Sex-Based Performance Disparities in Machine Learning Algorithms for Cardiac Disease Prediction: An Exploratory Study

Isabel Straw, Geraint Rees, Parashkev Nachev

Submitted to: Journal of Medical Internet Research
on: March 03, 2023

Table of Contents

Sex-Based Performance Disparities in Machine Learning Algorithms for Cardiac Disease Prediction: An Exploratory Study

Isabel Straw[1] BMBS, BMedSci, MPH, MRES; Geraint Rees[1]; Parashkev Nachev[1]

[1]University College London London GB

Corresponding Author:
Isabel Straw BMBS, BMedSci, MPH, MRES
University College London
222 Euston Road
London
GB

Abstract

BACKGROUND
The presence of bias in AI systems has garnered increased attention over the past decade, with inequities in algorithmic performance being exposed across the fields of criminal justice, education, and welfare services. In healthcare, the inequitable performance of medical algorithms across demographic groups may widen health inequalities. Here we identify and characterise bias in cardiology algorithms, looking specifically at algorithms used in the management of heart failure.

METHODS
Stage 1 involved a literature search of PUBMED and Web of Science for key terms relating to cardiac machine learning (ML) algorithms. Articles that built ML models to predict cardiac disease were evaluated for their focus demographic bias in model performance, and open-source datasets were retained for our own investigation. Two open-source datasets were identified; (i) UCI Heart Failure Dataset, (ii) UCI Coronary Artery Disease Dataset. We reproduced existing algorithms that have been reported for these datasets and tested them for sex biases in algorithm performance. Particular attention was paid to disparities in the False Negative Rate (FNR), due to the clinical significance of underdiagnosis and missed opportunities for treatment. A range of bias remediation techniques were implemented and assessed for their efficacy in reducing inequities, including dataset balancing, sex-specific feature selection and Fair Adversarial Gradient Tree Boosting.

RESULTS
In Stage 1, our literature search returned 127 articles of which 60 met the criteria for full review. Of these, only three papers highlighted sex differences in algorithm performance. In the papers that reported sex, there was a consistent underrepresentation of females in the datasets. No papers investigated racial or ethnic differences. In Stage 2, we reproduced algorithms reported in the literature achieving mean accuracies of 84.24% (3.51 SD) for Dataset 1, and 85.72% (1.75 SD) for Dataset 2 (Random Forest models). For Dataset 1, the FNR was significantly higher for females in 13 out of 16 experiments, meeting the threshold of statistical significance (-17.81% to -3.37%, $p<0.05$). A smaller disparity in the False Positive Rate (FPR) was significant for males in 13 out of 16 experiments (-0.48% to +9.77%, $p<0.05$)). We observed an overprediction of disease for males (higher FPR) and an underprediction of disease for females (higher FNR). Sex differences in feature importance suggests that feature selection needs to be demographically tailored.

DISCUSSION
Our research exposes a significant gap in cardiac ML research, highlighting that the underperformance of algorithms for female patients has been overlooked in the published literature. Our study quantifies sex disparities in the algorithmic performance and explores several sources of bias. We found an underrepresentation of females in the datasets used to train algorithms, identified sex biases in model error rates and demonstrated that a series of remediation techniques were unable to address the inequities present.

(JMIR Preprints 03/03/2023:46936)
DOI: https://doi.org/10.2196/preprints.46936

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Sex-Based Performance Disparities in Machine Learning Algorithms for Cardiac Disease Prediction: An Exploratory Study

Isabel Straw*[1], Geraint Rees[1], Parashkev Nachev[1]

[1]UCL Queen Square Institute of Neurology, University College London, London, UK

*Correspondence to Dr Isabel Straw, Institute of Health Informatics, University College London, London, UK.
Email: isabelstraw@doctors.org.uk

Word Count: 3496

ABSTRACT

BACKGROUND
The presence of bias in AI systems has garnered increased attention over the past decade, with inequities in algorithmic performance being exposed across the fields of criminal justice, education, and welfare services. In healthcare, the inequitable performance of medical algorithms across demographic groups may widen health inequalities. Here we identify and characterise bias in cardiology algorithms, looking specifically at algorithms used in the management of heart failure.

METHODS
Stage 1 involved a literature search of PUBMED and Web of Science for key terms relating to cardiac machine learning (ML) algorithms. Articles that built ML models to predict cardiac disease were evaluated for their focus demographic bias in model performance, and open-source datasets were retained for our own investigation. Two open-source datasets were identified; (i) UCI Heart Failure Dataset, (ii) UCI Coronary Artery Disease Dataset. We reproduced existing algorithms that have been reported for these datasets and tested them for sex biases in algorithm performance. Particular attention was paid to disparities in the False Negative Rate (FNR), due to the clinical significance of underdiagnosis and missed opportunities for treatment. A range of bias remediation techniques were implemented and assessed for their efficacy in reducing inequities, including dataset balancing, sex-specific feature selection and Fair Adversarial Gradient Tree Boosting.

RESULTS
In Stage 1, our literature search returned 127 articles of which 60 met the criteria for full review. Of these, only three papers highlighted sex differences in algorithm performance. In the papers that reported sex, there was a consistent underrepresentation of females in the datasets. No papers investigated racial or ethnic differences. In Stage 2, we reproduced algorithms reported in the literature achieving mean accuracies of 84.24% (3.51 SD) for Dataset 1, and 85.72% (1.75 SD) for Dataset 2 (Random Forest models). For Dataset 1, the FNR was significantly higher for females in 13 out of 16 experiments, meeting the threshold of statistical significance (-17.81% to -3.37%, $p<0.05$). A smaller disparity in the False Positive Rate (FPR) was significant for males in 13 out of 16 experiments (-0.48% to +9.77%, $p<0.05$)). We observed an overprediction of disease for males (higher FPR) and an underprediction of disease for females (higher FNR). Sex differences in feature importance suggests that feature selection needs to be demographically tailored.

DISCUSSION
Our research exposes a significant gap in cardiac ML research, highlighting that the underperformance of algorithms for female patients has been overlooked in the published literature. Our study quantifies sex disparities in the algorithmic performance and explores several sources of bias. We found an underrepresentation of females in the datasets used to train algorithms, identified sex biases in model error rates and demonstrated that a series of remediation techniques were unable to address the inequities present.

CONFLICT OF INTEREST: The authors have no conflict of interest to declare.

Key Words: Artificial Intelligence, Machine Learning, Cardiology, Healthcare, Health Equity, Medicine

INTRODUCTION
Artificial Intelligence (AI) has been proposed as an effective solution to many healthcare challenges and depends on the construction of Machine Learning (ML) algorithms from healthcare data. Recent

research has drawn attention to the possibility that algorithms may exhibit bias when applied to different demographic groups [1-6]. Such biases may widen health inequalities and negatively impact marginalised patients, such as women, minoritised racial and ethnic groups and other neglected subpopulations [1-7].

Over the past five years an increasing number of studies have quantified disparities in algorithmic performance for underserved populations [2-7]. Daneshjou and colleagues demonstrated that state-of-the art Dermatology algorithms tend to perform worse on darker skin tones [2]; Seyyed-Kalantari and colleagues exposed biases in radiology algorithms [3]; and Thompson and Colleagues reported increased false negative errors when classifying opioid misuse disorder for Black patients compared to White patients [4]. Beyond specific diagnoses, researchers have demonstrated that infrastructural AI systems used in hospital settings can be subject to referral bias, demonstrated by Obermeyer and colleagues who highlighted a hospital treatment allocation algorithm that overlooked the health needs of Black patients [5]. Yet despite the increasing number of papers describing this issue, most of the current uses of biomedical AI technologies do not account for the problem of bias [5-8]. Here, we evaluate algorithmic inequity in ML algorithms used for predicting cardiac disease, focusing on Heart Failure (HF).

Machine learning for Heart Failure (HF)

Heart failure (HF) is a clinical syndrome in which the heart is unable to maintain a cardiac output adequate to meet the metabolic demands of the body [9]. Traditionally, algorithmic tools capable of identifying at-risk patients have played a key role in informing decisions on heart failure management and end-of life care [10-12]. In recent years, ML algorithms that leverage biochemical data have been proposed as a superior alternative to traditional statistical models for identifying at risk patients with heart failure [13]. A range of ML techniques outperform traditional risk scores in forecasting heart-failure related events [13]. Yet given that existing medical research has described sex differences in both the presentation and management of heart failure, it is possible that algorithms trained on existing data will perform differently for males vs. females [14-15].

Sex differences in Heart Failure (HF)

Heart failure presents differently in female patients compared with male patients [14]. Females experience a wider range of symptoms, including higher fluid overload and lower health-related quality of life [14-15]. Moreover, females who present with heart failure are on average older, sustain a higher Ejection Fraction (EF) throughout later stages of disease and have a lower incidence of previous ischaemic heart disease [15]. Furthermore, the biochemical tests used to detect cardiac disease have been demonstrated to perform less well for female patients [16]. Troponin is one key biomarker used to predict disease, which has been demonstrated to be less sensitive in female patients [16]. Standard troponin criteria fail to detect one out of five acute myocardial infarcts occurring in females [16]. Historically, the neglect of sex-differences in cardiac pathophysiology has disadvantaged female patients and, if not considered during ML development, these inequities may manifest in the novel algorithms being integrated into cardiac care [14-19].

In our research we scope the published literature reporting algorithms that predict Heart Failure (HF) and investigate whether existing papers give attention to bias in ML algorithms. Furthermore, we examine the datasets of existing models for demographic representation, evaluate demographic inequities in algorithmic performance, and assess the efficacy of a series of bias-mitigation techniques.

METHODS

Our 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 1 provides an overview of our approach.

Figure 1: A flowchart detailing the steps of our methodology, including (1) the initial literature search and qualitative evaluation of identified studies, plus (2) the identification of datasets and interrogation of algorithms for demographic bias.

Stage 1 Literature Review: Qualitative Evaluation of Published Articles

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 (Figure 2, Supplementary Table 1-2). 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)

Of the retained papers, full texts were then reviewed to evaluate whether authors:

(1) Reported the demographic make-up of datasets.
(2) Evaluated demographic inequities in algorithm performance, meaning that the authors specifically examined differences in algorithmic performance by demographic groups defined by protected characteristics [17].

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

Figure 2: PRISMA 2020 flow diagram for new systematic reviews which included searches of databases and registers only (PRISMA templated obtained from PRISMA at https://prisma-statement.org/prismastatement/flowdiagram.aspx)

Stage 2: Quantitative Evaluation of Model Performance

Datasets

Two open-source datasets were uncovered in our literature review; (i) Dataset 1 – UCI for Heart Failure Prediction and (ii) Dataset 2 – UCI Cleveland Heart Disease dataset for identifying Coronary Artery Disease (CAD) [21-22]. Descriptive statistics were performed on both datasets, evaluating the mean and variance of the dataset variables for sexes separately, affected by disease/death (Table 1, Supplementary Table 3-5).

Using these datasets, we rebuilt the ML algorithms described in the published literature and performed an additional analysis exploring for inequities in algorithmic performance for demographic subgroups. As the only protected characteristic reported was Sex, we focus on sex disparities in performance. Despite our initial aim to focus on heart failure, we retained an uncovered CAD dataset to investigate whether trends identified for heart failure generalised to patients with CAD [22]. Supplementary Table 3 and Supplementary Table 4 provide details on Dataset 1 and Dataset 2 respectively.

Model Reproduction

We rebuilt the models described in the existing literature for these datasets, focusing on Random Forest (RF) algorithms which have been widely reported to be the most effective models [23]. For both datasets, data was split into test/training subsets (0.7/0.3), RF Models were built using SciKit

Learn, and RF parameters were tuned using GridSearch CV. We adopted a bootstrapping approach to quantify uncertainty, such that models were built, trained, and tested 100 times, from which average results were derived with standard deviation.

Statistical Analysis

Across the 100 runs, sex differences in each algorithm evaluation metric (Equations 1-4) were calculated and averaged, with accompanying statistical tests performed to evaluate for statistical significance of any identified sex disparities. Our method for examining differences in algorithmic error rates builds on the foundational work from Buolamwini and Gebru, who demonstrated that a range of ML algorithms for facial recognition performed poorly on darker skinned females [20]. 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 [24].

Variations in Model Development

We then introduced a variety of changes to the model development, to evaluate the impact on the identified sex disparities in performance.

(i) Changes to model training data

One widely proposed bias mitigation technique includes pre-processing the training data of a model to account for demographic representation, with previous research highlighting the benefit of training on demographically balanced or demographically stratified datasets [25]. We therefore created a range of datasets with varied sex representation and assessed for the impact on algorithm performance disparities. To form the sex-balanced dataset, we utilised the oversampling function of SMOTE(), which has been proposed as an effective method for improving representation of underserved populations in machine learning datasets [26]. The SMOTE package generates new minority data points based on existing minority samples through linear interpolation [25-26]. Models were rebuilt as per Section 2.2.2 using four different training datasets (sex-imbalanced, sex-balanced and sex specific, Supplementary Table 6-7):

- Original Sex-Imbalanced Training Data
- Sex-Balanced Training Data
- Female-only Training Data
- Male-only Training Data Experiments

(ii) 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 [25, 27]. To do this, Shapley Values have been widely accepted as a unified measure of feature importance since their proposal in 2017 [28].

In our experiments, we first perform an exploratory analysis, comparing feature importance for models trained on the male vs. female datasets. Secondly, we create four feature subsets from the original datasets, to evaluate the impact of changing the feature selection on performance disparities. As described in the introduction, existing clinical research has described demographic differences in the biochemical and clinical markers of HF disease (e.g., sex differences in Ejection Fraction and Troponin levels) [16]. Thus, we delineate four different feature subsets that vary in this information, to examine whether certain feature subsets perform better for different demographic groups. These four feature subsets are described in detail in Supplementary Tables 8-9, and include (i) Features with sex, (ii) Features without sex, (iii) Biochemical features, (iv) Clinical features.

Our final series of experiments are therefore performed across the four training datasets (sex-imbalanced, sex-balanced and sex specific), and the four feature sets giving 16 total experiments:

- Original Sex-Imbalanced Training Data Experiments (across four feature subsets)
- Sex-Balanced Training Data Experiments (across four feature subsets)
- Female Training Data Experiments (across four feature subsets)
- Male Training Data Experiments (across four feature subsets)

Model Evaluation & Identification of Performance Disparities

Models are evaluated using global evaluation metrics (e.g. Accuracy) and specific error rates (e.g. False Negative Rate [FNR]) (Equation 1 – 4). The difference between male and female scores are calculated to give a model's 'Sex performance disparity' (Equation 4). To evaluate for statistical significance, 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.

Our choice of evaluation metrics is guided by the clinical consequence of each of these scores.
The existing research on algorithmic bias has highlighted the importance of examining error rates, particularly in medicine where a false negative clinically translates to missed diagnoses or opportunities for treatment [3-6,25]. As described by Afrose and colleagues, focusing on global metrics of performance such as ROC_AUC scores can neglect subtler disparities arising from differences in error rates affecting subgroups [25]. When selecting a bias assessment metric, previous studies have chosen to focus on False Negative Rate and False Positive Rate, due to the clinical implications of these errors [4, 29-30]. Equation 2 places the error rates in their clinical context, demonstrating that the False Negative Rate (FNR) represents missed diagnoses and potentially missed treatment. For the error rates, we utilise the threshold of 0.5, as we are investigating for performance inequities in the existing reported models that utilised these default settings.

Equation 1: Error Rate Definitions

$$True\ Positive\ Rate = \frac{TP}{TP+FN} = 1 - FNR = Percentage\ of\ actual\ positives\ which\ are\ correctly\ identified$$

$$False\ Positive\ Rate = \frac{FP}{FP+TN} = 1 - TNR = Percentage\ of\ negative\ results\ falsely\ labelled\ as\ positive$$

$$True\ Negative\ Rate = \frac{TN}{TN+FP} = 1 - FPR = Percentage\ of\ negatives\ that\ are\ corretly\ identified$$

$$False\ Negative\ Rate = \frac{FN}{FN+TP} = 1 - TPR = Percentage\ of\ positive\ results\ falsely\ labelled\ as\ negative$$

Equation 2: Clinical Implications of Error Rates

$$True\ Positive\ Rate = Correct\ diagnosis\ that\ patient\ has\ disease$$

$$False\ Positive\ Rate = Misdiagnosis\ of\ disease\ when\ patient\ is\ healthy$$

$$True\ Negative\ Rate = Correct\ diagnosis\ that\ patient\ is\ healthy$$

$$False\ Negative\ Rate = Misdiagnosis\ that\ patient\ is\ healthy\ ,\ when\ patient\ has\ disease$$

Equation 3: Accuracy Evaluation Metric

$$Accuracy = \frac{True\ positives + True\ Negatives}{True\ positives + True\ Negatives + False\ Positives + False\ Negatives}$$

Equation 4: Sex Performance Disparity

$$Sex\ performance\ disparity = Score\ for\ males\ (mean) - Score\ for\ females\ (mean)$$

Fairness Techniques: Fair Adversarial Gradient Tree Boosting (FAGTB)

We implemented a recent fairness technique to evaluate whether these approaches were applicable to bias in HF algorithms. The Fair Adversarial Gradient Tree Boosting (FAGTB) is a recent technique proposed by Grari et al for mitigating bias in decision tree classifiers and the authors demonstrate the success of their technique on four datasets [8]. The authors focus on two definitions of fairness – Demographic Parity and Equalized Odds [8]. The Equalized Odds metric focuses on model FPR and FNR, and hence we highlight this for our paper. A summary of these fairness metrics is provided in Supplementary Section 10 for further interest.

Equation 5, Definition of Equalised Odds:

$$P(\hat{Y}=1 | S=0, Y=y¿ = P(\hat{Y}=1 | S=1, Y=y¿, \forall\, y \in \{0,1\}$$

To assess for the Equalised Odds the authors measure the Disparate Mistreatment (DM), which computes the absolute difference between FPR and the FNR for both demographics.

Equation 6, Disparate False Positive Rate:

$$D_{FPR}: |P(\hat{Y}=1 | Y=0, S=1) - P(\hat{Y}=1 | Y=0, S=0¿ \vee ¿$$

Equation 6, Disparate False Negative Rate:

$$D_{FNR}: |P(\hat{Y}=0 | Y=1, S=1) - P(\hat{Y}=0 | Y=1, S=0¿ \vee ¿$$

We compare the performance of the FAGTB algorithm to a standard Gradient Tree Algorithm. As per the original FAGTB paper, we repeat 10 experiments randomly sampling two subsets (0.8/0.2) and report evaluation metrics for the test set.

RESULTS

Literature Review Search Results

Our search returned 127 articles of which 60 met the criteria for full review and three highlighted sex differences in model performance. In the papers that reported sex, there was a consistent underrepresentation of females. No papers investigated racial or ethnic differences. One paper focused specifically on females with heart failure in which, Tison et al. highlight 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 [31].

Descriptive Statistics & Feature Importance

Dataset 1 (Heart Failure)

The mean descriptive statistics for each feature present in the Heart Failure (HF) dataset are provided in Table 1, which demonstrates subtle sex differences in the presentation of the disease. For 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 1: Descriptive statistics of the variables in Dataset 1 (Heart Failure) (n=299), stratified by Target (Death) and Sex

Sex	**Female (Sex = 0)** (n=105)				**Male (Sex = 1)** (n=194)			
Death (Target Variable)	**Survived (HF Death = 0)**		**Death (HF Death = 1)**		**Survived (HF Death = 0)**		**Death (HF Death = 1)**	
Total Count (N) & Event Rate (%)	71 (67.7%)		34 (32.4%)		132 (68.0%)		62 (32.0%)	
	Mean	Std. Dev	Mean	Std. Dev	Mean	Std. Dev	Mean	Std. Dev
Age (years)	58.6	10.6	62.2	12.3	58.8	10.7	66.9	13.5
Anaemia (boolean)	0.5	0.5	0.6	0.5	0.4	0.5	0.4	0.5
CPK (mcg/L)	462.0	517.7	507.7	779.7	582.8	853.2	759.3	1532.3
DM (boolean)	0.5	0.5	0.6	0.5	0.4	0.5	0.3	0.5
EF (percentage)	41.9	11.6	37.5	14.6	39.4	10.4	31.2	10.7
High BP (boolean)	0.4	0.5	0.5	0.5	0.3	0.5	0.4	0.5
Platelets (kiloplatelets/mL)	289757.6	98655.9	259512.7	107588.6	254232.4	94985.6	254663.7	94060.8
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 (boolean)	0.0	0.1	0.1	0.3	0.5	0.5	0.4	0.5

**For the Death variable, a value of 1 indicates mortality. CPK = Creatinine Phosphokinase, DM = Diabetes Mellitus, EF = Ejection Fraction, High BP = High Blood Pressure, Creatinine = Serum Creatinine, Sodium = Serum Sodium. Full details of dataset variables available at: https://archive.ics.uci.edu/dataset/519/heart+failure+clinical+records*

Our exploratory analysis identified further sex differences on examining feature importance. Figure 3 compares the rankings of feature importance for ML models built to predict HF built from the female dataset compared to the male dataset. These differences are important as existing ML algorithms built on mixed-sex cohorts suggest that Ejection Fraction (EF) can be used alone for modelling, an approach which may disadvantage females [23].

Figure 3: Comparison of feature rankings for male and female patients, ordered by SHAP values.

Dataset 2 (Coronary Artery Disease)

Supplementary Table 5 provides details of the CAD dataset and demonstrates that females with CAD have a higher resting BP and higher cholesterol compared to males. The categorical variable 'Resting ECG' is also higher for females, due to a higher incidence of Left Ventricular Hypertrophy.

Model Results and Performance Disparities

We replicated the algorithms described in the existing literature, reproducing the same previously reported mean predictive accuracies of 84.24% (3.51 SD) for Dataset 1, and 85.72% (1.75 SD) for Dataset 2 [23]. In Table 2 and Table 3 we present the disparity in performance for the sexes, where a positive value indicates higher value for males (see Equation 4).

For Dataset 1, Table 2 demonstrates that in 13 out of 16 experiments the False Negative Rate (FNR) is higher for females, meeting the threshold of statistical significance (mean difference of -17.81% to -3.37%, $p<0.05$). Figure 4 represents this disparity in performance graphically, providing the point estimates of FNR for the Sexes separately and highlighting that the disparity in FNR persisted across the variations in training data and selected features.

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%, $p<0.05$)). The sex performance disparities in Accuracy and ROC_AUC varied depending on the underlying shifts in the error rates for each sex (Table 2, Figure 5). On examining the individual error rates, we see consistencies in the sex disparities across feature sets, most notably an overprediction of disease for males (higher FPR) and an under prediction of disease for females (higher FNR – Table 2).

Our 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%, $p<0.05$ – Table 3) and a higher FPR for males (3.94% to 4.71%, $p<0.05$ – Table 3). Figure 6 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 6). Figure 7 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 (Figure 7).

Table 2: Sex performance disparities for models built from Dataset 1 (Heart Failure Disease) – Sex performance disparities are calculated as the performance for males, minus the performance for females (see Equation 4). Thus, a positive value indicates a higher score for males, a negative value indicates a higher score for females. All disparities are presented with alongside results of significance testing, where significant differences between the sexes highlighted with an asterix (p<0.05)

		Feature Subset used in Model Training			
Disparity in Model Performance *(Score for males – Score for females)*		**Features With Sex**	**Features Without Sex**	**Biochemical Features**	**Clinical Features**
Sex-Imbalanced Training Data					
	Accuracy Disparity (%)	*1.63 (0.03)	-0.72 (0.30)	0.10 (0.88)	-0.50 (0.49)
	ROC_AUC 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					
	Accuracy Disparity (%)	*-4.78 (<0.01)	*-7.25 (<0.01)	*-9.42 (<0.01)	*-3.63 (<0.01)
	ROC_AUC 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					
	Accuracy Disparity (%)	*-10.95 (<0.01)	*-9.75 (<0.01)	*-12.32 (<0.01)	*-9.64 (<0.01)
	ROC_AUC 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					
	Accuracy Disparity (%)	*-5.46 (<0.01)	*-5.73 (<0.01)	*-8.73 (<0.01)	*-2.46 (<0.01)
	ROC_AUC 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)

**Indicates a statistically significant difference (p<0.05) between the model's performance on male vs females.*

Figure 4: 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)
(See Additional Files)

Figure 5: 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)
(See Additional Files)

Table 3: Sex performance disparities for models built from Dataset 2 (Coronary Artery Disease) – Sex performance disparities are calculated as the performance for males, minus the performance for females (see Equation 4). Thus, a positive value indicates a higher score for males, a negative value indicates a higher score for females. All disparities are presented with alongside results of significance testing, where significant differences between the sexes highlighted with an asterix (p<0.05)

		Feature Subset used in Model Training			
Disparity in Model Performance *(Score for males – Score for females)*		**Features With Sex**	**Features Without Sex**	**Biochemical Features**	**Clinical Features**
Sex-Imbalanced Training Data					
	Accuracy Disparity (%)	0.32 (0.50)	0.64 (0.17)	0.13 (0.8)	0.25 (0.61)
	ROC_AUC 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					
	Accuracy Disparity (%)	*-4.01 (<0.01)	*-5.12 (<0.01)	*-7.32 (<0.01)	*-2.86 (<0.01)
	ROC_AUC 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					
	Accuracy Disparity (%)	*-9.25 (<0.01)	*-11.34 (<0.01)	*-11.49 (<0.01)	*-8.69 (<0.01)
	ROC_AUC 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					
	Accuracy Disparity (%)	*6.38 (<0.01)	*5.66 (<0.01)	*-1.66 (0.02)	*6.10 (<0.01)
	ROC_AUC Disparity (%)	*6.30 (<0.01)	*5.57 (<0.01)	1.52 (0.07)	*5.86 (0.01)
	FNR Disparity (%)	*-10.12 (<0.01)	*-10.10 (<0.01)	1.67 (0.17)	*-12.64 (<0.01)
	FPR Disparity (%)	*-2.48 (<0.01)	-1.04 (0.07)	1.38 (0.24)	0.92 (0.15)

**Indicates a statistically significant difference (p<0.05) between the model's performance on male vs females. To determine statistical significance, the Kolmogorov-Smirnov Tests were first run on the sex stratified results to determine the distribution of data (normal or not). Independent t-tests were used where data was normally distributed, and Mann-Whitney U tests were used when data was not normally distributed.*

Figure 6: 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 & Male). The Feature Set used is indicated within each training data group (Features with Sex, Features Without Sex, Biochemical Features & Clinical Features)
(See Additional Files)

Figure 7: 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 & Male). The Feature Set used is indicated within each training data group (Features with Sex, Features Without Sex, Biochemical Features & Clinical Features)

Variations in Training Data

(i) Sex-Balanced 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 4, Figure 5). The opposite trend was seen in Dataset 2, with models trained on sex-balanced data outperforming models trained on sex-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 4). The models trained on sex-balanced data in Dataset 2 reduced the FNR for both sexes when using the full feature set (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 4, Figure 6). The differences between the datasets may relate to underlying differences in the two cardiac conditions. Further, the failure to improve performance with sex-balanced training data may reflect the issues of mixing data that has conflicting indicators for disease.

Table 4 – Model results when trained on sex specific subsets for all patients and male/females separately, looking at the "Features Including Sex" subset.

	Dataset 1 (Heart Failure)				Dataset 2 (Coronary Artery Disease)			
	Sex-Imbalanced Training Data (n=209)	**Sex-Balanced Training Data (n=272)**	**Female Training Data (n=136)**	**Male Training Data (n=136)**	**Sex-Imbalanced Training Data (n=522)**	**Sex-Balanced Training Data (n=715)**	**Female Training Data (n=358)**	**Male Training Data (n=358)**
All Patients, Mean Accuracy (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 Mean Accuracy (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 Mean Accuracy (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 Mean, Mean 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 Mean, Mean 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)

(ii) Sex Specific Training Data

For Dataset 1, mean accuracy for all patients when trained on sex-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 sex-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). The overall improvement seen in the Dataset 2 models when trained on female data, relates to the increase in accuracy for females (90.69% 2.38 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, Figure 7).

Unsurprisingly, performance for each sex is lowest when trained on the opposing sex (Table 4, Figure 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 sex-imbalanced and sex-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 sex-imbalanced models (Table 4).

Variations in Feature Sets

Models built on the biochemical features subset gave the worst performance in terms of accuracy and FNR (Figures 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, Supplementary Table 5 highlights the different biochemical profiles for sick males and females, with sick females demonstrating a far higher Cholesterol level for than their male counterparts (mean values; 279.2 Female Sick vs. 247.5 Male Sick).

FAGTB Model

The DispFNR was consistently higher than the DispFPR. 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 5: Results of Bias Mitigation with Fair Adversarial Gradient Tree Boosting (FAGTB)

Dataset 1 (Heart Failure) **Experiments run on sex-imbalanced data with all features (averaged over 10 experiments)**		
Results on test set, averaged over 10 experiments	**Gradient Boosting Classifier**	**Fair Adversarial Gradient Tree Boosting (FAGTB)**
Accuracy	71.3	71.2
DispFPR	0.08	0.08
DispFNR	0.21	0.20
Dataset 2 (Coronary Artery Disease) **Experiments run on sex-imbalanced data with all features (averaged over 10 experiments)**		
Results on test set, averaged over 10 experiments	**Gradient Boosting Classifier**	**Fair Adversarial Gradient Tree Boosting (FAGTB)**
Accuracy	86.3	82.9
DispFPR	0.06	0.06
DispFNR	0.28	0.19

DISCUSSION

Our study sheds light on an important gap in existing cardiac machine learning research, with significant implications for digital health equity. We find that the majority of published ML studies predicting heart failure fail to acknowledge the underrepresentation of female patients in their datasets and do not perform stratified model evaluations, thus failing to assess sex disparities in algorithmic performance. Our 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 our approach we identified several potential sources of algorithmic bias.

First, we detected underrepresentation of females in training datasets that may produce inequalities in model fidelity. 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 underrepresentation [32]. When balancing the dataset, our methods did not include under sampling due to our small datasets, however this may be a potential avenue for future research.

Second, we considered featurisation and highlighted sex differences in the biochemical manifestation of disease. In current clinical practice, the diagnostic parameters used for identifying pathology are drawn from research trials dominated by male physiology – it is perhaps unsurprising therefore that algorithms built from this data tend to underperform in female disease. 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 [16].

There are further sources of inequitable performance that our evaluation 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 [33]. There are complex causal relationships between biological, environmental, and social factors that underpin the differences in disease rates seen across population subgroups [33]. 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 [33].

Our research was 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 [34-36].

Throughout this paper we 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 [34-36]. 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 patient's physiology [37].

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 [38]. In addition to improving representation, unsupervised techniques enable us to detect neglected subpopulations 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.

Conclusion and Limitations

In our paper we have identified inequities in the performance of cardiac ML algorithms. Our findings are limited by the small size of the uncovered datasets, reducing their potential generalisability, and hence we propose that larger studies focused on this issue are required. These datasets also came from the same source, as we found a limited number of open-access databases due to the confidential nature of patient data and issues of proprietary ownership. In addition, we focused on Random Forest (RF) models to replicate the papers uncovered in our 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.

In our paper we did not attempt to solve bias, instead we highlighted a problem that exists throughout cardiology that requires further attention. The issue we have identified in these ML models is a foundational problem across medical modelling, in any instance where the use of an 'average' is applied to a diverse population. It is possible that unsupervised machine learning and complex representational modelling may be a route forward for capturing heterogeneity in a previously unattainable manner and addressing issues of bias [38]. Our findings demonstrate that examining performance inequities across demographic subgroups is an essential approach for identifying biases in AI and preventing the perpetuation of inequalities into digital health systems.

Acknowledgments

The data sets analysed during this study are publicly available. Dataset 1 is available from the University of California Irvine Machine Learning Repository [21]. Dataset 2 is available from the IEEE Dataport Repository [22].

Funding: This work was supported by UK Research and Innovation (UKRI Grant Reference Number EP/S021612/1)

5.0 REFERENCES

1. O'neil C. Weapons of math destruction: How big data increases inequality and threatens democracy. Crown; 2017 Sep 5.

2. Daneshjou R, Vodrahalli K, Novoa RA, Jenkins M, Liang W, Rotemberg V, Ko J, Swetter SM, Bailey EE, Gevaert O, Mukherjee P. Disparities in dermatology AI performance on a diverse, curated clinical image set. Science advances. 2022 Aug 12;8(31):eabq6147. https://doi.org/10.1126/sciadv.abq6147

3. Seyyed-Kalantari L, Liu G, McDermott M, Chen IY, Ghassemi M. CheXclusion: Fairness gaps in deep chest X-ray classifiers. InBIOCOMPUTING 2021: proceedings of the Pacific symposium 2020 (pp. 232-243). https://doi.org/10.1142/9789811232701_0022.

4. Thompson HM, Sharma B, Bhalla S, Boley R, McCluskey C, Dligach D, Churpek MM, Karnik NS, Afshar M. Bias and fairness assessment of a natural language processing opioid misuse classifier: detection and mitigation of electronic health record data disadvantages across racial subgroups. Journal of the American Medical Informatics Association. 2021 Nov 1;28(11):2393-403. https://doi.org/10.1093/jamia/ocab148

5. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019 Oct 25;366(6464):447-53. https://doi.org/10.1126/science.aax2342

6. Cirillo D, Catuara-Solarz S, Morey C, Guney E, Subirats L, Mellino S, Gigante A, Valencia A, Rementeria MJ, Chadha AS, Mavridis N. Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. NPJ digital medicine. 2020 Jun 1;3(1):81. https://doi.org/10.1038/s41746-020-0288-5

7. Liu, X., Hu, P., Yeung, W., Zhang, Z., et al (2023). Illness severity assessment of older adults in critical illness using machine learning (ELDER-ICU): an international multicentre study with subgroup bias evaluation. The Lancet. Digital health, 5(10), e657–e667. https://doi.org/10.1016/S2589-7500(23)00128-0

8. Grari V, Ruf B, Lamprier S, Detyniecki M. Fair adversarial gradient tree boosting. In2019 IEEE International Conference on Data Mining (ICDM) 2019 Nov 8 (pp. 1060-1065). IEEE. https://doi.org/10.1109/ICDM.2019.00124

9. Savarese G, Becher PM, Lund LH, Seferovic P, Rosano GM, Coats AJ. Global burden of heart failure: a comprehensive and updated review of epidemiology. Cardiovascular research. 2022 Dec 1;118(17):3272-87. https://doi.org/10.1093/cvr/cvac013

10. Goldraich L, Beck-da-Silva L, Clausell N. Are scores useful in advanced heart failure?. Expert Review of Cardiovascular Therapy. 2009 Aug 1;7(8):985-97. https://doi.org/10.1586/erc.09.68

11. Treece J, Chemchirian H, Hamilton N, Jbara M, Gangadharan V, Paul T, Baumrucker SJ. A review of prognostic tools in heart failure. American Journal of Hospice and Palliative Medicine®. 2018 Mar;35(3):514-22. https://doi.org/10.1177/1049909117709468

12. Thorvaldsen T, Benson L, Ståhlberg M, Dahlström U, Edner M, Lund LH. Triage of patients with moderate to severe heart failure: who should be referred to a heart failure center?. Journal of the American College of Cardiology. 2014 Feb 25;63(7):661-71. https://doi.org/10.1016/j.jacc.2013.10.017

13. Garate Escamilla AK, Hajjam El Hassani A, Andres E. A comparison of machine learning

techniques to predict the risk of heart failure. Machine Learning Paradigms: Applications of Learning and Analytics in Intelligent Systems. 2019:9-26. https://doi.org/10.1007/978-3-030-15628-2_2

14. Sullivan K, Doumouras BS, Santema BT, Walsh MN, Douglas PS, Voors AA, Van Spall HG. Sex-specific differences in heart failure: pathophysiology, risk factors, management, and outcomes. Canadian Journal of Cardiology. 2021 Apr 1;37(4):560-71. https://doi.org/10.1016/j.cjca.2020.12.025

15. Walsh MN, Jessup M, Lindenfeld J. Women with heart failure: unheard, untreated, and unstudied. Journal of the American College of Cardiology. 2019 Jan 8;73(1):41-3. https://doi.org/10.1016/j.jacc.2018.10.041

16. Sobhani K, Nieves Castro DK, Fu Q, Gottlieb RA, Van Eyk JE, Noel Bairey Merz C. Sex differences in ischemic heart disease and heart failure biomarkers. Biology of sex differences. 2018 Dec;9:1-3. https://doi.org/10.1186/s13293-018-0201-y

17. Straw I. The automation of bias in medical Artificial Intelligence (AI): Decoding the past to create a better future. Artificial intelligence in medicine. 2020 Nov 1;110:101965. https://doi.org/10.1016/j.artmed.2020.101965

18. Hamberg K. Gender bias in medicine. Women's health. 2008 May;4(3):237-43. https://doi.org/10.2217/17455057.4.3.237

19. Krieger N, Fee E. Man-made medicine and women's health: the biopolitics of sex/gender and race/ethnicity. Women's health, politics, and power. 2020 Nov 25:11-29.

20: Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. InConference on fairness, accountability and transparency 2018 Jan 21 (pp. 77-91). PMLR.

21. Tanvir Ahmad, A.M., Sajjad Haider Bhatti, Muhammad Aftab, and Muhammad Ali Raza, Heart failure clinical records Data Set F. Government College University, Pakistan, Editor. 2020: University of California Irvine Machine Learning Repository. https://doi.org/10.24432/C5Z89R

22. Manu Siddhartha. (2020). Heart Disease Dataset (Comprehensive). IEEE Dataport. https://dx.doi.org/10.21227/dz4t-cm36

23. Davide Chicco, Giuseppe Jurman: "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone". BMC Medical Informatics and Decision Making 20, 16 (2020). https://doi.org/10.1186/s12911-020-1023-5

24. Mishra P, Pandey CM, Singh U, Gupta A, Sahu C, Keshri A. Descriptive statistics and normality tests for statistical data. Ann Card Anaesth. 2019 Jan-Mar;22(1):67-72. https://doi.org/10.4103%2Faca.ACA_157_18

25. Afrose, S et al. 'Subpopulation-Specific Machine Learning Prognosis for Underrepresented Patients with Double Prioritized Bias Correction'. Communications Medicine, vol. 2, no. 1, Sept. 2022, pp. 1–14. www.nature.com, https://doi.org/10.1038/s43856-022-00165-w.

26. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002 Jun 1;16:321-57. https://doi.org/10.1613/jair.953

27. Islam SR, Eberle W, Ghafoor SK, Ahmed M. Explainable artificial intelligence approaches: A survey. arXiv preprint arXiv:2101.09429. 2021 Jan 23. https://doi.org/10.48550/arXiv.2101.09429

28. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 4765–4774.

29. Borgese M, Joyce C, Anderson EE, Churpek MM, Afshar M. 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 2021 (Vol. 2021, p. 247). American Medical Informatics Association.

30. Allen A, Mataraso S, Siefkas A, Burdick H, Braden G, Dellinger RP, et al. A racially unbiased, machine learning approach to prediction of mortality: algorithm development study. JMIR Public Health Surveill 2020 Oct 22;6(4):e22400 https://doi.org/10.2196%2F22400

31. Tison, G.H., et al., Predicting Incident Heart Failure in Women With Machine Learning: The Women's Health Initiative Cohort. The Canadian journal of cardiology, 2021. 37(11): p. 1708-1714. https://doi.org/10.1016/j.cjca.2021.08.006

32. Pombo G, Gray R, Cardoso MJ, Ourselin S, Rees G, Ashburner J, Nachev P. Equitable modelling of brain imaging by counterfactual augmentation with morphologically constrained 3d deep generative models. Medical Image Analysis. 2023 Feb 1;84:102723. https://doi.org/10.48550/arXiv.2111.14923

33. McCradden MD, Joshi S, Mazwi M, Anderson JA. Ethical limitations of algorithmic fairness solutions in health care machine learning. The Lancet Digital Health. 2020 May 1;2(5):e221–3. https://doi.org/10.1016/S2589-7500(20)30065-0

34. Safer, J.D., et al., Barriers to healthcare for transgender individuals. Current opinion in endocrinology, diabetes, and obesity, 2016. 23(2): p. 168-171. https://doi.org/10.1097%2FMED.0000000000000227

35. Rutherford L, Stark A, Ablona A, Klassen BJ, Higgins R, Jacobsen H, Draenos CJ, Card KG, Lachowsky NJ. 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. PLoS One. 2021 Feb 11;16(2):e0246525. https://doi.org/10.1371/journal.pone.0246525

36. Beckwith, N., et al., Psychiatric Epidemiology of Transgender and Nonbinary Adult Patients at an Urban Health Center. LGBT health, 2019. 6(2): p. 51-61. https://doi.org/10.1089/lgbt.2018.0136

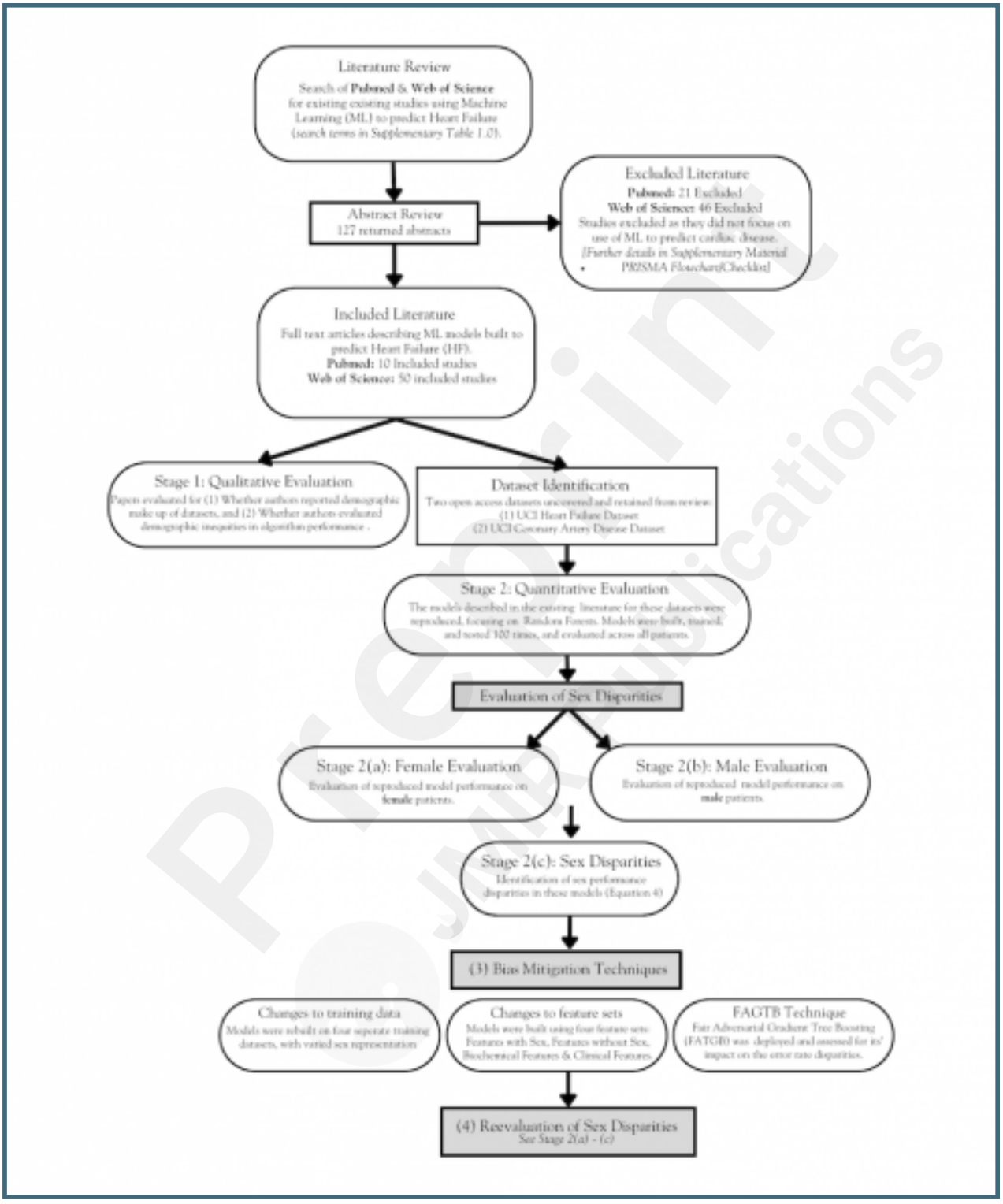
37. Vejjajiva A, Teasdale GM. Serum creatine kinase and physical exercise. British Medical Journal. 1965 Jun 6;1(5451):1653. https://doi.org/10.1136%2Fbmj.1.5451.1653

38. Carruthers R, Straw I, Ruffle JK, Herron D, Nelson A, Bzdok D, Fernandez-Reyes D, Rees G, Nachev P. Representational ethical model calibration. NPJ Digital Medicine. 2022 Nov 4;5(1):170. https://doi.org/10.1038/s41746-022-00716-4

Supplementary Files

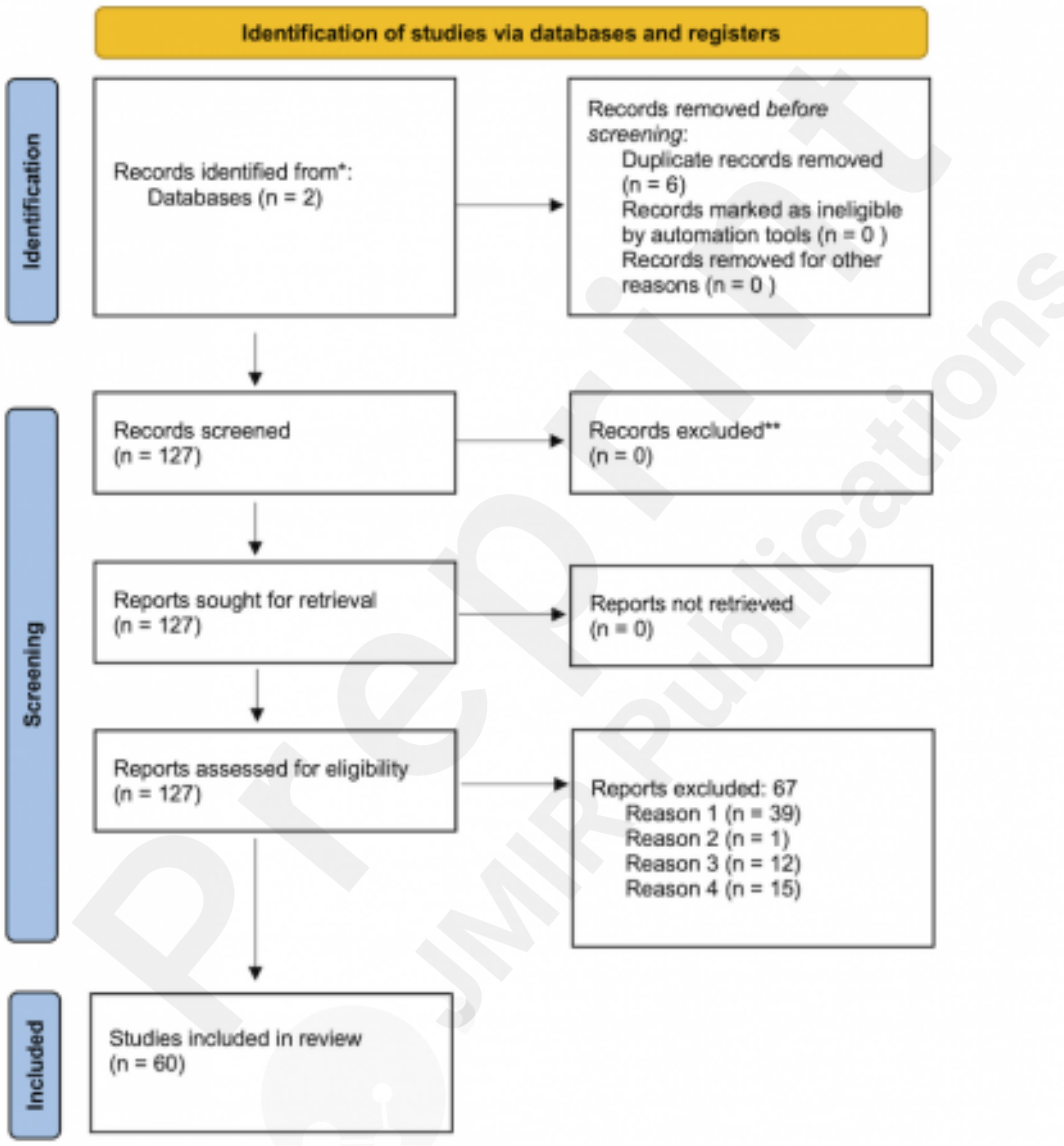
Figures

A flowchart detailing the steps of our methodology, including (1) the initial literature search and qualitative evaluation of identified studies, plus (2) the identification of datasets and interrogation of algorithms for demographic bias.



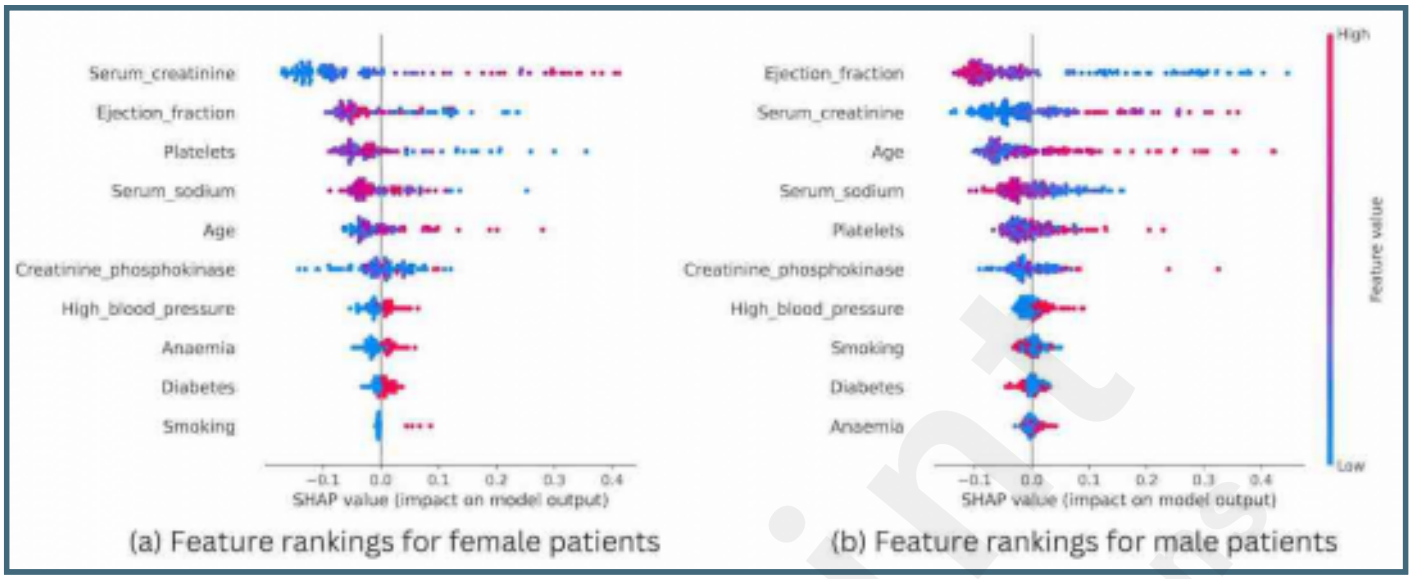
PRISMA 2020 flow diagram for new systematic reviews which included searches of databases and registers only (PRISMA templated obtained from PRISMA at https://prisma-statement.org/prismastatement/flowdiagram.aspx).

Figure 2: PRISMA 2020 flow diagram for new systematic reviews which included searches of databases and registers only (PRISMA templated obtained from PRISMA at https://prisma-statement.org/prismastatement/flowdiagram.aspx)

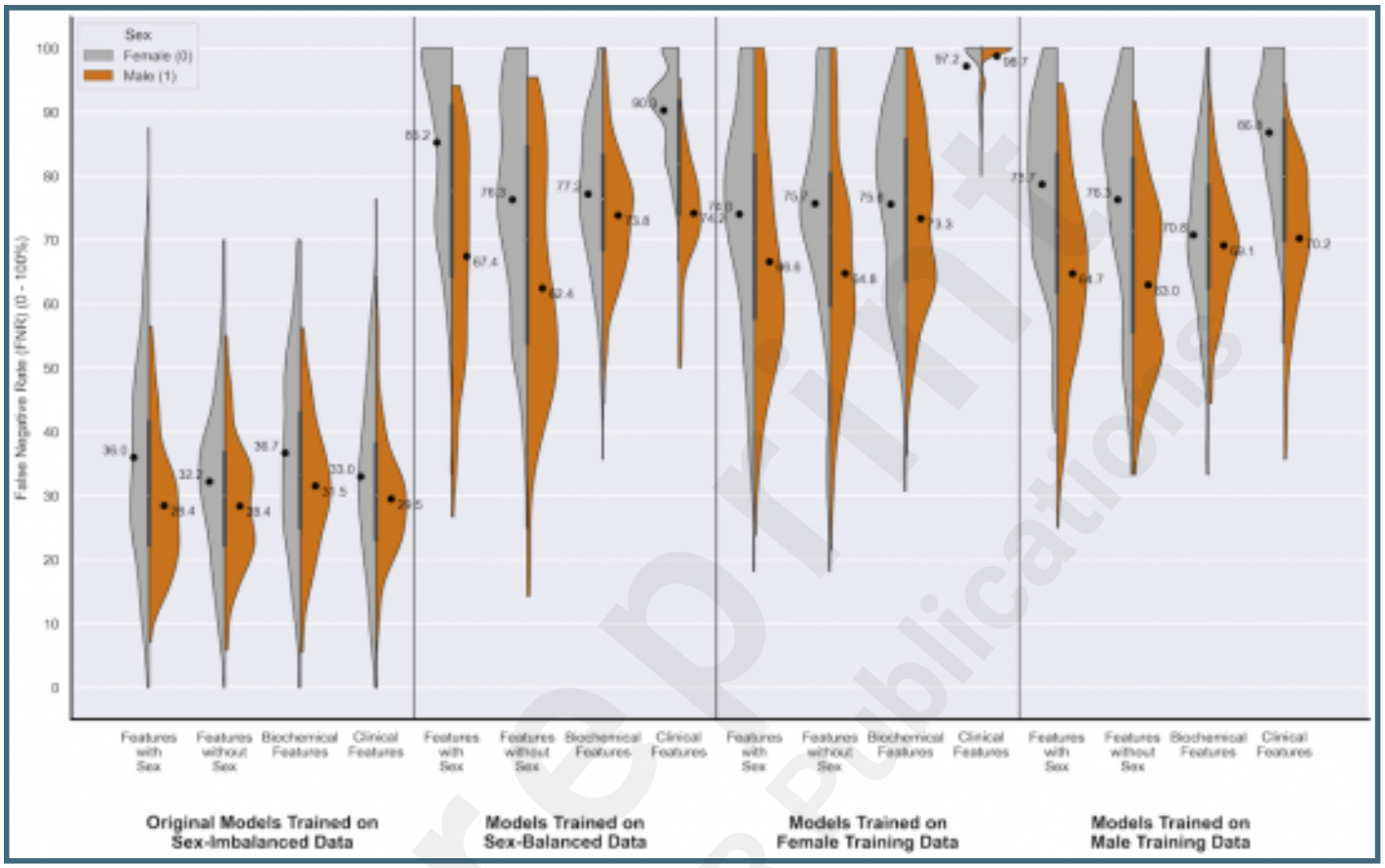


*Nb. *Reasons for exclusion: Reason 1: The study did not focus on biochemical data or laboratory tests, instead utilising different modalities (e.g., visual data from radiological scans); Reason 2: The study did not use machine learning techniques (e.g. it used traditional statistical methods). ; Reason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

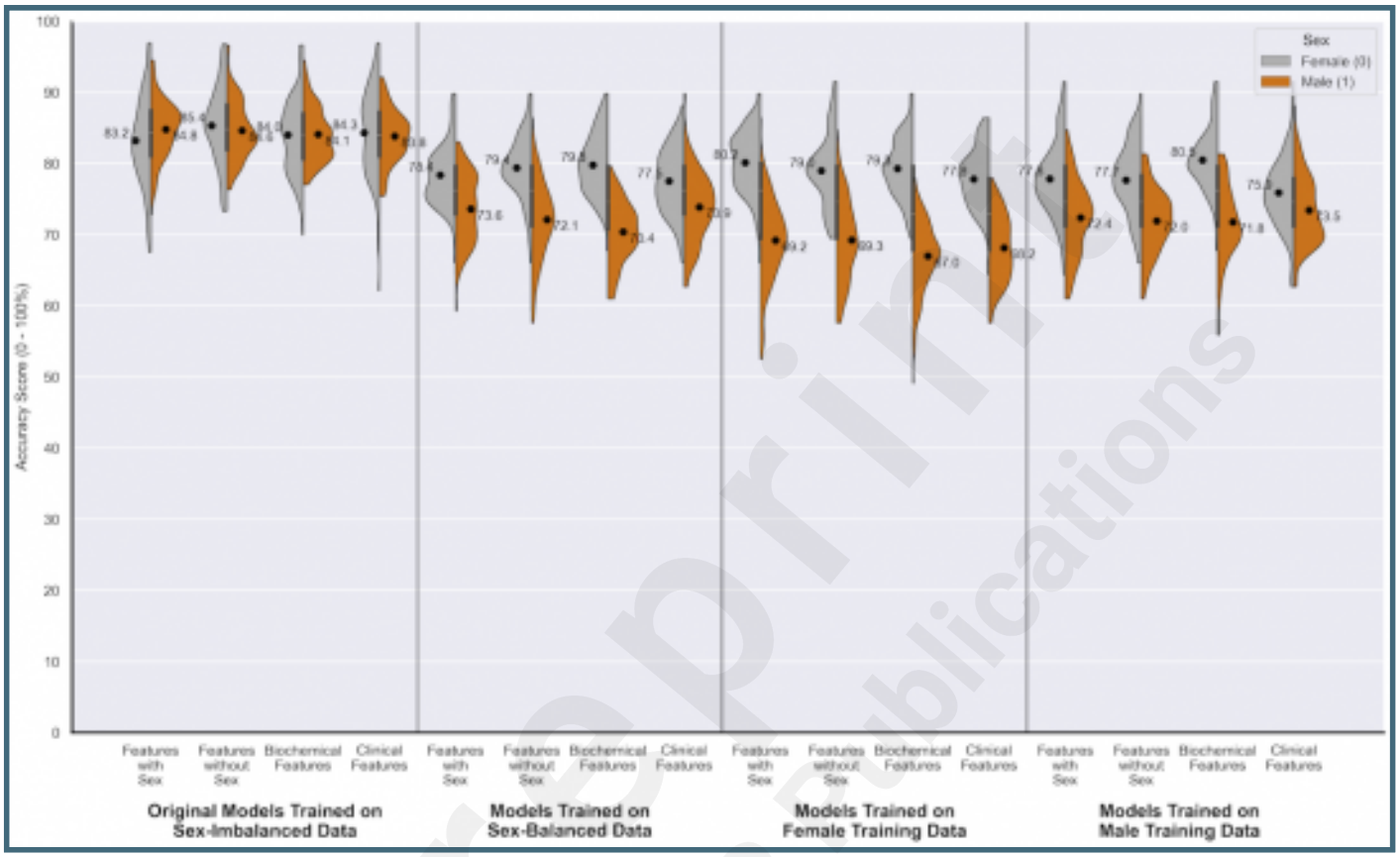
Comparison of feature rankings for male and female patients, ordered by SHAP values.



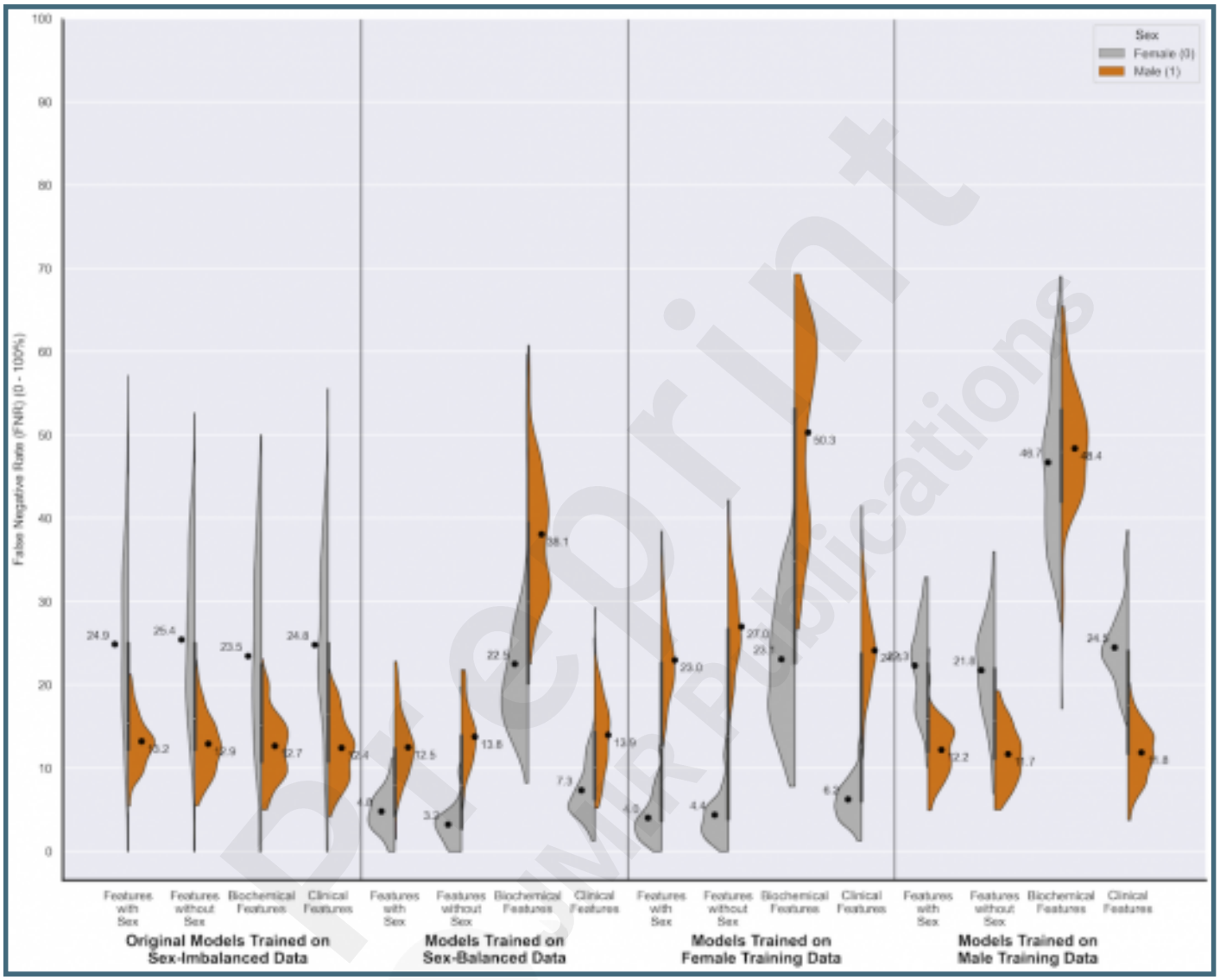
Dataset 1 (Heart Failure): A series of violin plots showing the sex stratified performance (False Negative Rate [0-100%]) of the Random Forests trained across the four feature sets, on the different variations in training data. The plots show male (orange) and female (grey) FNR alongside each other, 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).



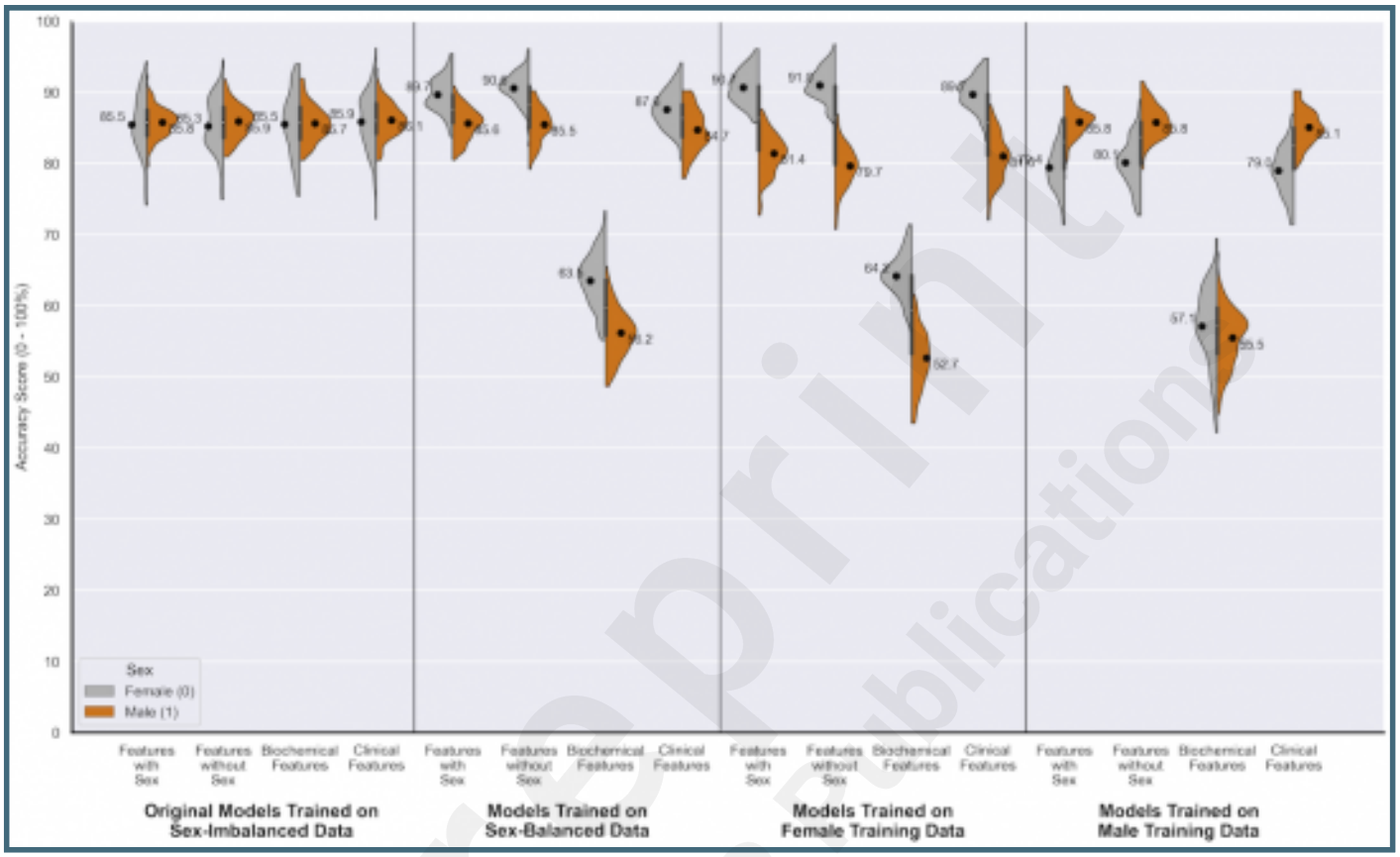
Dataset 1 (Heart Failure): A series of violin plots showing the sex stratified performance (Accuracy [0-100%]) of the Random Forests trained across the four feature sets, on the different variations in training data. The plots show male (orange) and female (grey) Accuracy alongside each other, in groups 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).



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 & Male). The Feature Set used is indicated within each training data group (Features with Sex, Features Without Sex, Biochemical Features & Clinical Features).



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 & Male). The Feature Set used is indicated within each training data group (Features with Sex, Features Without Sex, Biochemical Features & Clinical Features).



Multimedia Appendixes

Supplementary Material with supplementary tables.
URL: http://asset.jmir.pub/assets/e6c7bd8925d5d07e8d6059454f848a77.pdf