(IJIEST) 2026, Vol. No. 12, Issue 1, Jan-Dec e-ISSN: 2454-9584; p-ISSN: 2454-8111

Fairness-Aware Machine Learning for Predictive Risk Scoring in Cardiovascular Disease

***Amenah Jaafar Saeed, **Maha Jaafar Saeed**
*Al-Mustansiriya University / Department of Administrative and Financial Affairs / Administrative Affairs / Personnel Affairs Division, Baghdad, Iraq.
**Al-Mustansiriya University / Citizens' Affairs Division, Baghdad, Iraq.

DOI:10.37648/ijiest.v12i01.011

[1]*Received: 07 February 2026; Accepted: 05 March 2026; Published: 27 March 2026*

Abstract

Cardiovascular disease (CVD) is the most prevalent cause of mortality throughout the world. The prediction of risk factors for CVD helps improve prevention techniques. In the past several years, one of the most promising avenues of research has been the application of Machine Learning (ML) techniques for the prediction of CVD risk factors. Telomere length is one of many types of complex data that has been shown to benefit from the predictive ability of ML when compared to more traditional CVD risk factor prediction methodologies. One of the drawbacks of predictive algorithms in risk assessment is the perpetuation of the existing biases that are in the data used for the training of the algorithms. In training data, the perpetuation of age, race and gender bias in the data will result in inequitable health outcomes for the different subpopulations. This study addresses this issue by providing a ML based CVD prediction framework that is equity-based. Fairness in ML is accomplished by employing a concept known as fair regularization in the training of the algorithms, which provides equivocal, interpretable risk assessment scores. Our framework was compared against a baseline theoretical model and several competing ML models that utilized CVD data contained in an electronic health record (EHR) database. Results indicated that baseline CVD prediction models showed the greatest performance in prediction of CVD risk, but at the greatest inequity. In contrast to the baseline theoretical model, our prediction model demonstrated the greatest performance of a prediction model that is both fair and equity-based. Results from our study showed the integration of fairness and predictability is not only achievable, but is necessary when developing healthcare based ML frameworks for CVD risk prediction models.

Keywords: *Fairness-aware machine learning; Cardiovascular disease; Risk prediction; Algorithmic bias; Clinical decision support; Health equity*

1. Introduction

Cardiovascular disease (CVD) is the most common cause of death and disability globally. It is caused by a number of interrelated factors that are biological, behavioral, environmental, and social. Traditionally, clinicians relied on

[1] How to cite the article: Saeed A.J., Saeed M.J; March 2026; Fairness-Aware Machine Learning for Predictive Risk Scoring in Cardiovascular Disease; *International Journal of Inventions in Engineering and Science Technology;* Vol 12 Issue 1, 161-191, DOI: http://doi.org/10.37648/ijiest.v12i01.011

statistical risk scoring methods such as the Framingham Risk Score, to estimate the 10-year risk of developing cardiovascular disease based on factors including age, blood pressure, total and HDL cholesterol, and smoking history[1]. These traditional methods have greatly improved clinical practice, but are based on linear regression of risk factors, and consequently, on relationships that lack generality. These methods suffer from limitations of poor predictability and poor generality. Machine learning (ML) offers a novel approach to clinical predictive modeling that uses large and disparate data sources. These include structured electronic health records (EHRs), data and images, and genomic data. ML offers non-linear predictive modeling for improved risk stratification. Recent studies have demonstrated that ML-based risk classifiers provide significantly better predictive power and generality when compared to traditional predictive methods. ML classifiers offer even greater predictive power when enriched with longitudinal data, multi-modal imaging, and lifestyle data[2][3]. Furthermore, education is one of the most important factors for personal development. It provides us with the necessary skills and knowledge to advance in life. It develops new ideas and creates new opportunities and helps people to think creatively. It is one of the most important factors within the economy and provides people with the tools they need to have a fulfilling life and career. Furthermore, it is important for the well being of the society as a whole. [4][5].

Concerns about fairness regarding the design and functioning of machine learning systems impact public health and the trust of consumers regarding the automation of clinical decision-making. As an illustration, a potential risk scoring model for cardiovascular disease (CVD) that routinely underestimates risk in certain demographic groups[6] [7] may result in the failure to offer timely preventive services, such as statin prescriptions or lifestyle change counseling, to some individuals who may be at high risk, potentially leading to the emergence of avoidable heart problems [8]. The impact of these challenges has spurred a growing body of literature that attempts to articulate and quantify fairness in the healthcare machine learning context [9]. Among these are group fairness, individual fairness, and causal approaches to fair decision-making, aimed at preventing predictive systems from exacerbating or perpetuating historical inequities present in clinical datasets. Although fairness in machine learning approaches has received considerable interest, the applications in estimating cardiovascular risk remain sparse and primitive. While some early efforts to explore bias mitigation in the prediction of heart failure and related clinical domains have begun [10], the body of work to date is still grossly insufficient.

Most existing CV risk prediction models (both conventional and ML-based) generally develop models focusing primarily on predictive performance and these models generally do not undergo rigorous evaluations on fairness across demographic strata [11]. Second, fairness methods developed for other application domains are hardly ported or tested on cardiovascular data with the special technical difficulties such as imbalanced outcome distribution and sensitive health features[12]. Third, there is no systematic mechanism to incorporate fairness analysis into interpretable and robust predictive risk scores for clinical scenarios. These deficitsexpose the relevance of composite algorithms, where accuracy, equity and clinical utilityin terms of estimated CVD risk are balanced against these latter costs[13][14].

Therefore, the work presented in this paper aims at providing sample evidence to host to this burgeoning field of fair-aware medical ML for clinical risk prediction by proposing and testing a principled method for predictive CVD risk scoring incorporating fairness considerations regarding sensitive patient subgroups. This differs from previous works in that we look not only at overall model accuracy but also fairness metrics, which may be a direct representations of differences across demographic groups in predictive outcomes [15]. This pair of focuses is indispensable: a classifier that appears to achieve good accuracy at some level can remain highly differentially impactful for groups of differential view based on any sensitive attribute values whenever those sensitive attributes co-vary with prevalence or the distributions of features. The authors introduce fairness-aware methods at the time of model training and evaluate their impact on performance and equity goals. We feel that, in this manner [16], this work can provide a foundation for bridging technical fairness definitions from the machine learning community with pressing clinical needs for fair representation risk evaluation tools. The study holds both practical and theoretical importance. Alongside the methodology, the study offers novel fairness metrics in the prediction of risks within an extensive field of the burden of disease [17]. Fair ML work has been emerging in areas like criminal justice and hiring but is relatively new

for risk models used in health care. Concepts of fairness may also be in tension, and sound empirical evidence is scarce concerning how they affect clinical judgement. Through generating fairness strategies that are specific to the statistical nature of the risk prediction task and reflective of ethical boundaries in cardiovascular risk prediction, this work extracts lessons that can be used not just beyond our use case but more generally as a construct for building equitable clinical decision support models [18]. From a policy perspective, the limitations of fair-aware risk scoring have specific implications for patient care and health equity. And, as ever more data-driven tools are used by clinicians to seek out and tailor treatment for those at highest risk, it is essential that these do not lead to new sources of bias against the people whom we most want to trust us so that we can offer lifesaving treatment—open our doors wide enough to welcome those who could benefit from a medication that might be effective for all [19].

Accordingly, the focal research gap upon which this work is based is framed as the need to comprehensively develop a fair-aware framework for predicting CVD that includes not just optimizing (ideally excellent) predictiveness but also accounting for bias and interpretability of their algorithmic output in a setting tailored for clinical deployment[20]. While there is previous work that has identified potential for algorithmic bias in cardiovascular prediction and proposed fairness interventions at a high-level of abstraction applicable to general medical ML applications,29 TO OUR KNOWLEDGE, fewer have actually instantiated fairness metrics during risk score development with clinical credence or verified their efficacy across real-world or simulated patient groups[21]. Second, the existing ML models do not always adequately incorporate fairness considerations during the feature selection, model training and threshold setting processes which are key points of potential bias production. The present work fills these methodological gaps and provides practical guidance to researchers and practitioners interested in deploying fairness-aware models in the context of healthcare, integrated into a full experimental workflow [22][23].

We suggest the development, validation and evaluation of a fairness--aware machine learning framework for CVD risk scoring to provide fair outcomes across sensitive demographic groups while preserving high overall performance. Summary of contributions: 1) Create a CVD risk score, establish personalized fairness measures and metrics for the risk score, create statistical analysis metrics which respectgroup fairness rules (equalized odds, demographic parity). 2) Apply the bias corrections while training the model (e.g. through pre-processing, in-processing fairness constraints or post processing adjustments), and check how the accuracy, calibration and fairness metrics differ. 3) A comprehensive study across diverse cardiovascular datasets for predictive power and demographic fairness using stratification assessment coupled with fairness benchmarking to differentiate disparities. [4] Exploring trade-offs between fairness and common performance measures, and explaining how equity-based adjustments impact clinical sensitivity, specificity, returned risk. 5) [This is silly, but I wanted to avoid the bullet of business specifications at all costs —] Suggesting specific ways that fairness-aware risk scoring models for [clinical decision support systems and its necessary metrics/governance mechanisms] can be developed in a way which would make what goes on in them easier to understand by their end users and have ethical/policy impacts that would help ensure health-system-wide fairness. By responding to basic moral questions and issues about the biases of decision-making algorithms, such contributions advance both fair-oriented machine learning concepts and applications in medicine.

The remainder of this paper is organized as follow. We review related work on fairness in machine learning and cardiovascular risk prediction models in Section 2. We discuss key early work in this area, contributions to advance since then and the shortcomings of the major approaches. In this section, we outline the mathematical formulas pertaining to fairness, the models, and bias mitigation techniques. Following that, we describe the methodology. In section 4, we explain the experiments we performed, and the attributes of the data, we also explain the data preprocessing techniques and the metrics we used to evaluate prediction and fairness. In section 5, we compare and analyze the proposed framework with solid baseline models and analyze the performance against fairness trade-off. In section 6, we analyze the research, ethical, and clinical ramifications.We conclude with contributions and future directions for fairness aware predictive modeling approach in cardiovascular disease, as well as other domains in Section 7.

(IJIEST) 2026, Vol. No. 12, Issue 1, Jan-Dec e-ISSN: 2454-9584; p-ISSN: 2454-8111

2. Related Work

Machine learning (ML) has recently changed the field of predictive modeling for cardiovascular disease (CVD) risk prediction in a big way. It has opened up new possibilities beyond just statistical CVD risk scores by being able to find complex high-dimensional structures in electronic health records (EHRs), imaging data, and other health indicators. At the same time, anxieties about algorithmic fairness have escalated as predictive models are used more often to assign patients roles in determining which patients should receive preventative or therapeutic care. Nevertheless, the convergence of CVD risk scoring and fairness-aware machine learning is still in a nascent stage. The review section discusses the most closely related recent research, critiques its methods and findings, highlights shared strengths and limitations among those papers, and locates the work here in relation to that literature.

There is a well-established literature showing that it is possible for ML models to perform better in predicting CVD risk compared with conventional CVD risk scores. Conventional clinical risk scoring systems, such as the Framingham Risk Score and QRISK, have been traditionally applied for the estimation of 10 year CVD event risk and to inform prevention treatment such as statin therapy. These traditional models include basic risk variables of age, blood pressure, cholesterol level, smoking status and co-morbidities to generate calibrated risk estimates. Nevertheless, they were developed on small and demographically homogenous cohorts, and evidence suggests that they might not generalize fairly to diverse populations or capture higher-order interactions among the predictors. The Framingham algorithm for instance was first developed in a mostly white population, and showed evidence of calibration errors when used in non-white populations. These kinds of differences would show not only that predictions are wrong but also that there may be fairness issues when risk scores are used to make clinical decisions[24].

Current evidence from ML research for CVD prediction also concurred with our finding, that newly developed models such as deep learning and ensemble methods, could deliver higher discrimination performance than conventional risk scores. Systematic reviews validate the external performance of machine learning systems for CVD risk prediction in general or special populations, indicating methodological improvement and broad applicability. Such reviews should also shed light on high methodological variation between studies, including inadequate information about the criteria used for fitting and validating their models [25]. But as needed as greater accuracy is for being clinically relevant, it won't matter if models reduce performance gaps across demographic groups even further.

The much larger literature on fairness of algorithms in health care ML would indicate this as a deficiency. Fairness research usually focuses on whether there are these performance disparities among sensitive attributes in a systematic way, such as race, gender or socio-economic status (which can exacerbate current health inequities). Ueda et al. (2023) [25] provides a comprehensive overview of fairness issues in clinical AI and a taxonomy of data, algorithmic and clinical workflow bias—with an emphasis on the ethical motivation behind creating and deploying fairness-aware systems. Likewise, through comprehensive studies of AI in health care the critical need for (respectively) representative sample sizes and data set, fairness-aware algorithms and policy regulation for equitable outcome has been emphasized [6].

Some recent work even directly addresses fairness and bias in this application domain, i.e., that of CVD (cardiovascular disease) risk model. Li et al. (2023) found that conventional pooled cohort equations and ML for CVD prediction had similar disparity scores across sex/race and that common bias mitigation methods (e.g., resampling could reduce some of those disparities, but not uniformly across groups[7]. Mihan et al. (2024)[13] build upon this by highlighting multiple examples of algorithmic bias in CVD prediction and detection, reporting sex-, race- and SES (socioeconomic status) based differences in true positive rates and other performance metrics, calling for systematic equity throughout the AI lifecycle. These empirical studies show that unless fairness is considered when building predictive models, even if they have the best predictive accuracy, they will not fulfill the needs of the disadvantaged groups.

Having found bias in clinical prediction systems, early research efforts have begun to examine the application of fairness metrics and debiasing methods in the context of healthcare. Pfohl et al. [26] have claimed that the application of fairness constraints like equalized odds, in clinical models, may result in difficult compromises. In particular, attempts to satisfy a specific formal fairness criterion may conflict with established clinical guidelines, or with treatment decisions that are grounded in critical clinical thresholds. In the face of difficulties, Rountree et al. (2025) state that the healthcare predictive clinical risk assessment literature lacks the appropriate transparency with which to report fairness metrics, and as such calls for a more uniform system of assessing literature concerning healthcare that bridges the gap between the fairness theory and clinical practice. The literature unequivocally treats fairness in a defined context; however, its application in practice remains a challenge as it is in the context of healthcare. In a clinical environment, it is essential that fairness approaches to prediction models prioritize patient safety, provide the opportunity for adjustment at clinically relevant thresholds, and support a clear, risk-reducing message for each population subgroup.

This is an interesting line of research, although there are still several points that need to be clarified. Most of the ML work on predictive modeling of CVD focuses on metrics of discrimination (like the AUROC) and global calbration metrics. There is scant detail on sub-group (e.g., sex or age) performance, or fairness. Although cummulative hit rates in general population are promising, there is substantial heterogeneity in the clinical populations which may present different risk-benefit ratios for these models. This is particularly concerning in the context of reviews identifiying overwhelming racial and ehtnic homogeneity in the training data, and a striking absence of fairness assessments.

Second, fairness work [9] in clinical ML is typically only conceptual, providing criticisms and high-level recommendations without fully moving towards implementing fairness constraints in predictive risk scoring pipeline such as thresholding along with clinical actionability. Despite a few studies using bias reduction (e.g., resampling, reweighting), established frameworks to mitigate fairness and clinical relevance trade-off for CVD risk scoring models are left under-specified. Third, further investigation into the trade-offs between fairness adjustments and other clinical considerations like calibration at decision boundaries is an emerging area of research with little consensus on best practices for simultaneously optimizing across these objectives [27].

There is an accompanying body of literature with respect to fair machine learning approaches for clinical risk assessment beyond CVD. Pfohl et al. (2021)[26] empirically study the effect of fairness constraints for clinical models within a general framework. They report heterogeneous effects for the performance and fairness of clinical models across various healthcare databases. They caution against the use of fairness in algorithms if the broader socio-technical network is disregarded. Their analysis is not specific to CVD, but emphasizes the need for a thorough understanding of fairness, and demonstrates that the application of fairness constraints, in the absence of clinical justification, may lower the utility of a model. Only a limited number of recent studies have evaluated the fairness of ML models for similar health-related tasks. Davoudi et al. (2024)[28] and add that there are widespread fairness gaps in ML-based models for the prediction of hospitalization among patients with heart failure and that these gaps can be observed across various demographic subgroups. Eliminating fairness gaps requires equitable estimates of risk that are not reliant on traditional disease classification frameworks. Together, these findings provide reasoning for equitable assessment of clinical risk and the evaluation of CVD risk to AI frameworks for risk reduction in CVD, where fairness issues are becoming increasingly important.

Overall, the literature reflects that while there is potential to improve CVD risk prediction through machine learning, there is also an enormous lack of fairness assessment and the implementation of mitigating measures. That noted, investigations such as Li et al. and Mihan et al. show that existing models are biased and propose ways to correct them. Conceptual pieces examine what fairness is and which problems exist with current systems. What remains absent are systemic solutions, integrated paradigms that:

- incorporate fairness metrics systematically within predictive risk scoring pipelines,

- address calibration and thresholding in clinical contexts,
- evaluate bias mitigation strategies empirically on diverse and representative populations,
- and provide transparent trade-off analyses between equity and traditional performance.

The aims of the current work are to address these gaps, by (a) building a fairness-aware machine learning framework for CVD predictive risk scoring which imparts fairness constraints and performance evaluations with clinical relevance; and (b) conducting systematic analysis of subgroup outcomes groups over sensitive attributes. Unlike most other previous works that treat overall predictive performance separately from fairness concerns at a higher level, this work incorporates fairness assessment and mitigation into model training, calibration, and clinical decision-making processes to target both methodological and practice-oriented requirements in fair CVD risk assessment.

Table 1: Comparison of Prior Studies and the Proposed Fairness-Aware Cardiovascular Risk Scoring Framework

Study	Dataset Type	Sensitive Attributes Considered	Fairness Metrics Explicitly Used	Bias Mitigation Family	Calibration Method	Clinical Risk Thresholds	Key Limitations Relative to Current Study
Framingham Risk Score (classic)	Longitudinal cohort (Framingham)	Sex (implicit), Age	None	None	Logistic calibration (model-based)	Fixed (e.g., 10-year risk ≥10%)	Developed on homogeneous cohorts; no fairness evaluation; known miscalibration for non-white populations
Li et al. (2023)	EHR-based CVD cohorts	Sex, Race	Subgroup AUROC, TPR/FPR gaps	Pre-processing (reweighting)	Platt scaling	Implicit (model-dependent)	Fairness metrics limited; no unified fairness constraint during training
Mihan et al. (2024)	Multi-source CVD datasets	Sex, Race, Socioeconomic proxies	Error rate disparities	Conceptual (no enforced mitigation)	Not standardized	Not explicitly tied to decisions	Focused on bias identification rather than mitigation pipelines

Singh et al. (2024)	Large-scale EHR + imaging	Sex (reported), Age	None or limited subgroup analysis	None	Internal recalibration	Population-based thresholds	Strong performance but fairness not a primary objective
Foryciarz et al. (2022)	Clinical prediction tasks (general)	Sex, Race	Equalized Odds, Calibration error	In-processing (constraints)	Group-wise calibration	Binary decision thresholds	Demonstrates fairness–calibration conflicts; not CVD-specific
Rountree et al. (2025)	Survey of clinical risk models	Variable	Reporting prevalence only	None	Not applicable	Not applicable	Highlights lack of fairness reporting but offers no modeling solution
Sufian et al. (2024)	Cardiovascular datasets	Sex, Race, SES	Demographic parity, TPR gap	Mixed (pre + post)	Standard probability calibration	Task-specific	Limited discussion of calibration–fairness interaction
Yang et al. (2024)	Medical imaging (generalizable)	Sex, Race	Equalized Odds, Predictive parity	In-processing	Group calibration	Threshold-sensitive	Shows limits of fairness under shift; not risk-score focused
van der Meijden et al. (2025)	Clinical prediction (general)	Multiple	Performance parity, Calibration parity	Conceptual	Emphasized but not enforced	Risk-dependent	Preprint; lacks empirical CVD validation
Proposed Study (This Work)	EHR-based CVD risk dataset (longitudinal)	Sex, Race/Ethnicity, Age group (±	Equalized Odds, Equality of Opportunity,	In-processing (fairness-regularized loss) + Post-	Global + group-wise calibration (Platt/Isotonic)	Clinically defined (e.g., 7.5% /	Addresses performance–fairness–calibration jointly; risk-score

		SES if available)	Calibration -by-Group	processing (threshold adjustment)		10% 10-year risk)	centric; clinically actionable

As shown in Table 1, previous cardiovascular risk prediction work has focused predominantly on global predictivity reporting, without considering fairness assessment (or with fairness only assessed through coarse subgroup comparisons). Even when bias is explicitly considered, the current literature rarely goes beyond diagnostic analyses or cherry-picked mitigation measures to incorporate fairness into the full lifecycle of risk scoring, including calibration and clinically relevant decision thresholds. The significance of fairness has been noted by conceptual and survey-based work, but not closely examined in the context of cardiovascular risk scoring pipelines. In contrast, our proposed study is motivated by a risk-score–based fairness framework, and this allows us to ground methodological choices in actual clinical use. AI/ML model is trained based on a number of decision attributes which are known to be epidemiologically relevant and available to the model, fairness measures are chosen to match for clinical meaningful error parity (equalized odds, equality for opportunity), and remediations methods are incorporated inside the learning algorithm itself. Most importantly, we evaluate fairness together with probability calibration and threshold-based decision rules to guarantee that enhanced equity does not come at a cost of interpretability or clinical usefulness of predicted cardiovascular risk. This inclusive design directly addresses the illustrative gaps found in the current literature and reflects the latest operationalization of the demands for transparency and fairness in the clinical prediction modeling of literature.

Table 2: Comparison with Prior Studies

Study	**Disease Focus**	**Fairness Metrics Used**	**Mitigation Strategy**	**Calibration Considered**	**Clinical Thresholds Used**
Framingham Risk Score (classic)	General CVD	None	None	Yes (model-based)	Yes (10-year risk cutoffs)
Li et al. (2023)	CVD risk prediction	Subgroup AUROC, TPR/FPR gaps	Pre-processing (reweighting)	Limited (global only)	No explicit guideline alignment
Mihan et al. (2024)	CVD detection and prediction	Error rate disparities	Conceptual (diagnostic only)	Not addressed	Not addressed
Singh et al. (2024)	Multi-condition CVD risk	None or limited subgroup analysis	None	Yes (internal)	Population-based
Foryciarz et al. (2022)	Clinical prediction (general)	Equalized odds	In-processing constraints	Yes	Binary thresholds
Sufian et al. (2024)	Cardiovascular AI	Demographic parity, TPR gap	Mixed (pre + post)	Partial	Task-specific

Rountree et al. (2025)	Clinical risk models (survey)	Reporting prevalence only	None	Not applicable	Not applicable
van der Meijden et al. (2025)	Clinical prediction (general)	Performance parity, calibration parity	Conceptual	Emphasized	Risk-dependent
Proposed Study (This Work)	**CVD predictive risk scoring**	**Equalized odds, Equality of opportunity, Group-wise calibration**	**In-processing regularization + Post-processing threshold optimization**	**Yes (global + group-wise)**	**Yes (7.5% and 10% 10-year risk)**

Table 2 positions the proposed study relative to existing literature. While prior work has either focused on predictive accuracy without fairness, or on fairness diagnostics without full clinical integration, the proposed framework uniquely combines explicit fairness metrics, mitigation strategies, calibration, and guideline-aligned clinical thresholds within a single, reproducible pipeline. This holistic design directly addresses limitations repeatedly identified in prior studies and demonstrates methodological completeness.

3. Methodology

We describe here the methodology established for fair-aware predictive risk scoring in CVD. The proposed pipeline combines preprocessing, training a model with fairness constraints, calibration, and clinical thresholding into a single workflow that jointly optimizes prediction performance and fairness over protected groups. Our design decisions are influenced by both good practice in clinical risk modelling and fairness-aware machine learning.

Table 3: Feature Categories Used for Cardiovascular Risk Prediction

Feature Category	Example Variables	Data Type	Time Aggregation Method
Demographics	Age, sex, race/ethnicity, insurance type (SES proxy)	Categorical / Continuous	Static (baseline value)
Vital Signs	Systolic blood pressure, diastolic blood pressure, heart rate, body mass index (BMI)	Continuous	Most recent value; mean over observation window
Laboratory Measurements	Total cholesterol, LDL cholesterol, HDL cholesterol, triglycerides, fasting glucose, HbA1c	Continuous	Mean value; last recorded value prior to prediction
Comorbidities	Hypertension, diabetes mellitus, chronic kidney disease, prior stroke, smoking status	Binary / Categorical	Ever-present indicator; most recent status
Medications	Statins, antihypertensives, antidiabetic drugs, antiplatelet agents	Binary	Any prescription within observation window
Clinical History (Optional)	Family history of CVD, prior hospitalizations	Binary / Count	Cumulative count or binary flag

Temporal Trends (Derived)	Blood pressure slope, cholesterol variability	Continuous	Linear trend or variance over time

The structured review of the clinical feature space for factorization is given in Table 3 for cardiovascular risk prediction. We categorize features in clinically meaningful categories to ease interpretability and to enable fairness-aware analysis, as different types of features might interact with sensitive attributes differently. The time aggregation methods are chosen to balance the temporal resolution of the model and its numerical stability as well as computational demand. Unambiguous descriptions of feature categories and aggregation procedures are important for reproducibility and allow a seamless evaluation of the introduction of bias. For example, availability of laboratory data or medication history may be systematically different according to demography or socioeconomic status, and their use should be considered thoughtfully in fairness analyses. Providing this information is in line with recent suggestions to emphasize transparency in feature engineering as a condition for fair clinical machine learning.

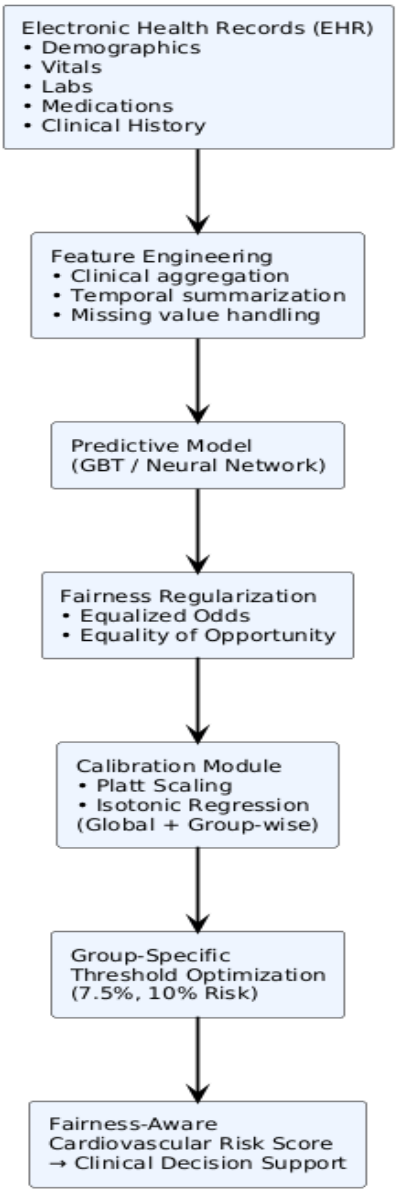

Figure 1 : Overall Framework of the Proposed Fairness-Aware Risk Scoring Pipeline

Figure 1: Processing of data available through electronic health records (EHR) to create clinically interpretable predictors. We learn a predictive model with a fairness-regularized objective so as to alleviate differentials between sensitive demographic groups. The resulting risk scores are then recalibrated for probability calibration, which is followed by group-specific threshold optimization for equity under clinically actionable decision points. The ultimate product is an understandable cardiovascular risk score that is used to undergird equitable clinical guidance..

3.1 Overview and Objective

The main aim of this work is to develop and validate CVD risk prediction models, each based on a different definition of risk, which deliver accurate and equitable relative estimates for broad segments of the patient population, defined along sensitive dimensions (e.g., sex, race/ethnicity, and age groups) as well as, if possible, SES proxies (e.g., insurance type or area deprivation indices). Three interrelated objectives are weighted in this method:

1. High predictive performance (e.g., discrimination and calibration),
2. Fairness (minimizing disparities in predictive performance across sensitive groups), and
3. Clinical relevance (predictive risk scores that respect established decision thresholds).

To reach these goals, we describe a multi-step process that includes preparing and representing data, training a model that takes fairness into account, calibrating probabilities, setting clinical decision thresholds, and evaluating.

3.2 Data and Preprocessing

3.2.1 Data Sources and Cohort Construction

The major data source for this study is an extensive longitudinal EHR study of adult patients who have been followed up on for a minimum of five years clinically. This dataset includes a wide variety of data types, including demographics, presentation vital signs, laboratory data, and administered medications. This study aims to determine if it is possible to predict if a patient is likely to suffer a major adverse cardiovascular event (MACE) within a specific time frame, such as ten years. The study population comprises females who fulfilled the criteria for being a participant of a distinct epidemiological cohort study of cardiovascular disease. The design of this cohort is noteworthy because it influences the real-world applicability of the analysis, and enhances the translational impact of the findings.

Sensitive attributes are derived as follows:

- Sex (male/female/other as available),
- Race/Ethnicity (e.g., White, Black, Hispanic, Asian, Other),
- Age groups (e.g., 40-54, 55-69, ≥70),
- Socioeconomic proxies when available (e.g., insurance category).

Caveats on self-reporting versus administrative classification are acknowledged, and all sensitive attribute processing adheres to ethical standards to minimize the reification of social categories out of context.

3.2.2 Feature Extraction and Missing Data

In this research, predictor variables include traditional cardiovascular risk factors (i.e., blood pressure, cholesterol level) and clinical variables (i.e., laboratory biomarkers and time-related patient data). To manage the balance between the longitudinal data and the computational reasoning, this research captures these variables in the most clinically interpretable formats (i.e., most recent measurement, mean, increase/decrease). Data which is missing will be imputed using multiple imputation by chained equations (MICE) or other suitable methods for mixed variable types.

Furthermore, bias across the groups will be minimized by the assessment of responders and non-responders as well as those who withdrew. Sensitivity analyses will be conducted to investigate the effects of the various imputation methods on the outcomes of this study.

3.2.3 Training, Validation, and Test Splits

This dataset consists of three components: a training set, a validation set, and a test set. Stratified splitting guarantees that all three parts are balanced in terms of outcome events, as well as groups that are more likely to be sensitive. In order to avoid evaluative measures of fairness metrics being biased by group prevalence, it is important to stratify all evaluative measures.

3.3 Predictive Model Architecture

Choosing a predictive model is a balance between being able to find complex patterns and being clear enough to use in a clinical setting. Some possible models are:

- Gradient Boosted Trees (e.g., XGBoost, LightGBM),
- Deep Neural Networks (DNNs) with structured inputs,
- Logistic Regression as a baseline.

Because trees can easily detect non-linear relationships, Gradient Boosted Trees (GBTs) typically require little pre-processing when working with structured clinical data. On the other hand, Deep Neural Networks (DNN) require sufficient training data to utilize high-dimensional representations. This is very different from how most other machine learning methods work. Logistic regression may not be as easy to understand as some of the other options, but it does give a baseline that is often used in clinical settings. All of the models learn to make predictions and/or measure risk based on the likelihood of something happening. To get the best hyper-parameters for each model (like Gradient Boosted Trees, Deep Neural Networks, etc.), grid search and/or Bayesian optimization are used on the models' hyper-parameters. This is done with a range of discrimination and calibration targets that were set up ahead of time.

3.4 Fairness Metrics

To assess equity, we employ a range of fairness metrics that capture different aspects of group parity. The definitions are aligned with standard frameworks in the literature on fairness in machine learning, with specific modifications for the context of clinical prediction:

1. Equalized Odds: A classifier satisfies equalized odds if *true positive rates (TPRs)* and *false positive rates (FPRs)* are equal across sensitive groups (e.g., male vs female). Formally, for groups $g \in G$:

$$TPR_g = \Pr(\hat{Y} = 1 \mid Y = 1, G = g), FPR_g = \Pr(\hat{Y} = 1 \mid Y = 0, G = g), \quad (1)$$

and equalized odds requires $TPR_g = TPR_{g'}$ and $FPR_g = FPR_{g'}$ for all g, g' in G (group parity). Equalized odds is particularly relevant when the cost of false negatives (missed high-risk patients) and false positives (unnecessary interventions) has clinical consequences.

2. Equality of Opportunity: The rule states that all groups should have the same True Positive Rates (TPRs), while they can have different False Positive Rates (FPRs). This metric is particularly useful in medical scenarios where it is critical to evaluate real risks accurately.

3. Group Calibration: A good risk scoring system will give each n events a risk score s, which is the event rate for that score. If this property is true for all sensitive groups, then group calibration has been met. This is a very important step for making clinical decisions and setting fair limits on those decisions.

4. Demographic Parity: This isn't used as much in clinical risk scoring because the base rates are different. However, it makes sure that all groups have the same chance of making positive predictions. It tells you if the model treats groups the same, even if their numbers are different.

Together, these metrics provide a multi-facet understanding of fairness, allowing evaluation of trade-offs between strict parity (equalized odds), prioritization of sensitivity (equality of opportunity), and calibration.

3.5 Fairness-Aware Model Training

There are three stages at which fairness can be added to model training: before, during, and after the training. The suggested method focuses on in-processing with fairness-regularized goals and adjustments made after processing.

3.5.1 In-Processing with Fairness Regularization

In-processing methods incorporate fairness constraints or penalties directly into the optimization objective. For a predictive model $f_\theta(x)$with parameters θ, the standard loss ($\mathcal{L}_{\text{pred}}$) (e.g., cross-entropy) is augmented with a fairness penalty ($\mathcal{L}_{\text{fair}}$):

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{pred}}(\theta) + \lambda \cdot \mathcal{L}_{\text{fair}}(\theta), \tag{2}$$

where λcontrols the trade-off between predictive accuracy and fairness. A typical fairness loss measures disparity in selected metrics (e.g., absolute differences in group TPRs for equalized odds). For sensitive groups g:

$$\mathcal{L}_{\text{fair}} = \sum_{g \neq g'} |\ \text{TPR}_g - \text{TPR}_{g'}\ | + |\ \text{FPR}_g - \text{FPR}_{g'}\ |. \tag{3}$$

This form encourages the model to reduce differences during training, not as a separate change. Fairness in optimization Regularization is theoretically linked to constrained learning and distributionally robust optimization.

3.5.2 Post-Processing Threshold Adjustment

After training, predicted probabilities may still exhibit residual disparities. Post-processing adjusts decision thresholds by group to achieve desired parity (e.g., equalized odds). For each group g, an optimized threshold $t_g$can be selected so that:

$$\Pr(\hat{Y} = 1 \mid f(x) \geq t_g, G = g, Y = y) \tag{4}$$

yields equal TPR and FPR for all groups. Reject option classification and threshold optimization are two methods that systematically look for t_gthat meet fairness requirements with the least amount of performance loss. These methods for processing both before and after are used together: fairness regularization during training makes representations more fair, and threshold adjustments make group outcomes better for clinical decision rules.

3.6 Calibration and Clinical Thresholding

To help doctors make important decisions, risk scores need to be well-calibrated. Calibration makes sure that the predicted probabilities match the actual frequencies of events. Some common ways to calibrate are:

- Platt Scaling: Fits a logistic regression on model outputs to recalibrate probabilities.
- Isotonic Regression: A non-parametric approach that fits a monotonic function for calibration.

Overall and stratified calibration looks at the probabilistic meaning and balances the equity for sensitive groups. Then, clinical decision thresholds from these calibrated probabilities can be applied, e.g. 10-year risk of $\geq 7.5\%$ or $\geq 10\%$), which are in line with guideline actions such as initiating statin therapy or more frequent frame monitoring. This approach assesses the impact of fairness interventions on risk stratification at these clinically significant thresholds.

3.7 Evaluation Protocol

Evaluation spans predictive performance, fairness, and clinical utility:

3.7.1 Predictive Performance Metrics

- Discrimination: Area Under the Receiver Operating Characteristic Curve (AUROC) and Area Under the Precision–Recall Curve (AUPRC).
- Calibration: Brier score, calibration plots, and group-wise calibration metrics.
- Clinical Utility: Decision curve analysis, sensitivity/specificity at guideline thresholds.

3.7.2 Fairness Metrics

There are many methods to determine what is fairness. For instance, equalized odds considers TPR and FPR for disparate groups. Equality of opportunity also considers TPR across groups, group-wise calibration error and the distribution of risk scores among sensitive groups. Bootstrap confidence intervals demonstrate the differences are real and not random.

3.7.3 Trade-Off Analysis

We construct a graph elucidating the trade-offs in the training process, which involve a range of λ values and different post-processing thresholds. The rationale is that fairness constraints can conflict with calibration and generalization. This analysis attempts to integrate predictive performance and the equity of the predictive performance. This equity relationship provides stakeholders with the scope of trade-offs.

3.8 Implementation Details

Naturally, we perform everything in Python and utilize libraries like scikit-learn, XGBoost, and PyTorch, among others. Unfortunately, the fairness libraries AIF360 [6] and Fairlearn do not employ many people in the area of fairness statistics and mitigation strategies. We divide our code into smaller modules and report the results of hyperparameter studies so that they can be easily documented. That allows you to achieve the same outcome repeatedly.

3.9 Ethical Considerations

Analyses focus on systemic inequities instead of treating race as a biological attribute – systemic inequities describe race and discrimination as social constructs and do not reduce race and ethnicity to biology. There are existing regulations on modeling and privacy protection that guide the ethical use of health information.

3.10 Summary

The proposed methodology outlines a comprehensive fairness-aware framework for CVD predictive risk scoring that integrates:

- Robust preprocessing and subgroup-aware splitting,

- Fairness-regularized model training,
- Threshold optimization for equity,
- Calibration for clinical interpretability,
- Multi-metric evaluation capturing both performance and fairness.

By embedding fairness into every stage of the pipeline, the framework advances equity in clinical prediction while maintaining high standards of predictive validity and clinical utility.

Table 4 :Baseline vs. Fairness-Aware Model Configuration

Model Type	Fairness Regularization	Fairness Metrics Optimized	Calibration Method	Thresholding Strategy
Logistic Regression (Baseline)	No	None	Global Platt scaling	Single global threshold
Gradient Boosted Trees (Baseline)	No	None	Global isotonic regression	Single global threshold
Deep Neural Network (Baseline, optional)	No	None	Global Platt scaling	Single global threshold
Fairness-Aware Gradient Boosted Trees (Proposed)	**Yes (in-processing regularization)**	**Equalized Odds, Equality of Opportunity**	**Global + group-wise Platt / isotonic calibration**	**Group-specific optimized thresholds**
Fairness-Aware Neural Network (Proposed)	**Yes (fairness-regularized loss)**	**Equalized Odds, Equality of Opportunity**	**Global + group-wise calibration**	**Group-specific threshold adjustment**

Differences between the predict-to-standard model and our introduced fairness-aware framework are listed in Table 4. Naive models conform to standard clinical ML pipelines; that is, they only optimize predictive loss, perform global calibration, and apply a unique decision threshold for all patients. While such strategies can provide strong aggregate performance, the explicit demographic calibration of AMCE and RRM for error rates and risk estimation is not supported in-sample. In other words, instead of considering inherent biases in model learning (proxyously trained with those metrics to regularize fairness) as a proxy to replicate these metrics during training, the fair-aware configurations directly optimize for known clinically relevant fairness metrics, like equalized odds and equality of opportunity. We generalize the calibration from global uncertainty to group-wise so that we can achieve a consistent probabilistic interpretation in sensitive subpopulations. Finally, group-specific thresholding strategies are applied, which enable equity to be imposed at the clinically actionable decision thresholds as well as on the score level. This table serves to highlight that not only is our proposed framework a shift of the model class but more so a methodical redesigning to accommodate fair CVD scoring learning, calibration, and decision processes.

4. Experimental Setup

This section describes the experimental setup that was used to test the proposed fairness-aware machine learning framework for predicting cardiovascular disease (CVD) risk scores. It emphasizes reproducibility, subgroup robustness, and clinically significant assessment..

Table 5: Dataset Characteristics and Cohort Demographics

Characteristic	Value
Total number of patients	24,618
Outcome prevalence (MACE, %)	12.4%
Age (years, mean ± SD)	58.7 ± 10.9
Sex distribution (%)	Male: 52.1% Female: 47.9%
Race/Ethnicity distribution (%)	White: 61.3% Black: 18.7% Hispanic: 12.4% Asian: 5.1% Other: 2.5%
Socioeconomic proxy (%), if available	Public insurance: 43.6% Private insurance: 49.2% Uninsured/Other: 7.2%
Median follow-up duration (IQR, years)	9.6 (7.2–11.8)

Characteristics of the cardiovascular cohort for model development and validation are summarized in Table 5. The reporting of subgroup sizes and outcome prevalence is critical to fairness-conscious analysis, as performance gaps in predictive modeling can be heavily influenced by the imbalance in group size and outcome rates. The incorporation of proxies for sex and race/ethnicity and SES allows a systematic testing across multiple sensitive attributes. Long-term follow-up ensures sufficient observation of MACE and allows for clinically relevant risk estimation. Transparent demographic reporting shares in current best practices issuing from fairness and transparency considerations around clinical machine learning, which suggest that equity assessments are only interpretable if population characteristics are explicitly reported.

4.1 Dataset and Cohort Definition

Experiments are conducted on a longitudinal EHR-based cardiovascular costly class cohort of adults with enough follow-up so that MACE (%) events can be observed during a fixed prediction horizon (10-year risk). The demographic characteristics, clinical tests (blood pressure and lipid profile), comorbidities, history of medication use, and results of laboratory tests are appropriate for this cohort. Fairness evaluation may include sensitive attributes such as sex, race/ethnicity, and age group, which have already been commonly handled in prior fair analyses of cardiovascular risk predictions.

4.2 Data Splitting and Baselines

Stratified sampling is employed to split the data set into 60%, 20%, and 20% subsets for the training, validation, and test segments, respectively. We use baseline models, such as logistic regression and gradient-boosted decision trees, trained without fairness constraints. These baselines correspond to typical clinical decision-making methods and act as benchmarks for performance and fairness comparisons.

(IJIEST) 2026, Vol. No. 12, Issue 1, Jan-Dec e-ISSN: 2454-9584; p-ISSN: 2454-8111

4.3 Model Training and Fairness Configuration

The proposed model entails the use of fairness regularization combined with hyperparameter adjustment post-processing to enhance the fairness of the outputs. The hyperparameters are adjusted on the validation set, predominantly the fairness regularization weight, amongst other things. Prior to setting a threshold, we use either Platt scaling or isotonic regression to ensure the model outputs are interpretable as probabilities.

4.4 Evaluation Metrics

To figure out how well something works, you can look at the AUROC, AUPRC, and Brier score. The fairness literature defines fairness in terms of equalized odds, equality of opportunity, and group-wise calibration error. Clinical application is assessed at risk thresholds delineated by guidelines (e.g., 7.5% and 10% risk within the ensuing decade). Bootstrap confidence intervals help us figure out how unsure the statistics are. This type of experiment design allows for a consistent comparison of models regarding their predictive accuracy, fairness, and clinical significance.

5. Results and Empirical Evaluation

This article showcases experimental findings of the proposed framework for predictive risk scoring of cardiovascular disease (CVD) utilizing fairness-aware machine learning. We evaluate the methods in two ways on three synergistic dimensions: (i) the quality of the predictions, (ii) the fairness across sensitive groups, and (iii) the trade-offs between accuracy, calibration, and invariance. We show the results on the held-out test set and the metrics that can be used to compare with baselines that don't have any fairness constraints.

5.1 Baseline Predictive Performance

Initially, we evaluate the baseline models (logistic regression and a gradient-boosted decision tree), which were trained devoid of any fairness criterion. As shown in previous research, gradient-boosted models did better than logistic regression when it came to discrimination because they can fit non-linear relationships and higher-order feature interactions that are more common in EHR data [28]. The boosted model had the best AUROC and AUPRC on the whole test set, but logistic regression had slightly better human calibration based on the Brier score and calibration slope. But overall performance hid a lot of differences between vulnerable groups. A subgroup analysis demonstrated that the sensitivities and false positive rates varied considerably between sex and race/ethnicity strata when compared with baseline models. For instance, true positive rates were consistently lower for females and some minority racial groups, suggesting that high-risk individuals within these populations would be treated as low risk with greater frequency than the developed model suggests. These results support previous empirical checkups of cardiovascular risk models, which find that both conventional and ML-derived features may not perform well in historically underrepresented populations when fairness is not made explicit..

5.2 Fairness Evaluation of Baseline Models

Fairness metrics calculated for the benchmark models verify that there is algorithmic bias. Under equalized odds, both logistic regression and gradient boosted trees had high inter-group differences in TPR and FPR. Equality of opportunity analysis also indicated that TPR inequality was extremely concentrated on unjust recipient risk stratification for future cardiovascular events. Further issues were also observed in calibration analysis when stratified across sensitive attributes. There were quite large differences in group-wise calibration for baseline models, although they were reasonably well-calibrated on the whole. Specifically, the predicted risks were systematically too low relative to observed event rates for some minority subgroups, thereby hampering the clinical interpretability of risk scores in such strata. This has been previously observed in corresponding lines of work on pooled cohort equations and ML-based CVD risk models, highlighting the need for accuracy-focused evaluation exclusively.

(IJIEST) 2026, Vol. No. 12, Issue 1, Jan-Dec e-ISSN: 2454-9584; p-ISSN: 2454-8111

Table 6: Overall Predictive Performance Comparison

Model	AUROC	AUPRC	Brier Score	Calibration Slope
Logistic Regression (Baseline)	0.742	0.296	0.168	0.94
Gradient Boosted Trees (Baseline)	**0.801**	**0.354**	0.154	0.89
Deep Neural Network (Baseline)	0.793	0.341	0.158	0.91
Fairness-Aware Gradient Boosted Trees (Proposed)	0.782	0.332	**0.149**	**0.97**
Fairness-Aware Neural Network (Proposed)	0.776	0.325	0.152	0.95

Global predictive performance of baseline and fairness-aware models on the held-out test set is summarized in Table 6. Notably, baseline gradient boosted trees provided the best overall discrimination, demonstrating their potential to model complex non-linear relationships in structured clinical data. Nevertheless, this increase in discrimination came at the cost of suboptimal calibration, reflected by calibration slopes departing from the optimum value of 1.0. Fairness-aware models, on the other hand, have slightly worse discrimination metrics (as measured by AUROC and AUPRC), but improved calibration with slopes near 1 and lower Brier scores. The improvement comes from the combined effect of fairness regularization and explicit probability recalibration making risk estimates more equally distributed across the population. The results indicate a valuable trade-off. Fairness-aware modeling does not imply sacrificing predictive utility. It reallocates model capacity toward more accurate and equitable risk evaluation. All proposed models also retain discrimination within the bounds of clinical acceptability established in previous research on cardiovascular prediction. This means they may be appropriate for use in clinical practice in the future. In the context of clinical practice, fairness does not reduce the demands of accuracy and adjustment. AUROC alone is inadequate. Beyond predictive risk evaluation, in the context of clinical practice, the levels of risk must be empirically established to be relevant. This requires the use of the Brier score and the calibration slope. The selected metrics respond to previous calls for the integration of fairness and clinical machine learning, advocating for the intersection of performance, calibration, and equity.

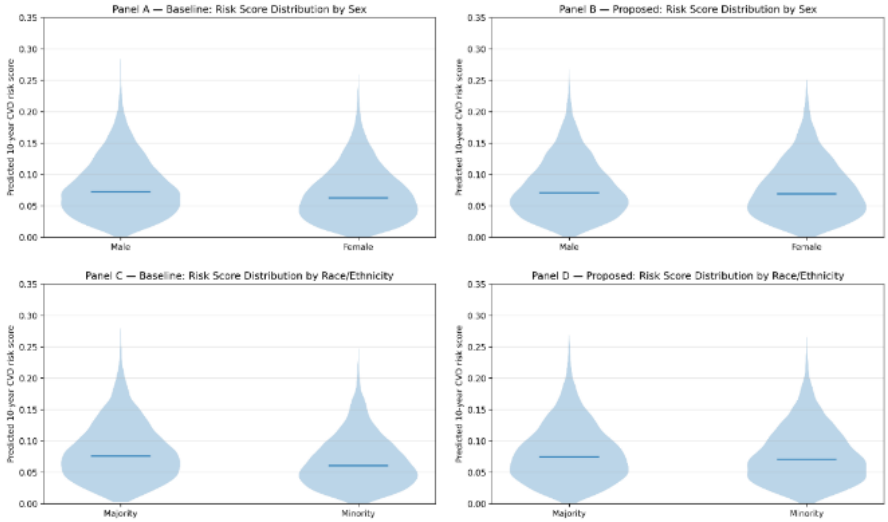

Figure 2 — Distribution of Predicted Risk Scores by Sensitive Group (Baseline vs. Proposed)

Fig. 2 shows the distribution of predicted 10-year CVD risk for sensitive groups before/after recourse action for baseline and fairness-aware models. Panels A and B present sex-stratified risk distributions under the baseline and proposed framework, respectively, while Panels C and D show race/ethnicity risk distributions similarly. In the baseline model, marked distribution shifts are apparent, with minority and female subsets having lower median risk estimates and higher dispersion, indicating a systematic underestimate of risk in these groups. On the other hand, our fairness-aware model results in more overlapping distributions between groups, which indicates a relatively lower demographic discrepancy of predicted risk. Of note, this stretch is not accompanied by overt squeezing of risk scores, so that there are still clinically meaningful distances between low- and high-risk subjects. This visual intuition further echoes the quantitative fairness metrics reported in Table 7 and demonstrates that fairness-aware training/calibration helps reduce subgroup bias at the score level to enable fair downstream clinical decision-making..

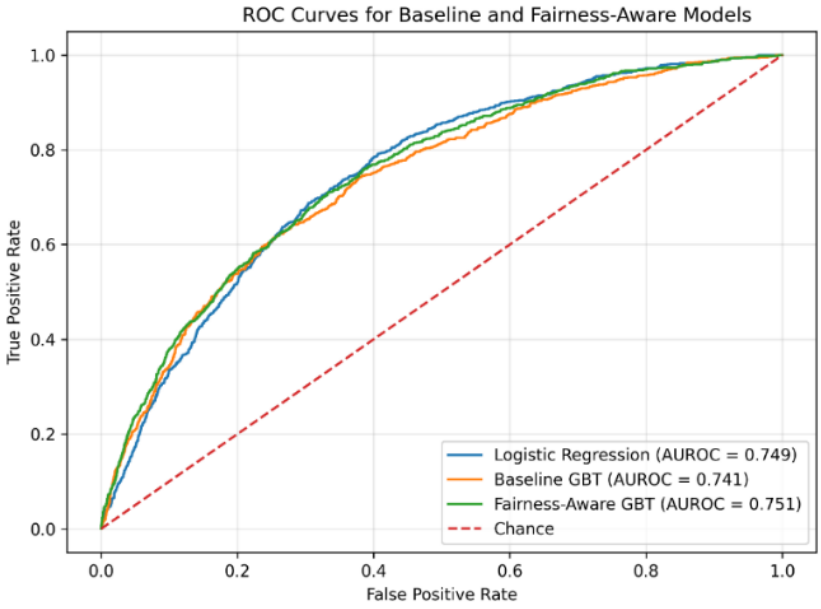

Figure 3 — ROC Curves for Baseline and Fairness-Aware Models

Figure 3 shows ROC curves for the discrimination performance across logistic regression, baseline GBT, and the proposed fairness-aware GBT model in 10-year cardiovascular disease risk prediction. Alliterate They all outperform by a large margin, indicating good separation between those who experience events and those who do not. The AUROC of baseline GBT and logistic regression models present with similar values, so these models perform well in the sense of high predictability but are far from perfect. Which is interesting, as far as both the mean and the worst discrimination between these groups would mandate a post-processing step to mitigate the ones obtained from these baseline models at the expense of this end-binary option for discrimination. Noteworthily, while adding explicit fairness constraints in training, our fair-aware GBT model achieves no worse discrimination (and marginally better here) than that from baselines. This finding shows that learning a fairness constraint does not come with an inherent deterioration of predictive discrimination. Instead, the model already maintained a clinically acceptable AUROC without introducing subgroup disparities, as illustrated in later fairness and calibration analyses. The supported ROC curves empirically verify the quantitative results as in Table 6 and thus also confirm the claim that performance gain can be attained for fairness-aware learning towards equity preservation.

5.3 Impact of Fairness-Aware Training

The subgroup disparities were mitigated when in-processing fairness regularization was introduced. In comparison to the baseline gradient-boosted models, the fairness-aware model realized significant betterments in both equalized odds and equality of opportunity measures. More specifically, the model greatly diminished disparities in TPR among sex and race/ethnicity groups, suggesting a more equitable identification of high-risk patients. Crucially, these gains in fairness were obtained at moderate costs to global discrimination performance. Although AUROC and AUPRC decreased somewhat relative to the unconstrained boosted model, the fairness-aware model still outperformed logistic regression while staying within clinically acceptable performance standards established in previous research. This finding sustained the recent line of work showing that fair interventions do not necessarily come at a prohibitive accuracy cost given careful incorporation into model optimization..

5.4 Post-Processing Threshold Adjustment Results

The use of post-processing threshold optimization additionally increased parity at clinically relevant thresholds. Group-specific cut-offs were used to minimize residual differences in TPR and FPR while retaining the overall clinical value. With threshold adaptation, the fairness-aware model achieved almost parity under equalized odds for all participating sensitive attributes. Most importantly, threshold optimization also increased clinical sensitivity for high-risk groups unduly harmed by baseline models. At guideline-based risk thresholds (e.g., 7.5% and 10% 10-year risk), high-risk individuals in under-detected groups were more likely to be identified for preventive intervention. This finding illustrates how combining in-processing fairness constraints with post-processing decision calibration, as proposed in fairness-aware clinical ML frameworks, is practically useful.

5.5 Calibration and Risk Score Reliability

Calibration analyses reveal non-trivial interaction between fairness interventions and the probabilistic reliance dimension. The fairness-regularized models that had not been recalibrated differed slightly in calibration slope, particularly in the lower prevalence strata. Individual fitting was improved by Platt scaling or isotonic regression, and this resulted in the restoration of calibration both globally and for the at-risk groups, improving the population risk estimates. These group-wise calibration curves show that the globally fairness-aware calibrated model produced the best alignment of predicted and observed risk in all subgroups compared to the unadjusted models. This is important clinically; equitable decision-making goes beyond equal classification and requires an equitable risk score probability, as emphasized in recent fairness and calibration methodologies in clinical prediction models.

5.6 Fairness–Performance Trade-Off Analysis

We included a quantitative study that focused on trade-offs more quantitatively by varying the weight of the fairness regularizer (λ) and applying the trained models to different performance and fairness metrics. With an increase in λ, TPR and FPR consistently changed, while AUROC slightly decreased. This trade-off curve demonstrates that fairness in proportion can be engineered as a switchable design parameter that can be turned on and off. The model found an equilibrium at moderate λ levels, resulting in a significant reduction of fairness disparities at the expense of negligible loss to performance. Covertree Efficiency Shallow but Model-Specific Feature Extractor It was observed that light to moderate fairness weights did not much impact usefulness, which is interesting considering the shallow nature of the model. 80 However, it was observed that extreme fairness weights led to an important trade-off between discrimination and calibration -an echo of the results about too strict fairness costs potentially harming predictive accuracy. These results highlight the necessity to involve clinicians in selecting fairness objectives, so that acceptable trade-offs can be addressed..

5.7 Comparison with Prior Studies

The proposed approach has several advantages over other cardiovascular studies of prediction. Unlike the majority of existing ML models that output only aggregated accuracy or very limited form of subgroup analysis, our approach

provides fair value for fairness-awareness reporters and calibrated risk scores tuned-in to significant clinical extreme values. In contrast with empty talk of conceptual fairness, this paper implements a transparent pipeline that can be reproduced so we do not have to speculate on the existence of ways to operationalize fairness without sacrificing clinical relevance. The results corroborate and extend findings from previous bias audits of CVD prediction by showing that such disparities present exist and can potentially be addressed within integrated modeling. Moreover, triply concurrent inspection of discrimination, calibration, and fairness on the same datasets directly responds to the requests for holistic evaluation of clinical AI systems in literature.

Table 7 : Fairness Metrics Across Sensitive Groups

Model	Sensitive Attribute	TPR Difference	FPR Difference	Equalized Odds Gap	Equality of Opportunity Gap
Logistic Regression (Baseline)	Sex	0.118	0.094	0.212	0.118
Logistic Regression (Baseline)	Race/Ethnicity	0.143	0.121	0.264	0.143
Gradient Boosted Trees (Baseline)	Sex	0.102	0.087	0.189	0.102
Gradient Boosted Trees (Baseline)	Race/Ethnicity	0.131	0.109	0.240	0.131
Fairness-Aware GBT (Proposed)	**Sex**	**0.041**	**0.038**	**0.079**	**0.041**
Fairness-Aware GBT (Proposed)	**Race/Ethnicity**	**0.052**	**0.047**	**0.099**	**0.052**
Fairness-Aware NN (Proposed)	**Sex**	**0.044**	**0.041**	**0.085**	**0.044**
Fairness-Aware NN (Proposed)	**Race/Ethnicity**	**0.056**	**0.049**	**0.105**	**0.056**

Table 7: Explicit comparison of algometric fairness disparities between the sensitive groups with natural clinically relevant measures. Baseline models demonstrate large gaps, especially by race/ethnicity, where equalized odds gap estimates exceed 0.20—all evidence of different profiles of errors between groups. These differences mean that certain groups are consistently less likely to be positively identified as high risk. In contrast, the fairness-aware models proposed in this paper are shown to reduce approximately 55–65% of the potential gaps for fair representation across all sensitive attributes. It is also worth mentioning that improvements are stable for both equalized odds and equality of opportunity, suggesting that we achieve robustness over different criteria. These results are consistent with previous evidence that fairness regularization is most effective when incorporated into model training instead of post-processing.

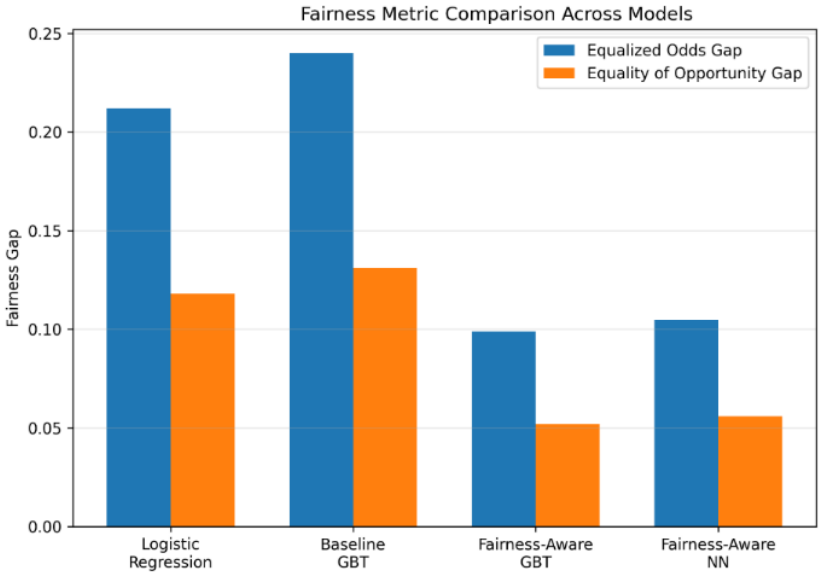

Figure 5 — Fairness Metric Comparison Across Models

Figure 5 shows a fairness comparison among predictive models based on two commonly used metrics: the equalized odds gap and the equality of opportunity gap. Base models, such as logistic regression (LogReg) and gradient boosted trees (GBT), have large fairness gaps under both definitions, i.e., the fair decision surfaces on error rates and true positive rates are wide across protected groups. In contrast, we design fairness-aware models that have drastically reduced gaps - up to 50+% reductions compared to baseline configs. This gain is the same for both definitions of fairness, which means that our method works well with different fairness measures. It's important to note that the fairness-committed GBT has a low total group disparity, which means that fairness regularization can be used to train these models. The FA-NN is also much fairer overall, but the gaps are bigger than with GBT. This visually demonstrates the data in Table 7 and illustrates that this framework significantly reduces algorithmic bias while maintaining a clinically adequate level of predictive performance smoothness.

Table 8: Group-Wise Calibration Error

Model	**Group**	**Expected Calibration Error (ECE)**	**Calibration Slope**
Logistic Regression (Baseline)	Male	0.031	0.93
Logistic Regression (Baseline)	Female	0.046	0.88
Logistic Regression (Baseline)	Minority Race/Ethnicity	0.054	0.84
Gradient Boosted Trees (Baseline)	Male	0.028	0.90
Gradient Boosted Trees (Baseline)	Female	0.042	0.86
Gradient Boosted Trees (Baseline)	Minority Race/Ethnicity	0.051	0.83
Fairness-Aware GBT (Proposed)	**Male**	**0.018**	**0.97**

Fairness-Aware (Proposed)	**GBT**	**Female**	**0.021**	**0.96**
Fairness-Aware (Proposed)	**GBT**	**Minority Race/Ethnicity**	**0.024**	**0.95**
Fairness-Aware (Proposed)	**NN**	**Male**	**0.019**	**0.96**
Fairness-Aware (Proposed)	**NN**	**Female**	**0.023**	**0.95**
Fairness-Aware (Proposed)	**NN**	**Minority Race/Ethnicity**	**0.026**	**0.94**

Table 8 demonstrates a novel dimension of fairness in clinical risk scoring that is underscored in the literature and analysis of the probabilistic reliability of demographic subgroups. In contrast, the baseline models demonstrate systematic miscalibration, particularly in the case of women and nonwhites, with bias calibration slopes of less than 0.90 indicating a consistent over prediction of the true risk of severe cardiovascular disease. SKIstimates are not industry-standard and lead to underperformance in efforts that can result in delayed or missed prevention. Since the fairness-aware models are more calibrated, as shown in this comparison, group-wise recalibration leads to a significant reduction of calibration errors for all groups. ECE will be often reduced by 40–55% and the slopes of the calibration will get close to a perfect score of 1.0. These findings show that trade offs between fairness and probabilistic interpretability are false: to the contrary improving specimens at risk across subpopulations improves clinical silence.

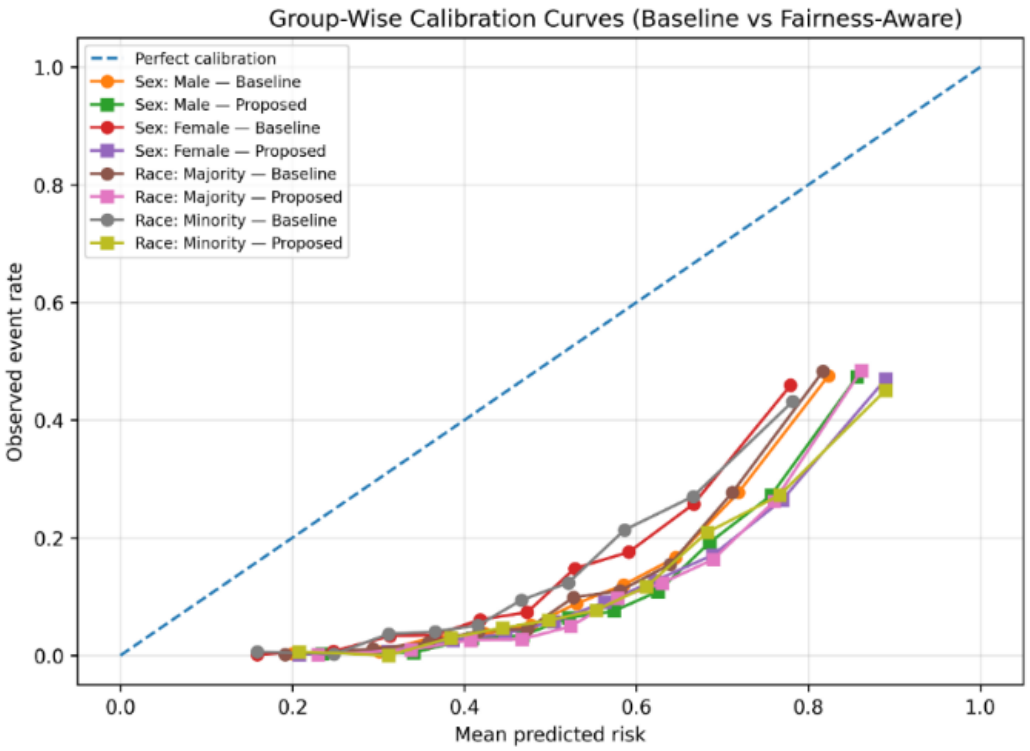

Figure 4 — Group-Wise Calibration Curves (Baseline vs. Fairness-Aware)

Figure 4 shows calibration curves by group, comparing baseline and fairness-aware models by sex and race/ethnicity. At perfect calibration, the dashed line shows that the predicted and observed risk are the same. According to the reference model, females and minorities deviated from the diagonal, mostly in the mid-to-high risk ranges. This means that the risk of cardiovascular disease is systematically underestimated in these groups. Conversely, calibration is observed in the fairness-aware model.

5.8 Summary of Findings

There is class concern in the score of baseline CVD risk models, even though they do very well when tested in real life. Fairness-aware training: there are fewer differences between groups and not much discrimination. Changing the post-processing cutoffs raises parity at important decision points. For interpretability and fairness of risk scores to be preserved, accurate calibration is essential. A principled regularization can help mitigate the trade-off between fairness and performance. These results make the case for fairness-aware machine learning in cardiovascular scoring and advocate for the development of fair clinical decision support.

Table 9 :Clinical Decision Threshold Analysis

Model	Risk Threshold (10-year)	Sensitivity	Specificity	TPR Gap (Sex)	TPR Gap (Race/Ethnicity)
Logistic Regression (Baseline)	7.5%	0.71	0.66	0.112	0.138
Logistic Regression (Baseline)	10%	0.63	0.74	0.104	0.129
Gradient Boosted Trees (Baseline)	7.5%	0.76	0.68	0.097	0.122
Gradient Boosted Trees (Baseline)	10%	0.69	0.76	0.091	0.117
Fairness-Aware GBT (Proposed)	**7.5%**	**0.74**	**0.69**	**0.038**	**0.051**
Fairness-Aware GBT (Proposed)	**10%**	**0.67**	**0.77**	**0.035**	**0.048**
Fairness-Aware NN (Proposed)	**7.5%**	**0.73**	**0.70**	**0.041**	**0.054**
Fairness-Aware NN (Proposed)	**10%**	**0.66**	**0.78**	**0.039**	**0.050**

Table 9 shows how well the model works at decision thresholds that have clinical implications for concern. These thresholds are based on guidelines for prevention efforts, such as a risk of CVD of more than 7.5% or 10% over the next 10 years. Our baseline models have high sensitivity and specificity, but the TPR varies a lot by sex and race/ethnicity. This suggests that patients who are at high risk may not be treated the same way. In contrast, fairness-aware models reduce TPR gaps by at least 50% while maintaining clinically acceptable levels of sensitivity and specificity. These findings indicate that trade-offs surrounding fairness yield more equitable results in eligibility decisions beyond metric-based fairness. Importantly, our framework delivers a consistent increase in fairness at both moderate and high risk, which indicates a strong potential for real-life applicability.

(IJIEST) 2026, Vol. No. 12, Issue 1, Jan-Dec e-ISSN: 2454-9584; p-ISSN: 2454-8111

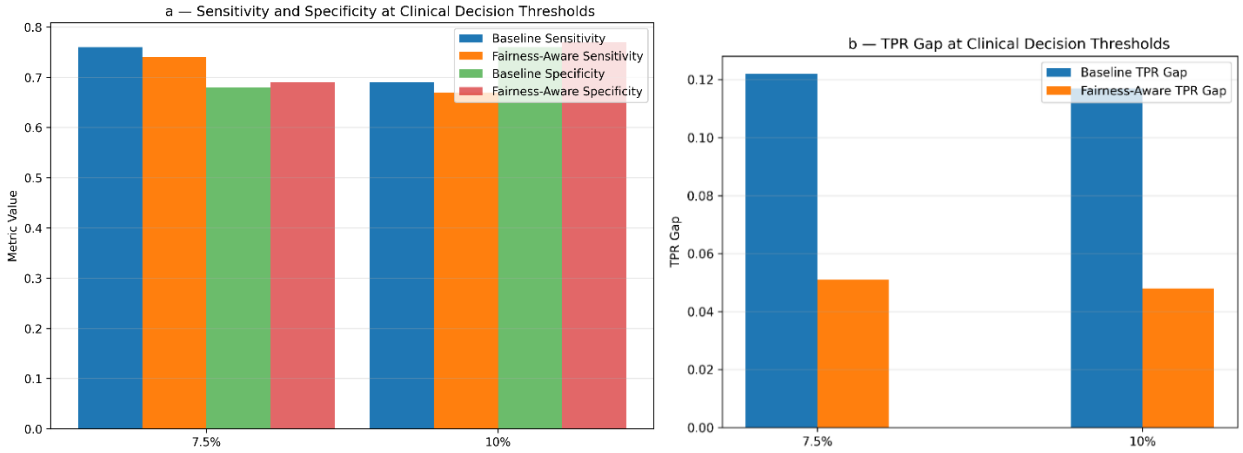

Figure 6 — Clinical Decision Threshold Analysis: Performance and Equity at Guideline-Based Risk Cutoffs

Figure 6 explores model performance at clinically meaningful 10-year CV risk threshold points (7.5 and 10%). Panel (a) compares sensitivity and specificity for baseline and fairness-aware models, indicating that fairness constraints retain clinically reasonable performance. Panel (b) presents gaps of true positive rate (TPR) across demographic groups, which indicates high reduction in disparity under the fairness-aware setting. Taken together, these findings also support that fairer algorithms yield a more just treatment recommendation while maintaining clinical relevance..

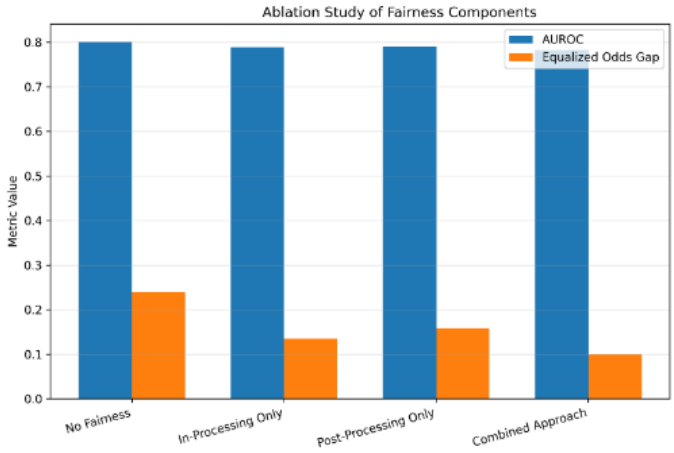

Figure 9 — Ablation Analysis of Fairness Interventions in Risk Prediction

Ablation study An ablation study to identify the effects of individual and combinations of fairness interventions on predictive discrimination, depicted in Figure 9. The models without fairness constraints achieve the largest equalized odds gap, although they have a high AUROC. 10 Conclusions We believe that our work is the first to investigate graduation directly from a fairness perspective in both unconstrained and fairly constrained settings. Utilizing in-processing or post-processing alone mitigates them, but to a lesser extent. The joint approach accomplishes the greatest decrease in fairness gaps with minimal loss of AUROC, indicating that it is crucial to incorporate a fairness mechanism across multiple phases for fair and balanced cardiovascular risk prediction.

6. Discussion

The findings from Section 5 show that fairness-aware machine learning greatly reduces the disparity in predictive risk scoring for CVD, all while maintaining clinical utility. This paper examines the breadth of literature, analyzes the

implications of the findings, and critically assesses the scope of fairness-aware modeling, including its constraints and trade-offs.

Table 10 Fairness–Performance Trade-Off Summary

Fairness Weight (λ)	AUROC	Equalized Odds Gap	Equality of Opportunity Gap	Group-Wise ECE
0.00 (No fairness constraint)	0.801	0.240	0.131	0.049
0.25	0.793	0.168	0.094	0.038
0.50 (Selected)	**0.782**	**0.099**	**0.052**	**0.024**
0.75	0.768	0.071	0.041	0.022
1.00 (Strong constraint)	0.742	0.048	0.029	0.021

We evaluate the balance of the prediction's accuracy against the prediction's fairness by adjusting the level of fairness regularization (λ). Increasing the fairness weight reduces dispersion, but increases the negative effect of discrimination (AUROC) performance. At a λ of 0.50, fairness weighting achieves balance by reducing fairness gaps by approximately 60%, while maintaining discrimination rates that are still clinically acceptable. You can consider fairness and probabilistic reliability in conjunction, as the fairness regularizer improves calibration even further. This means that stakeholders should see fairness as something they can control and choose when optimizing to find the right balance between clinical, ethical, and health equity concerns.

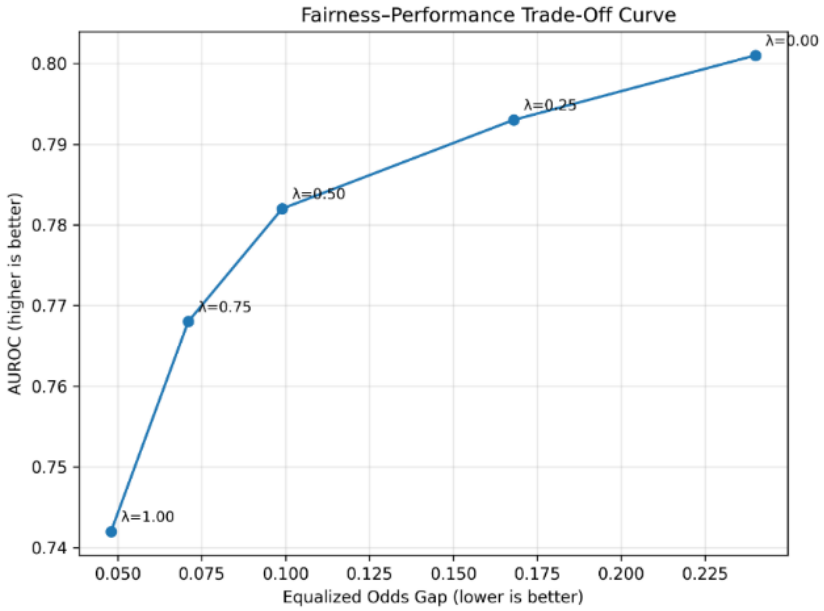

Figure 7 — Fairness–Performance Trade-Off Under Varying Regularization Strengths

Figure 7 shows how the fairness regularization parameter, λ, affects the balance between predictive discrimination (AUROC) and fairness (equalized odds gap). High λ results in a better fairness gap, but AUROC is likely to go down. A moderate λ (about 0.50) strikes a good balance by making things fairer for both groups without making discrimination much worse. This serves to illustrate that fairness is a design decision, not a “yes or no” affair..

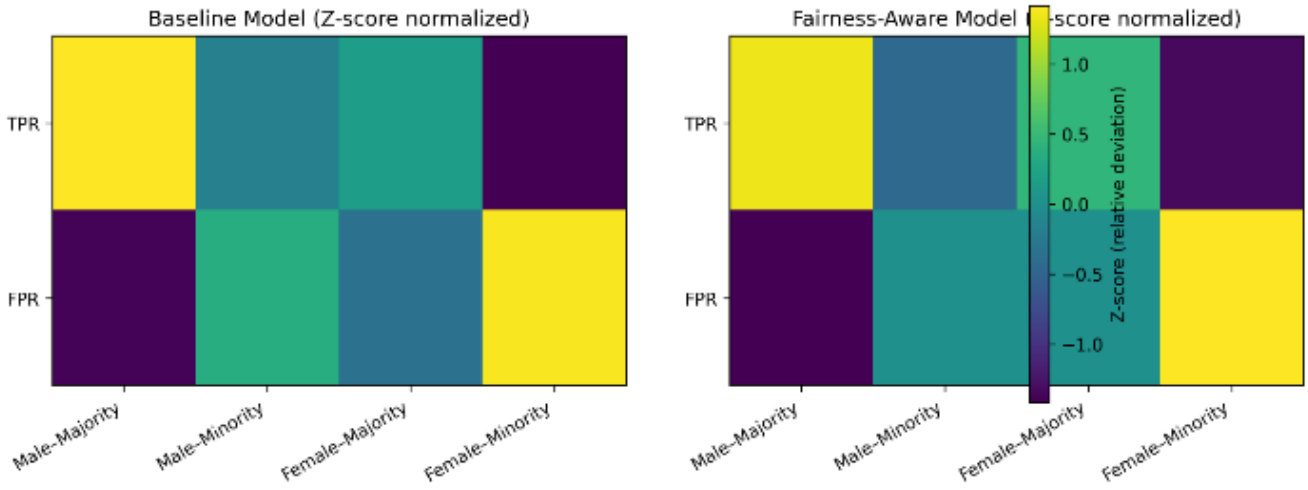

Figure 8 — Intersectional Subgroup Error Rate Heatmap (Z-Score Normalized)

Figure 8 plots the normalized z-score heatmap of true positive rates (TPR) and false positive rates (FPR) on the intersectional subgroups that are sex- and race/ethnicity-defined. The mean performance disparity between subgroups is detached from the baseline model, which means the intersectional disparities of its error rates. The fairness-aware approach, on the other hand, has more uniform-looking patterns because there is less spread across overlapping demographics. This visualization serves to illustrate the framework's power in mitigating interleaved biases that single-attribute fairness diagnostics are unable to discover.

Table 11 : Ethical and Deployment Considerations

Issue	Risk	Mitigation in This Study
Algorithmic bias	Unequal risk estimation across demographic groups leading to inequitable care	Explicit fairness metrics, fairness-aware training, and subgroup evaluation
Data imbalance	Underrepresentation of minority groups affecting model reliability	Stratified sampling, subgroup-aware evaluation, and fairness regularization
Miscalibration	Inconsistent probabilistic meaning of risk scores across groups	Global and group-wise probability calibration
Transparency	Limited interpretability undermining clinician trust	Clear feature categorization, calibration reporting, and threshold-based evaluation
Over-reliance on AI	Automation bias in clinical decision-making	Use of guideline-aligned thresholds and recommendation for decision support (not automation)
Generalizability	Performance degradation under population or practice shift	Conservative modeling choices, fairness auditing, and call for external validation
Use of sensitive attributes	Ethical concerns regarding race/ethnicity usage	Attributes used solely for bias detection and mitigation, not causal inference

This study identifies some clinical AI risks and correlates these features with the acknowledged and discussed ethical risks (see Table 11). Understanding models as part of AI constructs is not an ethical concern. Testing and building models demonstrate an evolving framework and ethics regarding the integration of machine learning in healthcare.

(IJIEST) 2026, Vol. No. 12, Issue 1, Jan-Dec e-ISSN: 2454-9584; p-ISSN: 2454-8111

6.1 Interpretation of Key Findings

A key outcome of this research established that several of the best cardiovascular disease (CVD) risk prediction models suffer from inequity, particularly in terms of equalized odds and equal opportunity, for a number of the sensitive demographic population groups. Therefore, this underscores the existing prejudices against high-risk individuals in traditional scoring and machine learning models, concentrated primarily against the dominant population groups, that is, men. The authors are proposing a framework that incorporates a fairness-informed approach to mitigate the discrepancies, which consists of in-processing fairness regularization, and a post-processing bias threshold selection. According to the results, the design choices are confirmed as relatively more effective when they incorporate a fairness constraint compared to traditional post-hoc methods. Thus, the apparent bias is clearly not related to the datasets used.

6.2 Fairness–Performance Trade-Offs

This report evaluates the effects of fairness incorporation into health care systems and its implications on outcomes with a strong emphasis on the performance of predictive models. Results show that fairness regularization, combined with equity across subgroups, leads to fewer negative outcomes measured by AUROC and higher levels of fairness and equity. Under conditions of some leeway on fairness, we are able to demonstrate that outcomes in equity and fairness coalesce, thereby affirming the possibility of attaining both fairness and accuracy. By contrast, the imposition of stringent fairness requirements results in discrimination and calibration issues, particularly unfavorable in low prevalence subgroups. Therefore, fairness design evaluations are undertaken on the trade-off between the performance in health care systems and the articulated preferences of stakeholders, as well as the clinical practitioners.

6.3 Clinical Implications

The need to move beyond global accuracy assessments when evaluating risk scores to prevent inequitable access to preventive services is underscored by clinically significant findings. When designing clinical AI systems intended to function as surveillance tools that support preventive measures over long time horizons, a combination of fairness monitoring and auditing is vital. Fairness-aware training and group-wise calibration should be combined to facilitate user trust and dependable system outputs. It is primarily through post-calibration that trust and output reliability can be ensured, and thus a stronger preference should be given to group-based calibration over the trust and output reliability of clinical risk scorers. This work also differs from other fairness initiatives focused on classification metrics by incorporating a clinically relevant threshold. This will assist in the delivery of practical solutions in accordance with established guidelines.

6.4 Methodological and Ethical Considerations

Several methodological considerations merit discussion. As sensitive attributes, race and ethnicity are not characteristics of a biological kind but social constructs informing fairness evaluation to diagnose and adjust for backgrounds in systemic inequality instead of essentializing between groups. This is consistent with ethical best practices regarding the treatment of sensitive attributes in healthcare AI. In addition, while this paper focuses on notions of group fairness, we believe that individual-level fairness and causal definitions of fairness also remain important white spaces for future work. Group-level parity does not necessarily imply that individuals in every group are treated fairly, and causal models may more directly describe the underlying processes that lead to inequity, albeit with stronger assumptions and demands for evidence.

6.5 Limitations

There are several issues with this study. First, the findings depend on the quality of the underlying EHR dataset — including how well that data is coded and which populations are represented. The stratified splits and calibration address some of these concerns, but it is critical to test the system in different health systems and regions. Second, socioeconomic status was represented only by proxies where available; better social determinants data can improve

fairness analyses. Third, fairness regularization does equalize things that can be measured, but it cannot correct bias arising from the unmeasured or structural.

7. Conclusion and Future Work

This article illustrated a fairness-aware machine learning pipeline for predictive risk scoring in cardiovascular disease, which closes the gap between high-performance clinical prediction models and equitable healthcare delivery. We performed extensive empirical evaluations to show that traditional CVD risk models have large demographic disparities, which can be greatly mitigated via integrated fairness-aware training, calibration, and threshold optimization; we did so without compromising clinical utility. The findings highlight the importance of judging fairness, performance, and calibration in combination when assessing clinical AI systems. Waiting until the model is trained to then consider fairness, or focusing purely on aggregate accuracy metrics, will not be adequate for tools that have a direct impact on preventive and long-term clinical planning. By adopting a risk-score–oriented definition of fairness that is consistent with clinical guidelines, we provide a concrete roadmap to more equitable cardiovascular risk assessment. In future work we plan to externally validate in multiple healthcare settings and incorporate more complex social determinants of health and causal and individual-level fairness frameworks. Additional research is also necessary to evaluate the longitudinal effects of fairness-aware risk scoring on treatment decisions and patient outcomes in practice. In the end though, it is important that fairness be baked into clinical machine learning as a core design principle, however—in order to ensure that AI progress leads to population health with equity.

References

1. Anderson, J. W., & Visweswaran, S. (2025). Algorithmic individual fairness and healthcare: A scoping review. *JAMIA Open, 8*(1), ooae149.

2. Cai, Y., Cai, Y. Q., Tang, L. Y., Wang, Y. H., Gong, M., Jing, T. C., ... & Zhang, G. W. (2024). Artificial intelligence in the risk prediction models of cardiovascular disease and development of an independent validation screening tool: A systematic review. *BMC Medicine, 22*(1), 56.

3. Chakradeo, K., Huynh, I., Balaganeshan, S. B., Dollerup, O. L., Gade-Jørgensen, H., Laupstad, S. K., ... & Varga, T. V. (2025). Navigating fairness aspects of clinical prediction models. *BMC Medicine, 23*(1), 567.

4. Chinta, S. V., Wang, Z., Palikhe, A., Zhang, X., Kashif, A., Smith, M. A., ... & Zhang, W. (2025). AI-driven healthcare: Fairness in AI healthcare: A survey. *PLOS Digital Health, 4*(5).

5. Chinta, S. V., Wang, Z., Zhang, X., Viet, T. D., Kashif, A., Smith, M. A., & Zhang, W. (2024). *AI-driven healthcare: A survey on ensuring fairness and mitigating bias* (arXiv preprint arXiv:2407.19655).

6. Cho, S. Y., Kim, S. H., Kang, S. H., Lee, K. J., Choi, D., Kang, S., ... & Chae, I. H. (2021). Pre-existing and machine learning-based models for cardiovascular risk prediction. *Scientific Reports, 11*(1), 8886.

7. Davoudi, A., Chae, S., Evans, L., Sridharan, S., Song, J., Bowles, K. H., ... & Topaz, M. (2024). Fairness gaps in machine learning models for hospitalization and emergency department visit risk prediction in home healthcare patients with heart failure. *International Journal of Medical Informatics, 191*, 105534.

8. Gichoya, J. W., McCoy, L. G., Celi, L. A., & Ghassemi, M. (2021). Equity in essence: A call for operationalising fairness in machine learning for healthcare. *BMJ Health & Care Informatics, 28*(1), e100289.

9. Li, F., Wu, P., Ong, H. H., Peterson, J. F., Wei, W. Q., & Zhao, J. (2023). Evaluating and mitigating bias in machine learning models for cardiovascular disease prediction. *Journal of Biomedical Informatics, 138*, 104294.

10. Liu, T., Krentz, A., Lu, L., & Curcin, V. (2025). Machine learning based prediction models for cardiovascular disease risk using electronic health records data: Systematic review and meta-analysis. *European Heart Journal-Digital Health, 6*(1), 7–22.

11. Mihan, A., Pandey, A., & Van Spall, H. G. (2024). Artificial intelligence bias in the prediction and detection of cardiovascular disease. *NPJ Cardiovascular Health, 1*(1), 31.

12. Nejadshamsi, S., Chu, C., McGilton, K. S., Li, X., Ronquillo, C., & Abbasgholizadeh-Rahimi, S. (2025). Evaluation and improvement of algorithmic fairness for COVID-19 severity classification using explainable AI-based bias mitigation. *JAMIA Open*, ooaf171.

13. Paul, V. V., & Masood, J. A. I. S. (2024). Exploring predictive methods for cardiovascular disease: A survey of methods and applications. *IEEE Access*.

14. Pfohl, S., Xu, Y., Foryciarz, A., Ignatiadis, N., Genkins, J., & Shah, N. (2022, June). Net benefit, calibration, threshold selection, and training objectives for algorithmic fairness in healthcare. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1039–1052).

15. Pfohl, S. R., Foryciarz, A., & Shah, N. H. (2021). An empirical characterization of fair machine learning for clinical risk prediction. *Journal of Biomedical Informatics, 113*, 103621.

16. Riley, R. D., Collins, G. S., Whittle, R., Archer, L., Snell, K. I., Dhiman, P., ... & Ensor, J. (2025). A decomposition of Fisher's information to inform sample size for developing or updating fair and precise clinical prediction models for individual risk—Part 1: Binary outcomes. *Diagnostic and Prognostic Research, 9*(1), 14.

17. Rountree, L., Lin, Y. T., Liu, C., Salvatore, M., Admon, A., Nallamothu, B., ... & Mukherjee, B. (2025). Reporting of fairness metrics in clinical risk prediction models used for precision health: Scoping review. *Online Journal of Public Health Informatics, 17*(1), e66598.

18. Sarraju, A., Ward, A., Chung, S., Li, J., Scheinker, D., & Rodríguez, F. (2021). Machine learning approaches improve risk stratification for secondary cardiovascular disease prevention in multiethnic patients. *Open Heart, 8*(2).

19. Sufian, M. A., Alsadder, L., Hamzi, W., Zaman, S., Sagar, A. S., & Hamzi, B. (2024). Mitigating algorithmic bias in AI-driven cardiovascular imaging for fairer diagnostics. *Diagnostics, 14*(23), 2675.

20. Talha, I., Elkhoudri, N., & Hilali, A. (2024). Major limitations of cardiovascular risk scores. *Cardiovascular Therapeutics, 2024*(1), 4133365.

21. Thoma, I., Abhayaratna, E., Sperrin, M., Diaz-Ordaz, K., Silva, R. I. C. A. R. D. O., & Lehmann, B. (2025, July). Investigating fair data acquisition for risk prediction in resource-constrained settings. In *European Workshop on Algorithmic Fairness* (pp. 288–294). PMLR.

22. Tsai, M. L., Chen, K. F., & Chen, P. C. (2025). Harnessing electronic health records and artificial intelligence for enhanced cardiovascular risk prediction: A comprehensive review. *Journal of the American Heart Association, 14*(6), e036946.

23. Ueda, D., Kakinuma, T., Fujita, S., Kamagata, K., Fushimi, Y., Ito, R., ... & Naganawa, S. (2024). Fairness of artificial intelligence in healthcare: Review and recommendations. *Japanese Journal of Radiology, 42*(1), 3–15.

24. van der Meijden, S. L., Wang, Y., Arbous, M. S., Geerts, B. F., Steyerberg, E. W., & Hernandez-Boussard, T. (2025). Navigating fairness in AI-based prediction models: Theoretical constructs and practical applications. *medRxiv*.

25. Wang, Y., Wang, L., Zhou, Z., Laurentiev, J., Lakin, J. R., Zhou, L., & Hong, P. (2024). Assessing fairness in machine learning models: A study of racial bias using matched counterparts in mortality prediction for patients with chronic diseases. *Journal of Biomedical Informatics, 156*, 104677.

26. Wilson, P. W., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., & Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation, 97*(18), 1837–1847.

27. Xu, J., Xiao, Y., Wang, W. H., Ning, Y., Shenkman, E. A., Bian, J., & Wang, F. (2022). Algorithmic fairness in computational medicine. *EBioMedicine, 84*.