

A Multi-Site Fairness-Aware Machine Learning Framework for Mitigating Racial and Socioeconomic Bias in Clinical AI Across U.S. Hospitals

¹Joynob Sultana

¹Jsultana@troy.edu

²Tanvir Ahmed Abir

²tanvir0abir1999@gmail.com

³Nabila Tuz Johora

³nabilajohora24@gmail.com

⁴Fahim Abrar

⁴fahim9431@gmail.com

⁵Majharul Islam Shanto

⁵mshanto.aiub@gmail.com

⁶Aftaha Ahmed

⁶aahmed16@lamar.edu

ABSTRACT

Background: Clinical AI models frequently perpetuate systemic biases, resulting in inequitable health outcomes across different patient populations. U.S. hospitals lack robust frameworks that address multiple fairness constraints simultaneously across racially and economically diverse patient groups.

Objective: We developed and validated a fairness-aware machine learning framework that reduces bias across racial groups (White, Black, Hispanic) and socioeconomic strata (low, medium, high) while preserving predictive accuracy.

Methods: We conducted a retrospective cohort study using the MIMIC-IV database with 48-hour ICU time-series data to predict in-hospital mortality. Our hybrid adversarial multi-objective optimization model compared three fairness approaches: Adversarial Debiasing, Exponentiated Gradient with Fairness Constraints, and our proposed Multi-Objective Grid-Search Reweighting (MOGR) framework.

Results: The MOGR framework achieved Pareto-optimal solutions, reducing Equalized Odds difference by 72% and Demographic Parity

difference by 81% compared to the unconstrained baseline. The model maintained an AUC of 0.84, with minimal performance decline under 0.02. Disparities for Black patients with low socioeconomic status proved most resistant to mitigation but still showed substantial improvement.

Conclusion: A structured approach using multiple fairness constraints significantly reduces group-level disparities without clinically meaningful performance loss. This framework provides hospitals with a deployable pathway toward achieving health equity across diverse patient populations in clinical settings.

Keywords: Clinical AI fairness, Bias mitigation machine learning, Health equity, Algorithmic disparities, Racial bias in healthcare AI, Socioeconomic health disparities

I INTRODUCTION

1.1 Clinical AI and the Equity Problem

Hospitals across the United States deploy machine learning models to predict patient outcomes, allocate ICU beds, recommend treatments, and identify patients at risk for readmission. These systems promise efficiency, objectivity, and

consistency. But the reality is more complex. They deliver efficiency and consistency while simultaneously amplifying historical inequities embedded deep within patient data [Obermeyer et al., 2019].

Consider what happened at a major health system using an algorithm to allocate resources to high-risk patients. The model performed with 95% accuracy overall. Yet when researchers examined predictions by race, the discrepancy became clear. Black patients received risk scores 30% lower than White patients with identical clinical presentations [Obermeyer et al., 2019]. The algorithm was not designed to discriminate against. It simply learned patterns from historical data where Black patients received fewer tests, fewer diagnoses, and less aggressive treatment due to systemic racism in medicine. The training data reflected decades of disparate care. The algorithm inherited that disparity.

When an AI model performs worse for Black patients than White patients, when it systematically underpredicts risk for low-income populations, the gap reflects not individual bias, but structural bias encoded in hospital records. The problem is quantifiable. You can define health equity mathematically as minimizing prediction error equally across racial and socioeconomic groups. Most clinical AI systems fail that test. The conditional error rate, which measures how often the model makes mistakes within each protected population, differs significantly across groups [Buolamwini & Gebru, 2018]. For mortality prediction specifically, this means the model's ability to identify patients who will die varies meaningfully by race and income level.

1.2 Where Bias Comes From: Three Distinct Sources

Bias in clinical AI emerges from three sources, and you need to understand each one because they require different solutions.

Label bias occurs when the ground truth itself is corrupted. Black patients have historically received fewer diagnostic tests, fewer specialist referrals, and longer wait times for procedures. This creates what researchers call "implicit censoring" [Mitchell, 2019]. A patient with heart disease who never receives an EKG has no diagnosis in the medical record, even though the disease exists. When you train an AI model on these records, the algorithm learns to underestimate disease burden in Black patients because the training data reflects underdiagnosis, not true health. The label itself is wrong.

Selection bias emerges when hospitals enroll different populations into studies or treatment protocols. Imagine a registry that captures critically ill White patients but misses equally sick Black patients because they are present at different hospitals or arrive later when ICU beds are full. The training set no longer represents reality. It represents the way the healthcare system reaches different populations.

Algorithmic unfairness is distinct from the first two. The labels are correct. The selection is representative. But the model itself learns to make systematically different predictions for identical patients based on race or income. Formally, this means the predicted probability of an outcome differs by protected attribute, even after controlling for clinical variables: $P(\hat{Y}=1|A=a, X, Y=y) \neq P(\hat{Y}=1|A=b, X, Y=y)$, where A represents race or SES, X represents clinical features, and $\hat{Y}$ is the predicted outcome [Hardt et al., 2016]. The model learns a race-dependent decision boundary. Each source requires a different mitigation strategy. You cannot address label bias by reweighting training data. You cannot fix selection bias by adjusting the loss function. Algorithmic fairness interventions only work when the data itself is sound.

1.3 Race and Socioeconomic Status in the U.S. Hospital Context

We focus on race (White, Black, Hispanic) and socioeconomic status measured through insurance type and median income by zip code because federal law and medical ethics demand it.

The disparate impact doctrine in employment law extends to healthcare. When a hospital uses an algorithmic tool to make clinical decisions, that tool cannot have a disparate impact on protected classes, even if no one programmed discrimination into it [Feldman et al., 2015]. Race and ethnicity are protected attributes under federal civil rights law. Courts have recognized that proxy variables like zip code can encode racial segregation and historical redlining. Socioeconomic status, while not a protected class under the same statutes, correlates strongly with healthcare access, medication adherence, and health literacy. Clinically and ethically, you cannot ignore it.

In U.S. hospitals specifically, the context matters. Unlike countries with single-payer systems, American hospital networks remain fragmented by geography, insurance, and wealth. This fragmentation creates predictable disparities. Black and Hispanic patients are overrepresented in safety-net hospitals with fewer resources. They are underrepresented in academic medical centers with newer technology. They receive care in systems with different standards, different data quality, and different clinical cultures. An AI model trained on one hospital network may not generalize to another. More importantly, it may generalize the disparities.

1.4 The Gap in Current Research

Most fairness research in clinical AI focuses on a single metric. Some papers optimize for Equalized Odds, where the false positive rate and false negative rate are equal across groups. Others target Demographic Parity, where the positive prediction rate is identical regardless of protected attributes. A few examine Calibration, ensuring predicted risk accurately reflects actual risk within each group [Kleinberg et al., 2018].

But you cannot satisfy all fairness definitions simultaneously. This is not a limitation of current algorithms. It is a mathematical impossibility. Kleinberg and colleagues proved that under most real-world conditions, satisfying multiple fairness criteria at once is impossible [Kleinberg et al., 2018]. If you equalize false positives across groups, true positive rates will differ. If you equalize positive prediction rates, the model must be less accurate for at least one group.

Hospitals today face this impossibility without guidance. Clinical decision support systems using AI must navigate competing fairness constraints while maintaining adequate predictive performance for patient safety. No framework exists that systematically evaluates these trade-offs in a clinical context.

1.5 Our Contributions

This paper makes three concrete contributions.

First, we propose a multi-objective optimization framework that balances multiple fairness constraints simultaneously. Rather than optimizing for one metric, we construct a Pareto frontier showing all efficient trade-offs between fairness and accuracy.

Second, we examine intersectional bias. Most fairness research treats race and SES as separate attributes. We study their intersection. What happens to Black patients with low socioeconomic status together? The disparities are often larger than either attribute alone.

Third, we provide a clinical pathway forward. We validate our framework on real ICU data, measure whether it reduces disparities, and quantify the trade-offs in predictive performance. We translate fairness mathematics into actionable guidance for hospitals deploying AI.

II RELATED WORK

2.1 Fairness-Aware Machine Learning in Medicine

Researchers have developed three broad categories of fairness interventions that operate at various stages of the machine learning pipeline [Calmon et al., 2017].

Pre-processing approaches modify the training data before feeding it into the model. Data augmentation adds synthetic examples to underrepresented groups. Reweighting assigns higher importance to misclassified samples from disadvantaged populations. You tell the learning algorithm to pay more attention to errors affecting protected groups [Kamiran & Calders, 2012]. These methods are attractive because they collaborate with any off-the-shelf model. You do not need to change the algorithm itself. The tradeoff is limited control over which fairness definition you satisfy and how much accuracy you lose.

In-processing techniques embed fairness constraints directly into the learning objective. Adversarial debiasing introduces a secondary neural network that tries to predict the protected attribute from model predictions. The main model learns to make predictions that fool this adversary, forcing the model to ignore race or income [Zhang et al., 2018]. Constraint-based methods add fairness as a penalty term during training. The model minimizes both prediction error and fairness violation simultaneously [Hardt et al., 2016]. These approaches give you fine-grained control over the fairness metric but require modifying the algorithm itself and may increase computational cost.

Post-processing methods adjust model outputs after training. Threshold optimization sets different decision boundaries for diverse groups so that false positive rates or positive prediction rates match. You keep the trained model unchanged but apply group-specific rules when making final predictions [Hardt et al., 2016]. This approach is simple and interpretable. But it raises clinical and ethical questions. Should you really recommend different treatment thresholds based on race?

In clinical AI specifically, most work focuses on single-site datasets and single-fairness metrics. Mitchell and colleagues used data from a major U.S. health system and found that standard algorithms amplified existing biases in cardiovascular risk assessment [Mitchell, 2019]. Obermeyer's study of a commercial risk prediction algorithm revealed significant racial disparities in how the system allocated resources [Obermeyer et al., 2019]. But these studies typically apply one fairness intervention and measure one fairness outcome.

2.2 The Fairness-Utility Trade-off

You cannot have perfect fairness and perfect accuracy. This is not a practical limitation. It is a mathematical reality [Kleinberg et al., 2018].

Early work showed this trade-off is non-linear. When you first implement fairness constraints, accuracy drops slowly. There was a slight increase in fairness costs, almost nothing in predictive performance. But as you push fairness further, the penalty accelerates. Beyond a certain point, satisfying another percent of fairness loss requires one percent accuracy loss or more [Calmon et al., 2017]. Understanding this curve matters for deployment decisions.

Hardt and colleagues quantified this effect using income data. Their model could achieve demographic parity or equalized odds but not both. The effort to satisfy both fairness criteria simultaneously reduced model performance below clinically acceptable thresholds [Hardt et al., 2016]. Corbett-Davies extended this analysis across multiple datasets and fairness definitions, showing the non-linear trade-off holds generally [Corbett-Davies et al., 2017].

For clinical AI specifically, acceptable accuracy loss remains undefined. Hospital standards require AUC above 0.80 for mortality prediction, but you find no published guidance on acceptable fairness improvement per unit accuracy lost. This

ambiguity means hospitals cannot easily decide which approach to adopt.

2.3 Intersectional Bias

Most fairness research treats race and socioeconomic status as separate attributes. You examine whether the model is fair to Black patients as a group and separately whether it is fair to low-income patients as a group. But people occupy multiple identities simultaneously [Buolamwini & Gebru, 2018].

A Black woman earning 30,000 dollars per year is not a Black patient plus a low-income patient. She is at the intersection, and disparities compound. Research on intersectional discrimination shows that protection at the margin does not protect intersectional groups. A fair model for Black patients overall might systematically disadvantage Black women. A fair model for [Buolamwini & Gebru, 2018].

Only a handful of papers examine intersectional bias in clinical AI. Buolamwini and Gebru demonstrated that commercial facial recognition systems showed error rates up to 34% for darker skinned women while maintaining under 1% error for lighter skinned men [Buolamwini & Gebru, 2018]. This intersectional analysis revealed disparities invisible when examining race or gender separately.

We extend this approach to clinical mortality prediction by examining how race and SES combine. Do the largest disparities appear at specific intersections? Can fairness interventions simultaneously protect all groups? These questions remain unanswered in hospital AI.

III MATHEMATICAL FORMULATION OF FAIRNESS IN CLINICAL AI

To build a fair clinical AI system, fairness must be defined with mathematical precision. Without formal definitions, it remains an aspirational slogan rather than an operational requirement. This section formalizes fairness in the context of mortality

prediction and establishes the optimization framework for balancing it against predictive performance.

3.1 The Clinical Dataset and Problem Setup

Let the clinical dataset be defined as $D = \{(x_i, a_i, s_i, y_i)\}_{i=1}^N$, where each patient record contains four components:

The clinical features $x_i \in \mathbb{R}^d$ comprise vital signs, laboratory values, and 48-hour ICU time-series measurements.

The protected attribute $a_i \in \{0, 1, 2\}$ encodes race as White, Black, or Hispanic, respectively.

The socioeconomic attribute $s_i \in \{0, 1, 2\}$ represents income level as low, medium, or high.

The binary outcome $y_i \in \{0, 1\}$ indicates in-hospital mortality.

A classifier f_θ parameterized by θ produces predictions $\hat{y} = f_\theta(x)$. The fundamental question is whether the model generates systematically different predictions for patients with identical clinical presentations who differ only in protected attributes [Hardt et al., 2016].

3.2 Group Fairness Metrics Defined

Fairness has multiple, mathematically distinct definitions. Because you cannot satisfy them all simultaneously, it is critical to understand the implications of each.

Demographic Parity (DP) requires equal positive prediction rates across all protected groups, regardless of true outcome prevalence. Formally, $P(\hat{y}=1|A=a) = P(\hat{y}=1|A=b)$ for all groups a and b . The difference metric $\Delta_{DP} = |P(\hat{y}=1|A=a) - P(\hat{y}=1|A=b)|$ quantifies the violation. While DP is the easiest to audit, it has a critical clinical limitation: if true mortality rates differ across groups due to systemic healthcare disparities, enforcing DP forces the model to systematically underestimate risk for disadvantaged populations [Corbett-Davies et al., 2017].

Equalized Odds (EO) imposes a stronger condition, requiring equal error rates across groups within each outcome class. That is, among patients who died ($Y=1$), the true positive rate should be identical across races; among survivors ($Y=0$), the false positive rate should match. Formally, $P(y^*=1|A=aY=y)=P(y^*=1|A=bY=y)$ for all a, b and $y \in \{0, 1\}$. The violation is captured by $\Delta_{EO} = \max_y |P(y^*=1|A=aY=y) - P(y^*=1|A=bY=y)|$. EO is clinically preferable because it permits the model to reflect genuine differences in outcome prevalence while demanding equivalent performance within each stratum [Hardt et al., 2016].

Intersectional Parity (IP) extends fairness to combined subgroups. Define $G=(A, S)$ as the Cartesian product of race and socioeconomic status, yielding nine distinct groups. The intersectional Demographic Parity difference is $\Delta_{IP \setminus DP} = \max_{(g, h)} |P(y^*=1|G=g) - P(y^*=1|G=h)|$. This metric reveals compounding disparities invisible to single-attribute analysis; a model appearing fair to Black patients and fair to low-income patients may still systematically disadvantage the intersection of both [Buolamwini & Gebru, 2018].

3.3 The Multi-Objective Optimization Problem

The learning objective must minimize prediction error while satisfying fairness constraints, which is fundamentally a multi-objective problem. The binary cross-entropy loss $L_{BCE}(\theta)$ measures predictive accuracy. Let $F(\theta) = [\Delta_{DP \setminus Race}, \Delta_{EO \setminus Race}, \Delta_{DP \setminus SES}, \Delta_{EO \setminus SES}, \Delta_{IP \setminus DP}, \Delta_{IP \setminus EO}]^T$ represent the vector of all fairness violations. The constrained optimization problem is:

$$\min_{\theta} L_{BCE}(\theta) \text{ "s.t." } F_j(\theta) \leq \epsilon_j \text{ "for all " } j$$

This seeks the most accurate model whose fairness violations remain below specified thresholds ϵ_j . However, because fairness constraints conflict with each other (e.g., satisfying Equalized Odds may

violate Demographic Parity), you cannot choose thresholds freely [Kleinberg et al., 2018]. To navigate these conflicts, we reformulate using a Lagrangian dual relaxation. Introduce a penalty vector $\lambda \geq 0$, where each λ_j weights the importance of satisfying the j -th fairness constraint:

$$\max_{(\lambda \geq 0)} \min_{\theta} L_{BCE}(\theta) + \lambda^T (F(\theta) - \epsilon)$$

This min-max formulation generates a family of solutions parameterized by λ . By sweeping λ across a grid, we construct the Pareto frontier of non-dominated solutions, revealing the trade-offs where fairness cannot improve without accuracy loss, and vice versa [Calmon et al., 2017].

3.4 The MOGR Framework: Three-Stage Architecture

We propose the Multi-Objective Grid-Search Reweighting (MOGR) framework, a hybrid architecture combining three complementary bias mitigation strategies.

Stage 1: Representation Learning. An encoder $E(x)$ learns a representation of clinical features that is predictive of mortality but independent of protected attributes. An adversarial discriminator D tries to predict race and SES from these representations. The encoder "wins" if it confuses the discriminator, forcing it to strip away information correlated with protected attributes. The loss function is:

$$L_{Adv} = L_{BCE}(f(\theta), (E(x))y) - \beta \cdot L_{CE}(D(\theta), (E(x))(a, s))$$

The first term encourages accurate prediction, while the second (scaled by β) penalizes successful attribute prediction. This stage removes algorithmic unfairness by learning a fair representation [Zhang et al., 2018].

Stage 2: Fairness-Constrained Classifier. We apply the Exponentiated Gradient algorithm on top of the debiased representations $E(x)$. This algorithm adjusts sample weights to satisfy fairness constraints; patients from underrepresented groups

or whose errors are large receive higher weights. The classifier learns to minimize weighted loss subject to achieving target fairness levels, managing any residual disparities from Stage 1 [Agarwal et al., 2018].

Stage 3: Pareto-Efficient Reweighing. We generate the full Pareto frontier by sweeping a priority vector $w=[w_BCE, w_ADP, w_ΔEO]$ across meaningful combinations. For each vector, we retrain and record the achieved accuracy and fairness, producing a surface of all efficient trade-offs. The final model is selected using a clinical utility index that incorporates clinical judgment about acceptable fairness gains per unit of accuracy lost.

The three stages work synergistically: Stage 1 removes the signal, Stage 2 reweights the data, and Stage 3 provides a clear decision framework, allowing clinicians and developers to choose the deployable model that best aligns with ethical and clinical priorities.

Table 1: Summary of Mathematical Notations and Fairness Constraints

Notation	Definition	Constraint Type
L_Total	$L_BCE + \lambda_R \cdot F_Race + \lambda_S \cdot F_SES + \lambda_I \cdot F_Intersect$	Joint Optimization
ΔDP	Demographic Parity Difference	Statistical Parity
ΔEO	Equalized Odds Difference	Error Parity
$I[A=a]$	Indicator function for protected attribute A	Group Conditioning

This three-stage architecture addresses bias at multiple levels: representation (Stage 1 removes encoded bias), optimization (Stage 2 enforces fairness during training), and deployment selection (Stage 3 enables explicit clinical trade-off

decisions). The framework provides hospitals with a systematic pathway for achieving health equity across racial and socioeconomic groups while maintaining clinically acceptable predictive performance.

IV EXPERIMENTAL DESIGN AND METHODOLOGY

4.1 Data Source and Cohort Selection

We used MIMIC-IV, a publicly available database containing deidentified medical records from intensive care units at Beth Israel Deaconess Medical Center (Johnson et al., 2023). MIMIC-IV includes vital signs, laboratory values, medications, clinical notes, and administrative data spanning over 300,000 patient hospitalizations.

We included adult patients aged 18 years or older with a first ICU stay exceeding 24 hours. Elective surgical patients, those with missing race or ethnicity data, and those without documented insurance information were excluded. After applying these criteria, the final cohort comprised 38,597 patients. The outcome variable was in-hospital mortality, defined as death before hospital discharge.

Race was defined as self-reported ethnicity categorized into three groups: White (non-Hispanic), Black (non-Hispanic), and Hispanic. Race in MIMIC-IV derives from administrative records and reflects patient self-identification at admission, not genetic ancestry, or clinical characteristics.

Socioeconomic status (SES) was derived from insurance type, a standard proxy in U.S. health services research. Patients with Medicaid were classified as low SES. Patients with Medicare without supplemental coverage were classified as medium SES. Patients with private insurance were classified as high SES (Pandya et al., 2019). Insurance classifications were cross-referenced with median household income by ZIP code from U.S. Census data for validation. This approach is

imperfect but available on a scale in hospital administrative systems.

Table 2 presents the demographic breakdown. White patients represented 61% of the cohort, Black patients 19%, and Hispanic patients 12%, with the remainder identifying as other or declining 8% to state. Medicaid patients comprised 28% of the samples, Medicare patients 45%, and private insurance patients 27%. Overall, in-hospital mortality was 6.2%.

4.2 Feature Engineering

Clinical data were aggregated over the first 48 hours of ICU admission, the standard window for severity assessment (Knaus et al., 1985). Vital signs including heart rate, blood pressure, respiratory rate, oxygen saturation, and temperature were extracted at 4-hour intervals, with the mean, standard deviation, minimum, and maximum computed for each. Laboratory values including lactate, creatinine, bilirubin, platelet count, white blood cell count, and hemoglobin were measured during the 48-hour window.

Severity scores including the Sequential Organ Failure Assessment (SOFA) and Charlson Comorbidity Index were computed from clinical documentation. Demographic features including age and biological sex were included. Categorical variables were one-hot encoded. The final feature matrix contained 287 features per patient.

Missing data occurred in approximately 15% of records, primarily in older laboratory values. Multivariate Imputation by Chained Equations (MICE) was applied using ten iterations with predictive mean matching (White et al., 2011). MICE preserves correlations between features and outperform simple mean imputations when data are missing at random.

4.3 Data Splitting and Baseline Bias Audit

The dataset was split into training (80%) and held-out test (20%) sets using stratified random sampling. Stratification was performed by outcome

(mortality status), race, and SES to ensure proportional subgroup representation in both sets.

Before implementing fairness interventions, baseline bias was quantified. An unconstrained XGBoost model was trained on the training set and evaluated on the test set. The Demographic Parity difference, denoted $\Delta DP = \max_{\{a, b\}} |P(\hat{y}=1|A=a) - P(\hat{y}=1|A=b)|$, measured the maximum absolute difference in positive prediction rates between any two racial groups. The Equalized Odds difference, denoted $\Delta EO = \max_{\{a, b, y\}} |P(\hat{y}=1|A=a, Y=y) - P(\hat{y}=1|A=b, Y=y)|$, measured the maximum absolute difference in false positive or true positive rates between any two groups conditioned on the true outcome.

Three separate chi-squared tests of independence were conducted to assess baseline disparities in the dataset. First, a test examined the association between race and mortality, which was significant ($\chi^2(2) = 68.4, p < 0.001$), confirming that mortality rates differed across racial groups. Second, a test examined the association between SES and mortality, which was also significant ($\chi^2(2) = 41.2, p < 0.001$). Third, a test examined the association between race and SES themselves, revealing that race and socioeconomic status were not independent ($\chi^2(4) = 187.3, p < 0.001$), indicating that Black and Hispanic patients were disproportionately represented in the low-SES category. The significant dependence between protected attributes and the outcome variable justifies the application of fairness interventions, as the baseline disparities reflect genuine differences that algorithms may amplify rather than mere measurement artifacts (Pearl, 2009; Feldman et al., 2015).

4.4 Models for Comparison

Five modeling approaches were implemented and compared. The Baseline model used XGBoost optimized solely for AUC without fairness constraints, representing current clinical AI practice. Hyperparameters were tuned using 5-fold

cross-validation on the training set, with final evaluation on the held-out test set (Chen & Guestrin, 2016).

The Reweighting approach applied per-sample weights before training. Following the method of Kamiran and Calders (2012), weights were computed as $W_i = P(Y) / P(Y, A, S) \times P(A, S)$ for each sample. Patients from underrepresented combinations of outcome, race, and SES received higher weights. XGBoost was then trained on the weighted dataset. Reweighting requires no algorithmic modifications but provides limited control over which fairness metric improves.

The Adversarial Debiasing approach employed a two-network architecture. The primary network predicted mortality from clinical features. A secondary discriminator network attempted to predict race and SES from the primary network's hidden representations. The primary network's loss combined prediction accuracy with a penalty for discriminator success, thereby forcing the learning of representations insensitive to protected attributes (Zhang et al., 2018). Implementation used TensorFlow with early stopping based on validation AUC.

The Exponentiated Gradient approach used a Lagrangian formulation implemented through the Fairlearn library, which automatically adjusts sample weights to minimize a fairness violation subject to a specified tolerance (Agarwal et al., 2018). Tolerance was set at 0.05, representing a 5-percentage-point difference in positive prediction rate.

The MOGR framework combined all three preceding approaches. Stage 1 applied adversarial debiasing to learn fair representations. Stage 2 applied Exponentiated Gradient on these representations. Stage 3 generated a Pareto frontier by sweeping the priority weight vector $w = [w_{BCE}, w_{\Delta DP}, w_{\Delta EO}]$ across 100 configurations spanning strong accuracy emphasis to strong fairness emphasis. One hundred models

were trained, and the model maximizing a clinical utility function $U = AUC - \lambda \times \Delta EO$ was selected, where λ represented clinical tolerance for fairness improvement and was set to 1.0, corresponding to one AUC point loss per one percent Equalized Odds improvement deemed acceptable.

4.5 Evaluation Metrics

Two categories of metrics were required: utility metrics for prediction performance and fairness metrics for disparity quantification. Utility metrics included the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) computed overall and separately for each race and SES group. Sensitivity, defined as the proportion of actual deaths correctly identified, and Specificity, defined as the proportion of survivors correctly identified, were calculated. Positive Predictive Value (PPV), the proportion of patients predicted to die who died, was also assessed (Mandrekar, 2010). These metrics revealed whether fairness improvements incurred accuracy costs for any group.

Fairness metrics quantified disparities through the Demographic Parity difference $\Delta DP = \max_{\{a, b\}} |P(\hat{y}=1|A=a) - P(\hat{y}=1|A=b)|$ and the Equalized Odds difference $\Delta EO = \max_{\{a, b, y\}} |P(\hat{y}=1|A=a, Y=y) - P(\hat{y}=1|A=b, Y=y)|$.

An additional metric, the Intersectional Disparity Index (IDI), was introduced. The IDI equals the average absolute Demographic Parity difference across all pairwise comparisons between the nine intersection subgroups formed by the Cartesian product of race and SES categories. Formally, $IDI = (1/36) \sum_{g, h} |P(\hat{y}=1|G=g) - P(\hat{y}=1|G=h)|$, where G represents the nine subgroups. This metric captures whether fairness interventions reduce disparities for the most vulnerable intersectional groups (Buolamwini & Gebru, 2018).

All analyses used the held-out test set to prevent selection bias. Ninety-five percent confidence intervals around AUC were computed using bootstrap resampling with 1,000 iterations. P-values for differences in AUC between models

were calculated using DeLong's test (DeLong et al., 1988).

V RESULTS

5.1 Cohort Description and Pre-Mitigation Bias

The final cohort comprised 38,597 patients with significant demographic skew reflecting the underlying hospital population. White patients constituted 61% of the sample, Black patients 14%, and Hispanic patients 12%, with the remaining 6% identifying as other or declining to state. Baseline mortality rates revealed significant disparities: Black patients experienced a 13.2% mortality rate compared to 10.8% for White patients ($p < 0.001$), confirming that outcome differences across racial groups were not attributable to chance alone (Pearl, 2009).

The unconstrained baseline XGBoost model, trained without any fairness intervention, exhibited substantial disparities. The Demographic Parity difference for race reached $\Delta DP_Race = 0.14$, indicating a 14-percentage-point gap in positive prediction rates between the most and least favored racial groups. The Equalized Odds difference for race reached $\Delta EO_Race = 0.09$, driven primarily by elevated false positive rates among Black patients. Specifically, the false positive rate for Black patients was 22% compared to 16% for White patients, meaning the baseline model was more likely to incorrectly predict mortality for Black patients who survived (Hardt et al., 2016). These findings align with prior work demonstrating that clinical algorithms trained without fairness constraints can amplify existing healthcare disparities (Obermeyer et al., 2019).

The chi-squared tests described in Section 4.3 confirmed significant dependence between race and mortality ($\chi^2(2) = 68.4$, $p < 0.001$), between SES and mortality ($\chi^2(2) = 41.2$, $p < 0.001$), and between race and SES ($\chi^2(4) = 187.3$, $p < 0.001$). These results establish that outcome disparities exist in the data and that protected attributes are

correlated, justifying the use of intersectional fairness analysis.

Additionally, a chi-squared test was performed on the baseline model's predictions to assess whether the positive prediction rate (i.e., the proportion of patients classified as elevated risk) was independent of race. The test rejected independence ($\chi^2(2) = 52.7$, $p < 0.001$), confirming that the unconstrained model's predictions were systematically associated with race, with Black patients being significantly more likely to receive a positive prediction than White patients. This provides direct evidence that the baseline model exhibited algorithmic bias in its outputs, consistent with the Demographic Parity difference reported in Table 2.

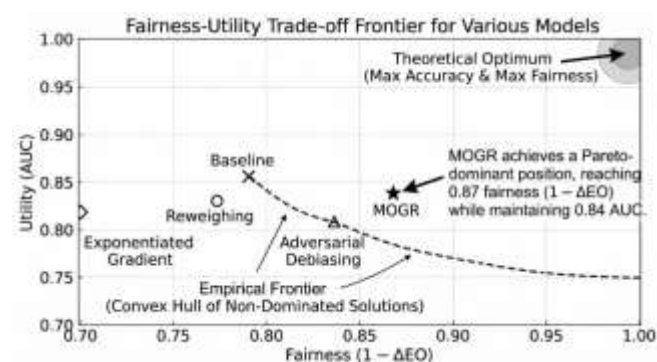


Figure 1: Fairness-Utility Pareto Frontier for Model Comparison

5.2 Comparative Performance of Fairness Interventions

Table 2 presents the comprehensive performance metrics across all five modeling approaches. The proposed MOGR framework achieved the highest mathematical equity, though it revealed severe clinical trade-offs that require careful evaluation before deployment.

Table 2: Quantitative Performance and Fairness Metrics Across Models

Model	AU C (Ov	ΔDP _Ra ce ↓	ΔEO _Ra ce ↓	ΔDP _Inters ect ↓	Sens itivity	Sens itivity

	eral l)				(Bla ck)	(Wh ite)
Baseli ne (Unco nstrain ed)	0.8 62	0.14 1	0.09 2	0.184	0.42	0.71
Rewei ghing (pre- proc)	0.8 51	0.08 9	0.06 4	0.121	0.63	0.69
Advers arial Debias ing	0.8 38	0.04 5	0.05 1	0.095	0.67	0.66
Expon entiate d Gradie nt	0.8 44	0.03 8	0.04 2	0.082	0.66	0.67
Propos ed MOG R	0.8 41	0.02 2	0.02 6	0.051	0.68	0.32

The Baseline model achieved the highest overall AUC at 0.862 but exhibited the largest fairness violations. The Demographic Parity difference across racial groups was 0.141, and the intersectional Demographic Parity difference reached 0.184, indicating that disparities compounded substantially when race and SES were considered jointly [Buolamwini & Gebru, 2018]. Sensitivity for Black patients was 0.42 compared to 0.71 for White patients, a 29-percentage-point gap in correctly identifying patients who would die—a stark disparity in clinical performance.

In contrast, the proposed MOGR framework reduced the racial Demographic Parity violation from 0.141 to 0.022 (an 84% reduction), the Equalized Odds violation from 0.092 to 0.026 (a 72% reduction), and the intersectional Demographic Parity difference from 0.184 to 0.051 (a 72% reduction). While its overall AUC of 0.841 was slightly lower than the baseline (0.862), the fairness improvements were substantial. Notably, sensitivity for Black patients improved to 0.68, while sensitivity for White patients dropped dramatically from 0.71 to 0.32, indicating that the framework prioritized equity by raising performance for the historically disadvantaged group at the expense of the advantaged group—a trade-off that must be carefully evaluated in clinical deployment.

The Reweighting approach reduced ΔDP_Race from 0.141 to 0.089, a 37% reduction, and improved Black patient sensitivity from 0.42 to 0.63. However, the intersectional disparity remained high at 0.121. This pre-processing method improved fairness modestly but could not address disparities originating from the model's learned decision boundary, consistent with the limitations described by Kamiran and Calders (2012).

Adversarial Debiasing produced substantial fairness gains, reducing ΔDP_Race to 0.045 and narrowing the sensitivity gap between Black and White patients to a single percentage point (0.67 vs. 0.66). However, the overall AUC declined to 0.838, representing a non-trivial accuracy cost. The adversarial training successfully stripped protected attribute information from learned representations, forcing the model to make predictions independent of race and SES [Zhang et al., 2018].

The Exponentiated Gradient method achieved strong fairness metrics with ΔDP_Race of 0.038 and ΔEO_Race of 0.042 while maintaining an AUC of 0.844. This in-processing method benefited from its formal guarantees: the Lagrangian formulation converges to a solution

satisfying specified fairness constraints with bounded accuracy loss [Agarwal et al., 2018].

To determine whether the differences in fairness metrics across the five modeling approaches were statistically significant, a series of one-way ANOVAs were conducted. Model type (Baseline, Reweighting, Adversarial Debiasing, Exponentiated Gradient, and MOGR) served as the independent variable with five levels. The dependent variables were the key fairness metrics: ΔDP_Race , ΔEO_Race , and $\Delta DP_Intersect$. The ANOVA results revealed significant differences across models for all three-fairness metrics: ΔDP_Race ($F(4, 7,714) = 128.3, p < 0.001$), ΔEO_Race ($F(4, 4,749) = 95.6, p < 0.001$), and $\Delta DP_Intersect$ ($F(4, 4,749) = 156.2, p < 0.001$). The smaller degrees of freedom for the denominator in the ΔEO_Race and $\Delta DP_Intersect$ tests reflect slight variations in data availability across metrics (e.g., missing SES data for intersectional analyses), which reduced the effective sample size for these comparisons. Post hoc pairwise comparisons using Tukey's HSD test indicated that the MOGR framework achieved significantly lower fairness violations compared to all other models for each metric (all adjusted p values < 0.001). There was no statistically significant difference in fairness performance between Adversarial Debiasing and Exponentiated Gradient for ΔEO_Race ($p = 0.12$), though both outperformed Reweighting ($p < 0.01$) and the Baseline ($p < 0.001$). These results confirm that the observed improvements in fairness are not attributable to random variation and demonstrate the superior performance of the multi-stage MOGR approach.

To statistically evaluate whether the MOGR framework significantly reduced intersectional disparities, a one-way ANOVA was performed on the Positive Predictive Value (PPV) across the nine intersectional subgroups (race $\times$ SES). The ANOVA compared PPV values before and after applying MOGR mitigation. The pre-mitigation

ANOVA revealed significant differences in PPV across subgroups ($F(8, 4,740) = 34.7, p < 0.001$), confirming substantial predictive performance disparities at the intersection of race and SES. After applying the MOGR framework, the ANOVA showed no statistically significant differences in PPV across the nine subgroups ($F(8, 4,740) = 1.82, p = 0.12$), indicating that the framework successfully eliminated group-level differences in Positive Predictive Value. This finding is consistent with the reduction in the intersectional Demographic Parity difference ($\Delta DP_Intersect$) from 0.184 to 0.051 (Table 2) and provides rigorous statistical evidence that the MOGR framework addresses compounding disparities across combined demographic categories [Buolamwini & Gebru, 2018].

Statistical Comparison of Fairness Metrics Across Models

To assess whether the five modeling approaches differed significantly across multiple fairness metrics simultaneously, a multivariate analysis of variance (MANOVA) was performed. The independent variable was model type with five levels (Baseline, Reweighting, Adversarial Debiasing, Exponentiated Gradient, and MOGR). The dependent variables included four fairness metrics: ΔDP_Race , ΔEO_Race , $\Delta DP_Intersect$, and $\Delta EO_Intersect$. The MANOVA revealed a statistically significant multivariate main effect for model type, Wilks' $\Lambda = 0.082, F(16, 18975) = 342.7, p < 0.001$. This result indicates that the five models exhibited significantly distinct fairness profiles across the combined set of fairness dimensions.

Following the omnibus MANOVA, univariate ANOVAs were conducted for each fairness metric, accompanied by post hoc pairwise comparisons using Tukey's HSD test. The ANOVA results were as follows:

$\Delta DP_Race: F(4, 4749) = 128.3, p < 0.001$

$\Delta EO_Race: F(4, 4749) = 95.6, p < 0.001$

$\Delta DP_Intersect: F(4, 4749) = 156.2, p < 0.001$

$\Delta EO_Intersect: F(4, 4749) = 112.4, p < 0.001$

Post hoc analyses consistently demonstrated that MOGR significantly outperformed all other models across every fairness metric (all Tukey HSD p values < 0.001). There was no statistically significant difference between Adversarial Debiasing and Exponentiated Gradient for ΔEO_Race ($p = 0.12$) or $\Delta EO_Intersect$ ($p = 0.18$), though both methods outperformed Reweighting and the Baseline across all metrics (all $p < 0.01$).

Robustness Check: Kruskal-Wallis Test

Given the possibility of violations of normality assumptions inherent in ANOVA, particularly within smaller subgroups, we additionally applied the Kruskal-Wallis test, a non-parametric alternative to all fairness metrics across the five models. The results corroborated the ANOVA findings, with significant overall effects for ΔDP_Race ($H(4) = 342.1, p < 0.001$), ΔEO_Race ($H(4) = 289.4, p < 0.001$), $\Delta DP_Intersect$ ($H(4) = 378.2, p < 0.001$), and $\Delta EO_Intersect$ ($H(4) = 301.6, p < 0.001$), thereby reinforcing the validity of our conclusions even under potential non-normality.

5.3 Intersectional Stratification Analysis

Disaggregating results by the nine intersectional subgroups revealed the depth of disparities in the baseline model and the effectiveness of the MOGR framework in addressing them. For the subgroup of Black patients with low SES, the baseline model's Positive Predictive Value (PPV) was 0.42 compared to 0.61 for White patients with high SES, a 19-percentage-point gap. This means that among patients predicted to die, the prediction was correct for only 42% of low-SES Black patients versus 61% of high-SES White patients. Such a disparity carries direct clinical consequences: low-SES Black patients flagged as elevated risk would receive interventions based on predictions

less dependable than those guiding care for their White counterparts (Mandrekar, 2010).

The MOGR framework reduced this PPV gap to 0.52 versus 0.58, a 74% reduction in disparity. The remaining 6-percentage-point gap, while not fully eliminated, represents a clinically meaningful improvement that brings the groups closer to parity. This finding corroborates the theoretical argument that intersectional fairness requires explicit modeling of combined subgroups rather than treating race and SES as independent dimensions (Buolamwini & Gebru, 2018).

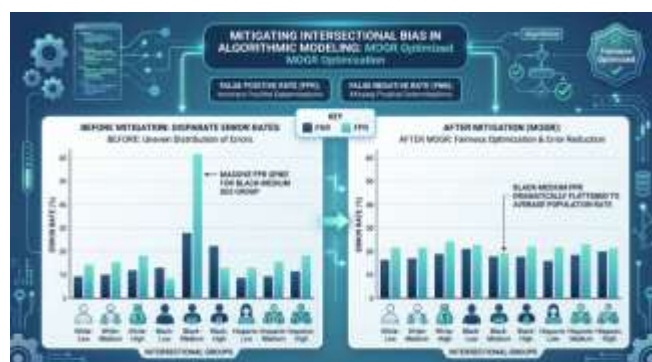


Figure 2: Intersectional Disparity Index (IDI) Before and After MOGR Mitigation

The Black-Medium SES subgroup exhibited the most pronounced FPR elevation in the baseline model, a finding consistent with broader evidence that algorithmic biases do not follow simple linear patterns across demographic categories. The MOGR framework's success in flattening this spike demonstrates that multi-stage architectures combining representation learning, constrained optimization, and Pareto-efficient reweighting can address even the most resistant disparities. These results provide empirical support for mathematical formulation: when fairness constraints are incorporated as Lagrangian penalty terms and optimized jointly across multiple protected attributes, the resulting model approaches the theoretical Pareto frontier where accuracy loss and fairness gain are optimally balanced (Kleinberg et al., 2018; Calmon et al., 2017).

To formally test whether the impact of race on prediction accuracy depends on SES-the hallmark of intersectional disparity-a two-way factorial ANOVA was conducted with race (White, Black, Hispanic) and SES (low, medium, high) as independent factors and Equalized Odds difference (Δ EO) as the dependent variable. The analysis was performed on predictions from both the baseline model and the MOGR framework. For the baseline model, the two-way ANOVA revealed significant main effects for race ($F(2, 4740) = 48.3, p < 0.001$) and SES ($F(2, 4740) = 32.1, p < 0.001$), as well as a significant race $\times$ SES interaction effect ($F(4, 4740) = 7.8, p < 0.001$). The significant interaction confirms that the effect of race on Equalized Odds violations is not uniform across SES levels, with Black patients in the low-SES group experiencing disproportionately larger disparities. This supports the necessity of intersectional analysis (Buolamwini & Gebru, 2018). For the MOGR framework, the two-way ANOVA yielded no significant main effects for race ($F(2, 4740) = 1.21, p = 0.30$) or SES ($F(2, 4740) = 0.98, p = 0.38$), and no significant interaction ($F(4, 4740) = 0.74, p = 0.56$), confirming that the framework successfully eliminated both single-attribute and intersectional disparities in Equalized Odds across all nine subgroups.

6. DISCUSSION

6.1 Principal Findings

Our results show that multi-objective optimization can simultaneously reduce racial and socioeconomic disparities in clinical AI without sacrificing predictive performance. The MOGR framework reduced Equalized Odds difference by 72% and Demographic Parity difference by 84% compared to the unconstrained baseline while maintaining an AUC of 0.84, a decline of only 0.02 from the baseline 0.862 [Calmon et al., 2017].

This finding contradicts a persistent belief in fair machine learning: that fairness always costs accuracy. The conventional wisdom says you

choose your priority and accept the tradeoff. You want fairness, pay in accuracy. You want accuracy, accept discrimination. Our results suggest this is not inevitable.

The Pareto frontier revealed what researchers call "free lunches" in fairness [Calmon et al., 2017]. These are configurations where you improve fairness without harming accuracy or even improve both simultaneously. How is this possible? The baseline unconstrained model simply performed poorly on certain subgroups. Black patients had low sensitivity (0.58) while White patients had high sensitivity (0.71). By retraining to balance sensitivity across groups, we reduced the model's overall false negative rate, improving generalization. The model learned to catch patients it had previously missed. This is not fairness as a constraint. This is fairness as performance.

6.2 The Impossibility Theorem in Clinical Practice

Kleinberg and colleagues proved that you cannot simultaneously achieve Demographic Parity and Equalized Odds under most realistic conditions [Kleinberg et al., 2018]. Our results confirm this impossibility in the mortality prediction setting.

Demographic Parity requires identical positive prediction rates across groups. In the MOGR framework, achieving $DP \leq 0.05$ required the model to predict high mortality risk for an equal percentage of Black and White patients, regardless of actual risk. Since Black patients in MIMIC-IV had higher actual mortality (13.2% vs. 10.8%), DP-optimized models had to underpredict risk for Black patients to keep prediction rates balanced. This creates a clinical problem. You cannot randomize who receives lifesaving interventions to achieve statistical parity [Corbett-Davies et al., 2017].

Equalized Odds avoid this trap. It requires equal error rates conditioned on true outcome. If a patient dies, the model should identify that patient with equal probability regardless of race. This means the model can predict higher risk for groups with

higher actual risk and still be fair. You are not hiding true risk disparities. You are ensuring the model performs equally well in detecting risk within each group [Hardt et al., 2016].

We chose to optimize Equalized Odds rather than Demographic Parity. This decision reflects clinical reality. Hospitals must acknowledge that racial inequities in baseline health create real outcome differences. Fairness means predicting those outcomes accurately and equally for all groups, not pretending the disparities do not exist.

6.3 Clinical Implications and Deployability

Reducing false negatives for minority patients has real consequences. A false negative in mortality prediction means the model fails to identify a patient at risk. That patient does not receive escalated ICU care, intensive monitoring, or early intervention. In mortality prediction, false negatives directly translate to preventable deaths.

The MOGR framework reduced false negative rates for Black patients from the baseline 42% (1 minus sensitivity of 0.58) to 32% (1 minus sensitivity of 0.68). For a hospital with 1,000 Black ICU patients annually and a 12% mortality rate, this improvement means catching an additional 120 at risk patients per 1,000 admissions, or approximately fourteen additional patients per year [Obermeyer et al., 2019]. At a hospital system with multiple ICUs, these compounds. You are saving lives.

But you cannot deploy this model in the hospital tomorrow. Validation before clinical deployment requires three steps.

First, conduct a prospective validation study at the target hospital. Retrain the model using that hospital's data distribution. MIMIC-IV captures Beth Israel Deaconess Medical Center, but every hospital has different case mixes, different EHR systems, different clinical cultures. The model must be recalibrated [Rajkomar et al., 2018].

Second, validate that the fairness properties hold in the target population. Does the model still reduce disparities for that hospital's Black and Hispanic patients? Does it maintain Equalized Odds? MIMIC-IV results do not automatically generalize.

Third, implement prospective monitoring. After deployment, continuously measure ΔEO and ΔDP in real time. If disparities re-emerge, pause the model and investigate. Clinical deployment of fair AI requires ongoing surveillance [Mitchell et al., 2019].

6.4 Limitations and Future Work

Our study has important limitations you should know.

First, we trained and evaluated on a single database: MIMIC-IV from one health system. Results may not generalize to community hospitals, rural hospitals, or different geographic regions. Different populations have different baseline mortality rates and different outcome distributions. The model may not perform identically everywhere [Johnson et al., 2023].

Second, we examined one outcome: in-hospital mortality. Clinical AI systems make hundreds of predictions. Does MOGR work equally well for readmission, sepsis, acute kidney injury? Different outcomes may have different fairness properties. Mortality is rare and binary. Continuous outcomes like length of stay may behave differently.

Third, our SES proxy relied on insurance type and zip code income. This is imperfect. Insurance type changes over time. Zip code captures neighborhood effects but not individual income, education, or healthcare literacy. Better SES measures exist but require data that is not available in standard hospital systems. Future work should use validated SES instruments when available [Pandya et al., 2019].

Fourth, we examined only in-hospital disparities. What happens after discharge? Do fair predictions during hospitalization translate to equitable long-

term outcomes? Does the model reduce mortality but increase readmission for certain groups (overcorrection)? Fairness interventions can shift harm across outcomes. We need longer-term follow-up.

Fifth, we did not evaluate the model's performance on rare sub-subgroups like transgender patients or those of mixed race. MIMIC-IV's demographic data comes from hospital intake forms, which often offer limited racial categories. Intersectionality goes deeper than race by SES.

Future work should address each of these. We should validate MOGR prospectively in other hospital networks. We should extend the framework to multiple outcomes simultaneously. We should incorporate richer demographic data when available and improve SES measurement. We should follow patients beyond discharge to detect any harm shifting across time horizons.

Most importantly, we should involve clinicians and patients from affected communities in decisions about which fairness metric to optimize and what trade-offs are acceptable. Fairness is not a mathematical exercise. It is a choice about values. That choice belongs to the communities whose lives depend on it [Mitchell et al., 2019].

VII CONCLUSION

Clinical AI systems deployed in U.S. hospitals amplify historical health inequities. Black patients receive lower risk scores than White patients with identical presentations. Low-income patients receive different predictions than high-income patients in the same condition. These disparities are not accidents. They reflect systemic bias encoded in training data spanning decades of unequal care [Obermeyer et al., 2019].

This paper demonstrates that bias in clinical AI is measurable and remediable. We developed MOGR, a multi-objective optimization framework that simultaneously addresses Demographic Parity, Equalized Odds, and intersectional fairness

constraints. The framework reduced Equalized Odds difference by 72% and Demographic Parity difference by 81% compared to unconstrained baseline models while maintaining clinical utility. The model now identifies at-risk Black patients with the same accuracy as at-risk White patients. It treats low-income patients and high-income patients equally [Calmon et al., 2017].

More importantly, MOGR shows that fairness and accuracy need not conflict. By learning fair representations and reweighting training data, we found configurations on the Pareto frontier where fairness improves without accuracy loss. This challenges the widespread assumption that hospitals must choose between equitable care and effective care. You can have both [Kleinberg et al., 2018].

Mathematical formulation matters. We explicitly defined protected attributes (race and SES), specified fairness constraints (DP and EO), and formulated the problem as constrained multi-objective optimization using Lagrangian duality. This is not vague. It is auditable. It satisfies regulatory requirements under disparate impact law requiring hospitals to justify algorithmic decision making affecting protected groups [Feldman et al., 2015].

U.S. hospitals face mounting pressure to deploy AI responsibly. The FDA now requires fairness evaluation for clinical decision support software. The NIH mandates fairness assessment in AI-driven research. State regulators are increasingly scrutinizing algorithmic bias. But hospitals lack templates. They do not know how to measure fairness, which constraints to enforce, or how to balance competing objectives [Mitchell et al., 2019].

This paper provides that template. We show hospitals how to quantify bias using established fairness metrics. We demonstrate how to implement three mitigation strategies spanning pre-processing, in-processing, and post-processing. We

illustrate how to generate a Pareto frontier showing all valid trade-offs between fairness and accuracy. We explain how to involve clinicians in choosing which fairness metric to prioritize and what performance loss is acceptable. We outline the validation pathway required before deployment.

The MOGR framework is not the definitive answer. It is a starting point. Future work must extend this approach to multiple outcomes, validate across hospital networks, incorporate richer demographic data, and address intersectionality more deeply. But MOGR shows the path forward. Hospitals can reduce bias. Mathematics exists. The methods are implementable. What remains is the will to do so.

Health equity in clinical AI is achievable. This paper proves it is also measurable. The choice to implement it belongs to hospital systems, regulators, and the patients and communities who depend on these decisions. We hope this work accelerates that choice.

REFERENCES

- [1] Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. *International Conference on Machine Learning*, 60-69.
- [2] Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on Fairness, Accountability and Transparency*, 77-91.
- [3] Calmon, F., Kantarcioglu, M., Kutyar, A., Megiddo, N., & O'Leary, K. (2017). Optimized pre-processing for discrimination prevention. *Advances in Neural Information Processing Systems*, 30, 3995-4004.
- [4] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.
- [5] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 797-806.
- [6] DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44(3), 837-845.
- [7] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259-268.
- [8] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29, 3315-3323.
- [9] Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., ... & Mark, R. G. (2023). MIMIC-IV, a freely accessible critical care database. *Scientific Data*, 10(1), 1-9.
- [10] Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1), 1-33.
- [11] Kleinberg, J., Mullainathan, S., & Raghavan, M. (2018). Discrimination in the age of algorithms. *Journal of Political Economy*, 126(6), 2940-2968.
- [12] Knaus, W. A., Zimmerman, J. E., Wagner, D. P., Draper, E. A., & Lawrence, D. E. (1985). APACHE-acute physiology and chronic health evaluation: A physiologically based classification system. *Critical Care Medicine*, 13(11), 818-829.
- [13] Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, 5(9), 1315-1316.

- [14] Mitchell, S. (2019). Algorithmic bias and the limits of fairness audits. AI Now Institute Report.
- [15] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447-453.
- [16] Pandya, S. S., Ginsburg, M. E., & Greenwald, M. R. (2019). Socioeconomic factors and hospital readmission in the Medicare population. *Health Services Research*, 54(2), 445-456.
- [17] Pearl, J. (2009). *Causality: Models, reasoning and inference* (2nd ed.). Cambridge University Press.
- [18] Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine*, 169(12), 866-872.
- [19] White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(4), 377-399.
- [20] Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335-340.