Machine learning is more accurate and biased than risk scoring tools in the prediction of postoperative atrial fibrillation after cardiac surgery

Joyce C Ho *[1], Shalmali Joshi †[2], Eduardo Valverde ‡[3], Kathryn Wood §[4], Kendra Grubb ¶[5], Miguel Leal ⊠[5], and Vicki Stover Hertzberg **[4]

[1]Department of Computer Science, Emory University

[2]Department of Biomedical Informatics, Columbia University

[3]Department of Computer Science, Georgia Institute of Technology

[4]School of Nursing, Emory University

[5]School of Medicine, Emory University

Abstract

Incidence of postoperative atrial fibrillation (POAF) after cardiac surgery remains high and is associated with adverse patient outcomes. Risk scoring tools have been developed to predict POAF, yet discrimination performance remains moderate. Machine learning (ML) models can achieve better performance but may exhibit performance heterogeneity across race and sex subpopulations. We evaluate 8 risk scoring tools and 6 ML models on a heterogeneous cohort derived from electronic health records. Our results suggest that ML models achieve higher discrimination yet are less fair, especially with respect to race. Our findings highlight the need for building accurate and fair ML models to facilitate consistent and equitable assessment of POAF risk.

Keywords: POAF, ML, Fairness

*joyce.c.ho@emory.edu
†sj3261@cumc.columbia.edu
‡evalverde3@gatech.edu
§kathryn.wood@emory.edu
¶kendra.janel.grubb@emory.edu
⊠miguel.a.leal@emory.edu
**vhertzb@emory.edu

1 Introduction

Although there have been advancements in cardiac surgery techniques, the incidence of postoperative atrial fibrillation (POAF) following cardiac surgery has not decreased significantly and still ranges from 15% to 50% [1, 2]. Unfortunately, there are short- and long-term adverse outcomes associated with POAF including morbidity, mortality, and longer, more expensive hospitalizations [3, 4, 5, 6, 7]. Early identification of patients at risk for developing POAF has long been desired to guide preventative and treatment strategies. To this end, more than a dozen POAF risk scoring algorithms have been introduced encompassing a variety of risk factors including patient demographics and clinical characteristics as well as surgical characteristics. Yet a recent review found only patient age had no conflicting evidence across existing studies [8]. Moreover, these scoring systems offer moderate discrimination with area under the receiver operating characteristic curve (AUROC) scores ranging between 0.55 and 0.87 and may not generalize broadly as the performance is assessed on relatively small, homogeneous patient populations.

Machine learning (ML) has been proposed as an alternative to achieve better predictive performance [9]. A recent scoping review found that support vector machines (SVM), gradient boosting machines (GBM), and random forests (RF) using clinical characteristics can predict POAF risk more accurately than existing risk scores with promising specificity, sensitivity, and AUROC scores [9]. Three existing works compared multiple ML algorithms with Lu et al [10] and Parise et al. [11] concluding that SVM achieved the best performance while GBM performed the best in Karri et al. [12]. Despite their promise, indiscriminate application of ML models can exacerbate existing health disparities if they are not trained on a representative sample [13].

Unfortunately, significant race and sex disparities exist as the number of patients undergoing cardiac surgery procedures and the outcomes for these patients [14]. Incidence of POAF after coronary artery bypass graft (CABG) surgery is higher in White patients [15]. It has also been suggested males are more likely to experience POAF following CABG [16, 17] although there exists conflicting evidence [18]. However, only 2 studies utilizing ML report the ethnicity composition of the underlying dataset and both studies assessed the performance in populations with less than 4% Black patients [12, 19]. Thus, a crucial unanswered question is whether the better performance of ML

algorithms may exacerbate existing disparities.

The objective of this study is to assess both the predictive performance and fairness of existing POAF risk scoring tools with popular ML algorithms on a heterogeneous population, with more than 20% of the patients identifying as Black. We assess the fairness of the predictive models in both race and sex subpopulations. We also restrict our evaluation to common structured data found within electronic health records (EHRs) as such algorithms can provide quicker (and hopefully more accurate) management strategies [9].

2 Methods

2.1 Data Source

Our study was conducted using de-identified EHRs from the Emory Healthcare clinical data warehouse. Secondary data analysis was approved by the Emory University Institutional Review Board. Adult patients who received cardiac surgery in the outpatient or inpatient setting between January 1, 2013 and December 31, 2017 were included. Cardiac surgery was defined using the Current Procedural Terminology (CPT) codes as either venous grafting for CABG or surgical procedures on cardiac valves (see Supplemental Material for full list). For security purposes, patient identifiers were omitted and certain records were excluded based on the date shifting logic. Patients who had a prior history of atrial fibrillation (AF), defined by the International Classification of Diseases codes of '427.31' for the 9th revision (ICD-9) or 'I48.XX' for the 10th revision (ICD-10) were excluded from the study. We used the presence of the AF ICD-9 or ICD-10 code following the cardiac surgery procedure date to identify cases of POAF. The value of 0 was assigned to patients that did not experience POAF and had at least 1 encounter after the cardiac surgery.

All the clinical variables including age, sex, race, height, weight, and blood pressure were extracted from the EHR. We used the most recent value collected within the 1 year prior to the cardiac surgery date. The presence of clinical comorbidities for the risk scoring systems was determined using diagnostic (ICD-9 or ICD-10), procedural (CPT), and medication codes. For the ML clinical variables, we grouped the diagnostic codes using the single-level Clinical Classifications Software (CCS) system and medication codes using Anatomical Therapeutic Chemical (ATC) Level 3 classification

codes.

2.2 Risk Scores

We evaluated POAF risk scoring systems and incident AF risk scoring systems that utilize commonly collected measures in structured EHR data. Although a recent review identified at least 12 distinct POAF scoring systems, [8] several used echocardiographic measurements such as left atrial dilation, left atrial diameter, and left ventricular ejection fraction which are often captured in unstructured text and are not easily accessible broadly. As such, we focused on the following 8 risk scores: (1) $CHADS_2$,[20] (2) CHA_2DS_2-VASc,[20] (3) HATCH, [21] (4) COM-AF, [22] (5) C_2HEST, [23] (6) mC_2HEST, [24] (7) AFRI [25], and (8) CHARGE-AF [26]. The Python code for the scoring systems is openly available as a GitHub repository (`https://github.com/joyceho/afib`). The predictor variables for each model can be found in Supplemental Table 3.

2.3 Machine Learning Models

Six commonly used ML algorithms were explored that have been previously benchmarked from previous existing studies: (1) logistic regression (LR), (2) decision tree (DT), (3) SVM, (4) RF, (5) GBM, and (6) multi-layer perceptron (MLP). The ML models were constructed using the popular Python open-source software library, `scikit-learn` version 1.5.0 [27]. The ML models were supplied with age, race, gender, CCS, and ATC codes. CCS and ATC codes that were not present in at least 20% of the patients were excluded. A total of 71 variables were supplied as input to the models. Exhaustive hyperparameter optimization was performed using 5-fold cross validation on the training dataset (see Supplemental Table 5 for the parameter search space for each model). The optimal hyperparameter for each model was identified using AUROC.

2.4 Training and Evaluation

We used stratified Monte Carlo cross validation to randomly split the data into 70-30% train-test. This process was repeated 10 times to assess model performance. Data imputation was required for age, height, weight, and blood pressure. Mean imputation from the training data was used. Predictive performance was measured using AUROC and area under the precision recall curve

(AUPRC) on the test set. AUROC and AUPRC were calculated using the `scikit-learn` package on the test data. We also assessed the fairness of the models on the sex (female and male) and race (White and Black/Other) subgroups. Two popular group fairness metrics were used, demographic parity ratio (DPR) and equalized odds ratio (EQR). DPR measures whether the predictive proportion of POAF across the subgroups are equal (i.e., the prediction risk should be independent of sex or race). EQR ensures the true positive rate and false positive rate of predictions are the same across the subgroups. Both DPR and EQR range from 0 to 1 with 1 indicates fairness across the subgroups. DPR and EQR were computed using the `fairlearn` package version 0.10.0.[28] All analyses were performed using Python version 3.9.7.

3 Results

3.1 Patient Characteristics

Out of the final study population of 4961 patients, 1953 (39.4%) experienced POAF following cardiac surgery with an average onset of xxx. Baseline characteristics of the overall study population and the 2 outcome groups (no POAF and POAF) are reported in Table 1. The incidence of POAF experienced in males (40.1%) and Whites (42.5%) was statistically higher than in females (37.9%) and Blacks (32.3%), respectively.

3.2 Performance Comparison

Table 2 summarizes the discrimination and fairness performance of the 14 models (8 risk scoring algorithms and 6 ML models). For each performance metric, the value represents the mean across the 10 test splits. Statistical significance in discrimination performance between any 2 models was assessed using a one-tailed paired t-test that the difference is greater than 0 (i.e., one model consistently outperforms the other).

The ML model that achieved the best discrimination was RF with AUROC and AUPRC of 0.671 and 0.558, respectively. Only GBM yielded a p-value above 0.001 for AUROC (0.03) and AUPRC (0.06) during the one-tailed paired t-test between RF and the other 5 ML models. Among the risk scoring systems, CHARGE-AF achieved the best performance with AUROC and AUPRC of 0.585 and 0.449,

Table 1: Baseline characteristics in patients with and without POAF.

	Missing	Overall (n=4961)	No POAF (n=3008)	POAF (n=1953)	P-Value
Age, mean (SD)	19	64.6 (12.4)	63.3 (13.1)	66.5 (10.9)	<0.001
Male, n (%)	0	3300 (66.5)	1976 (65.7)	1324 (67.8)	0.133
Ethnicity, n (%)	0				<0.001
White		3426 (69.1)	1970 (65.5)	1456 (74.6)	
Black		1085 (21.9)	735 (24.4)	350 (17.9)	
Other		450 (9.1)	303 (10.1)	147 (7.5)	
CHF, n (%)	0	2066 (41.6)	1226 (40.8)	840 (43.0)	0.123
HTN, n (%)	0	2999 (60.5)	1888 (62.8)	1111 (56.9)	<0.001
DM, n (%)	0	1382 (27.9)	864 (28.7)	518 (26.5)	0.098
Stroke, n (%)	0	959 (19.3)	560 (18.6)	399 (20.4)	0.123
Vascular Disease, n (%)	0	2004 (40.4)	1253 (41.7)	751 (38.5)	0.027
COPD, n (%)	0	509 (10.3)	307 (10.2)	202 (10.3)	0.914
CAD, n (%)	0	4041 (81.5)	2503 (83.2)	1538 (78.8)	<0.001
SHF, n (%)	0	379 (7.6)	228 (7.6)	151 (7.7)	0.887
Hyperthyroid, n (%)	0	18 (0.4)	14 (0.5)	4 (0.2)	0.211
Valvular, n (%)	0	3099 (62.5)	1783 (59.3)	1316 (67.4)	<0.001
PVD, n (%)	0	1265 (25.5)	723 (24.0)	542 (27.8)	0.004
Obesity, n (%)	0	420 (8.5)	255 (8.5)	165 (8.4)	1.000
Height (m), mean (SD)	65	67.7 (4.6)	67.4 (4.6)	68.0 (4.5)	<0.001
Weight (kg) , mean (SD)	41	192.0 (45.8)	189.9 (44.5)	195.3 (47.5)	<0.001
SBP, mean (SD)	40	135.3 (22.3)	135.1 (22.2)	135.5 (22.4)	0.559
DBP, mean (SD)	41	74.4 (12.9)	74.8 (12.9)	73.8 (12.8)	0.015

Table 2: Average discrimination and fairness performance of the prediction models across 10 Monte Carlo cross-validation splits.

			Race		Sex	
Model	AUROC	AUPRC	DPR	EQR	DPR	EQR
RF	0.671	0.558	0.824	0.816	0.967	0.946
GBM	0.666	0.553	0.837	0.835	0.975	0.948
SVM	0.658	0.540	0.828	0.827	0.976	0.958
LR	0.645	0.521	0.659	0.618	0.949	0.903
DT	0.630	0.534	0.970	0.950	0.962	0.942
MLP	0.622	0.494	0.789	0.772	0.953	0.923
CHARGE-AF	0.578	0.443	0.896	0.895	0.941	0.926
AFRI	0.568	0.474	1.000	1.000	1.000	1.000
COM-AF	0.550	0.416	0.993	0.980	0.884	0.858
HAVOC	0.527	0.401	0.999	0.999	0.999	0.998
HATCH	0.527	0.400	1.000	1.000	1.000	1.000
mC_2HEST	0.526	0.408	1.000	1.000	1.000	1.000
CHA_2DS_2-VASc	0.525	0.408	1.000	1.000	1.000	1.000
$CHADS_2$	0.517	0.401	1.000	1.000	1.000	1.000
C_2HEST	0.509	0.401	1.000	1.000	1.000	1.000

respectively. Notably, all 6 ML models outperformed CHARGE-AF and the other risk scoring tools at statistically significant levels (p-value < 0.001) for both discrimination metrics.

The risk scoring systems generally yielded the best group fairness concerning race as all but CHARGE-AF, COM-AF, and HAVOC resulted in both DPR and EQR of 1. Notably, CHARGE-AF is the only risk scoring system incorporating race as a variable (see Supplemental Table 3), yet achieves the worst group fairness. In contrast, all the ML models except DT perform worse in terms of DPR and EQR to CHARGE-AF (DPR = 0.896 and EQR = 0.895).

A similar group fairness trend is observed for sex in terms of risk scoring systems again as CHARGE-AF, COM-AF, and HAVOC do not yield DPR and EQR of 1. However, the ML models achieve slightly better performance than CHARGE-AF and COM-AF in terms of DPR and EQR for race. Surprisingly, COM-AF which incorporates sex as a variable yields the worst DPR and EQR performance with values of 0.884 and 0.858, respectively.

4 Discussion

In this study, we evaluated the performance of 6 ML models and 8 risk scoring algorithms to predict POAF. We demonstrated that RF outperformed the other ML methods and all of the risk scores considered in terms of AUROC and AUPRC. Furthermore, there were statistical differences between the discrimination performance of RF and the other models except for the GBM algorithm. The AUROC and AUPRC of these 14 models were all under 0.671 and 0.558 respectively. Compared to the existing ML studies, the discrimination performance is lower as they achieved an AUROC of at least 0.72. However, these models used indicators related to cardiac surgery which are not commonly available in the structured EHR data.

In contrast to the discrimination performance, six of the risk scores outperformed all of the ML methods and three of the risk scores with respect to metrics of fairness. In fact, the results indicate the ML models exacerbated race and sex differentials when used for POAF prediction, which is consistent with existing evidence for other outcomes such as cardiovascular risk [29], dermatology [30], and population health [31]. Thus, better discrimination performance may not always be desired as it might exacerbate existing race and sex disparities. This suggests further investigation is

necessary to holistically assess the efficacy of ML algorithms for POAF prognostication in real clinical contexts, [32] and whether bias mitigation mechanisms should be adopted to minimize disparities in outcomes and interventions.

Abbreviations

POAF: post-operative atrial fibrillation; **AUROC**: area under the receiver operating characteristic curve; **ML**: machine learning; **SVM**: support vector machines; **GBM**: gradient boosting machines; **RF**: random forests; **CABG**: coronary artery bypass graft

5 Supplemental Information

5.1 Risk scoring algorithms

The variables used for each of the 8 risk scores are summarized in Table 3. While other risk scoring systems have been developed using POAF as an event of interest[8], they are not benchmarked in our study as they use variables such as left atrial dilatation, left atrial diameter, left ventricular ejection fraction, and length of stenosis. The scores in Table 3 can be computed from demographic information, diagnosis tables (ICD-9/ICD-10), vital signs commonly collected, and medication tables.

Table 3: Risk scoring systems, the original event of interests, and their associated variables. The events of interest are thromboembolic events in patients with atrial fibrillation (VTE), incident AF (AF), and POAF.

Score	Event	Age	Gender	Race	CHF	HTN	DM	Stroke	Vascular	COPD	CAD	SHF	Hyperthyroid	PVD	MI	Height	Weight	SBP	DBP	Smoking	HTN Meds
$CHADS_2$ [20]	VTE	x			x	x	x	x													
CHA_2DS_2-VASc [20]	VTE	x	x		x	x	x	x	x												
HATCH [21]	AF	x			x	x				x											
COM-AF [22]	POAF	x	x		x	x	x	x													
C_2HEST [23]	AF	x				x				x	x	x	x								
mC_2HEST [24]	AF	x				x				x	x	x	x								
AFRI [25]	POAF	x												x		x	x				
CHARGE-AF [26]	AF	x		x	x	x	x								x	x	x	x	x	x	x

5.2 ML for POAF

A recent scoping review identified 7 papers that used ML for predicting POAF after cardiac surgery.[9] Of the 7 studies, 3 relied on electrocardiogram data while the remaining 4 used clinical documentation, administrative data, or Holter monitoring. The sample size, ethnicity composition, and model performance of the 4 ML studies using administrative data are summarized in Table 4. As can be seen, none of the patient populations contains more than 3.4% Black.

Table 5 summarizes the hyperparameter search space for each of the ML models. For each train-test split and ML model, GridSearchCV in scikit-learn was performed using 5-folds on the train

Table 4: Previous ML studies for POAF prediction.

Authors	**Dataset**	**n**	**White-Black (%)**	**Best ML (AUROC)**
Magee et al. [19]	Cardiopulmonary Research Science and Technology Institute	19620	90.7-3.2	LR (0.72)
Karri et al. [12]	MIMIC-III	6040	74.0-3.4	GBM (0.74)
Lu et al. [10]	Second Affiliated Hospital of Zhejiang University School of Medicine	1400	Unknown	SVM (0.78)
Parise et al. [11]	Maastricht University Medical Center+	394	Unknown	SVM (0.95)

Table 5: Hyperparameter search space for the different ML models.

Model	**Parameter Space**
LR	C: [0.001, 0.01, 0.1, 1]
DT	Criterion: [gini, entropy]
	Max depth: [3, 4, 5, 6, 7, 8]
RF	Max depth: [10, 15, 20]
	Min leaf samples: [5, 10]
	Num estimators: [50, 100, 200, 300]
GBM	Max depth: [5, 8, 10]
	Learning rate: [0.1, 0.01]
	Min leaf samples: [5, 10]
	Num estimators: [50, 100, 150, 200]
SVM	C: [0.1, 1, 10, 100]
	Kernel: [linear, rbf]
MLP	Hidden layers: [(15, 10), (15,), (25,), (50,), (25, 25,)]

split to find the optimal hyperparameter values. The ML model is then retrained using the optimal hyperparameter values and the performance is evaluated on the test set.

References

[1] Giovanni Filardo, Ralph J Damiano, Gorav Ailawadi, Vinod H Thourani, Benjamin D Pollock, Danielle M Sass, Teresa K Phan, Hoa Nguyen, and Briget Da Graca. Epidemiology of new-onset atrial fibrillation following coronary artery bypass graft surgery. *Heart*, 104(12):985–992, 2018.

[2] Orlando R Suero, Ahmed K Ali, Lauren R Barron, Matthew W Segar, Marc R Moon, and Subhasis Chatterjee. Postoperative atrial fibrillation (poaf) after cardiac surgery: clinical practice review. *Journal of Thoracic Disease*, 16(2), 2024.

[3] Rachel Eikelboom, Rohan Sanjanwala, Me-Linh Le, Michael H Yamashita, and Rakesh C Arora. Postoperative atrial fibrillation after cardiac surgery: a systematic review and meta-analysis. *The Annals of Thoracic Surgery*, 111(2):544–554, 2021.

[4] Ben O'Brien, Peter S. Burrage, Jennie Yee Ngai, Jordan M. Prutkin, Chuan-Chin Huang, Xinling Xu, Sanders H. Chae, Bruce A. Bollen, Jonathan P. Piccini, Nanette M. Schwann, Aman Mahajan, Marc Ruel, Simon C. Body, Frank W. Sellke, Joseph Mathew, and J. Daniel Muehlschlegel. Society of cardiovascular anesthesiologists/european association of cardiothoracic anaesthetists practice advisory for the management of perioperative atrial fibrillation in patients undergoing cardiac surgery. *Journal of Cardiothoracic and Vascular Anesthesia*, 33(1):12–26, 2019.

[5] Ahmed AlTurki, Mariam Marafi, Riccardo Proietti, Daniela Cardinale, Robert Blackwell, Paul Dorian, Amal Bessissow, Lucy Vieira, Isabelle Greiss, Vidal Essebag, Jeff S. Healey, and Thao Huynh. Major adverse cardiovascular events associated with postoperative atrial fibrillation after noncardiac surgery: a systematic review and meta-analysis. *Circulation: Arrhythmia and Electrophysiology*, 13(1):e007437, 2020.

[6] Peter S Burrage, Ying H Low, Niall G Campbell, and Ben O'Brien. New-onset atrial fibrillation in adult patients after cardiac surgery. *Current anesthesiology reports*, 9:174–193, 2019.

[7] Michael K. Wang, Pascal B. Meyre, Rachel Heo, P.J. Devereaux, Lauren Birchenough, Richard Whitlock, William F. McIntyre, Yu Chiao Peter Chen, Muhammad Zain Ali, Fausto Biancari, Jawad Haider Butt, Jeff S. Healey, Emilie P. Belley-Côté, Andre Lamy, and David Conen. Short-

term and long-term risk of stroke in patients with perioperative atrial fibrillation after cardiac surgery: systematic review and meta-analysis. *CJC open*, 4(1):85–96, 2022.

[8] Hugh Fleet, David Pilcher, Rinaldo Bellomo, and Tim G Coulson. Predicting atrial fibrillation after cardiac surgery: a scoping review of associated factors and systematic review of existing prediction models. *Perfusion*, 38(1):92–108, 2023.

[9] Adham H El-Sherbini, Aryan Shah, Richard Cheng, Abdelrahman Elsebaie, Ahmed A Harby, Damian Redfearn, and Mohammad El-Diasty. Machine learning for predicting postoperative atrial fibrillation after cardiac surgery: A scoping review of current literature. *The American Journal of Cardiology*, 209:66–75, 2023.

[10] Yufan Lu, Qingjuan Chen, Hu Zhang, Meijiao Huang, Yu Yao, Yue Ming, Min Yan, Yunxian Yu, and Lina Yu. Machine learning models of postoperative atrial fibrillation prediction after cardiac surgery. *Journal of Cardiothoracic and Vascular Anesthesia*, 37(3):360–366, 2023.

[11] Orlando Parise, Gianmarco Parise, Akshayaa Vaidyanathan, Mariaelena Occhipinti, Ali Gharaviri, Cecilia Tetta, Elham Bidar, Bart Maesen, Jos G. Maessen, Mark La Meir, and Sandro Gelsomino. Machine learning to identify patients at risk of developing new-onset atrial fibrillation after coronary artery bypass. *Journal of Cardiovascular Development and Disease*, 10(2):82, 2023.

[12] Roshan Karri, Andrew Kawai, Yoke Jia Thong, Dhruvesh M Ramson, Luke A Perry, Reny Segal, Julian A Smith, and Jahan C Penny-Dimri. Machine learning outperforms existing clinical scoring tools in the prediction of postoperative atrial fibrillation during intensive care unit admission after cardiac surgery. *Heart, Lung and Circulation*, 30(12):1929–1937, 2021.

[13] Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. Ethical machine learning in healthcare. *Annual review of biomedical data science*, 4:123–144, 2021.

[14] Mohamad Alkhouli, Fahad Alqahtani, David R Holmes, and Chalak Berzingi. Racial disparities in the utilization and outcomes of structural heart disease interventions in the united states. *Journal of the American Heart Association*, 8(15):e012125, 2019.

[15] David W Yaffee, Raymond G McKay, Jeffrey Mather, Scott Vella Sorensen, Andrew Kehm, Sean McMahon, Trevor Sutton, and Sabet W Hashim. Racial disparities in atrial fibrillation after coronary artery bypass: Impact of left atrial volume. *Annals of Thoracic Surgery Short Reports*, 1(4):631–634, 2023.

[16] Mariana Fragão-Marques, Jennifer Mancio, João Oliveira, Inês Falcão-Pires, and Adelino Leite-Moreira. Gender differences in predictors and long-term mortality of new-onset postoperative atrial fibrillation following isolated aortic valve replacement surgery. *Annals of Thoracic and Cardiovascular Surgery*, 26(6):342–351, 2020.

[17] Giovanni Filardo, Gorav Ailawadi, Benjamin D Pollock, Briget da Graca, Teresa K Phan, Vinod Thourani, and Ralph J Damiano Jr. Postoperative atrial fibrillation: sex-specific characteristics and effect on survival. *The Journal of thoracic and cardiovascular surgery*, 159(4):1419–1425, 2020.

[18] Giovanni Filardo, Gorav Ailawadi, Benjamin D Pollock, Briget Da Graca, Danielle M Sass, Teresa K Phan, Debbie E Montenegro, Vinod Thourani, and Ralph Damiano. Sex differences in the epidemiology of new-onset in-hospital post–coronary artery bypass graft surgery atrial fibrillation: a large multicenter study. *Circulation: Cardiovascular Quality and Outcomes*, 9(6):723–730, 2016.

[19] Mitchell J Magee, Morley A Herbert, Todd M Dewey, James R Edgerton, William H Ryan, Syma Prince, and Michael J Mack. Atrial fibrillation after coronary artery bypass grafting surgery: development of a predictive risk algorithm. *The Annals of thoracic surgery*, 83(5):1707–1712, 2007.

[20] Gregory YH Lip, Robby Nieuwlaat, Ron Pisters, Deirdre A Lane, and Harry JGM Crijns. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the Euro Heart Survey on atrial fibrillation. *Chest*, 137(2):263–272, 2010.

[21] Cees B De Vos, Ron Pisters, Robby Nieuwlaat, Martin H Prins, Robert G Tieleman, Robert-Jan S Coelen, Antonius C van den Heijkant, Maurits A Allessie, and Harry JGM Crijns. Progression

from paroxysmal to persistent atrial fibrillation: clinical correlates and prognosis. *Journal of the American College of Cardiology*, 55(8):725–731, 2010.

[22] Lucrecia M Burgos, Andreína Gil Ramírez, Leonardo Seoane, Juan F Furmento, Juan P Costabel, Mirta Diez, and Daniel Navia. New combined risk score to predict atrial fibrillation after cardiac surgery: Com-af. *Annals of cardiac anaesthesia*, 24(4):458–463, 2021.

[23] Yan-Guang Li, Daniele Pastori, Alessio Farcomeni, Pil-Sung Yang, Eunsun Jang, Boyoung Joung, Yu-Tang Wang, Yu-Tao Guo, and Gregory YH Lip. A simple clinical risk score (c2hest) for predicting incident atrial fibrillation in asian subjects: derivation in 471,446 chinese subjects, with internal validation and external application in 451,199 korean subjects. *Chest*, 155(3):510–518, 2019.

[24] Yan-Guang Li, Jin Bai, Gongbu Zhou, Juan Li, Yi Wei, Lijie Sun, Lingyun Zu, and Shuwang Liu. Refining age stratum of the c2hest score for predicting incident atrial fibrillation in a hospital-based chinese population. *European journal of internal medicine*, 90:37–42, 2021.

[25] Mikhael F El-Chami, Patrik D Kilgo, K Miriam Elfstrom, Michael Halkos, Vinod Thourani, Omar M Lattouf, David B Delurgio, Robert A Guyton, Angel R Leon, and John D Puskas. Prediction of new onset atrial fibrillation after cardiac revascularization surgery. *The American journal of cardiology*, 110(5):649–654, 2012.

[26] Alvaro Alonso, Bouwe P Krijthe, Thor Aspelund, Katherine A Stepas, Michael J Pencina, Carlee B Moser, Moritz F Sinner, Nona Sotoodehnia, João D Fontes, A Cecile JW Janssens, Richard A Kronmal, Jared W Magnani, Jacqueline C Witteman, Alanna M Chamberlain, Steven A Lubitz, Renate B Schnabel, Sunil K Agarwal, David D McManus, Patrick T Ellinor, Martin G Larson, Gregory L Burke, Lenore J Launer, Albert Hofman, Daniel Levy, John S Gottdiener, Stefan Kääb, David Couper, Tamara B Harris, Elsayed Z Soliman, Bruno H C Stricker, Vilmundur Gudnason, Susan R Heckbert, and Emelia J Benjamin. Simple risk model predicts incidence of atrial fibrillation in a racially and geographically diverse population: the charge-af consortium. *Journal of the American Heart Association*, 2(2):e000102, 2013.

[27] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vander-

plas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[28] Hilde Weerts, Miroslav Dudík, Richard Edgar, Adrin Jalali, Roman Lutz, and Michael Madaio. Fairlearn: Assessing and improving fairness of ai systems. *Journal of Machine Learning Research*, 24(257):1–8, 2023.

[29] Uri Kartoun, Shaan Khurshid, Bum Chul Kwon, Aniruddh P Patel, Puneet Batra, Anthony Philippakis, Amit V Khera, Patrick T Ellinor, Steven A Lubitz, and Kenney Ng. Prediction performance and fairness heterogeneity in cardiovascular risk models. *Scientific Reports*, 12(1):12542, 2022.

[30] Adewole S Adamson and Avery Smith. Machine learning and health care disparities in dermatology. *JAMA dermatology*, 154(11):1247–1248, 2018.

[31] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[32] Melissa Mccradden, Oluwadara Odusi, Shalmali Joshi, Ismail Akrout, Kagiso Ndlovu, Ben Glocker, Gabriel Maicas, Xiaoxuan Liu, Mjaye Mazwi, Tee Garnett, Lauren Oakden-Rayner, Myrtede Alfred, Irvine Sihlahla, Oswa Shafei, and Anna Goldenberg. 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*, pages 1505–1519, 2023.