examine potential AI bias across a range of subgroups, moving beyond the focus of Sex in the previous sections. The descriptive statistics for each of these demographic features are provided in Tables 5.2 to Table 5.4, and the balance of the target variable within each group is provided in Table 5.5.

Clinical Variables

Medical Conditions: The dataset contained details on the patient's co-existing and previous medical conditions, including: Diabetes, Hypertension, Angina, Atopy, Asthma, Heart attack, Chronic Obstructive Pulmonary Disease (COPD) and Stroke. These conditions were included in our dataset, due to the existing research that describes the impact of chronic disease on a patient's psychological state [210, 211]. Due to both the biological and sociological effects of these chronic illnesses, many patients suffering these conditions also develop psychiatric complaints [210, 211]. I therefore chose to retain these variables within the dataset, to explore for their effects on the target variable of "Psychiatric Care" (See Tables 5.2 to Table 5.3).

Clinical Measurements: Several clinical measurements were included in the feature set including: (i) Reaction time, (ii) Weight and (iii) Body Fat (Tables 5.2 to Table 5.4.). Existing research has described how these features (e.g. Weight) may contribute to deleterious mental health effects, while further research has detailed the negative impact that poor mental health may have on these clinical measurements [212–214]. By including these measurements we are able to explore these potential pathways that contribute to the evolution of psychiatric disease. The details of these measurements are provided in Tables 5.2 to Table 5.4.

Neuroticism Score: Neuroticism score is a widely used tool for measuring neurotic traits, which is defined as a personality dimension consisting of components such as mood instability, worry, anxiety and irritability [215, 216]. It's use in medicine has been criticised, with researchers questioning it's validity, however it's use in psychiatric profiling and ML modelling is widespread - thus we opt to include it in our own study to explore its effects on issues of health and model equity [115, 215–218].

Table 5.2: Biobank Dataset: Summary statistics for **continuous** variables used to predict Psychiatric Care

Feature	Mean (SD)	Range	Data Type
Age	54.78 (7.44) years	40.00 - 70.00 years	float64
Body_fat	30.08 (8.20) percentage	5.50 - 58.97 percentage	float64
Bmi	26.54 (4.21) kg/m$\hat{2}$	14.68 - 56.60 kg/m$\hat{2}$	float64
Reaction_time	537.39 (100.38) milliseconds	297.00 - 1726.00 milliseconds	float64

Table 5.3: Biobank dataset: Summary statistics for **categorical** variables used to predict Psychiatric Care

Feature	Possible Categories	Most Frequent (%)	Data Type
Sex	0 (Female), 1 (Male)	Female, 53.27%	float64
Handedness	0 (Right), 1 (Left)	Right, 89.08%	float64
Gp visits for Mental health	0 (No), 1 (Yes)	No, 67.85%	float64
Angina	0 (No), 1 (Yes)	No, 98.51%	float64
Stroke	0 (No), 1 (Yes)	No, 99.35%	float64
Insulin Treatment	0 (No), 1 (Yes)	No, 99.94%	int64
Neuroticism Score	0 (Low) to 12 (High)	Low, 14.68%	float64
Copd	0 (No), 1 (Yes)	No, 99.39%	float64
Smoking	0 (Non-smoker), 1 (Smoker)	Non-smoker, 61.36%	float64
Atopy	0 (No), 1 (Yes)	No, 76.97%	float64
Asthma	0 (No), 1 (Yes)	No, 90.13%	float64
Hypertension	0 (No), 1 (Yes)	No, 81.95%	float64

Table 5.4: Summary statistics for continuous variables converted to categorical variables (Age Group and BMI Binary).

Variable	Category	Count (%)
Age Group	1 (40-50 years)	8,551 (30.25%)
	2 (50-60 years)	12,329 (43.62%)
	3 (60-70 years)	7,385 (26.13%)
BMI Binary	0 (BMI <= 25)	11,296 (39.61%)
	1 (BMI > 25)	17,221 (60.39%)

5.2.1 Data pre-processing and feature engineering

The Biobank dataset was imported into Jupyter Notebook, null values were removed, the target variable of Psychiatric care was selected, and features relevant to the question of bias were identified from the dataset (Age, BMI, Sex etc, as listed above). Features were evaluated for collinearity through correlation metrics and were then ranked for their importance in predicting the target using:

1. (i) Shapley values,
2. (ii) Mutual information and
3. (iii) Recursive Feature Elimination.

In calculating these metrics, 50 runs were performed to account for instability. The final set of selected features are summarised in Tables 5.2 to Table 5.4, with the results of the Feature Ranking presented in Figures 5.6 to Figure 5.8. The overall descriptive features of the dataset, with the attack rate of the target variable and details of class balance are provided in Table 5.5.

5.2.2 Model development and evaluation of bias

The next section of the methodology is divided into five stages.

1. Stage 1: Model development and feature analysis
2. Stage 2: Are there any disparities in model performance with regards to the included sensitive attributes?
3. Stage 3: For identified disparities, is there a causal relationship between the sensitive attribute and the prediction?
4. Stage 4: Is the causal path fair (via an explanatory variable) or unfair (via a proxy)?
5. Stage 5: Causal Fairness Adjusted Model

5.2.3 Stage 1: Model development and feature analysis

Our first step involves the construction of a series of ML models, similar to those described in Chapters 3 and 4. A series of models were built and compared for their respective performance across all patients in the dataset including:

- Random Forest Models
- XG Boost Models
- Logistic Regression Models
- Support Vector Machines
- Deep Learning / Neural Network Models

Models were built using the Scikit Learn package, with hyperparameter turning performed using the GridSearch CV package. In training the model, the dataset was split 80% training and 20% test, the class weight set as balanced, and random state at 42. The

models were then evaluated according the traditional performance metrics set out in Chapter 3, namely: ROC AUC Score, Accuracy, False Negative Rate and False Positive Rate (See Table 3.3). Similarly to the previous chapters, a bootstrapping approach was adopted to quantify uncertainty in the consistency of model predictions. Reflecting the methods of Chapters 3 and 4, I built separate models 100 times, with a different split of data each time, and evaluated the mean performance metrics with variance over these 100 runs.

5.2.4 Stage 2: Is there a disparity in Model Performance?

The next step was to evaluate for any possible disparities in algorithmic performance affecting a particular demographic group. For this section I focused on the Logistic Regression (LR) Models, as these gave the highest performance in Stage 1 (Table 5.6 in the results details this further). The performance metrics of the LR models across the 100 runs were broken down by subgroup, to evaluate subgroup-specific performance and any algorithmic disparities. As detailed in the last section, model performance was therefore considered across the following subgroups:

1. Sex (Male and Female)
2. BMI (Low vs. High [>25])
3. Age (Young, Middle and Old)
4. Handedness (Left and Right)

For each subgroup, the mean scores were calculated for each performance metric (Accuracy, ROC AUC, FNR and FPR), and the mean difference between the subgroups with regards to each sensitive attribute was calculated with statistical significance - mirroring the methods set out in Chapter 3.

5.2.5 Stage 3: Is there a causal relationship between the sensitive attribute and the prediction?

Once disparities in algorithmic performance were identified in Stage 2, I went a step further and examined the causal structures that may underpin these differences. In keeping with the expected flow of research articles, the results of our subgroup analysis are detailed in the results section below, however it is worth mentioning here that the greatest disparity in algorithmic performance was again seen between males and females. Thus our focus for the causal analysis, and from this point on, is on the sex disparity in algorithmic performance.

Due to the magnitude and significance of the Sex disparity described below, we focus specifically on causal pathways to sex disparities in model fidelity. To do this, I make use of the methods detailed in the background section of:

1. Counterfactual methods, and
2. Causal effects

Firstly, to calculate counterfactual effects, one must consider what would happen to the outcome of interest if you changed a specific variable, while holding everything else constant. To do this the following steps were taken to calculate the counterfactual outcomes in terms of predicting psychiatric care, for each Sex:

1. **Definition of Causal Question:** What would be the effect on predictions of psychiatric care, if all females were treated as male?
2. **Creation of Counterfactual Dataset:** We create a dataset for the hypothetical scenario in which the original females are changed to males.
3. **Counterfactual predictions:** Models are then built to predict the target outcome (psychiatric care) for the counterfactual dataset.
4. **Comparison of results:** The performance of the models overall, and for the subgroups, is compared between the original and counterfactual worlds.

For this stage I created a function that replicated the previous equity analysis, but performed this across both the original and counterfactual worlds. The same bootstrapping technique was deployed, running 100 experiments in which different models were built for each run (with a different training split of data), and predictions are made in the original world and counterfactual world datasets. The mean performance metrics were then printed for the original males and females, and for the counterfactual males (originally females) in the counterfactual world. These measures give us an idea of the causal influence of Sex on the model's prediction, as we are addressing the question of *"What prediction would these females have received, if they were treated as males, all other factors being the same?"*.

5.2.6 Stage 4: Is the causal path fair or unfair?

In Stage 3 I utilised counterfactual methods to begin the causal evaluation, however as detailed in the introduction, identifying a causal effect does not necessarily mean one identifies "unfairness", as there may be "fair" causal reasons for this effect. Thus to dive deeper into the identified sex disparity in model performance, I next examined the path specific effects that may contribute to errors observed in the model prediction. Here, we consider the question: **Is the causal path fair (via an explanatory variable) or unfair (via a proxy for the sensitive attribute)?**

To do this, I performed further feature analysis and examined potential causal relationships between variables within the dataset using Directed Acyclic Graphs (DAGs). The following steps were taken:

1. **Error-aware feature examination:** I compared the distribution of features be-

tween cases that were "correctly predicted" and those that were "incorrectly predicted".

2. **Average Causal Mediating Effect (ACME):** I calculated the ACME of all features on Sex and the target variable of Psychiatric care, to identify which may be acting as a mediators and involved in causal pathways.
3. **Causal Paths:** I drew out causal graphs, taking into account the previously calculated ACME scores, to identify both fair and unfair paths that may have contributed to the observed sex disparity in algorithmic performance.

To quantify the effect of the paths present in our causal diagrams, I used causal mediation analysis to evaluate the direct effect of Sex on Psychiatric Care, and the indirect effects (via other variables). To do this, the Mediation package from the statistics library "statsmodels" was used. The code included a mediation analysis loop, iterating over each mediator variable (which is all of the other features used to predict the outcome), creating two regression models:

1. **The outcome model:** The outcome model predicted the dependent variable using both the independent variable and the mediator
2. **The mediator model:** The mediator model predicted the mediator, using the independent variable, assessing how much the independent variable influenced the mediator.

Both models were fitted and the following effects were calculated for the features in the dataset:

- Total Effect (TE)
- Average Direct Effect (ADE)
- Average Causal Mediation Effect (ACME)

5.2.7 Stage 5: Causal Fairness Adjusted Model

In the final stage I used the information derived from the first four stages, to adjust model development to account for mediating pathways that may be influencing the identified algorithmic disparities. For features were the ACME was identified as a potential contributor to model disparity, this ACME score was used to create an "adjustment metric" to downplay its effect. Thus, I used the adjustment metric to downregulate or upregulate specific features, to mediate their effect on the model prediction, in an attempt to reduce the observed demographic disparity in performance.

5.3 Results

5.3.1 Stage 1: Model development and feature analysis

Descriptive Statistics

The original dataset consisted of 28519 instances with 47 features, which was narrowed down to a the selected features detailed in Tables 5.2 to Table 5.4. I examined algorithmic performance differences across demographic subgroups defined by Sex, Age Group, BMI and Handedness. Table 5.5 below provides the breakdown of the population by these features and the target variable, and was used to identify any issues of class imbalance within the subgroup categories.

Table 5.5: Biobank Dataset: Count of patients within each demographic subgroup, stratified by the target variable (Psychiatric care). The percentage of positive instances with respect to the target variable are provided for comparison across subgroups

		Psychiatric Care	
Demographic Feature		Yes	No
Sex	Female	1561 (10.3%)	13630
	Male	1130 (8.5%)	12196
Handedness	Left	309 (10.0%)	2772
	Right	2356 (9.4%)	22828
Age Group	Group 1	844 (9.9%)	7707
	Group 2	1139 (9.2%)	11190
	Group 3	682 9.2%)	6703
BMI Binary	Healthy	1005 (9.0%)	10160
	Overweight	1660 (9.7%)	15440

Feature Evaluation

The features in the dataset were evaluated for their importance in predicting psychiatric care, from which our series of "selected features" were chosen for model development (Tables 5.2 to Table 5.4). The relative importance of each of these features are presented in Figures 5.6 to 5.7, which highlight their rankings according to (i) SHAP values, (ii) Mutual Information and (iii) Recursive Feature Elimination.

Figure 5.6: Biobank Dataset: Feature rankings for the dataset, based on SHAP values and ordered by magnitude.

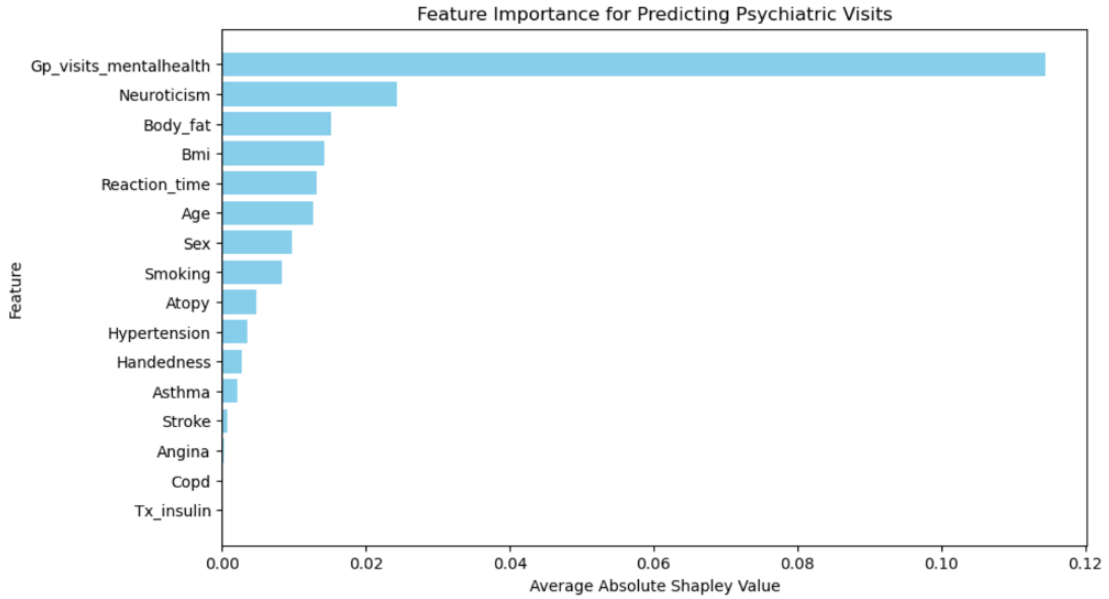


Figure 5.7: Biobank Dataset: Feature Rankings based on Average Mutual Information (n=50)

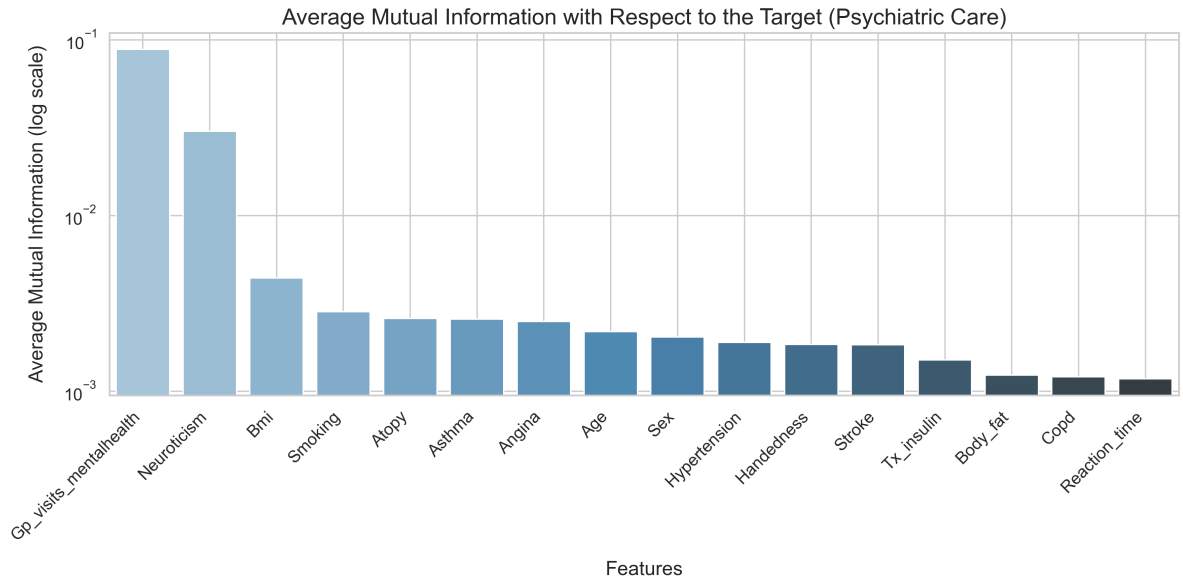
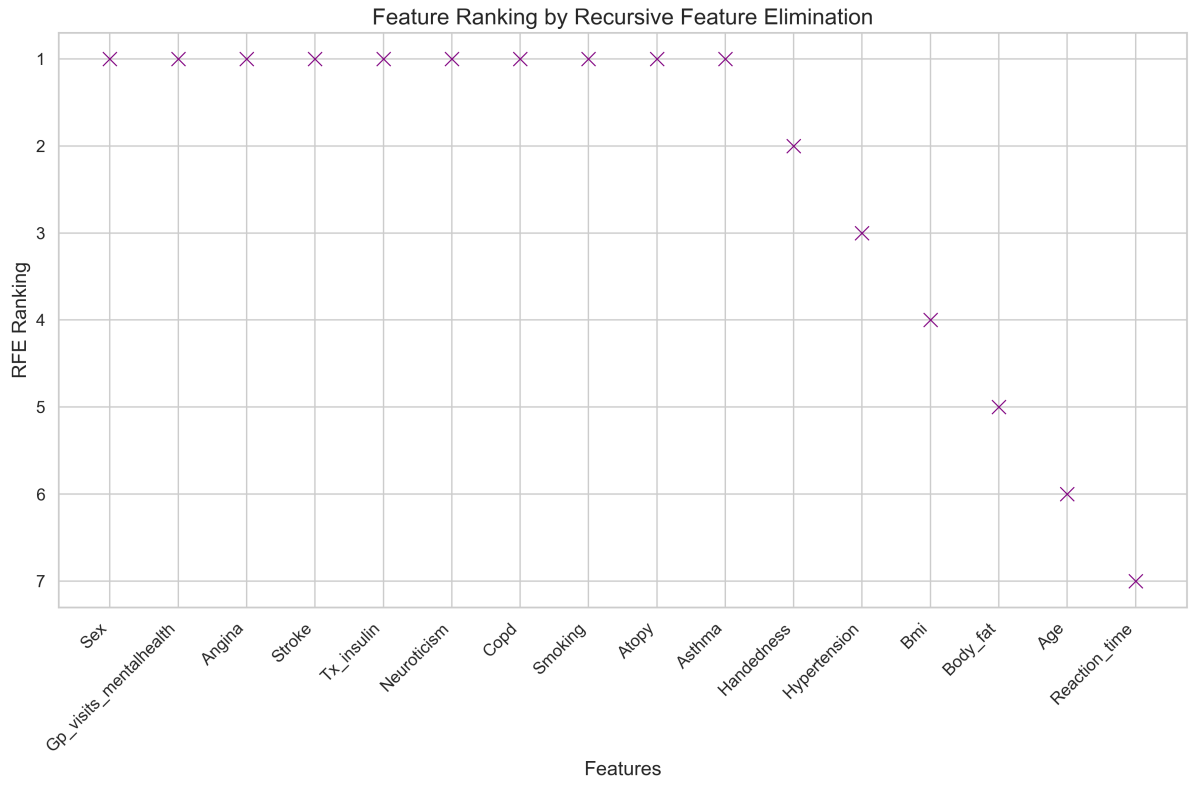


Figure 5.8: Biobank Dataset: Feature rankings for the dataset, based on values from Recursive Feature Elimination (RFE) evaluation (Position 1 being the greatest contributor)



Stage 1: Model Performance

The features described in the section above were used to build our models predicting psychiatric care, for which I deployed a series of algorithms: L*ogistic Regression, XG Boosts, Random Forest Classifiers, Support Vector Machines and Neural Networks.* I first examined the model performance overall for all patients, and then broke this down by subgroups to identify any disparities, using the methods described in the previous chapter. The evaluative metrics describing performance across all patients in the dataset are provided in Table 5.6, alongside Table 5.7 which details the best parameters identified and used for each model. For the following stages below, I then chose to focus on the Logistic Regression Model, as this algorithm gave the best overall performance when taking into account the Accuracy, F1 and ROC Scores (Table 5.6). The Logistic Regression model had an equivalent ROC Score to the Neural Network, with much better error rates and F1 Score (LR F1 Score 0.43 vs NN F1 Score 0.11).

Table 5.6: Predicting psychiatric care: Model performance on test set, inclusive of all patients

Performance on test set	**Logistic Regression**	**XG Boost**	**Random Forest Classifier**	**Support Vector Machine**	**Neural Network**
Accuracy	0.84	0.79	0.79	0.82	0.91
ROC_AUC	0.88	0.87	0.88	0.88	0.88
F1 Score	0.43	0.43	0.44	0.43	0.11
FNR	31.90	0.14	0.13	0.29	0.94
FPR	64.41	0.22	0.21	0.16	0.01

Table 5.7: Predicting psychiatric care: Details of Parameters for Each Machine Learning Model

Logistic Regression	XG Boost	Random Forest Classifier	Support Vector Machine	Neural Network
C: 0.01 (Class weight: {0: 0.55, 1: 5.23}, Solver: liblinear)	Learning rate: 0.1 (Max depth: 5, N Estimators: 200)	Min samples left: 4 (Min samples split: 20, N Estimators: 100)	svm_estimator_C: 0.1	Model: sequential_3 (Total params: 3905)

5.3.2 Stage 2: Is there a disparity in model performance?

My next step was to re-examine these performance metrics for the subgroups separately in order to identify any performance inequities. I adopted the same approach to that detailed in Chapter 4, in which I printed the results of the bootstrapping techniques,

detailing the performance metrics of the models for each subgroup over 100 experimental runs. The mean performance metrics were calculated across the runs (n=100), and mean scores are presented alongside standard deviation in Table 5.8. Table 5.9 to Table 5.10 then provide the mean difference in performance for each metric, for each subgroup with accompanying p-values for statistical significance. As per the pre-ceeding chapters, the sex-performance disparity was calculated as per Equation 3.3 (males minus females), thus a positive result indicates a higher value for males.

The most significant performance difference was found between the female and male subgroups, in terms of both overall performance and the error rates (Table 5.9). On comparing males and females, we see that the **performance of the models for males is 12.56% higher in terms of accuracy (p<0.01), with a significantly lower false positive rate (18.01% Males vs 32.41% females, mean difference -14.39%, p<0.01)** (See Table 5.10). The clinical context here, of the model's high false positive rate, would manifest in the real world as an increased pathologisation of healthy females, which mirror existing concerns of medical sociologists who have highlighted the historic and ongoing the misdiagnosis and pathologisation of women and non binary patients in psychiatry [115, 123].

In Figures 5.9 to 5.12 I have presented each of the the global performance metrics (ROC and Accuracy) and the specific error rates (FNR and FPR) for each subgroup. Figure 5.9 highlights that the accuracy for most subgroups fall close to the overall mean accuracy of 76.28%, with the exception of the female patienta where the mean accuracy sits at 70.33%. The same pattern is observed when measuring by ROC score, as illustrated in Figure 5.10. Figure 5.11 and 5.12 dive into the disparities existing in the error rates, where Figure 5.12 demonstrates the most dramatic difference between the females (FPR 32.4% +- 0.9) and males (FPR 18.019% +- 0.8).

Table 5.8: Predicting Psychiatric Care: Mean performance metrics by demographic subgroup

Group	Accuracy	ROC	FNR	FPR
Overall	76.208 ± 0.617	87.909 ± 0.683	6.289 ± 1.148	25.608 ± 0.662
Male	82.889 ± 0.755	90.319 ± 0.935	7.343 ± 1.822	18.019 ± 0.809
Female	70.331 ± 0.826	85.446 ± 0.915	5.522 ± 1.214	32.410 ± 0.896
Low BMI	76.218 ± 0.921	87.447 ± 1.219	7.666 ± 1.858	25.387 ± 0.982
High BMI	76.200 ± 0.773	88.223 ± 0.849	5.452 ± 1.358	25.755 ± 0.833
Young	75.743 ± 1.132	87.151 ± 1.297	6.984 ± 1.954	26.136 ± 1.204
Middle aged	74.967 ± 0.906	87.675 ± 1.055	6.023 ± 1.758	26.969 ± 0.960
Old	78.823 ± 1.023	89.218 ± 1.239	5.923 ± 2.112	22.725 ± 1.119
Left-handed	76.186 ± 0.623	87.874 ± 0.765	6.252 ± 1.260	25.620 ± 0.666
Right-handed	76.391 ± 1.691	88.160 ± 1.900	6.560 ± 3.437	25.509 ± 1.856

Table 5.9: Predicting Psychiatric Care: Group Differences in Model Global Performance Metrics, *nb. the sex-performance disparity was calculated as per Equation 3.3 (males minus females), thus a positive result indicates a higher value for males.*

	Accuracy		ROC	
	Mean Difference (%)	P Value	Mean Difference (%)	P Value
Sex	12.5580	0.0000	4.8733	0.0000
BMI	−0.0183	0.8792	0.7759	0.0000
Handedness	0.2049	0.2568	0.2860	0.1642
Young vs. Middle	0.7761	0.0000	−0.5239	0.0020
Young vs. Old	−3.0801	0.0000	−2.0663	0.0000
Middle vs Old	−3.8562	0.0000	−1.5424	0.0000

Table 5.10: Predicting Psychiatric Care: Group Differences in model error rates (FNR and FPR), *nb. the sex-performance disparity was calculated as per Equation 3.3 (males minus females), thus a positive result indicates a higher value for males*

	FNR		FPR	
Group	Mean Difference (%)	P Value	Mean Difference (%)	P Value
Sex	1.8206	0.0000	−14.3908	0.0000
BMI	−2.2139	0.0000	0.3680	0.0047
Handedness	0.3076	0.4018	−0.1108	0.5748
Young vs. Middle	0.9612	0.0003	−0.8331	0.0000
Young vs. Old	1.0604	0.0003	3.4108	0.0000
Middle vs Old	0.0992	0.7185	4.2439	0.0000

Figure 5.9: Predicting Psychiatric Care: Violin plot demonstrating difference in model performance across the demographic subgroups, measured by Accuracy

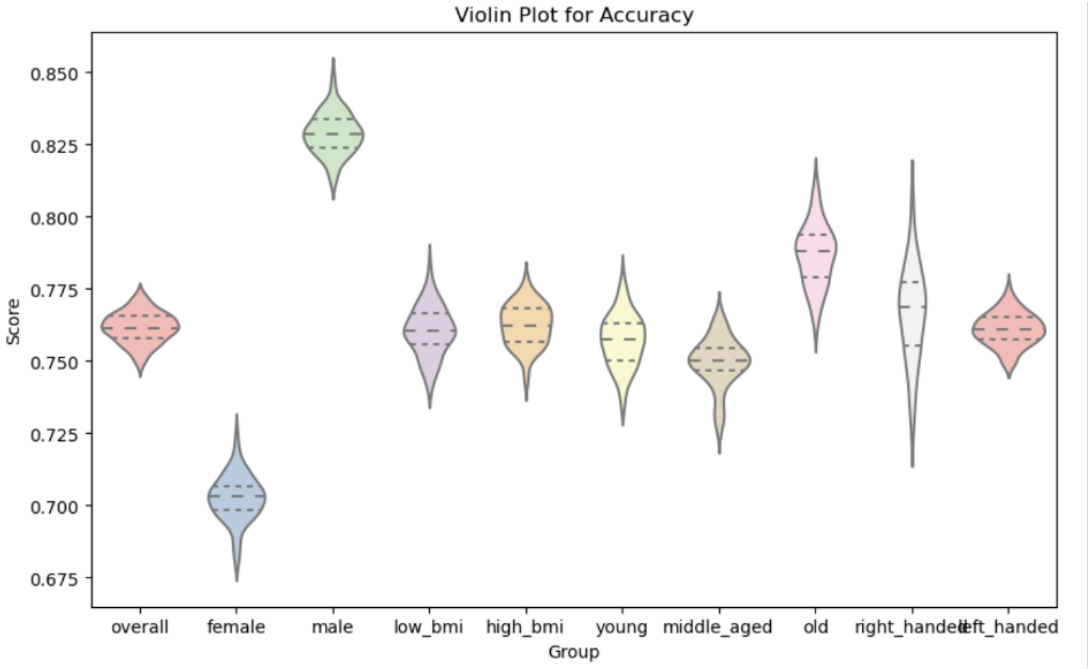


Figure 5.10: Predicting Psychiatric Care: Violin plot demonstrating difference in model performance across the demographic subgroups, measured by ROC Score

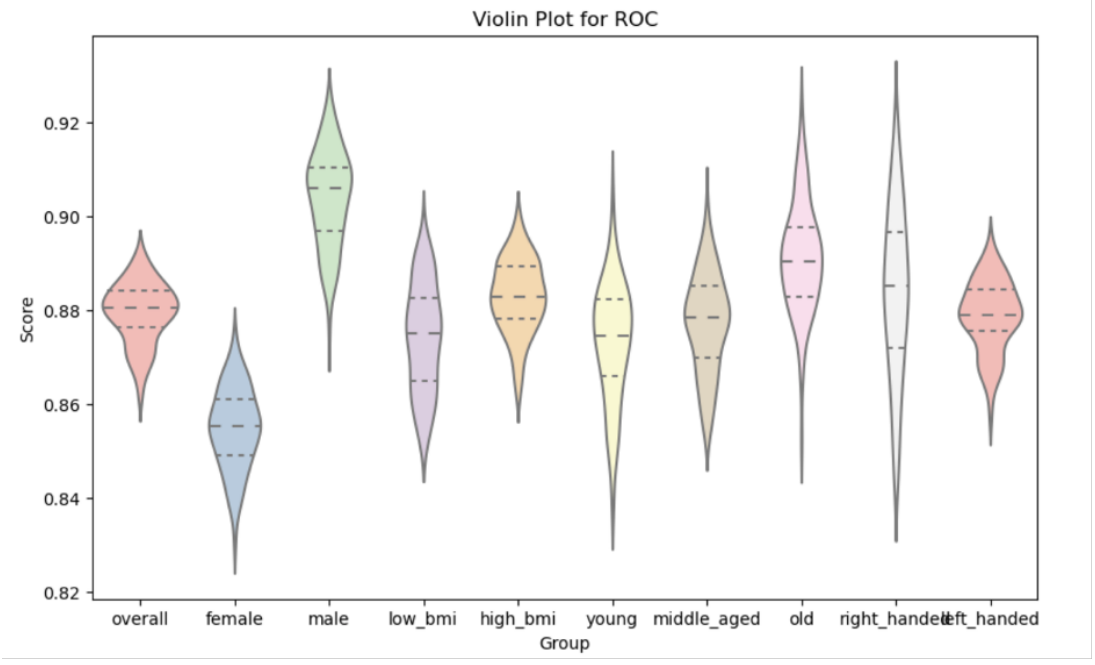


Figure 5.11: Predicting Psychiatric Care: Violin plot demonstrating difference in model performance across the demographic subgroups, measured by False Negative Rate

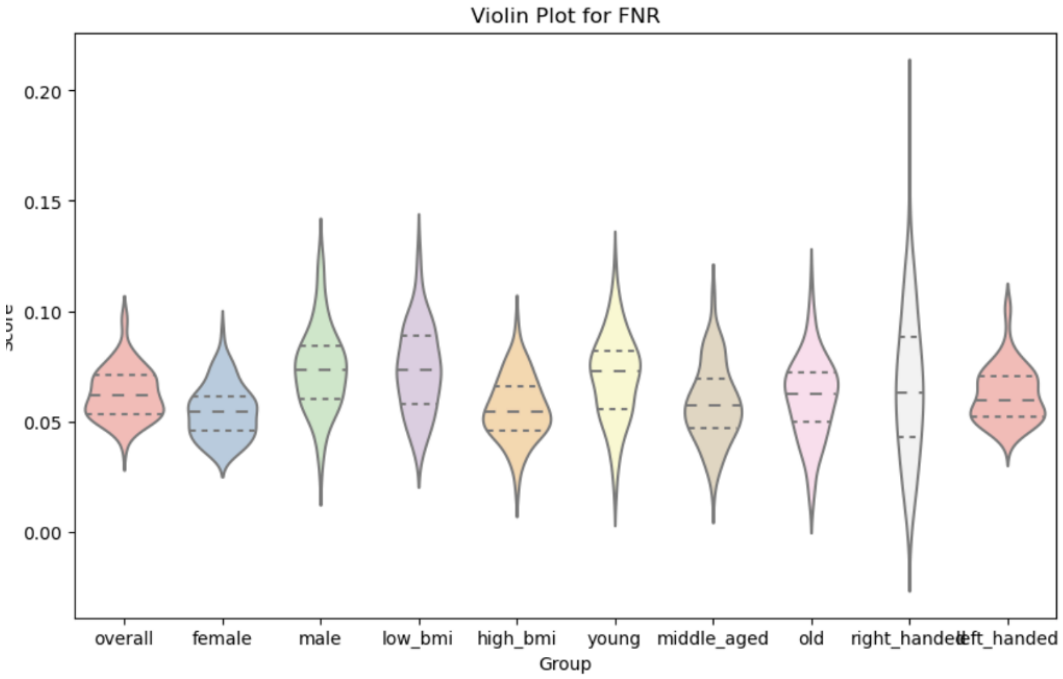
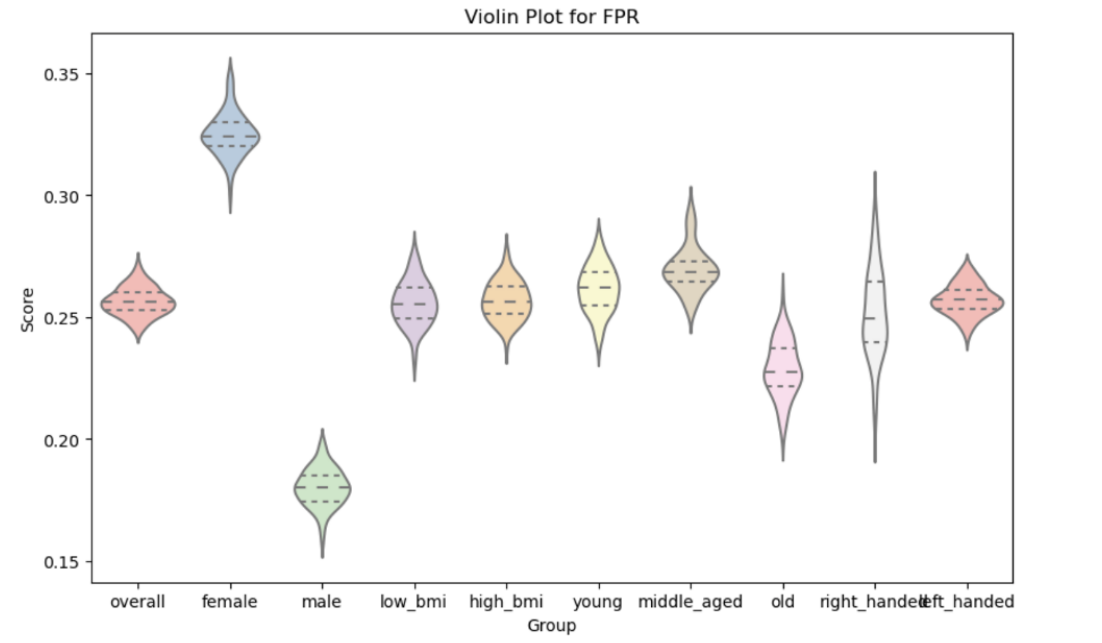


Figure 5.12: Predicting Psychiatric Care: Violin plot demonstrating difference in model performance across the demographic subgroups, measured by False Positive Rate



5.3.3 Stage 3: Is there a causal relationship between the sensitive attribute and the prediction?

In this section we dive into the causal mechanisms that may underpin the algorithmic sex disparity identified in the last stage. First, I quantified the total effect (TE) of all variables in the dataset, including Sex, on the Target Variable (Table 5.11). The confidence interval gives us the range of values in which we can be confident that the true parameter lies, thus it is important to note here that the confidence intervals for Sex includes zero, hence this required more investigation. The variables that ranked top in terms of Total Effect were:

1. (i) GP Visits for Mental Health (TE 1.65)
2. (ii) Neuroticism Score (TE 0.386)
3. (iii) BMI Score (TE 0.125)
4. (iv) Smoking (TE 0.081) (See Table 5.11.

It appeared that "GP visits for mental health" were the most influential predictor of psychiatric care, with higher GP visits being strongly associated with an increased likelihood of having secondary psychiatric care. Similarly, the Neuroticism score (TE 0.386) indicates that higher neuroticism scores were also associated with a higher likelihood of visiting the psychiatrist. BMI, Smoking and Age demonstrate significant but smaller effects, where an increase in their value increased the likelihood of the patient needing psychiatric care.

Table 5.11: Predicting psychiatric care: Estimated total effects with standard errors and confidence intervals

Feature	Total Effect (Coefficient)	Standard Error	95% CI Lower	95% CI Upper
GP Visits Mental Health	1.651	0.039	1.575	1.727
Neuroticism	0.386	0.023	0.341	0.431
BMI	0.125	0.041	0.046	0.205
Sex	0.087	0.049	-0.009	0.182
Smoking	0.081	0.023	0.036	0.125
Age	0.053	0.026	0.003	0.104
Atopy	0.047	0.022	0.003	0.091
Tx Insulin	0.038	0.021	-0.003	0.078
Reaction Time	0.029	0.023	-0.016	0.074
Angina	0.028	0.020	-0.011	0.068
Stroke	0.024	0.019	-0.013	0.062
Handedness	0.024	0.023	-0.020	0.069
Asthma	0.023	0.022	-0.020	0.065
COPD	-0.006	0.020	-0.046	0.033
Hypertension	-0.021	0.023	-0.066	0.025
Body Fat	-0.121	0.057	-0.232	-0.010

Stage 2: Counterfactual World

To explore this in greater depth, I next created our counterfactual world, in which all variables were kept the same, except the females were treated as males. Once the counterfactual dataset was created, I ran the same analysis described in Stage 2, calculating the performance metrics for the counterfactual males (females treated as male in the counterfactual world). Table 5.12, and Figure 5.13 to Figure 5.16, demonstrate that when females are treated as males, the performance of the model improves in terms of both global performance metrics and the sub-group error rates (Table 5.12).

Examining this in greater depth, Table 5.12 shows the mean performance metrics for the females in the original and counterfactual worlds. The column detailing the **"Counterfactual Difference"** provides the change in each performance metric when the original females are treated as males, summarised in Equation 5.6. From these results we see that the model performance for female improves in the counterfactual world, with the greatest change seen in the False Positive Rate (CF Difference 6.9%, Original Females 32.5%, Counterfactual Males 25.7%, $p<0.01$) and Accuracy Score (CF Difference -5.9%, Original Females 70.3 %, Counterfactual Males 76.2%, $p<0.01$). These results are presented graphically in Figures 5.13 to 5.16, where we see the original sex disparity in performance closing in the counterfactual world.

$$\text{CF Difference} = \text{Original Females Mean}(\%) - \text{Counterfactual Males Mean}(\%) \quad (5.6)$$

Equation 5.6: Equation for calculating the "Counterfactual Difference" presented in Table 5.12, representing the change in scores for females, when treated as males

Table 5.12: Comparison of Original and Counterfactual Means with Statistical Significance. The Counterfactual difference refers to the difference between the scores for Original Females and Counterfactual Males, this representing the change in score when females are treated as males

Metric	Counterfactual Difference	P Value	Overall Original Mean (%)	Overall Counterfactual Mean (%)	Original Females Mean (%)	Counterfactual Males (Originally Females) Mean (%)	Original Males Mean (%)
Acc	-5.894	0.000	76.173	76.172	70.278	76.172	82.900
ROC	-2.599	0.000	87.554	87.514	84.916	87.514	90.119
FNR	-0.900	0.000	6.178	6.188	5.287	6.188	7.382
FPR	6.828	0.000	25.667	25.667	32.495	25.667	18.011

Figure 5.13: Counterfactual effects: Performance of model predicting psychiatric care for all patients, original females and males, and counterfactual males, measured by Accuracy

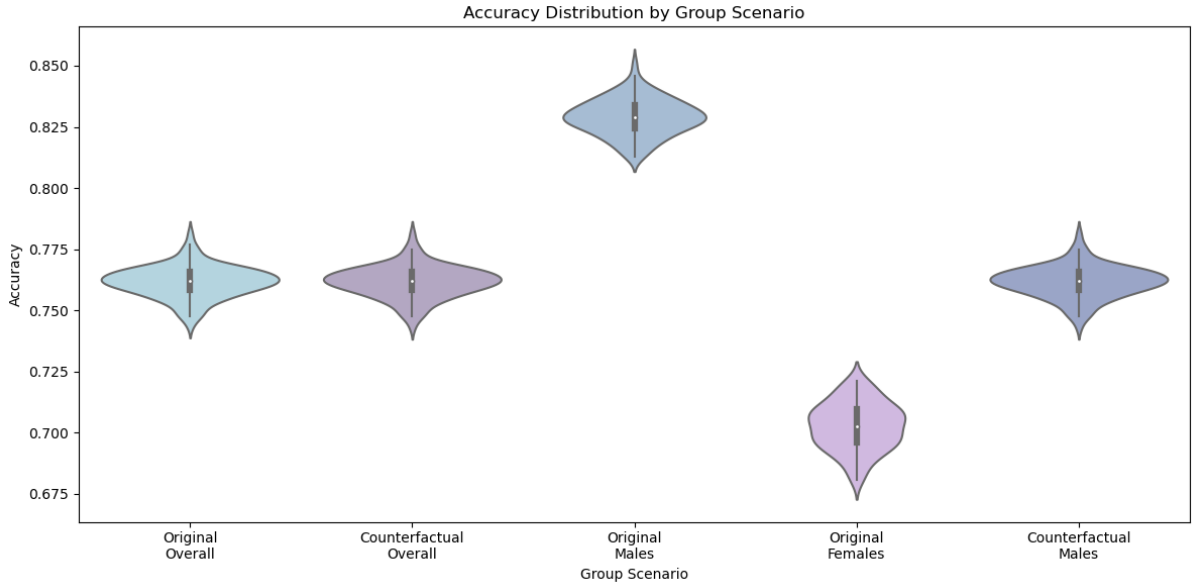


Figure 5.14: Counterfactual effects: Performance of model predicting psychiatric care for all patients, original females and males, and counterfactual males, measured by ROC Score

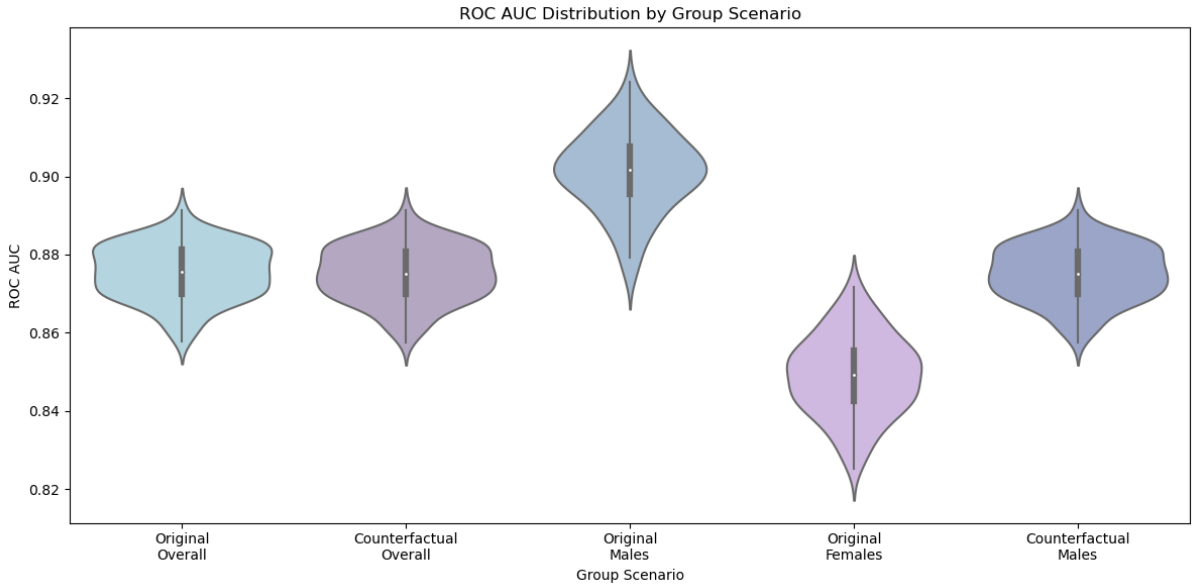


Figure 5.15: Counterfactual effects: Performance of model (False Negative Rate) for all patients, original females and males, and counterfactual males

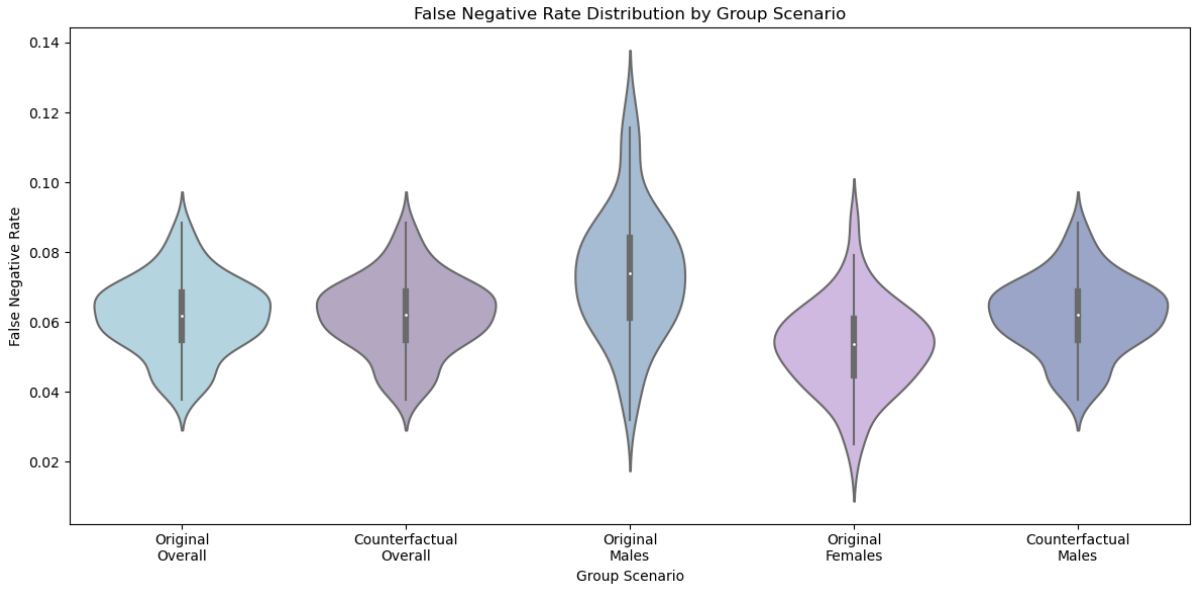
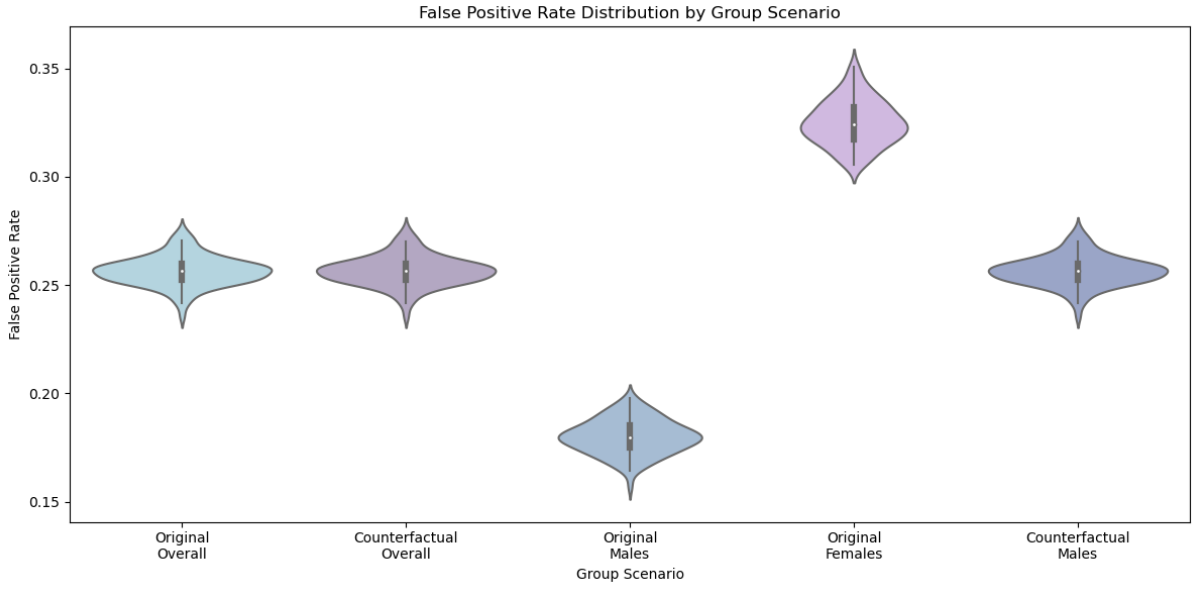


Figure 5.16: Counterfactual effects: Performance of model (False Positive Rate) for all patients, original females and males, and counterfactual males



5.3.4 Stage 4: Is the causal path fair or unfair?

The findings from our counterfactual analysis indicates we need to examine the causal pathway between sex and the outcome more closely. The initial findings suggest that there could be a causal path here between Sex and the model prediction, that is affecting the inequity observed. However, we cannot truly determine this without examining the causal paths involved and the potential influence of mediating factors. While the counterfactual methods have allowed us to exposed sex-specific differences, we still need to answer the question as to whether this observation is due the patients sex, or another factor that co-varies with it. To examine this element, I now turn to our causal diagrams and the evaluation of causal mediating effects.

Our next step was to understand the causal path, in order to determine whether it is "fair", or "unfair" - for example, is the path mediated by a fair explanatory variable as was the case in the Berkley admissions case described in the Introduction. To map out the causal path and better understand the other variables in the dataset, I took the following steps:

1. **Error aware feature examination:** I compared the distribution of features between cases that were "correctly predicted" and "incorrectly predicted" instances, to identify which features may be relevant to prediction errors.
2. **Causal Paths:** Causal graphs were drawn out, taking into account the ACME, to identify both fair and unfair paths that may contribute to the observed disparity.
3. **Average Causal Mediating Effect (ACME):** I calculated the ACME of all features, to examine which may be acting as a mediator in the relationship between Sex and the Target of psychiatric care.

(i) Feature Examination and Incorrect Predictions

To delve into the underlying reasons why females were experiencing a greater error rate, I performed descriptive statistics on the "correctly predicted", and "incorrectly predicted" females - with the intention of understanding how these cohorts differed in terms of the other features in the dataset (e.g. were women with a high BMI more commonly "mis-predicted"). Figures 5.17 to 5.21 demonstrate the different distribution in characteristics between these groups, for example, there seems to be a much greater proportion of females with a history of GP visits for their mental health, who are incorrectly predicted. Further, there is a noticeable difference in the Neuroticism score, with females who receive a wrong prediction having a higher neuroticism score than their correctly predicted counterparts (Fig 5.17). The noticeable difference in the distribution of these two variables - GP Mental Health visits and Neuroticism Score - drew our attention to these factors as potential mediators of the sex disparity, and are explored in greater depth below.

Figure 5.17: Predicting psychiatric care: A comparison of the distribution of the Neuroticism Score between the correctly predicted, and incorrectly predicted, females in the dataset

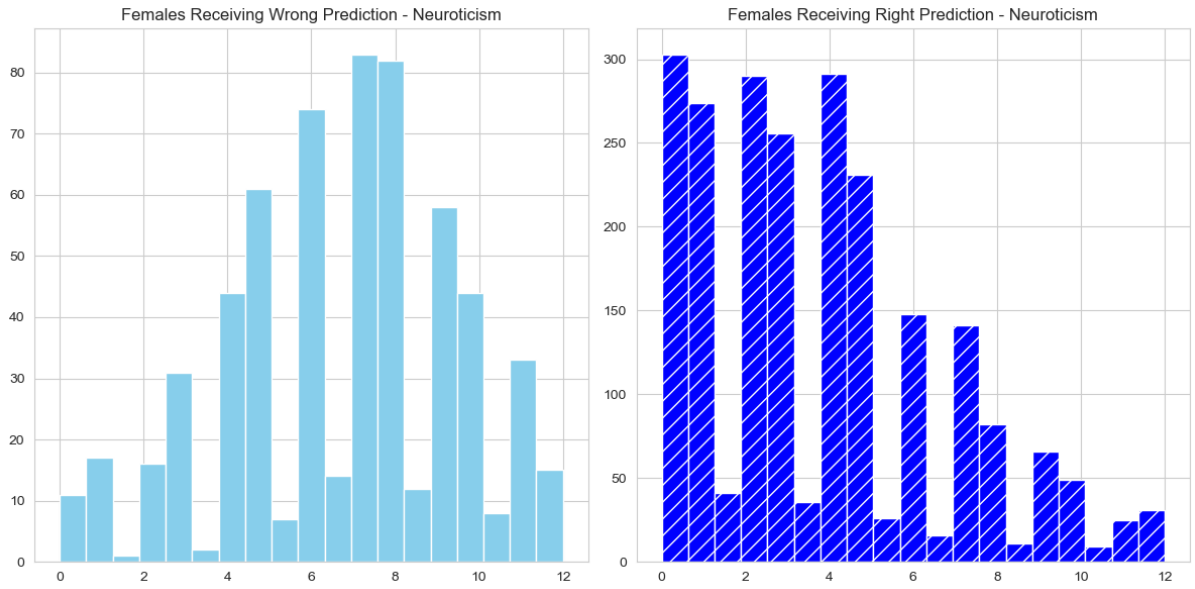


Figure 5.18: Predicting psychiatric care: A comparison of the distribution of GP visits for mental health between the correctly predicted, and incorrectly predicted, females in the dataset

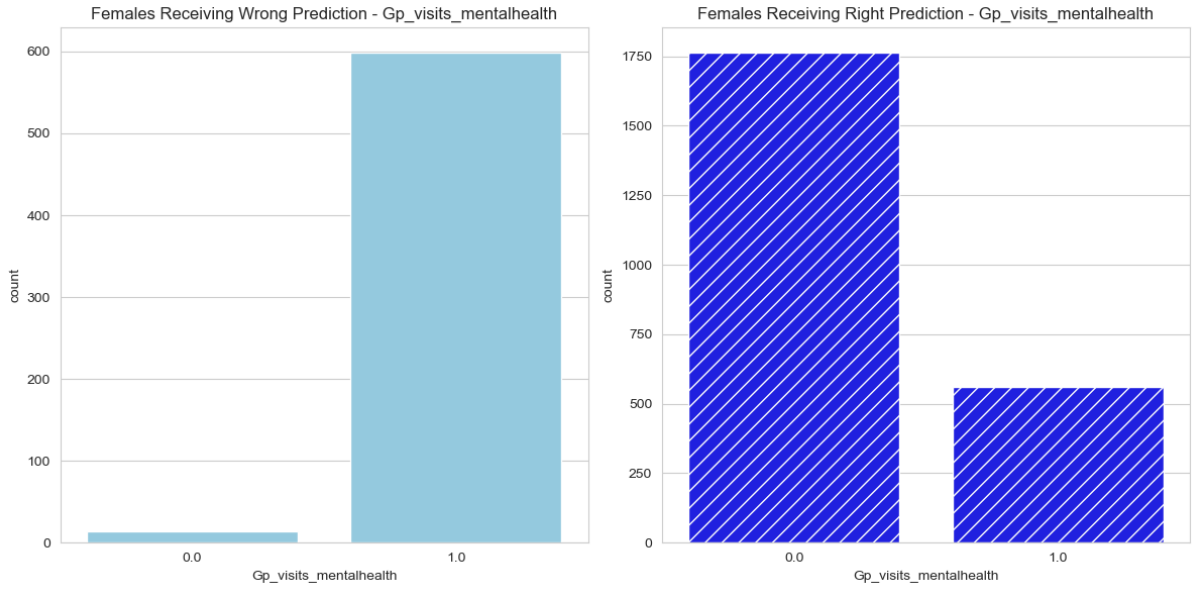


Figure 5.19: Predicting psychiatric care: A comparison of the distribution of age between the correctly predicted, and incorrectly predicted, females in the dataset

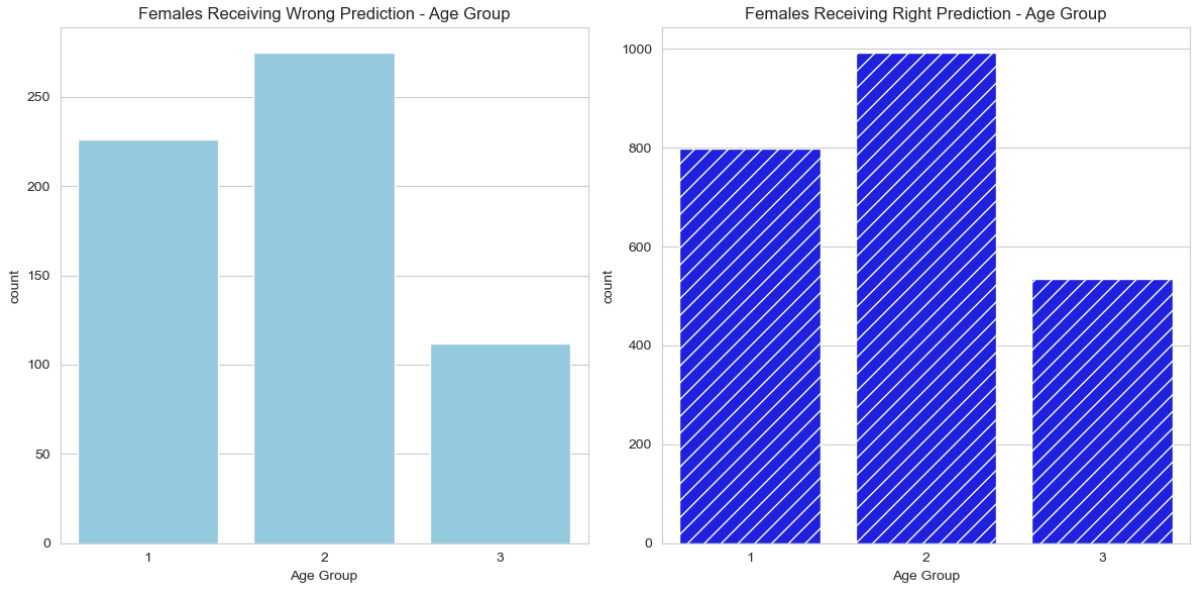


Figure 5.20: Predicting psychiatric care: A comparison of the distribution of the Smoking variable between the correctly predicted, and incorrectly predicted, females in the dataset

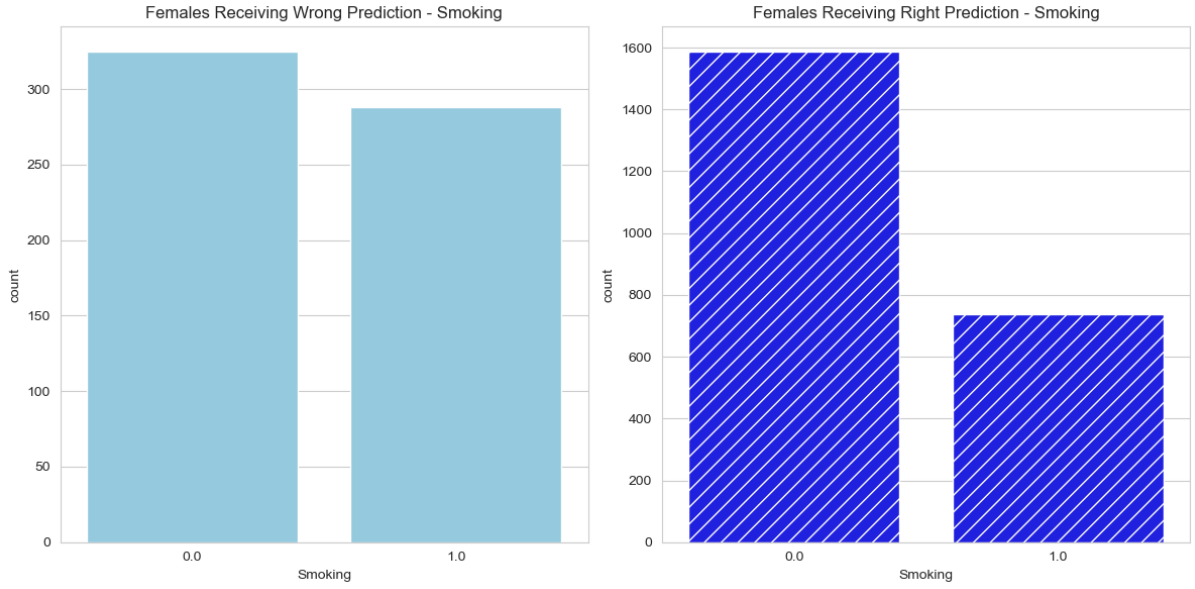
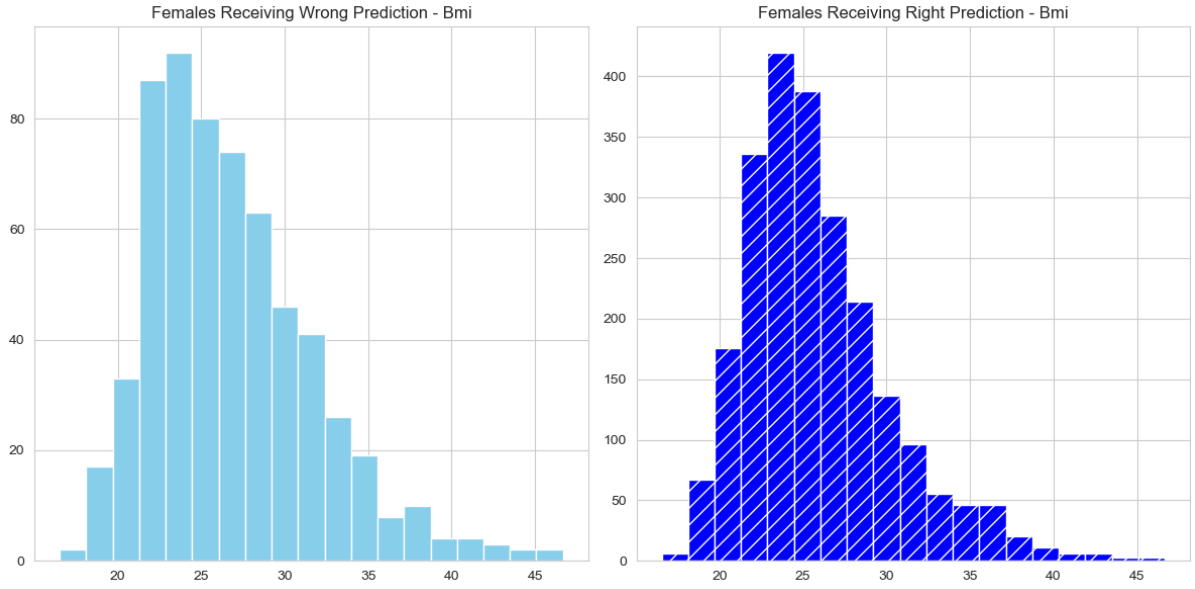


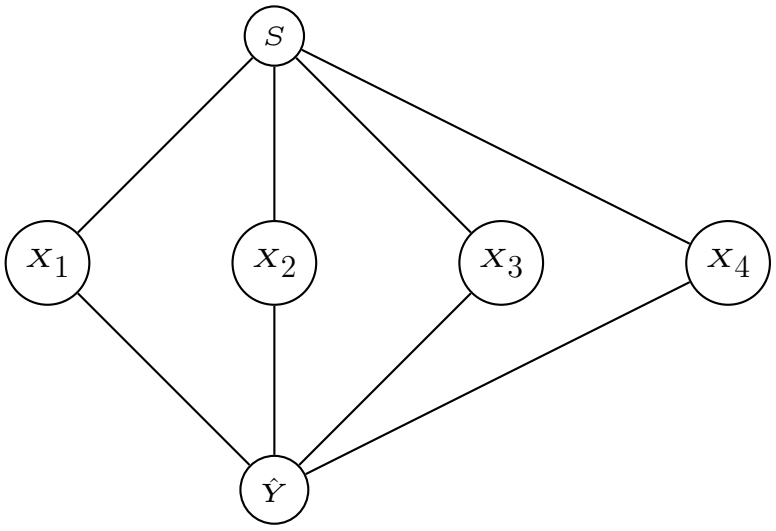
Figure 5.21: Predicting psychiatric care: A comparison of the distribution of BMI between the correctly predicted, and incorrectly predicted, females in the dataset



(ii) Causal Paths: Directed Acyclic Graphs (DAGs)

In order to understand the worst performance for female patients I explored the potential relationship between Sex (S), Psychiatric Care (Y) and the remaining variables which may act as mediators (Figure 5.22). The relationships between the features and the target variable were first visualised using Directed Acyclic Graphs (DAGs). The relations illustrated with the graphs were then quantified using average causal mediation effects.

Figure 5.22: Predicting psychiatric care: Causal graph for the relationship between Sex (S), psychiatric care (Y), and the potential mediating variables ($X_1..X_4$, e.g. Age, BMI, Neuroticism score)



(iii) Average Causal Mediating Effects

To understand which of the paths presented in Figure 5.22 were most likely to be contributing to the algorithmic sex disparity, I performed a mediation analysis on the relationship between the independent variable Sex, and the dependent variable of Psychiatric Care. The analysis allows us to understand whether there is a direct relationship between the independent and dependent variable, and how it may be mediated through the other variables. The results of our mediation analysis are presented in Table 5.13, which details the results for features with the greatest magnitude of effect. For this section it is helpful to note the following definitions:

1. **Average Causal Mediating Effect (ACME):** The change in outcome attributed to the indirect path through the mediator, reflecting the mediator's effect size in transmitting the influence of Sex onto the target.
2. **Average Direct Effect (ADE):** The change in the outcome attributed to the independent variable (sex) not through the mediator i.e. the direct effect of Sex on psychiatric visits.

3. **Total Effect (TE):** The total effect of the independent variable on the outcome, which is the sum of ADE and ACME.
4. **Proportion Mediated:** The proportion of the total effect that is mediated by the mediator.

Table 5.13 demonstrates the results of our causal mediation analysis, where **features are ranked by the significance of their mediating effect on the relationship between Sex and Psychiatric Care**. Here we see that the direct effects (ADE) of sex on psychiatric care were significant in most cases, but the value of the ADE differs between variables. This first appears counterintuitive as the value represents the direct effect of Sex on Psychiatric care, which might be assumed to stay constant across the mediators. However, this variation can emerge due to:

1. Differences in each mediation model built for each variable
2. Unobserved confounder variables that influence both the mediator and the outcome, and
3. Interaction effects between the independent variable and the mediator which are not captured when evaluated separately.

Table 5.13: Predicting psychiatric care: Summary of Causal Mediation Analysis Results

Variable	Measure	Estimate	Lower CI Bound	Upper CI Bound	P-value
GP Visits	ACME (average)	−0.0392	−0.0432	−0.0354	<0001
	ADE (average)	0.0219	0.0161	0.0282	<0001
	Total Effect	−0.0172	−0.0246	−0.0099	<0001
	Prop. Mediated (average)	2.2660	1.6828	3.7091	<0001
Neuroticism	ACME (average)	−0.0226	−0.0248	−0.0203	<0001
	ADE (average)	0.0051	−0.0014	0.0115	0.1320
	Total Effect	−0.0174	−0.0240	−0.0105	<0001
	Prop. Mediated (average)	1.2870	0.9377	2.0605	<0001
Smoking	ACME (average)	0.0025	0.0019	0.0033	<0001
	ADE (average)	−0.0196	−0.0267	−0.0133	<0001
	Total Effect	−0.0171	−0.0243	−0.0106	<0001
	Prop. Mediated (average)	−0.1489	−0.2545	−0.0914	<0001
Body Fat	ACME (average)	−0.0157	−0.0222	−0.0098	<0001
	ADE (average)	−0.0015	−0.0101	0.0074	0.7480
	Total Effect	−0.0172	−0.0240	−0.0106	<0001
	Prop. Mediated (average)	0.9135	0.5104	1.6171	<0001
BMI	ACME (average)	0.0028	0.0019	0.0037	<0001
	ADE (average)	−0.0200	−0.0267	−0.0128	<0001
	Total Effect	−0.0172	−0.0240	−0.0102	<0001
	Prop. Mediated (average)	−0.1585	−0.2827	−0.0980	<0001

The features of "GP Visits for mental health" and "Neuroticism" showed high proportions of mediation, suggesting these are significant pathways through which sex influences psychiatric care (Table 5.13). On examining the role of Neuroticism, we see that the ACME was significant and negative (-0.0226), with a very high proportion of the effect being mediated (128.7%). This indicates that the direct effect of being female on psychiatric care is significantly influence by Neuroticism, as per the path presented in Figure 5.23.

Figure 5.23: Predicting psychiatric care: Causal pathway from Sex (S) to Psychiatric Care (P), mediated by Neuroticism (N), which has an Average Causal Mediating Effect (ACME) of -0.023 [Table 5.13]

$$S \longrightarrow N \longrightarrow P$$

The negative Neuroticism ACME reveals an important finding in the original data, suggesting that higher Neuroticism scores reduce the likelihood of psychiatric care after

controlling for sex - indicating that higher Neuroticism scores do not necessarily translate to higher care needs for females. These findings may explain the inflated false positive rate for females that we have observed. The females in the dataset have higher baseline scores on the Neuroticism scale, hence if the predictive model overestimates the impact of Neuroticism on psychiatric care needs, females with higher Neuroticism scores may be over-predicted, leading to more false positives.

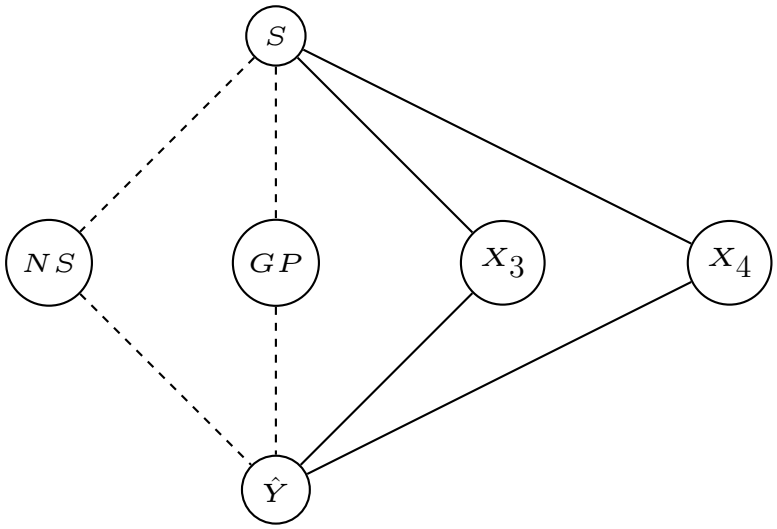
The Average Causal Mediation Effect (ACME) for "GP Visits for Mental Health" was also significant and negative, with an ACME of -0.0391 ($p<0.01$). The effect suggests that as GP visits for mental health increases, the likelihood of requiring psychiatric care decreases after controlling for sex. One explanation for this may be that increased engagement with the GP for mental health might reduce the need for psychiatric care. However, similarly to the issue observed with Neuroticism scores above, if the model does not account for this mediation pathway, it may overestimate psychiatric care for females who frequently visit the GP. Given that the females have a higher baseline rate of engaging with the GP in the original dataset, this may lead to the higher false positive rate impacting females.

Smaller mediation effects are seen for the variables of Smoking, Body Fat and BMI. Smoking has a positive ACME of 0.0025 ($p<0.01$), suggesting that smoking slightly increases the likelihood of psychiatric care amongst females (Table 5.13). Interestingly, the variables of BMI and Body Fat have mediating effects in opposing directions, with ACMEs of 0.0028 ($p<0.01$) and -0.0157 ($p<0.01$) respectively. This is surprisingly, as often BMI and Body Fat correlate, with Body Fat increasing as BMI increases. Thus if one was a proxy for the other, their causal effects might be expected to be align in the same direction. However, BMI and Body Fat can diverge under certain conditions, especially for individuals with higher muscle mass, with sex (as women tend to have higher body fat percentages than men), with ethnicity, and with height. Thus, these variables may be acting as proxies for other features in the dataset or unobserved confounders.

5.3.5 Stage 5: Causal Fairness Adjusted (CFA) Model

In the final stage I used the insights from the causal analysis to adjust the original Logistic Regression model, to see whether this could reduce the sex-based disparity in performance. In the previous sections I identified that the effect of Sex on psychiatric outcome was significantly mediated by (i) Neuroticism Score and (ii) GP Visits for Mental Health - represented in the now adjusted causal graph below (Figure 5.24).

Figure 5.24: Predicting Psychiatric Care: Adjusted causal graph for the relationship between Sex (S) and psychiatric care (Y), highlighting the significant paths mediated by (i) Neuroticism score (NS) and (ii) GP Visits (GP)



To adjust for the mediating effects of these variables, a new model was built that brought in sex-specific adjustments to these to variables. As detailed in the methods, the ACME scores were used for this, giving the following adjustment factors:

1. **Neuroticism Adjustment Factor (0.02257):** The value is derived from the results of Table 5.13. If a female had a non-zero value for Neuroticism, the prediction probability was multiplied by 1 - ACME, giving an adjusted score.
2. **GP Visits Adjustment Factor (0.03916):** For a female with a non-zero value fro GP Visits, the prediction probability was adjusted by multiplying it by (1 - ACME).

Table 5.14 presents the results of this approach, in which we see our Causal fairness Adjusted (CFA) Model produces a much lower sex disparity in algorithmic performance. We both maintain the performance overall for all patients (in terms of both accuracy and ROC scores), and reduce the disparity between the subgroups for FPR, however not for FNR (Tables 5.15 to Table 5.18). Table 5.15 shows that the mean difference (%) in Accuracy between males and females falls from 12.55% in the original model, to 1.66% in the CFA Model. Further, Table 5.17 shows that the CFA Model exhibits a far lower mean difference in False Positive Rate between the sexes (1.47%), compared to the original model (-14.39%). In contrast however, the False Negative Rate increases significantly for the CFA model, with a mean difference of -28.35% between the sexes, compared to 1.82% in the original model. These findings demonstrate that while the CFA model maintains overall performance and shrinks the sex disparity in the FPR, the error rate manifests instead as a higher FNR affecting the female patients.

Group	Accuracy	ROC	FNR	FPR
Overall	84.716 ± 0.510	87.591 ± 0.624	36.381 ± 2.541	13.098 ± 0.599
Male	85.597 ± 0.899	89.983 ± 0.913	20.056 ± 3.429	13.876 ± 1.091
Female	83.943 ± 0.787	85.067 ± 0.939	48.403 ± 3.510	12.402 ± 0.883
Low BMI	85.720 ± 0.774	86.902 ± 1.061	43.518 ± 4.009	11.415 ± 0.934
High BMI	84.058 ± 0.747	88.024 ± 0.785	32.080 ± 2.820	14.210 ± 0.855
Young	84.492 ± 1.119	86.921 ± 1.375	36.628 ± 4.830	13.194 ± 1.328
Middle Aged	84.147 ± 0.835	87.378 ± 0.929	36.153 ± 3.474	13.784 ± 0.845
Old	85.661 ± 0.906	88.486 ± 1.024	36.453 ± 4.509	12.108 ± 1.082
Left handed	84.819 ± 0.561	87.549 ± 0.709	37.217 ± 2.747	12.918 ± 0.688
Right handed	83.866 ± 1.739	87.921 ± 1.796	29.785 ± 7.832	14.584 ± 2.211

Table 5.14: Performance Metrics by Demographic Subgroup of the Causal Fairness Adjusted (CFA) Model

Table 5.15: Comparison of the Mean Difference in **Accuracy Scores**, for the CFA Model and Original Unadjusted Model

Group Comparison	CFA Model		Original Unadjusted Model	
	Mean Difference (%)	P Value	Mean Difference (%)	P Value
Sex	1.655	0.000	12.5580	0.0000
BMI	−1.661	0.000	−0.0183	0.8792
Handedness	−0.9537	0.0000	0.2049	0.2568
Young vs Middle	0.3450	0.0143	0.7761	0.0000
Young vs Old	−1.1687	0.0000	−3.0801	0.0000
Middle Vs Old	−1.5137	0.0000	−3.8562	0.0000

Table 5.16: Comparison of the **ROC Scores** for the CFA Model and Original Unadjusted Model

Group Comparison	CFA Model		Original Unadjusted Model	
	Mean Difference (%)	P Value	Mean Difference (%)	P Value
Sex	4.9165	0.0000	4.8733	0.0000
BMI	1.1213	0.0000	0.7759	0.0000
Handedness	0.3723	0.0553	0.2860	0.1642
Young vs Middle	−0.4574	0.0064	−0.5239	0.0020
Young vs Old	−1.5649	0.0000	−2.0663	0.0000
Middle Vs Old	−1.1075	0.0000	−1.5424	0.0000

Table 5.17: Comparison of the **False Positive Rate** for the CFA Model and Original Unadjusted Model for FPR

Group Comparison	CFA Model		Original Unadjusted Model	
	Mean Difference (%)	P Value	Mean Difference (%)	P Value
Sex	1.4739	0.0000	−14.3908	0.0000
BMI	2.7956	0.0000	0.3680	0.0047
Handedness	1.6661	0.0000	−0.1108	0.5748
Young vs Middle	−0.5899	0.0002	−0.8331	0.0000
Young vs Old	1.0865	0.0000	3.4108	0.0000
Middle Vs Old	1.6764	0.0000	4.2439	0.0000

Table 5.18: Comparison of the **False Negative Rate** for the CFA Model and Original Unadjusted Model

Group Comparison	CFA Model		Original Unadjusted Model	
	Mean Difference (%)	P Value	Mean Difference (%)	P Value
Sex	−28.3466	0.0000	1.8206	0.0000
BMI	−11.4378	0.0000	−2.2139	0.0000
Handedness	−7.4324	0.0000	0.3076	0.4018
Young vs Middle	0.4748	0.4258	0.9612	0.0003
Young vs Old	0.1743	0.7922	1.0604	0.0003
Middle Vs Old	−0.3005	0.5982	0.0992	0.7185

5.4 Discussion

In this chapter I have introduced unique applications of causal methods and counterfactual fairness to the field of psychiatry algorithms, introducing new techniques for uncovering and mitigating biases in AI models in healthcare. I have adapted causal inference techniques for our purpose, deploying these methods to identify mediating variables that influence disparities in algorithmic performance. Our findings build on the work of previous chapters, again demonstrating significant sex disparities in algorithmic performance and diving into the role of other variables in the dataset that influence the model performance for different subgroups. I have expanded the scope of included subgroups, examining algorithmic performance disparities across additional groups defined by BMI, Age and Handedness. The most significant disparity in model performance was found between the sexes, with female patients experiencing lower accuracy rates and an inflated rate of false positives. Our construction of counterfactual worlds revealed that models performed better for females when they were treated as males, and our examination of path-specific effects exposed the mediating variables contributing to this phenomena.

On examining causal effects within the dataset, initially the direct effect of sex on psychiatric care did not appear significant, however on breaking this down into path-specific effects the mediating pathways were uncovered. Here, it was found that the variables of Neuroticism and GP Visits for Mental Health, were mediating the relationship between Sex and Psychiatric Care. On identifying these pathways the next question becomes - *are these "fair" or "unfair" causal pathways.*

In fair pathways, the mediator is considered to be an "explanatory variable", such as in the Berkley admissions case where the variable of departmental choice was informing the difference in rejection rates affecting male vs. female students (See Chapter 5 - Introduction). Alternatively, in an unfair pathway the mediator may simply be acting as a proxy for the sensitive attribute and thus facilitating unfair discrimination via proxy variables. Proxy variables have been described in depth by Cathy O'neil, who gives the examples of AI models that predict crime using a patient's address as a proxy for race, demonstrating cases where algorithmic discrimination may be missed if one doesn't consider the impact of mediating variables [2].

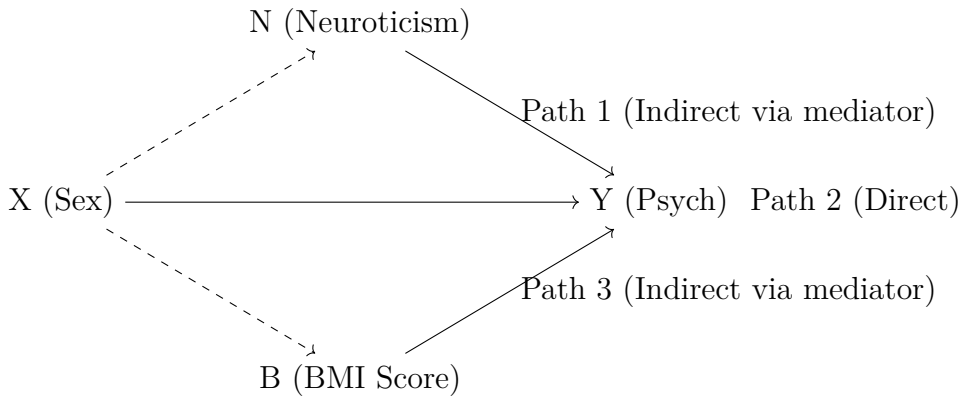
Neuroticism Scores and Sex

In our experiments the Neuroticism score played a major role in the error rate affecting female patients, which is an interesting finding given the heavy critique that this diagnostic framework has received in recent years [115, 205, 218–221]. This scoring system

has been criticised for pathologising women by focusing on typically feminine traits (e.g. worrying) as opposed to masculine traits (e.g. anger). Bauermeister and Gallacher state that the scoring system is prone to gender bias, reflecting wider issues described in the psychiatry domain whereby psychiatric frameworks are biased towards female traits and overly pathologise healthy women [115, 205, 219–221]. In contrast, other researchers have proposed that the higher neuroticism scores observed amongst female patients is a biological phenomenon. For example, Djudiyah and colleagues argued have argued that women are biologically more prone to neuroticism due to "hormonal changes, menstruation, pregnancy and breastfeeding" [222]. Thus, here we see that the philosophical and sociological considerations explored in Chapters 1 and 2 are particularly relevant, as the conceptualisation of what is algorithmically "fair" in this context, is underpinned by our understanding of the roots and validity of these diagnostic frameworks. In one camp sit researchers questioning the validity of these diagnostic tools, while in parallel we see scientists proposing that these tools are accurate means for evaluating disease.

The proponents of either argument will likely fall on opposing sides when answering the question of whether including the "Neuroticism" variable in ML model development is "fair", or whether the causal path in which Neuroticism acts as a mediator is "fair". To provide a similar example that assists in understanding this issue, we can consider other clinical scoring frameworks that have been contested for their biased effects, such as BMI [223]. A growing body of research has extensively criticised the use of the BMI Metric in medicine, due to it's misleading assessments of female patients and those with African heritage [223]. Yet, at the same time we see a continuation of BMI-based medical research that advocates for the utilty of this framework [224]. Thus in both cases, BMI and Neuroticism, when they appear as mediators in a causal path, determining questions of "fairness" require an exploration of the underlying sociological perspectives on these tools. In the section below, Figure 5.25 illustrates the paths from Sex to Psychiatric care, both direct, and indirect via the clinical scores of Neuroticism and BMI, illustrating how they may be considered either "fair" or "unfair". Beneath this, I have listed the opposing philosophical arguments, by which scholars may argue that these pathways are either "fair" or "unfair".

Figure 5.25: Causal graph of potential fair or unfair mediated paths from Sex (S) to psychiatric care (Y)



Argument with path is fair	Argument with path is unfair
Path 1: Sex causes high neuroticism, high neuroticism scores indicate psychiatric disease	Neuroticism scoring frameworks are biased & pathologise female traits. High scores ≠ disease.
Path 2: Sex causes psychiatric disease	Sex does not cause psychiatric disease
Path 3: Sex causes higher BMI, higher BMI causes psych disease	BMI scoring frameworks are sexist/racist, high scores ≠ disease

Our research findings provide a useful contribution to this debate. The causal mediating effect of Neuroticism demonstrated that - when adjusting for sex - higher neuroticism scores did not translate to higher utilisation of psychiatric care services. This finding indicates that while females may score higher on the Neuroticism framework, this does not translate to mental illness, supporting the research that these scores are not indicative of disease.

Bias in the target variable

In this chapter we have focused on psychiatric care as our target variable, which may be even more prone to bias than other medical conditions. As described by Barocas and colleagues, biases in the definition of a target variable are especially critical in questions of fair ML, as these cases are guaranteed to bias the predictions [43]. The authors give the example of "credit-worthiness", which is a construct created to assist in decisions of extending credit to consumers [43]. In itself however, "credit-worthiness" is not an intrinsic property that people possess or lack and its construction was prone to intersecting demographic biases.

In psychiatry, while the issue of mental illness affecting patients is undeniable, the means by which we have constructed the terms, diagnostic criteria and labels of mental disease are fraught with subjective biases and historic discrimination. Herein lies a deeper

issue of the "Ground Truth" on which we rely on in medicine. The historic construction of psychiatric labels to pathologise women and racial minorities may influence a significant bias in the target variable, that means the higher rates of false positives for these groups is hard to avoid.

The converse of this can also result in missed care for groups who have not been targeted by the psychiatry domain in a pathologising manner. For example, it is known that male suicide is a significant issue, and often these patients do not engage with healthcare services or support before ending their lives [225]. Thus, there is a sparsity of data on severe mental illness, particularly in men [225]. The political and social context from which the psychiatry field emerged resulted in a pathologisation of healthy cis and trans women, and racial minorities, but also the neglect of men who were suffering. These biases in the domain, that emerge as a bias in the target variable, are reflected in the false negatives seen affecting men and false positives affecting women.

GP Visits for Mental Health and Sex

The differential engagement of males and females with GP services also influenced the model's prediction of Psychiatric Care. Female patients had a higher baseline rate of GP Visits, resulting in higher false positives for females when the model associated higher GP visits with a greater need for psychiatric care. Our research is focused on utilising these mediation findings to improve model fairness, thus we do not dive into the what these findings mean for fairness in the real world context, which would involve consider the following opposing hypotheses:

1. **GP Visits as a fair mediator:** One could argue that the increased utilisation of GP services by female patients is protective for their mental health, as the GP is able to manage symptoms without the need for a referral to specialist care.
2. **GP Visits as an unfair mediator:** Alternatively, one could consider the research that demonstrates female patients are less likely to be referred by their GP to specialist services due to assumptions that their issues are less severe, thus seeing this is an unfair causal pathway by which female patients are missing out on referrals which are needed for psychiatric input.

In this research we do not explore these deeper questions, as we are focused on improving the performance of models in simulated environments, however it is likely that there is no one simple answer and these circumstances vary significantly across real world contexts.

Algorithmic Equity and Causal Fairness Modelling

The best results in terms of equity were derived from the Counterfactual model, as opposed to the Causal Fairness Adjusted (CFA) model. The counterfactual model maintained the

same overall scores for all patients in the dataset, while improving the Accuracy and ROC scores for females, and reducing the FPR scores (Table 5.12). The CFA model successfully improved the FPR disparity, however this occurred with a parallel worsening of the False Negative Rate (Table 5.17 to Table 5.18). The use of ACME Scores to account for the effect of mediating variables on disparities in algorithmic performance was effective, however the approach requires fine tuning to ensure model errors aren't reversed or displaced onto another subgroup in the dataset.

Limitations of the causal approach

There are several limitations to our causal approach that must be acknowledged. Firstly, the methods of causal inference assume that there are **no unmeasured confounders**, which is particular hard to determine in our context of psychiatric ill health. The causes of mental illness are multi-factorial, with new evidence constantly emerging, thus capturing all potential features within one model would be extremely challenging. It is unlikely that our models account for all confounders that may inform psychiatric care, and may neglect important confounders that influence the relationship between sex and psychiatric symptoms. For example, the co-morbid conditions of pregnancy and menopause are known to affect mental health amongst female patients and the presence of these conditions may play an important mediating role in the relationship between sex and accessing psychiatric care services. Our approach is limited in it's inability to account for unobserved confounding effects.

Common to all causal modelling is the challenge of **model generalisability**, as causal models developed in one setting may not generalise well to other populations. Our dataset was pulled from the UK Biobank and thus is relevant to the UK population, however rates of psychiatric illness and the contributing factors vary significantly between regions and nations around the globe. As such, the generalisiability of our findings may be significantly limited [91, 95, 189].

With regards to the other assumptions of causal inference, our approach is more robust. For example, a common issue in causal modelling is that of **Temporal precedence**, where by it can be difficult to establish the temporal order of events (e.g. the effect of an environmental exposure on an outcome). In our instance however, we can be confident that determination of Sex predates psychiatric care. The only exception may be for transgender individuals, if Sex is conflated with gender and individuals are categorised differently following transition.

Furthermore, the positivity (overlap) assumption of causal modelling assumes that each individual must have a non-zero probability of receiving each level of treatment,

given every combination of covariates in the dataset. In our experiments our "treatment" was Sex, and thus our examination of the balance of males and females in the dataset, and the sex-stratified distribution of other dataset features was particular important. Adopting these methods ensured that there was sufficient representation of both sexes in our dataset, and that there were no covariate spaces in which only males or females were present [91, 226].

Finally, we are ultimately limited by the lack of ground truth in the psychiatry domain, where diagnostic structures, traditional frameworks and assumptions that underpin clinical practice are under ongoing review [115]. In the creation of our models, we are reliant on the labelling of the target variable, yet it is possible that the males and females in this dataset suffer from misdiagnosis themselves. With the psychiatric field evolving at such a pace, it is challenging to determine whether we really have ground truth in our target variable.

Conclusion

In this chapter I have exposed algorithmic performance biases in the domain of psychiatry, demonstrating differential performance on the basis of sex and exposing mediating variables that influence this relationship. I have demonstrated the value of causal methods for going beyond the simpler evaluations of group fairness performed in the previous chapters, to unearth path specific effects and proxy/explanatory mediating variables. I demonstrate that the results of causal evaluation and causal effects can be integrated into models, to make fairness-based adjustments that reduce disparities in model error rates, but that true conclusions regarding "fairness" often require a deeper philosophical and anthropological evaluation of how mediating variables themselves were formed.

Chapter 6

Discussion

> The mathematization and formalization of social issues brings with it a veneer of objectivity and positions its operations as value-free, neutral, and amoral. The intrinsically political tasks of categorizing and predicting things such as "acceptable behavior", "ill" health, and "normal" body type then pass as apolitical technical sorting and categorizing tasks. Unjust and harmful outcomes, as a result, are treated as side effects that can be treated with technical solutions such as debiasing datasets, rather than problems that have deep roots in the mathematization of ambiguous and contingent issues, historical inequalities, and asymmetrical power hierarchies or unexamined problematic assumptions that infiltrate data practices.
>
> *Birhane (2021) [227]*

Achieving fairness in healthcare AI presents a critical challenge, necessitating a rich interdisciplinary approach that examines the ethical, methodological, and practical dimensions of mitigating demographic biases in predictive models. In this research I have examined the historical and anthropological roots of bias in healthcare and, through a series of experiments, exposed the persistence of these issues in evolving AI systems and evaluated measures for addressing these harms.

Chapter 1 and 2 provided a comprehensive overview of historic research into healthcare disparities and the emerging work relating issues if equity in AI models. In Chapter 2, I honed in on the anthropological and sociological roots of bias in healthcare, exploring the perspectives of key philosophers and researchers in this space. In Chapter 3, I exposed inequities in the performance of AI models used in cardiology, and in Chapter 4 I demonstrated the (in)applicability of a range of fairness notions for addressing these harms, identifying specific challenges inherent to the field of medical modelling. In Chapter 5 I introduced novel techniques to untangling the complex relationships that underpin bias in healthcare AI, looking specifically at psychiatry. Here, I deployed causal frameworks for understanding the roots of model bias and evaluated the utility of Counterfactual models and a "Causal-Fairness Adjusted" (CFA) models for reducing critical errors affecting disadvantaged patient groups in the population.

6.1 Summary of findings

The research of this thesis has been intentional interdisciplinary, tying technical methods with a deep sociological understanding of the roots of AI Bias. Throughout this text, I have demonstrated why this approach is essential, as often the results of the computational analysis required a philosophical interpretation. We saw this starkly in the conclusive remarks of Chapter 5, where we found that one could adopt opposing stances on whether a "path", or cause, of algorithmic inequity is fair or unfair based on the interpretation of our causal graphs for psychiatric algorithms (Figure 5.25). My findings are therefore well aligned with the latest guidance on ML fairness research, which makes clear that interdisciplinary socio-technical solutions are required for the true resolution of AI bias and inequity. Here, I have demonstrated how this manifests in applications of causal fairness, and how one can use causal modelling approaches to tie technical findings to anthropological evaluations of contributing factors to model bias. I will now review the main themes identified throughout this thesis, discuss the key lessons learnt, and provide a comprehensive critique of the research performed.

6.1.1 Roots of bias in healthcare AI

In the conceptual analysis of Chapter 2, I explored the roots of medical bias, drawing on domain knowledge from medical anthropology and sociology, distinguishing between unintentional and intentional harms in healthcare. I described how unintentional harms may emerge from a lack of knowledge regarding the physiology of minoritised groups who have been poorly represented in research samples, where as intentional harms may result from biased medical tools and frameworks that were built in oppressive contexts (e.g. psychiatric models that pathologise specific patient groups [51, 115, 118, 123]). Thus understanding the nuances by which power operates within and throughout the medical domain is essential to understanding how difference features present in medical AI development may contribute to disparities in algorithmic performance.

In my critique of the medical domain in the introductory and conceptual chapters, I exposed the means by which so much of medical knowledge is socially constructed and affected by power relationships within society - echoing the historic works of Foucault [25]. We saw the impact of this manifest in the experimental chapters, in which the neglect of female physiology emerged in cardiac ML models less capable of predicting disease and psychiatric models prone to pathologise healthy female patients. A key lesson here for ML researchers in healthcare is to understand that the ground truth of medicine cannot be taken as certain.

In the conceptual chapter, I have described limitations of Evidence based medicine, where the application of an average to heterogeneous patient groups ultimately limits care for some. The insights gained from this traditional research base and its respective clinical trials, cannot be assumed to provide optimal information for all patients, who are not equally represented in the original research samples. Furthermore, I have reviewed the emerging research that criticises the diagnostic metrics, thresholds and frameworks that are used in healthcare today [48, 49, 115, 134]. With this in mind, researchers would be wise to avoid seeing medical dogma as a true ground truth during model development, and instead consider the dynamism of the healthcare field which is constantly being critiqued and updated.

6.1.2 Challenges of achieving ML fairness in healthcare

In this thesis I have applied the latest computational techniques for tackling AI bias to the fields of computational cardiology and psychiatry, identifying core challenges that are unique to medicine. In Chapter 4, I deployed a range of pre-processing and in-processing techniques that failed to resolve inequities in algorithmic performance, which I related to the inappropriate treatment of the sensitive attribute by these methodologies. Given that sensitive attributes (SA) such as Sex or Age have biological effects, their presence in the causal pathway to a disease, means that fairness approaches built to minimise the information provided by the SA may worsen model performance (perhaps more so for the disadvantaged group). As a result, such fairness methods may not be appropriate in the medical context.

I have identified additional challenges that have been discussed throughout the wider ML fairness literature, including the absence of important demographic features in data collection (e.g. Race), precluding the evaluation of their effects on fairness [228]. Further, as detailed by Wan and Colleagues, the rigidity of many fairness approaches fail to capture the complex interplay between multiple fairness needs (e.g. interactions between race, sex and age). For example it is possible that a model that has been designed to be fair for women, may still be unfair for black women [228]. In Chapter 4 I referenced the work of Carruther and Colleagues, investigating the role of high-dimensional representation learning for addressing these intersectional issues [7]. The capability of these models to capture rich, complex data, may be a route forward for addressing inter-sectional biases and not limiting oneself to a focusing on a single group or demographic [7]

6.1.3 Context specific fairness notions

Throughout this thesis I have identified that fairness approaches need to be tailored to the clinical context at hand and the technical specifics of the algorithm being developed. The analysis of sex-specific feature rankings with different methods in Chapter 3 (e.g. Correlation coefficients vs. Gini Importance) illuminated subtle differences that may be overlooked if the approach isnt tailored to one's specific model. The sex-specific feature rankings were different depending on whether one used correlation coefficients or the RF-model specific approach of Gini Importance, hence when evaluating feature importance in the context of fairness, there is no one size fits all approach.

In addition, at the beginning of this thesis we discussed some classic texts in the ML fairness space, that have described the "Zoo of fairness metrics" and the challenge of picking one fairness notion when you cannot satisfy all of them [90]. In particular, as detailed by Narayanan, when the prevalence of the target outcome differs between demographic groups it may be impossible to equalise error rates across sub-populations [89]. In medicine, where disease rates are known to often differ across demographic groups, this point is particularly important, and was explored in greater depth in the causal chapter (see below). The choice of fairness metric and a fairness approach must instead by guided by domain knowledge and the clinical context. For example, minimising the FPR may be prioritised in cases of potential interventional harm (e.g. prescribing a potent drug), where as FNR may be prioritised in cases where the neglect of need would be highly consequential (e.g. diagnosis of an aggressive disease that is time critical).

6.1.4 Causal fairness in healthcare

In Chapter 5, I demonstrated the value of causal modelling for unpacking algorithmic biases in healthcare, where demographic differences in disease rates can complicate our understanding of algorithmic inequity. Due to the complex contributions of demographic factors to disease outcomes, along social, environmental, and biological pathways, differences in model outcomes or model performance may not always indicate algorithmic discrimination [14, 103]. This is a particular issue in healthcare, where, if we are to understand whether a model is acting "unfairly", we require nuanced techniques capable of teasing out the pathways along which demographic features inform model predictions. In this research, I have demonstrated the value of both counterfactual modelling and causal mediation analysis for identifying potential causal relationships, and unpacking the mediating effects of sensitive attributes, and their associated variables, on a target outcome.

6.1.5 Implications for policy and clinical practice

The implications of this research extend beyond the machine learning space, with relevance to both practising healthcare professionals and those working in policy. For clinicians, I have demonstrated the differential value that certain features (e.g. blood tests) provide for males and females, illustrating how computational methods can be used to evaluate the relative utility of various clinical tests for differing clinical groups. These findings may enhance clinical practice, by empowering doctors to consider results in the more personalised context of the patient in front of them. Further, by identifying the anthropological roots of these different forms of bias and inequity, and showing how these manifest in AI models, clinicians may derive insights regarding their own practice and the existing policies of their departments, which may depend on outdated ideas and knowledge.

For those working in policy, this research offers several key contributions. Firstly, it is clear that there is no one size fits all approach to ML fairness, and the selection of the computational method for resolution must be tailored to the specific scenario. In choosing methods for fairness resolution, truly interdisciplinary teams are required that integrate both sociological knowledge and domain skills, as we have seen that this underlying context informs the efficacy of different ML fairness techniques. Finally, in my review of the literature of this domain, Ive demonstrated neglected areas in healthcare AI that warrant further attention, particularly in the fields of cardiology and psychiatry.

6.2 Critique of Research

6.2.1 Group fairness & causal paths

One area that this thesis has not focused on is that of individual fairness [43]. Authors have challenged the approach of group fairness, questioning why we should be concerned with group-level differences and not individual-level differences [43]. In the introduction I detailed the reasons for this approach, highlighting the historic texts in public health that argue why certain inequities are considered more morally objectionable than others.

In "Concerns for Equity in Health" Sudhir Anand argues that demographic group inequalities may be considered less tolerable than individual ones [20]. Anand argues that in health we may be more adverse to certain inter-group inequalities, such as racial or gender inequalities, than to inequalities where the groups are randomly defined (say by the first letter of a persons surname) [20]. Further, Anand argues that we may be more averse to socio-economic inequalities in health, than inter-individual inequalities

that are unconditional on information about individuals [20]. The argument made for this differentiation is that group inequalities give rise to the suspicion that they derive from social (as opposed to natural e.g. genetic) factors and are therefore avoidable with the right public intervention [20, 21]. Sen expands on this idea further, describing the seriousness of particular injustices that result from social arrangements, as opposed to say a personal decision in a health-harming behaviour [20, 21]. The disadvantage we suffer as the result of an unfairly arranged world, appear less tolerable to us than the disadvantage we suffer as a result of our own making (e.g. sex-based inequalities, vs. non/smoker inequalities [20, 21]).

Such an argument suggests that we should be less tolerant of socially constructed inequalities, as opposed to those from underlying physiological processes, which introduces the challenge of distinguishing between the total contribution of these two arms. Sen uses the example of sex to illustrate the point [20, 21]. There is biological evidence that females tend to have better survival chances than males, indeed even female foetuses have a lower probability of spontaneous miscarriage [21]. This is why, despite the fact more boys are born than girls (even a higher proportion of male foetuses are conceived), females tend to predominate in societies [21]. Marmot states that in the developed world, for mortality, the biological advantage is with women [21, 27]. This statement sits in contrast to the works of medical feminists such as Emily Cleghorn, who in "Unwell Women" identifies the multifaceted challenges faced by women as a result of the androcentric medical system which affects their health and lifespan [48]. It would appear that the truth sits somewhere between the two. Females have some biological advantages, yet their biological outcomes are mediated by the systematic barriers they face in society and the healthcare system [20, 21, 48]. Thus, when we observe sex-based disparities in the population, how should we interpret the inequality that is undoubtedly formed of both social and biological phenomena? To do this, we must move beyond observational methods which have been the mainstay of fairness notions so far, and apply the causal methods that have been explored in this thesis.

6.2.2 Society and fairness in flux

The relationship between the society and the individual is in constant flux, dependent on local factors and political shifts. In a society of flux, with changing politics, competing population needs, shifting roles and evolving power relations - the social position of the individual is fluid. Hence, examining the injustices of history and how entrenched power imbalances in society materialise in our daily lives provides a powerful direction for uncovering inequities in any discipline; however the approach is limited if we truly wish to gain a full picture of equity. To truly measure equity within a community, we need to

describe advantage and disadvantage at the individualised level, accounting for shifting political and social norms.

Presently, the prevailing method to evaluate algorithmic bias is to stratify an outcome of interest (e.g. access to healthcare) by specific demographic features (e.g. socioeconomic class) and investigate the relationship between these variables. Such an approach requires the pre-selection of attributes of interest, and a resulting division of the population. A problem arises however when we become overly reliant on historic group parameters and neglect the heterogeneity of populations and ultimately the experience of the individual. Power shifts over time, between communities and is mediated by an array of factors [21, 25, 29]. The mediation of one element of identity with another, introduces nuance that cannot be captured by group categorisations.

The common approach to health disparities is limited in its simple stratification of demographic groups which fails to capture the heterogeneous and fluid nature of advantage/disadvantage within a society. For example, if we consider the lens of identity, how do we compare the relative disadvantage between a wealthy, black, gay, abled-bodied man who speaks the native language of his country; and a middle income, asian, well-educated man with a disability and a chronic health condition. The purpose of this analogy is to demonstrate that the process of simple stratification cannot account for the complexity of identity and experiences of marginalisation that affect patients. The lenses of race, sexuality, ableism, class and overall health cannot capture the nuances of each individual's experience in a society. Our examination of sociopolitical discourse provided by Foucault, Marmot, Sen etc illuminate clear divisions in society where disadvantage falls along a line of identity e.g. educational outcomes for low income children who attend overburdened schools. Yet while illuminating the needs of groups, the demographic based technique may fail to capture individuals who fall between the gaps that traditionally define disadvantage, which have previously been described as "faultlines".

6.2.3 Fairness approaches & neglected groups

In addition to the challenges described above, we must also consider the influence of factors that may be "unknowable" in our datasets. In "Fairness for Unobserved Characteristics: Insights from Technological Impacts on Queer Communities", Tomasev and colleagues explore the limitations of existing approaches to fairness in machine learning (ML) [229]. Current methods that explore demographic parity in algorithmic performance rest on the inherent assumption that protected characteristics (e.g., Sex) are knowable and available within datasets [229]. Such methods are limited in their ability to identify group disadvantage that results from unobserved characteristics e.g., gender

identity [229]. Unobserved characteristics are omitted from datasets either because the information isnt recorded, or because the attribute cannot be quantified e.g. gender identity is a fluid cultural construct that may change over time and social contexts [25, 229]. Despite the very real disadvantage that a community defined by an unobserved characteristic may face, their invisibility within the data renders the evaluation of respective inequity challenging.

We can consider such unobserved characteristics as "latent features", of which another example that has been proposed is "perceived control". Researchers have argued that individuals who report low perceived control over their health, have worse health outcomes [27–29]. Marmot argues that it may be possible to characterise societies on degree of control and demonstrates that the higher the mean level of control of individuals in a society, the lower the rates of coronary heart disease [27–29]. Intelligence has also been proposed as an unobserved latent features, with several researchers suggesting that cognitive ability is the greatest causative factor of health inequities - a controversial opinion given the conflicting research on the validity of intelligence measures and the multiple barriers that different individuals may face in manifesting their intelligence [21, 27–29]. Cognitive ability is rarely documented in healthcare datasets, and may represent another unobserved feature, the identification of which would facilitate a new perspective on healthcare inequalities [21, 27–29].

The traditional approach in ML fairness research to pre-select a demographic attribute of interest, ultimately restricts the focus of concern toward one disadvantaged group, potentially limiting the discovery of other underserved populations. Epistemological research has previously demonstrated the role that scholars play in the construction of knowledge, including our understanding of identity, and how the application of inherited sociological frameworks in our work can restrict our research findings.

In "Queer Data" Kevin Guyan explores these challenges through a queer lens, and examines how data collection can shape our understanding of "normal" in the context of queerness [230]. Guyan contrasts the proportion of the population that place themselves in an LGBTQ+ category on the UK Census (2.5 percent), with the 33 percent of people who identify as not completely heterosexual when presented with a scale of 0 (completely heterosexual) to 6 (completely homosexual) [230]. The format of the question relating to queerness informed the insights that the statisticians could obtain on how common it is to be queer. These "limitations of method" are relatively under-explored in the ML fairness literature, yet when our technique for evaluating algorithmic fairness stems from a preconception of where to look, we remain blind to other possibilities. Herein lies the key limitation of current fairness approaches that stratify model biases by demographic

groups, based on historically known disadvantages

6.2.4 Chosen data and medical records

In their review of Worldwide AI Ethics regulations, Correa and colleagues higlight the lack of global representation has been described in the context of datasets used for building machine learning (ML) models. We are now seeing this same pattern emerging in the international instruments developed to guide AI use [5]. In this thesis I have only focused on electronic health information that includes biochemical data, clinical scores and the outcomes of diagnostic frameworks (e.g. the Neuroticism score). This neglects the wider array of data sources on which healthcare AI models may be built, described in greater detail by Jones and colleagues [94].

In their article examining algorithmic bias in medical imaging systems, the authors discuss a wide range of sources for AI bias that have not been covered in this thesis. For example, medical equipment may play a particularly pertinent role for AI bias in imaging algorithms. As detailed by the authors, it is not uncommon for different groups, in different parts of the world, to be scanned with different equipment or to have natural variations in their physical characteristics, causing the illness to manifest differently [94]. Jones et al provide the example of datasets obtained from diagnostic ultrasound, where patients in different geographical areas are referred for scans at different stages in disease progression due to local policies [94]. As a result, the disease appears systematically different on scans, dependant on location [94]. Location specific artefacts may then be picked up in the imaging dataset and affect model performance and generalisibility.

6.2.5 Human Bias, AI bias & AI potential

One area that was beyond the scope of this thesis was a comparison of algorithmic bias with existing human bias. Researchers such as Mehrabi have stated that *"there are clear benefits to algorithmic decision-making; as unlike people, machines do not become tired or bored"*. Humans are fallible, and medical practitioners are known to carry both conscious and unconscious biases that impact their treatment of patients [48, 115, 123]. In one sense, evaluating for bias and discrimination in computational systems is easier than evaluating clinicians, as we can utilise simulated environments, control parameters, and evaluate behaviour over multiple experimental runs. The same cannot be done for clinicians operating in the real world, which limits our ability to understand the impact of existing clinician bias on health inequalities. It could be that an algorithmic system may exhibit a small bias, however this could be less than the current state of play in healthcare.

Building further on this idea, other researchers have examined whether it might

be easier to debias AI systems compared to human practitioners. At the simplest level, compared to humans AI models may be more easily evaluated for the data (knowledge) that they rely on, the rules that they use, and the means by which they make decisions. In a sense they could be more interpretable than a human, and these concepts of model transparency and interpretability are examined in depth in the Explainable AI domain.

Beyond explainability, researchers have also proposed that advanced AI methods may be the best avenue for improving health inequities in the future, contrasting significantly with scientists concerned about the impact of discriminatory AI bias [2, 7, 15, 50]. In a paper from our research lab, titled "Representational Ethical Calibration", our team explored for this possibility, examining the means by which representational models could account for population heterogeneity and achieve individuation of treatment, thus eradicating issues of bias emerging due to group-to-individual level inference. However, it is important to note that representation isnt everything. In this thesis I have reviewed the issues of both a lack of information/representation regarding marginalised groups, but also the intentional harms relating to political constructed medical tools. No matter the changes we make to representation in datasets, AI will always perform worse for marginalised groups if it is using medical tools designed to pathologise or stigmatise them, as opposed to treat and support their needs [115, 121, 123].

6.3 Conclusion & Research Contribution

In conclusion, the research of this thesis underscores the deep complexity of addressing AI bias in healthcare, revealing that truly equitable solutions must transcend technical fixes and engage deeply with socio-political structures that have shaped medical knowledge and praxis. Through this work, I have examined the sociological and anthropological roots of AI bias in healthcare, and identified two major sources of harm: unintentional harms related to a lack of representation in research, and intentional harms that stem from historically oppressive medical tools. To address these harms, I have proposed distinct solutions, including high-dimensional modelling for addressing issues of representation, and causal modelling for dismantling biases embedded in socially constructed frameworks. Furthermore, I argue that existing fairness notions that rely on "blinding" a model to a sensitive attribute, or training against it, are largely inappropriate in healthcare where demographic features play a role in the manifestation of disease. Instead, "attribute-aware" approaches, that factor in causal pathways between sensitive attributes and target outcomes, are vital for ensuring the equitable development of medical AI.

My findings illustrated that both counterfactual fairness and causal mediation analysis

can be used to untangle the complex relationships between a sensitive attribute and the target outcome, ensuring that sensitive information is leveraged only when biologically relevant and beneficial to marginalised groups. This socio-technical approach provides a route for ML researchers to move beyond simple technical fixes and create models that truly account for past harms, and mitigate future healthcare inequities. As we continue to develop AI technologies in healthcare, the commitment to socially-aware and technically rigorous methods, will be paramount in determining whether our new digital systems perpetuate historic harms or act as a vehicle for change.

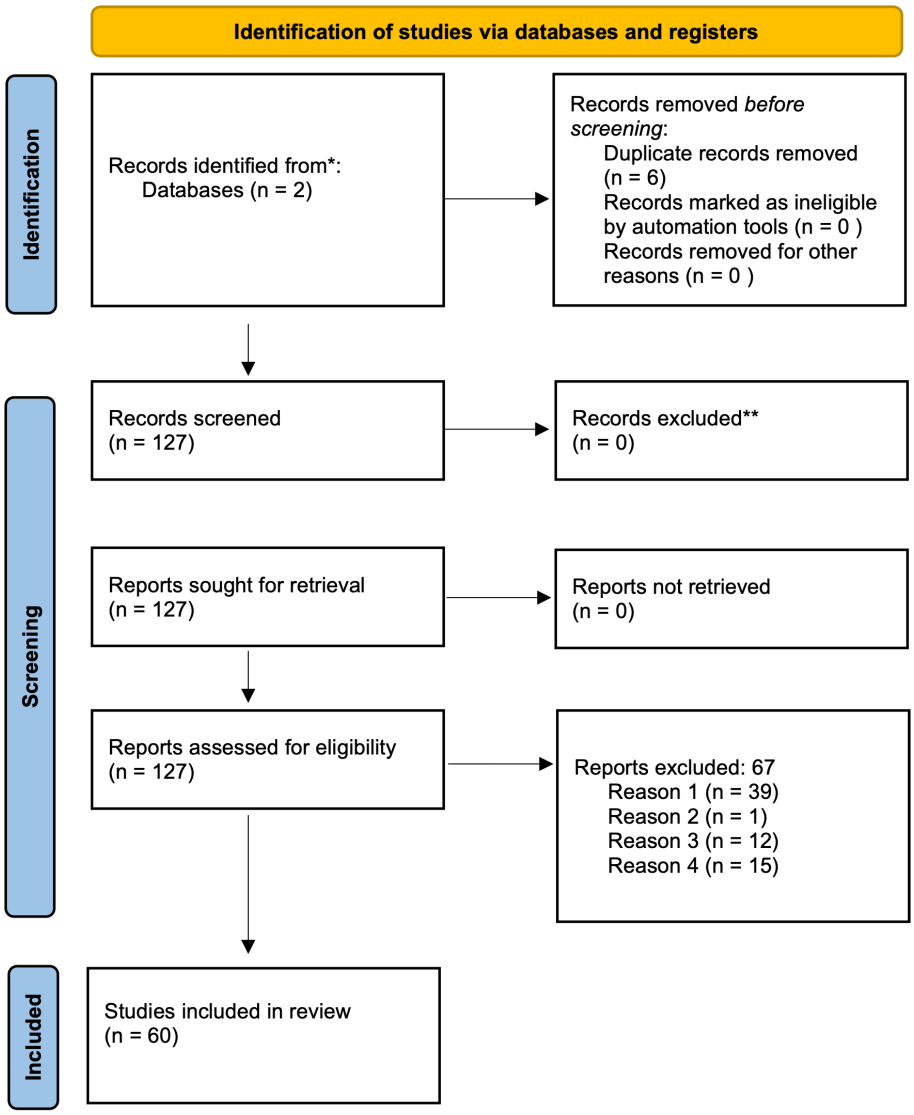
Chapter 7

Supplementary Material

Table 7.1: **Supplementary Table 7.1:** Literature Review Details and MESH Terms for search carried out between 1st April 2022 and 22nd May 2022 (time-span of search: 1900-01-01 to 2022-05-22). Nb. the "article type" was restricted to full research papers, and did not include isolated abstracts

Academic Database with MESH Terms	**Number of Results**
PubMed	**35 Results**
((((artificial intelligence[MeSH Major Topic]) OR (machine learning[MeSH Major Topic]) OR (deep learning[MeSH Major Topic]) OR (unsupervised learning[MeSH Major Topic]) OR (supervised learning[MeSH Major Topic])) AND ((heart failure [MeSH Major Topic]) or (cardiac failure[MeSH Major Topic])) AND ((predic*[MeSH Terms]))))	
Web of Science	**98 Results**
(((TI=(artificial intelligence) OR TI=(machine learning) OR TI=(unsupervised machine learning) OR TI=(supervised machine learning)) AND ((TI=(cardiac failure) OR TI=(heart failure)) AND (AB=(predict*)))))	

Figure 7.1: **Supplementary Figure 7.1:** PRISMA 2020 flow diagram for new systematic reviews which included searches of databases and registers only. PRISMA templated obtained from PRISMA at urlhttps://prisma-statement.org/prismastatement/flowdiagram.aspx



Reasons for Exclusion:

1. Reason 1: The study did not focus on biochemical data or laboratory tests, instead utilising different modalities (e.g., visual data from radiological scans).
2. Reason 2: The study did not use machine learning techniques (e.g. it used traditional statistical methods).
3. Reason 3: The study did not describe empirical research that involved the development of ML models for prediction of cardiac disease (e.g., the paper was a review or commentary).
4. Reason 4: The retrieved study was not a full paper, instead it was a conference or meeting abstract.

Bibliography

[1] Ansh Bhatnagar and Devyani Gajjar. "Policy implications of artificial intelligence (AI)". In: *UK Parliament POST* (2021). URL: https://researchbriefings.files.parliament.uk/documents/POST-PN-0708/POST-PN-0708.pdf.

[2] Cathy O'neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2017.

[3] Virginia Eubanks. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.

[4] Daniel S Schiff et al. "Global AI ethics documents: What they reveal about motivations, practices, and policies". In: *Codes of Ethics and Ethical Guidelines: Emerging Technologies, Changing Fields* (2022), pp. 121–143.

[5] Nicholas Kluge Corrêa et al. "Worldwide AI Ethics: a review of 200 guidelines and recommendations for AI governance". In: *arXiv e-prints* (2022), arXiv–2206.

[6] Isabel Straw and Honghan Wu. "Investigating for bias in healthcare algorithms: a sex-stratified analysis of supervised machine learning models in liver disease prediction". In: *BMJ health & care informatics* 29.1 (2022).

[7] Robert Carruthers et al. "Representational ethical model calibration". In: *NPJ Digital Medicine* 5.1 (2022), p. 170.

[8] Isabel Straw, Geraint Rees, and Parashkev Nachev. "Sex-Based Performance Disparities in Machine Learning Algorithms for Cardiac Disease Prediction: Exploratory Study". In: *Journal of Medical Internet Research* 26 (2024), e46936.

[9] Daniel Van Niekerk et al. "I dont have a gender, consciousness, or emotions. Im just a machine learning model"". In: *UNESCO Digital Library* (2023). URL: https://unesdoc.unesco.org/ark:/48223/pf0000387189.

[10] Daniel Van Niekerk et al. "Challenging Systematic Prejudices: An Investigation into Bias Against Women and Girls". In: *UNESCO Digital Library* (2024). URL: https://ircai.org/project/challenging-systematic-prejudices/.

[11] Isabel Straw et al. "Insights From a Clinically Orientated Workshop on Health Care Cybersecurity and Medical Technology: Observational Study and Thematic Analysis". In: *Journal of Medical Internet Research* 26 (2024), e50505.

[12] Richard L Kravitz. "Personalized medicine without the omics". In: *Journal of general internal medicine* 29 (2014), pp. 551–551.

[13] Centers for Disease Control and Prevention (CDC). *CDC - Attaining Health Equity - Healthy Communities Program*. en-us. May 2019. URL: https://www.cdc.gov/nccdphp/dch/programs/healthycommunitiesprogram/overview/healthequity.htm (visited on 11/21/2023).

[14] Melissa D McCradden et al. "Ethical limitations of algorithmic fairness solutions in health care machine learning". In: *The Lancet Digital Health* 2.5 (2020), e221–e223.

[15] Joy Buolamwini and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification". In: *Conference on fairness, accountability and transparency*. PMLR. 2018, pp. 77–91.

[16] Robert Challen et al. "Artificial intelligence, bias and clinical safety". In: *BMJ quality & safety* 28.3 (2019), pp. 231–237.

[17] Laleh Seyyed-Kalantari et al. "CheXclusion: Fairness gaps in deep chest X-ray classifiers". In: *BIOCOMPUTING 2021: proceedings of the Pacific symposium*. World Scientific. 2020, pp. 232–243.

[18] Sharmin Afrose et al. "Subpopulation-specific machine learning prognosis for underrepresented patients with double prioritized bias correction". In: *Communications medicine* 2.1 (2022), p. 111.

[19] Drew Roselli, Jeanna Matthews, and Nisha Talagala. "Managing bias in AI". In: *Companion proceedings of the 2019 world wide web conference*. 2019, pp. 539–544.

[20] Sudhir Anand, Fabienne Peter, and Amartya Sen. *Public Health, Ethics, and Equity*. Oxford University Press, 2004.

[21] Amartya Sen. "Why Health Equity?" In: *Health Economics* 13.2 (2004), pp. 109–125.

[22] René Descartes and Gustav Gröber. *Discours de la méthode: 1637*. Heitz Bonita Springs, FL, USA, 1905.

[23] Aristotle Aristotle. *The complete works*. Harvard University Press Cambridge, MA, 1968.

[24] Paula Gottlieb. *Aristotle's Ethics: Nicomachean and Eudemian Themes*. Cambridge University Press, 2022.

[25] Michel Foucault. "The subject and power". In: *Critical inquiry* 8.4 (1982), pp. 777–795.

[26] Lisa Downing. "After Foucault: Culture, Theory and Criticism in the 21st Century". In: (2018).

[27] Michael G Marmot, Martin J Shipley, and Geoffrey Rose. "Inequalities in deathspecific explanations of a general pattern?" In: *The Lancet* 323.8384 (1984), pp. 1003–1006.

[28] M Marmot et al. "Social causes of inequity in health". In: *Public Health and Ethics* (2004), p. 37.

[29] Michael Marmot. "The health gap: the challenge of an unequal world". In: *The Lancet* 386.10011 (2015), pp. 2442–2444.

[30] Sergi Albert-Ballestar and Anna García-Altés. "Measuring health inequalities: a systematic review of widely used indicators and topics". In: *International journal for equity in health* 20 (2021), pp. 1–15.

[31] Roy A Carr-Hill et al. *The public health observatory handbook of health inequalities measurement*. South East Public Health Observatory Oxford, 2005.

[32] Gareth H Williams. "The determinants of health: structure, context and agency". In: *Sociology of Health & Illness* 25.3 (2003), pp. 131–154.

[33] WHO Commission on Social Determinants of Health and World Health Organization. *Closing the gap in a generation: health equity through action on the social determinants of health: Commission on Social Determinants of Health final report*. World Health Organization, 2008.

[34] "Health Equity: Overview". In: *Online* (2010). URL: `https://www.who.int/health-topics/health-equity#tab=tab_1`.

[35] Shoshana Zuboff. "The age of surveillance capitalism". In: *Social theory re-wired*. Routledge, 2023, pp. 203–213.

[36] Corinne Cath. "Governing artificial intelligence: ethical, legal and technical opportunities and challenges". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376.2133 (2018), p. 20180080.

[37] Ziad Obermeyer et al. "Dissecting racial bias in an algorithm used to manage the health of populations". In: *Science* 366.6464 (2019), pp. 447–453.

[38] Kate Crawford et al. "The AI now report". In: *The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term* 2 (2016).

[39] John McCarthy. "Artificial intelligence, logic and formalizing common sense." In: *Philosophical logic and artificial intelligence* (1989), pp. 161–190.

[40] Turing AM. "Computing machinery and intelligence". In: *Comput Mach Intell* 49 (1950), pp. 433–60.

[41] Sarah Graham et al. "Artificial intelligence for mental health and mental illnesses: an overview". In: *Current psychiatry reports* 21 (2019), pp. 1–18.

[42] Sebastian Raschka and Vahid Mirjalili. *Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2*. Packt Publishing Ltd, 2019.

[43] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT Press, 2023.

[44] Jeffrey A Claridge and Timothy C Fabian. "History and development of evidence-based medicine". In: *World journal of surgery* 29.5 (2005), pp. 547–553.

[45] Iqbal Ratnani et al. "Evidence-based medicine: history, review, criticisms, and pitfalls". In: *Cureus* 15.2 (2023).

[46] Margaret M Lock and Vinh-Kim Nguyen. *An anthropology of biomedicine*. John Wiley & Sons, 2018.

[47] Guilherme Pombo et al. "Equitable modelling of brain imaging by counterfactual augmentation with morphologically constrained 3d deep generative models". In: *Medical Image Analysis* 84 (2023), p. 102723.

[48] Elinor Cleghorn. *Unwell women: misdiagnosis and myth in a man-made world*. Penguin, 2022.

[49] Angela Saini. *Superior: The return of race science*. Beacon Press, 2019.

[50] Amy Nelson and Parashkev Nachev. "Machine Learning in PracticeClinical Decision Support, Risk Prediction, Diagnosis". In: *Clinical Applications of Artificial Intelligence in Real-World Data*. Springer, 2023, pp. 231–245.

[51] Isabel Straw and Chris Callison-Burch. "Artificial Intelligence in mental health and the biases of language based models". In: *PloS one* 15.12 (2020), e0240376.

[52] Madhumita Pal and Smita Parija. "Prediction of heart diseases using random forest". In: *Journal of Physics: Conference Series*. Vol. 1817. 1. IOP Publishing. 2021, p. 012009.

[53] Scientific The United Nations Educational and Cultural Organization (UNESCO). "Recommendation on the Ethics of Artificial Intelligence". In: *UNESCO Digital Library* (2021), p. 21. URL: https://unesdoc.unesco.org/ark:/48223/pf0000380455.

[54] J. Ross Quinlan. "Induction of decision trees". In: *Machine learning* 1 (1986), pp. 81–106.

[55] Corinna Cortes and Vladimir Vapnik. "Support-vector networks". In: *Machine learning* 20 (1995), pp. 273–297.

[56] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. "Learning representations by back-propagating errors". In: *nature* 323.6088 (1986), pp. 533–536.

[57] Jan Kietzmann et al. “Deepfakes: Trick or treat?” In: *Business Horizons* 63.2 (2020), pp. 135–146.

[58] Silvana Secinaro et al. “The role of artificial intelligence in healthcare: a structured literature review”. In: *BMC medical informatics and decision making* 21 (2021), pp. 1–23.

[59] Nishita Mehta, Anil Pandit, and Sharvari Shukla. “Transforming healthcare with big data analytics and artificial intelligence: A systematic mapping study”. In: *Journal of biomedical informatics* 100 (2019), p. 103311.

[60] Fei Jiang et al. “Artificial intelligence in healthcare: past, present and future”. In: *Stroke and vascular neurology* 2.4 (2017).

[61] Daniel T Hogarty et al. “Artificial intelligence in dermatologywhere we are and the way to the future: a review”. In: *American journal of clinical dermatology* 21 (2020), pp. 41–47.

[62] Adwitiya Ray et al. “Artificial intelligence and Psychiatry: An overview”. In: *Asian Journal of Psychiatry* 70 (2022), p. 103021.

[63] Davide Chicco and Giuseppe Jurman. “Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone”. In: *BMC medical informatics and decision making* 20 (2020), pp. 1–16.

[64] Babak Ehteshami Bejnordi et al. “Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer”. In: *Jama* 318.22 (2017), pp. 2199–2210.

[65] Pritam Mukherjee et al. “A shallow convolutional neural network predicts prognosis of lung cancer patients in multi-institutional computed tomography image datasets”. In: *Nature machine intelligence* 2.5 (2020), pp. 274–282.

[66] Lars Lau Raket et al. “Dynamic ElecTronic hEalth reCord deTection (DETECT) of individuals at risk of a first episode of psychosis: a case-control development and validation study”. In: *The Lancet Digital Health* 2.5 (2020), e229–e239.

[67] Sarah Parisot et al. “Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimers disease”. In: *Medical image analysis* 48 (2018), pp. 117–130.

[68] Narjice Chafai et al. “Emerging applications of machine learning in genomic medicine and healthcare”. In: *Critical Reviews in Clinical Laboratory Sciences* 61.2 (2024), pp. 140–163.

[69] Mikkel Brabrand et al. “Risk scoring systems for adults admitted to the emergency department: a systematic review”. In: *Scandinavian journal of trauma, resuscitation and emergency medicine* 18 (2010), pp. 1–8.

[70] Zach Rozenbaum et al. “CHA2DS2-VASc score and clinical outcomes of patients with acute coronary syndrome”. In: *European Journal of Internal Medicine* 36 (2016), pp. 57–61.

[71] Areti Sofogianni et al. “Cardiovascular risk prediction models and scores in the era of personalized medicine”. In: *Journal of Personalized Medicine* 12.7 (2022), p. 1180.

[72] Kelly Joyce et al. “Toward a sociology of artificial intelligence: A call for research on inequalities and structural change”. In: *Socius* 7 (2021), p. 2378023121999581.

[73] Safiya Umoja Noble. “Algorithms of oppression: How search engines reinforce racism”. In: *Algorithms of oppression*. New York university press, 2018.

[74] Zhisheng Chen. “Ethics and discrimination in artificial intelligence-enabled recruitment practices”. In: *Humanities and Social Sciences Communications* 10.1 (2023), pp. 1–12.

[75] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. “Identifiability of causal-based fairness notions: A state of the art”. In: *arXiv preprint arXiv:2203.05900* (2022).

[76] Dan Heaton et al. “The algorithm will screw you: Blame, social actors and the 2020 A Level results algorithm on Twitter”. In: *Plos one* 18.7 (2023), e0288662.

[77] Helen Smith. “Algorithmic bias: should students pay the price?” In: *AI & society* 35.4 (2020), pp. 1077–1078.

[78] Anthony Kelly. “A tale of two algorithms: The appeal and repeal of calculated grades systems in England and Ireland in 2020”. In: *British Educational Research Journal* 47.3 (2021), pp. 725–741.

[79] Bruno Mallett. “Reviewing the impact of OFQUALs assessment algorithmon racial inequalities”. In: *COVID-19 and Racism*. Policy Press, 2023, pp. 187–198.

[80] Richard J Chen et al. “Algorithmic fairness in artificial intelligence for medicine and healthcare”. In: *Nature biomedical engineering* 7.6 (2023), pp. 719–742.

[81] Davide Cirillo et al. “Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare”. In: *NPJ digital medicine* 3.1 (2020), pp. 1–11.

[82] Roxana Daneshjou et al. “Disparities in dermatology AI performance on a diverse, curated clinical image set”. In: *Science advances* 8.31 (2022), eabq6147.

[83] Lu Tang, Jinxu Li, and Sophia Fantus. “Medical artificial intelligence ethics: A systematic review of empirical studies”. In: *DIGITAL HEALTH* 9 (2023). PMID: 37312939, p. 20552076231186064. DOI: 10.1177/20552076231186064.

[84] Marissa Borgese et al. “Bias assessment and correction in machine learning algorithms: a use-case in a natural language processing algorithm to identify hospitalized patients with unhealthy alcohol use”. In: *AMIA Annual Symposium Proceedings*. Vol. 2021. American Medical Informatics Association. 2021, p. 247.

[85] Hossein Estiri et al. “Predicting COVID-19 mortality with electronic medical records”. In: *NPJ digital medicine* 4.1 (2021), p. 15.

[86] Agostina J Larrazabal et al. “Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis”. In: *Proceedings of the National Academy of Sciences* 117.23 (2020), pp. 12592–12594.

[87] Hale M Thompson et al. “Bias and fairness assessment of a natural language processing opioid misuse classifier: detection and mitigation of electronic health record data disadvantages across racial subgroups”. In: *Journal of the American Medical Informatics Association* 28.11 (2021), pp. 2393–2403.

[88] Ninareh Mehrabi et al. “A survey on bias and fairness in machine learning”. In: *ACM computing surveys (CSUR)* 54.6 (2021), pp. 1–35.

[89] Arvind Narayanan. “Fairness Definitions and Their Politics.(2018)”. In: *Conference on Fairness, Accountability, and Transparency, NYC*. 21.

[90] Alessandro Castelnovo et al. “The zoo of fairness metrics in machine learning”. In: (2021).

[91] Judea Pearl. “Causal Inference in Statistics: An Overview”. In: *Statistics Surveys* 3 (2009), pp. 96–146.

[92] Pedro Sanchez et al. “Causal machine learning for healthcare and precision medicine”. In: *Royal Society Open Science* 9.8 (2022), p. 220638.

[93] Tobias Hatt. “Causal AI in Personalised Healthcare”. In: *Dimensions of Intelligent Analytics for Smart Digital Health Solutions*. Chapman and Hall/CRC, 2024, pp. 62–77.

[94] Charles Jones et al. “A causal perspective on dataset bias in machine learning for medical imaging”. In: *Nature Machine Intelligence* (2024), pp. 1–9.

[95] Judea Pearl. *Causality*. Cambridge university press, 2009.

[96] Drago Pleko, Elias Bareinboim, et al. “Causal Fairness Analysis: A Causal Toolkit for Fair Machine Learning”. In: *Foundations and Trendső in Machine Learning* 17.3 (2024), pp. 304–589.

[97] George L Engel. “The clinical application of the biopsychosocial model”. In: *The Journal of medicine and philosophy* 6.2 (1981), pp. 101–124.

[98] Vinay Chamola et al. “A review of trustworthy and explainable artificial intelligence (xai)”. In: *IEEE Access* (2023).

[99] Xiaoxuan Liu et al. “Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension”. In: *The Lancet Digital Health* 2.10 (2020), e537–e548.

[100] Shaswath Ganapathi et al. “Tackling bias in AI health datasets through the STANDING Together initiative”. In: *Nature Medicine* 28.11 (2022), pp. 2232–2233.

[101] Craig E Kuziemsky et al. “AI Quality Standards in Health Care: Rapid Umbrella Review”. In: *Journal of Medical Internet Research* 26 (2024), e54705.

[102] Anne AH de Hond et al. “Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review”. In: *NPJ digital medicine* 5.1 (2022), p. 2.

[103] Melissa Mccradden et al. “What’s fair is fair? Presenting JustEFAB, an ethical framework for operationalizing medical ethics and social justice in the integration of clinical machine learning: JustEFAB”. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 2023, pp. 1505–1519.

[104] Carol Jonann Bess. “Gender bias in health care: a life or death issue for women with coronary heart disease”. In: *Hastings Women’s LJ* 6 (1995), p. 41.

[105] Andrzej ytkowski et al. “Anatomical normality and variability: Historical perspective and methodological considerations”. In: *Translational Research in Anatomy* 23 (2021), p. 100105.

[106] Ruth Chadwick. “Normality as convention and as scientific fact”. In: *Handbook of the Philosophy of Medicine. Dordrecht: Springer Netherlands* (2017), pp. 17–28.

[107] Robert Wachbroit. “Normality as a biological concept”. In: *Philosophy of Science* 61.4 (1994), pp. 579–591.

[108] Eugenia Cheng. *X+ Y: A Mathematician’s Manifesto for Rethinking Gender*. Hachette UK, 2020.

[109] Audre Lorde. *The master’s tools will never dismantle the master’s house*. 2018.

[110] Kimberlé Williams Crenshaw. “Mapping the margins: Intersectionality, identity politics, and violence against women of color”. In: *The public nature of private violence*. Routledge, 2013, pp. 93–118.

[111] Sylvia Federici. “Caliban and the Witch: Women, the Body and Primitive Accumulation, Autonomedia”. In: *New York* (2004).

[112] Daniel Fryer, Inga Strümke, and Hien Nguyen. “Shapley values for feature selection: The good, the bad, and the axioms”. In: *Ieee Access* 9 (2021), pp. 144352–144360.

[113] Caroline Criado Perez. *Invisible women: Data bias in a world designed for men*. Abrams, 2019.

[114] Ginelle Wolfe and Nicole Fogwell. “DSM Discrimination and the LGBT Community: Using the History of Diagnostic Discrimination Against Sexual Minorities to Contextualize Current Issues in Transgender and Gender Diverse Mental Healthcare”. In: *Psychology from the Margins* 4.1 (2022), p. 2.

[115] Jessica Taylor. *Sexy But Psycho: How the Patriarchy Uses Womens Trauma Against Them*. Hachette UK, 2022.

[116] Saul V Levine, Louisa E Kamin, and Eleanor Lee Levine. “Sexism and psychiatry.” In: *American Journal of Orthopsychiatry* 44.3 (1974), p. 327.

[117] Roberta Satow. “Where has all the hysteria gone?” In: *Psychoanalytic Review* 66.4 (1979), p. 463.

[118] Jonathan M Metzl. *The protest psychosis: How schizophrenia became a black disease*. Beacon Press, 2010.

[119] Stuart C Gilman. “Degeneracy and race in the nineteenth century: the impact of clinical medicine”. In: *The Journal of ethnic studies* 10.4 (1983), p. 27.

[120] Andrew Bank. “Of native skulls and noble caucasians: phrenology in colonial South Africa”. In: *Journal of Southern African Studies* 22.3 (1996), pp. 387–403.

[121] Vanessa Northington Gamble. “Under the shadow of Tuskegee: African Americans and health care”. In: *Health Psychology* (2016), pp. 434–441.

[122] Lundy Braun. *Breathing race into the machine: The surprising career of the spirometer from plantation to genetics*. U of Minnesota Press, 2014.

[123] Katarina Hamberg. “Gender bias in medicine”. In: *Womens health* 4.3 (2008), pp. 237–243.

[124] Irving Zucker and Brian J Prendergast. “Sex differences in pharmacokinetics predict adverse drug reactions in women”. In: *Biology of sex differences* 11 (2020), pp. 1–14.

[125] Natasha A Karp and Neil Reavey. “Sex bias in preclinical research and an exploration of how to change the status quo”. In: *British journal of pharmacology* 176.21 (2019), pp. 4107–4118.

[126] Cat Bohannon. *Eve: How the Female Body Drove 200 Million Years of Human Evolution*. Random House Canada, 2023.

[127] Lucy Cooke. *Bitch: a revolutionary guide to sex, evolution and the female animal*. Random House, 2022.

[128] Bryan D Haughom et al. “Do complication rates differ by gender after metal-on-metal hip resurfacing arthroplasty? A systematic review”. In: *Clinical Orthopaedics and Related Researchő* 473.8 (2015), pp. 2521–2529.

[129] Sarah Gauci et al. “Biology, bias, or both? The contribution of sex and gender to the disparity in cardiovascular outcomes between women and men”. In: *Current Atherosclerosis Reports* 24.9 (2022), pp. 701–708.

[130] Kristen Sullivan et al. “Sex-specific differences in heart failure: pathophysiology, risk factors, management, and outcomes”. In: *Canadian Journal of Cardiology* 37.4 (2021), pp. 560–571.

[131] Mary Norine Walsh, Mariell Jessup, and JoAnn Lindenfeld. *Women with heart failure: unheard, untreated, and unstudied*. 2019.

[132] Gianluigi Savarese et al. “Global burden of heart failure: a comprehensive and updated review of epidemiology”. In: *Cardiovascular research* 118.17 (2022), pp. 3272–3287.

[133] The British Heart Foundation. “Bias and Biology: The Heart Attack Gender Gap”. In: *Online* (). URL: https://www.bhf.org.uk/-/media/files/what-we-do/wales/bias-and-biology-report-bhf-cymru-english-for-web.pdf?rev=3575994d3706401dacbf83ebcf34c2f9&hash=CA66D8CF01421D6A2E014AED141648CB.

[134] Kimia Sobhani et al. “Sex differences in ischemic heart disease and heart failure biomarkers”. In: *Biology of sex differences* 9 (2018), pp. 1–13.

[135] Jo-Ann Eastwood et al. “Anginal symptoms, coronary artery disease, and adverse outcomes in Black and White women: the NHLBI-sponsored Women’s Ischemia Syndrome Evaluation (WISE) study”. In: *Journal of women’s health* 22.9 (2013), pp. 724–732.

[136] Sujoya Dey et al. “Sex-related differences in the presentation, treatment and outcomes among patients with acute coronary syndromes: the Global Registry of Acute Coronary Events”. In: *Heart* 95.1 (2009), pp. 20–26.

[137] Rita F Redberg. “Gender, race, and cardiac care: why the differences?” In: *Journal of the American College of Cardiology* 46.10 (2005), pp. 1852–1854.

[138] Julia Stehli et al. “Sex differences persist in time to presentation, revascularization, and mortality in myocardial infarction treated with percutaneous coronary intervention”. In: *Journal of the American Heart Association* 8.10 (2019), e012161.

[139] Ehsan Khan et al. “Differences in management and outcomes for men and women with ST-elevation myocardial infarction”. In: *Medical Journal of Australia* 209.3 (2018), pp. 118–123.

[140] Tor Melberg, Bjørg Kindervaag, and Jan Rosland. “Gender-specific ambulance priority and delays to primary percutaneous coronary intervention: a consequence of the patients’ presentation or the management at the emergency medical communications center?” In: *American heart journal* 166.5 (2013), pp. 839–845.

[141] Claire Raphael et al. “Limitations of the New York Heart Association functional classification system and self-reported walking distances in chronic heart failure”. In: *Heart* 93.4 (2007), pp. 476–482.

[142] Stephen J Greene et al. “Comparison of New York Heart Association class and patient-reported outcomes for heart failure with reduced ejection fraction”. In: *JAMA cardiology* 6.5 (2021), pp. 522–531.

[143] Tanvir Ahmad et al. “Survival analysis of heart failure patients: A case study”. In: *PloS one* 12.7 (2017), e0181001.

[144] C. S. Eke et al. “Identification of Optimum Panel of Blood-based Biomarkers for Alzheimers Disease Diagnosis Using Machine Learning”. In: (2018), pp. 3991–3994. DOI: 10.1109/EMBC.2018.8513293.

[145] Ioannis Kavakiotis et al. “Machine learning and data mining methods in diabetes research”. In: *Computational and structural biotechnology journal* 15 (2017), pp. 104–116.

[146] Jing Xiao et al. “Comparison and development of machine learning tools in the prediction of chronic kidney disease progression”. In: *Journal of translational medicine* 17.1 (2019), pp. 1–13.

[147] Anna Karen Garate Escamilla, Amir Hajjam El Hassani, and Emmanuel Andres. “A comparison of machine learning techniques to predict the risk of heart failure”. In: *Machine Learning Paradigms: Applications of Learning and Analytics in Intelligent Systems* (2019), pp. 9–26.

[148] C Beulah Christalin Latha and S Carolin Jeeva. “Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques”. In: *Informatics in Medicine Unlocked* 16 (2019), p. 100203.

[149] Kathleen H Miao, Julia H Miao, and George J Miao. “Diagnosing coronary heart disease using ensemble machine learning”. In: *International Journal of Advanced Computer Science and Applications* 7.10 (2016).

[150] *Heart failure clinical records*. UCI Machine Learning Repository. DOI: https://-doi.org/10.24432/C5Z89R. 2020.

[151] *Heart Disease Dataset*. UCI Machine Learning Repository. DOI: https://-doi.org/10.24432/C52P4X. 1988.

[152] Faisal Maqbool Zahid et al. “Gender based survival prediction models for heart failure patients: A case study in Pakistan”. In: *PloS one* 14.2 (2019), e0210602.

[153] Saba Bashir et al. “Improving heart disease prediction using feature selection approaches”. In: *2019 16th international bhurban conference on applied sciences and technology (IBCAST)*. IEEE. 2019, pp. 619–623.

[154] *Cleveland Heart Disease Dataset*. IEEE Dataport. DOI: https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive. 2020.

[155] Rahma Atallah and Amjed Al-Mousa. "Heart disease detection using machine learning majority voting ensemble method". In: *2019 2nd international conference on new trends in computing sciences (ictcs)*. IEEE. 2019, pp. 1–6.

[156] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30 (2017).

[157] Sheikh Rabiul Islam et al. "Explainable artificial intelligence approaches: A survey". In: *arXiv preprint arXiv:2101.09429* (2021).

[158] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[159] Prabhaker Mishra et al. "Descriptive statistics and normality tests for statistical data". In: *Annals of cardiac anaesthesia* 22.1 (2019), pp. 67–72.

[160] Angier Allen et al. "A racially unbiased, machine learning approach to prediction of mortality: algorithm development study". In: *JMIR public health and surveillance* 6.4 (2020), e22400.

[161] Alvin Rajkomar et al. "Ensuring fairness in machine learning to advance health equity". In: *Annals of internal medicine* 169.12 (2018), pp. 866–872.

[162] Tsehay Admassu Assegie. "Heart disease prediction model with k-nearest neighbor algorithm". In: *International Journal of Informatics and Communication Technology (IJ-ICT)* 10.3 (2021), p. 225.

[163] Fahd Saleh Alotaibi. "Implementation of machine learning model to predict heart failure disease". In: *International Journal of Advanced Computer Science and Applications* 10.6 (2019).

[164] Maryam Panahiazar et al. "Using EHRs and machine learning for heart failure survival analysis". In: *Studies in health technology and informatics* 216 (2015), p. 40.

[165] Joon-myoung Kwon et al. "Artificial intelligence algorithm for predicting mortality of patients with acute heart failure". In: *PloS one* 14.7 (2019), e0219302.

[166] Kenichi Nakajima et al. "Machine learning-based risk model using 123I-metaiodobenzylguanidine to differentially predict modes of cardiac death in heart failure". In: *Journal of Nuclear Cardiology* 29.1 (2022), pp. 190–201.

[167] Eric D Adler et al. "Improving risk prediction in heart failure using machine learning". In: *European journal of heart failure* 22.1 (2020), pp. 139–147.

[168] Geoffrey H Tison et al. “Predicting incident heart failure in women with machine learning: the Womens Health Initiative Cohort”. In: *Canadian Journal of Cardiology* 37.11 (2021), pp. 1708–1714.

[169] Jonathan Huang et al. “Evaluation and mitigation of racial bias in clinical machine learning models: scoping review”. In: *JMIR Medical Informatics* 10.5 (2022), e36388.

[170] Rachel KE Bellamy et al. “AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias”. In: *IBM Journal of Research and Development* 63.4/5 (2019), pp. 4–1.

[171] Bahar Memarian and Tenzin Doleck. “Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI), and higher education: A systematic review”. In: *Computers and Education: Artificial Intelligence* (2023), p. 100152.

[172] Aditya Singhal et al. “Toward Fairness, Accountability, Transparency, and Ethics in AI for Social Media and Health Care: Scoping Review”. In: *JMIR Medical Informatics* 12.1 (2024), e50048.

[173] Max Hort et al. “Bias mitigation for machine learning classifiers: A comprehensive survey”. In: *ACM Journal on Responsible Computing* (2023).

[174] Faisal Kamiran and Toon Calders. “Data preprocessing techniques for classification without discrimination”. In: *Knowledge and information systems* 33.1 (2012), pp. 1–33.

[175] Ryan Poplin et al. “Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning”. In: *Nature biomedical engineering* 2.3 (2018), pp. 158–164.

[176] Imon Banerjee et al. “Reading race: AI recognises patient’s racial identity in medical images”. In: *arXiv preprint arXiv:2107.10356* (2021).

[177] Md Rahat Shahriar Zawad and Peter Washington. “Evaluating Fair Feature Selection in Machine Learning for Healthcare”. In: *arXiv preprint arXiv:2403.19165* (2024).

[178] Nima Shahbazi et al. “Representation bias in data: a survey on identification and resolution techniques”. In: *ACM Computing Surveys* 55.13s (2023), pp. 1–39.

[179] Harini Suresh and John Guttag. “A framework for understanding sources of harm throughout the machine learning life cycle”. In: *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 2021, pp. 1–9.

[180] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. "Mitigating unwanted biases with adversarial learning". In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 2018, pp. 335–340.

[181] Nitesh V Chawla et al. "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16 (2002), pp. 321–357.

[182] Vincent Grari et al. "Fair adversarial gradient tree boosting". In: *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE. 2019, pp. 1060–1065.

[183] Leo Rutherford et al. "Health and well-being of trans and non-binary participants in a community-based survey of gay, bisexual, and queer men, and non-binary and Two-Spirit people across Canada". In: *PLoS One* 16.2 (2021), e0246525.

[184] Joshua D Safer et al. "Barriers to healthcare for transgender individuals". In: *Current Opinion in Endocrinology, Diabetes and Obesity* 23.2 (2016), pp. 168–171.

[185] Noor Beckwith et al. "Psychiatric epidemiology of transgender and nonbinary adult patients at an urban health center". In: *LGBT health* 6.2 (2019), pp. 51–61.

[186] Athasit Vejjajiva and Graham M Teasdale. "Serum creatine kinase and physical exercise". In: *British Medical Journal* 1.5451 (1965), p. 1653.

[187] Silvia Chiappa. "Path-specific counterfactual fairness". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 7801–7808.

[188] Weishen Pan et al. "Explaining Algorithmic Fairness Through Fairness-Aware Causal Path Decomposition". In: KDD '21 (2021), pp. 1287–1297. DOI: 10.1145/3447548.3467258. URL: https://doi.org/10.1145/3447548.3467258.

[189] Stefan Feuerriegel et al. "Causal machine learning for predicting treatment outcomes". In: *Nature Medicine* (2024).

[190] Julius von Kügelgen, Luigi Gresele, and Bernhard Schölkopf. "Simpson's paradox in Covid-19 case fatality rates: a mediation analysis of age-related causal effects". In: *IEEE transactions on artificial intelligence* 2.1 (2021), pp. 18–27.

[191] Catherine Gebhard et al. "Impact of sex and gender on COVID-19 outcomes in Europe". In: *Biology of sex differences* 11 (2020), pp. 1–13.

[192] Sergio E Chiarella, Christina Pabelick, and YS Prakash. "Sex differences in the coronavirus disease 2019". In: *Sex-based differences in Lung physiology* (2021), pp. 471–490.

[193] Garima Sharma, Annabelle Santos Volgman, and Erin D Michos. "Sex differences in mortality from COVID-19 pandemic: are men vulnerable and women protected?" In: *Case Reports* 2.9 (2020), pp. 1407–1410.

[194] Ricardo De La Vega et al. "Could attitudes toward COVID-19 in Spain render men more vulnerable than women?" In: *Global public health* 15.9 (2020), pp. 1278–1291.

[195] Natalia Díaz-Rodríguez et al. “Gender and sex bias in COVID-19 epidemiological data through the lens of causality”. In: *Information Processing & Management* 60.3 (2023), p. 103276.

[196] Jean C Digitale et al. “Key concepts in clinical epidemiology: collider-conditioning bias”. In: *Journal of Clinical Epidemiology* 161 (2023), pp. 152–156.

[197] Thanaphong Phongpreecha et al. “Multivariate prediction of dementia in Parkinsons disease”. In: *npj Parkinson's Disease* 6.1 (2020), p. 20.

[198] Roland M Atkinson, Linda Ganzini, and Michael J Bernstein. “Alcohol and substance-use disorders in the elderly”. In: *Handbook of mental health and aging*. Elsevier, 1992, pp. 515–555.

[199] Sangmin Byeon and Woojoo Lee. “An Introduction to Causal Mediation Analysis With a Comparison of 2 R Packages”. In: *Journal of Preventive Medicine and Public Health* 56.4 (2023), p. 303.

[200] Daniel Major-Smith. “Exploring causality from observational data: An example assessing whether religiosity promotes cooperation”. In: *Evolutionary Human Sciences* 5 (2023), e22.

[201] Matt J Kusner et al. “Counterfactual fairness”. In: *Advances in neural information processing systems* 30 (2017).

[202] Nina de Lacy et al. “Predicting individual cases of major adolescent psychiatric conditions with artificial intelligence”. In: *Translational psychiatry* 13.1 (2023), p. 314.

[203] Lonnie R Snowden. “Bias in mental health assessment and intervention: Theory and evidence”. In: *American journal of public health* 93.2 (2003), pp. 239–243.

[204] Dana Becker and Sharon Lamb. “Sex bias in the diagnosis of borderline personality disorder and posttraumatic stress disorder.” In: *Professional Psychology: Research and Practice* 25.1 (1994), p. 55.

[205] Jane Serrita Jane. *Gender bias in diagnostic criteria for personality disorders: An item response theory analysis*. University of Virginia, 2001.

[206] Meng-Chuan Lai, Simon Baron-Cohen, and Joseph D Buxbaum. “Understanding autism in the light of sex/gender”. In: *Molecular autism* 6 (2015), pp. 1–5.

[207] Hannah L Belcher et al. “Gender bias in autism screening: measurement invariance of different model frameworks of the Autism Spectrum Quotient”. In: *BJPsych Open* 9.5 (2023), e173.

[208] Cathie Sudlow et al. “UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age”. In: *PLoS medicine* 12.3 (2015), e1001779.

[209] James K Ruffle et al. “Computational limits to the legibility of the imaged human brain”. In: *NeuroImage* 291 (2024), p. 120600.

[210] Labanté Outcha Daré et al. “Co-morbidities of mental disorders and chronic physical diseases in developing and emerging countries: a meta-analysis”. In: *BMC public health* 19 (2019), pp. 1–12.

[211] Nirupama Putcha et al. “Comorbidities and chronic obstructive pulmonary disease: prevalence, influence on outcomes, and management”. In: *Seminars in respiratory and critical care medicine*. Vol. 36. 04. Thieme Medical Publishers. 2015, pp. 575–591.

[212] Christine Emmer, Michael Bosnjak, and Jutta Mata. “The association between weight stigma and mental health: A meta-analysis”. In: *Obesity Reviews* 21.1 (2020), e12935.

[213] Barton Willage. “The effect of weight on mental health: New evidence using genetic IVs”. In: *Journal of Health Economics* 57 (2018), pp. 113–130.

[214] A Skurvydas et al. “Relationship between simple reaction time and body mass index”. In: *Homo* 60.1 (2009), pp. 77–85.

[215] Rudy Bowen et al. “Mood instability is the distinctive feature of neuroticism. Results from the British Health and Lifestyle Study (HALS)”. In: *Personality and Individual Differences* 53.7 (2012), pp. 896–900.

[216] Thomas A Widiger and Joshua R Oltmanns. “Neuroticism is a fundamental domain of personality with enormous public health implications”. In: *World psychiatry* 16.2 (2017), p. 144.

[217] Morgane Evin et al. “Personality trait prediction by machine learning using physiological data and driving behavior”. In: *Machine Learning with Applications* 9 (2022), p. 100353.

[218] Johan Ormel, Judith Rosmalen, and Ann Farmer. “Neuroticism: a non-informative marker of vulnerability to psychopathology”. In: *Social psychiatry and psychiatric epidemiology* 39 (2004), pp. 906–912.

[219] Sarah Bauermeister and John Gallacher. “A psychometric evaluation of the 12-item epq-r neuroticism scale in 384,183 uk biobank participants using item response theory (irt)”. In: *BioRxiv* (2019), p. 741249.

[220] Jane M Ussher. “The Madness of Women: Myth and Experience”. In: *The Palgrave Handbook of the History of Human Sciences*. Springer, 2022, pp. 1853–1876.

[221] Paula J Caplan. *They say you’re crazy: How the world’s most powerful psychiatrists decide who’s normal.* Addison-Wesley/Addison Wesley Longman, 1995.

[222] Marina Sulastiana Djudiyah, Diana Harding, and Suryana Sumantri. "Gender differences in neuroticism on college students". In: *Dalam Asean Conference. 2nd Psychology & Humanity. Psychology Forum UMM*. Vol. 18. 2016, pp. 1432–1451.

[223] Harold Edward Bays et al. "Obesity pillars roundtable: body mass index and body composition in black and female individuals. Race-relevant or racist? Sex-relevant or sexist?" In: *Obesity Pillars* 4 (2022), p. 100044.

[224] Stephen R Daniels. "The use of BMI in the clinical setting". In: *Pediatrics* 124.Supplement_1 (2009), S35–S41.

[225] Silvia Sara Canetto and Isaac Sakinofsky. "The gender paradox in suicide". In: *Suicide and Life-Threatening Behavior* 28.1 (1998), pp. 1–23.

[226] Maya L Petersen et al. "Diagnosing and responding to violations in the positivity assumption". In: *Statistical methods in medical research* 21.1 (2012), pp. 31–54.

[227] Abeba Birhane. "Algorithmic injustice: a relational ethics approach". In: *Patterns* 2.2 (2021).

[228] Mingyang Wan et al. "Modeling techniques for machine learning fairness: A survey". In: *arXiv preprint arXiv:2111.03015* (2021).

[229] Nenad Tomasev et al. "Fairness for unobserved characteristics: Insights from technological impacts on queer communities". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 2021, pp. 254–265.

[230] Kevin Guyan. *Queer Data: Using Gender, Sex, and Sexuality Data for Action*. Bloomsbury Publishing, 2022.