ISBN: 978-628-96613-3-0. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2026.1.1.1229

Algorithmic Auditing and Bias Mitigation in Clinical AI Models for Rural Honduran Populations: A Simulation-Based Study

Isaac Zablah[1]; Edwin Hernandez[2]; Fiama Garcia[3]; Antonieta Zuniga[4]; Antonio Garcia Loureiro[5]
[1,3]Faculty of Medical Sciences, National Autonomous University of Honduras, Honduras,
jose.zablah@unah.edu.hn, fiama.garcia@unah.edu.hn
[2]EGLA Corp., USA, *edwinhm@eglacorp.com*
[4]Dirección de Medicina Forense, Ministerio Público, Honduras, *antonietazuniga2311@yahoo.com*
[5]Department of Electronics and Computing, University Santiago de Compostela, Spain, *antonio.garcia.loureiro@usc.es*

Abstract*– The use of clinical artificial intelligence (AI) models in healthcare systems has prompted concerns about algorithmic fairness, especially for marginalized populations who have been underrepresented in medical data. Simulated data reflecting rural Honduran demographics and illness incidence patterns is used to examine demographic and distributional biases in clinical prediction models. Gradient boosting (XGBoost) and neural network classifiers (Multilayer Perceptron with 3 hidden layers) were trained for cardiovascular risk prediction using synthetic clinical datasets generated by Monte Carlo simulation. Fairness metrics including equalized odds difference, demographic parity ratio, and calibration error assessed model performance discrepancies across protected factors like geographic location, socioeconomic status, and indigenous heritage. Sample reweighting, adversarial debiasing, and decision threshold optimization were evaluated. Baseline models showed significant performance differences, measuring AUROC differences of up to 0.089 between urban and remote rural populations ($p < 0.001$) and equalized odds differences reaching 0.15 across socioeconomic groupings. Adversarial debiasing reduced equalized odds differences by 62.4% while maintaining clinically acceptable discrimination (AUROC = 0.823) after mitigation. Sample reweighting increased demographic parity by 41.3% without degrading predictive performance. These findings demonstrate that clinical AI models developed without explicit fairness constraints can exacerbate existing healthcare disparities, and that simulation-based auditing frameworks combined with bias mitigation techniques can substantially enhance algorithmic equity while maintaining predictive utility. The findings recommend required algorithmic auditing measures before deployment in resource-limited healthcare settings.

Keywords*-- Algorithmic fairness; bias mitigation; clinical decision support; rural health; Honduras.

I. Introduction

Artificial intelligence in clinical decision-making is one of the biggest changes in healthcare. Diagnostic imaging interpretation, risk stratification, treatment selection, and resource allocation across medical disciplines are being assisted by machine learning algorithms [1]. The FDA approved 882 AI-enabled medical devices in radiology (76%), cardiology (10%), and neurology (4%), as of May 2024 [2]. This rapid expansion highlights the promise and importance of ensuring these technologies work equally across patient populations.

Clinical AI systems can embed and exacerbate healthcare inequities, according to growing evidence. Obermeyer and colleagues found that a widely used algorithm affecting millions of US patients underestimated Black patients' health requirements relative to White patients with equal illness severity [3]. The algorithm evaluated healthcare expenditures to estimate health needs, but Black patients received less care owing to systemic hurdles, so it mistakenly determined they were healthier. Similar biases have been seen in dermatological AI systems trained on light-skinned people [4], chest radiograph interpretation algorithms that underdiagnose minorities [5], and sepsis prediction models that perform worse in minorities [6].

Low- and middle-income countries (LMICs) with limited healthcare facilities and disadvantaged populations may be disproportionately at risk from biased algorithms. These issues are evident in Honduras, a 10-million-person Central American nation. The country has acute physician shortages (0.37 physicians per 1,000 population compared to 2.6 in the US), with healthcare resources concentrated in metropolitan centers and 45% of the population in rural areas with restricted access [7]. Nearly 90% of Hondurans receive health care from the Ministry of Health [8]. Rural communities, frequently indigenous, face the greatest challenges due to geography, economics, and culture [9].

AI-assisted healthcare technologies could improve care gaps by supporting decision-making and resource optimization, or they could worsen inequalities if algorithms fail underrepresented populations. We are unaware of any comprehensive algorithmic fairness review for clinical AI applications in Central America. The lack of large-scale electronic health record (EHR) data from the region and auditing algorithm frameworks in resource-constrained settings contribute to the research shortage.

This study uses simulation to create synthetic clinical datasets on rural Honduran demographics, disease prevalence, and healthcare utilization to fill these gaps. We wanted to quantify baseline performance disparities in machine learning models for cardiovascular risk prediction across demographic subgroups, evaluate bias mitigation strategies, and develop

practical algorithmic auditing recommendations for similar LMIC contexts.

We hypothesized that clinical prediction models trained without explicit fairness constraints would show significant performance disparities across geographic, socioeconomic, and ethnic subgroups, but mitigation techniques could meaningfully reduce these disparities while maintaining clinically acceptable predictive performance.

II. Materials and Methods

A. Study Design and Simulation Framework

Virtual rural Honduran clinical datasets were created using Monte Carlo simulation. The lack of large-scale EHR data from the target population, the ability to precisely control demographic distributions and establish ground truth for bias assessment, the reproducibility and transparency of simulation studies, and the ethical concerns about using real patient data for algorithm development in vulnerable populations without established governance frameworks led to this methodology.

Pan American Health Organization (PAHO), Honduras Demographic and Health Survey (ENDESA), World Health Organization (WHO) statistics, and published literature on Central American disease prevalence and risk factor distributions were used to parameterize the simulation framework [10]-[12]. Population parameters were adjusted to meet Honduran adult hypertension (22.6%), diabetes (7.4%), and cardiovascular disease risk factor prevalence [8].

B. Synthetic Data Generation

A synthetic cohort of N = 15,000 adult patients aged 18–80 was stratified by geography: urban (35%, Tegucigalpa and San Pedro Sula metropolitan areas), accessible rural (40%, within 2 hours of secondary healthcare facilities), and remote rural (25%). Population distributions from national surveys are shown here [13], most of the Honduran population is mestizo (about 92% of the total), with indigenous and Afro-descendant groups living in well-defined rural geographical areas [14]-[16].

We generated 24 clinical features for each simulated patient, including demographics (age, sex, geographic location, indigenous heritage, socioeconomic quintile), vital signs (systolic and diastolic blood pressure, heart rate, body mass index), laboratory values (fasting glucose, HbA1c, total cholesterol, LDL cholesterol, HDL cholesterol, triglycerides, creatinine, estimated glomerular filtration rate), and lifestyle factors. Models for feature distributions used multivariate normal distributions with correlation structures from epidemiological studies and appropriate modifications for non-normal variables [17].

The binary outcome variable (10-year cardiovascular event: myocardial infarction, stroke, or cardiovascular death) was generated using a logistic model with Framingham Risk Score and SCORE2 risk factors and calibration adjustments to match Honduras' 140 per 100,000 cardiovascular mortality rate [18]. Importantly, we established systematic data quality changes by geographic location: urban patients had 95% full data for 95% of characteristics, accessible rural patients had 82%, and remote rural patients had 68%. Laboratory access and clinical documentation are uneven in resource-limited settings [19].

C. Machine Learning Model Development

We trained two classifier architectures for 10-year cardiovascular risk prediction:

1) Gradient Boosting Machine (GBM): XGBoost implementation with hyperparameters optimized via 5-fold cross-validation (learning rate: 0.05, maximum depth: 6, minimum child weight: 3, subsample: 0.8, colsample by tree: 0.8, number of estimators: 200). Missing values were handled using the algorithm's native support for sparse data.

2) Neural Network (NN): Multilayer perceptron with three hidden layers (128, 64, 32 neurons), ReLU activation functions, dropout regularization (rate: 0.3), and Adam optimizer (learning rate: 0.001). Missing values were imputed using multiple imputations by chained equations (MICE) prior to training [20].

The dataset was partitioned into training (60%), validation (20%), and test (20%) sets using stratified sampling to preserve outcome and subgroup proportions. All model development and hyperparameter tuning were performed on training and validation sets; test set results are reported for final evaluation.

D. Fairness Metrics

We assessed algorithmic fairness using multiple complementary metrics, recognizing that no single measure captures all dimensions of equity [21]:

1) Equalized Odds Difference (EOD): The maximum absolute difference in true positive rates (TPR) and false positive rates (FPR) between protected groups. A classifier satisfies equalized odds when TPR and FPR are equal across groups [22]. EOD = max(|TPRA - TPRB|, |FPRA - FPRB|).

2) Demographic Parity Ratio (DPR): The ratio of positive prediction rates between the disadvantaged and advantaged groups. Values close to 1.0 indicate parity [23].

3) Calibration Error by Group: Expected calibration error (ECE) computed separately for each protected group, measuring whether predicted probabilities match observed frequencies [24].

Protected attributes examined included: geographic location (urban vs. accessible rural vs. remote rural), socioeconomic status (lowest two quintiles vs. upper three quintiles), and self-reported indigenous heritage (indigenous vs. non-indigenous).

E. Bias Mitigation Strategies

We tested three bias mitigation methods:

1) Sample Reweighting: Preprocessing that gives underrepresented or disadvantaged groups more weight during training. Weights are the inverse of the product of group prevalence and outcome prevalence within each group [25].

2) Adversarial Debiasing: Given that gradient boosting machines operate through non-differentiable tree-based splits that preclude direct gradient-based adversarial training, adversarial debiasing was implemented exclusively on the neural network (MLP) architecture, where backpropagation supports in-processing fairness constraints. The adversarial debiasing configuration used the equalized odds constraint, providing both predictor outputs and true outcome labels to the adversary network. For comparability across mitigation strategies in Table IV, GBM baseline performance is reported alongside MLP adversarial debiasing results, reflecting the complementary roles of each architecture in the experimental design [26]. We used the equalized odds constraint to give the adversary predictions and true labels.

3) Threshold Optimization: A post-processing method that finds group-specific categorization thresholds to equalize error rates. We optimized criteria to minimize equalized odds difference with a minimum acceptable sensitivity of 0.70 [27].

F. Statistical Analysis

Using 1,000 bootstrap resamples, the area under the receiver operating characteristic curve (AUROC) with 95% confidence intervals assessed model discrimination. Calibration plots and Hosmer-Lemeshow tests assessed calibration. DeLong's test for AUROC comparisons and chi-square testing for categorization measures were used to compare subgroup performance [28].

We ran 100 independent simulations with different random seeds and reported mean metrics with 95% confidence intervals to assure robustness. Statistical significance was set at $p < 0.05$. All analyses were done in Python 3.10 with scikit-learn 1.3.0, XGBoost 1.7.6, TensorFlow 2.13.0, and AI Fairness 360 [29].

G. Ethical Considerations

As this study employed entirely synthetic data without any real patient information, institutional review board approval was not required. The simulation parameters were derived exclusively from aggregated, publicly available epidemiological statistics. No identifiable or potentially re-identifiable data were used at any stage.

III. Results

A. Simulated Cohort Characteristics

The synthetic cohort comprised 15,000 patients with mean age 48.7 years (SD 15.2), 54.3% female. The patient distribution was 5,250 urban (35.0%), 6,000 accessible rural (40.0%), and 3,750 remote rural (25.0%). A higher proportion of 2,850 patients (19.0%) identified indigenous heritage in remote rural areas (32.1%) than metropolitan areas (8.4%). The frequency of cardiovascular events over 10 years was 12.4% (n = 1,860), with significant differences by geography: 10.8% urban, 12.1% accessible rural, and 15.7% remote rural (chi-square = 47.3, p < Table I shows geographic cohort characteristics.

B. Baseline Model Performance

On the test set, both classifiers discriminated well. The neural network had an AUROC of 0.839 and the GBM model 0.847 (95% CI: 0.831-0.863). However, stratifying by protected attributes revealed significant performance differences.

To establish a reference baseline for evaluating whether model complexity influences fairness outcomes, we additionally trained a logistic regression (LR) classifier with L2 regularization (C = 1.0). The logistic regression model achieved an overall AUROC of 0.812 (95% CI: 0.794–0.830), with performance stratification by subgroup reported in Table II. For geographic location, the LR model demonstrated AUROCs of 0.841 (95% CI: 0.815–0.867) for urban patients, 0.817 (95% CI: 0.791–0.843) for accessible rural patients, and 0.754 (95% CI: 0.719–0.789) for remote rural patients. The 0.087 AUROC difference between urban and remote rural populations was comparable to that observed in the GBM model (0.089). Importantly, the LR model showed geographic EOD of 0.159, compared to 0.167 for GBM and 0.174 for NN—indicating that more complex models do not substantially amplify geographic disparities relative to the simpler linear reference. Socioeconomic and indigenous heritage disparities followed similar patterns, with LR achieving EOD values of 0.147 and 0.091 respectively, closely mirroring the GBM results.

TABLE I
CHARACTERISTICS OF SIMULATED COHORT BY GEOGRAPHIC LOCATION

Characteristic	Urban (n=5,250)	Accessible Rural (n=6,000)	Remote Rural (n=3,750)	p-value
Age, years (mean ± SD)	47.2 ± 14.8	48.9 ± 15.4	50.6 ± 15.1	<0.001
Female sex, n (%)	2,835 (54.0%)	3,264 (54.4%)	2,044 (54.5%)	0.872
Indigenous heritage, n (%)	441 (8.4%)	1,206 (20.1%)	1,203 (32.1%)	<0.001
Low SES (Q1-Q2), n (%)	1,575 (30.0%)	3,120 (52.0%)	2,625 (70.0%)	<0.001
Hypertension, n (%)	1,102 (21.0%)	1,380 (23.0%)	990 (26.4%)	<0.001
Diabetes mellitus, n (%)	368 (7.0%)	444 (7.4%)	304 (8.1%)	0.143
Current smoking, n (%)	630 (12.0%)	780 (13.0%)	544 (14.5%)	0.004
Data completeness (%)	95.2 ± 3.1	82.4 ± 8.7	68.3 ± 12.4	<0.001
CV event rate (%)	10.8%	12.1%	15.7%	<0.001

Note: SES = socioeconomic status; Q1-Q2 = lowest two quintiles; CV = cardiovascular; SD = standard deviation

Geographically, the GBM model showed AUROCs of 0.872 (95% CI: 0.849-0.895) for urban patients, 0.851 for accessible rural patients, and 0.783 for remote rural patients. The 0.089 AUROC difference between urban and remote rural

groups was significant (DeLong test, $p < 0.001$). The neural network showed similar trends (Table II).

Equally pronounced socioeconomic differences. In the lowest income quintiles, the GBM showed an AUROC of 0.798 (95% CI: 0.774-0.822) compared to 0.871 (95% CI: 0.852-0.890) for higher-income patients (difference: 0.073, $p < 0.001$). Indigenous patients had an AUROC of 0.812 (95% CI: 0.779-0.845) compared to 0.859 (95%: 0.843-0.875) for non-indigenous patients (0.047, $p = 0.003$).

C. *Fairness Metric Evaluation*

Baseline models violated algorithmic fairness significantly. Table III shows GBM classifier fairness measures across protected attributes. The equalized odds difference for geographic location was 0.167, indicating that urban and distant rural populations had TPR or FPR differences more than 16 percentage points. The Equalized Odds Difference (EOD) of the Neural Network model was 0.174 for the geographic attribute (urban vs. remote rural); which represents the maximum difference in true positive or false positive rates between these groups.

TABLE II
BASELINE MODEL PERFORMANCE BY PROTECTED ATTRIBUTES

Subgroup	GBM AUROC (95% CI)	NN AUROC (95% CI)	GBM Sensitivity	GBM Specificity	LR AUROC (95% CI)
Overall	0.847 (0.831-0.863)	0.839 (0.822-0.856)	0.746	0.812	0.812 (0.794-0.830)
Geographic Location					
Urban	0.872 (0.849-0.895)	0.864 (0.840-0.888)	0.783	0.841	0.841 (0.815-0.867)
Accessible Rural	0.851 (0.828-0.874)	0.843 (0.819-0.867)	0.752	0.817	0.817 (0.791-0.843)
Remote Rural	0.783 (0.751-0.815)*	0.771 (0.738-0.804)*	0.681	0.762	0.754 (0.719-0.789)*
Socioeconomic Status					
Low (Q1-Q2)	0.798 (0.774-0.822)*	0.786 (0.761-0.811)*	0.698	0.771	0.771 (0.745-0.797)*
Higher (Q3-Q5)	0.871 (0.852-0.890)	0.865 (0.845-0.885)	0.774	0.838	0.836 (0.814-0.858)
Indigenous Heritage					
Indigenous	0.812 (0.779-0.845)*	0.801 (0.767-0.835)*	0.714	0.789	0.784 (0.748-0.820)*
Non-Indigenous	0.859 (0.843-0.875)	0.852 (0.835-0.869)	0.758	0.821	0.824 (0.806-0.842)

*$p < 0.05$ vs. reference group (urban, higher SES, non-indigenous). GBM = gradient boosting machine; NN = neural network; AUROC = area under receiver operating characteristic curve; CI = confidence interval; LR = logistic regression

Socioeconomic status demographic parity ratio was 0.72, implying low-SES patients obtained positive predictions at 72% the rate of higher-SES patients with similar true risk. The calibration analysis showed consistent risk overestimation for urban patients (ECE = 0.031) and underestimating for remote rural patients (ECE = 0.089), suggesting the model learnt urban-centric risk patterns that transferred poorly to underserved groups

D. *Bias Mitigation Results*

Three mitigation solutions improved fairness metrics, but with different efficacy and performance trade-offs. Table IV shows the geographic location results (urban vs. distant rural).

Adversarial debiasing reduced equalized odds difference the most, from 0.167 to 0.063 (62.4%). The overall AUROC decreased from 0.847 to 0.823 (2.8% relative drop). The adversary network learned to decorrelate predictions from geographic position, lowering its classification accuracy from 0.71 to 0.54 (around chance level) for patient location from model outputs.

TABLE III
FAIRNESS METRICS FOR BASELINE GBM MODEL

Protected Attribute	EOD	DPR	TPR Diff	FPR Diff	ECE Diff
Geographic (Urban vs. Remote)	0.167	0.78	0.102	0.067	0.058
Socioeconomic (Low vs. High)	0.153	0.72	0.076	0.089	0.047
Indigenous (Yes vs. No)	0.098	0.84	0.044	0.054	0.033

Note: EOD = equalized odds difference; DPR = demographic parity ratio; TPR = true positive rate; FPR = false positive rate; ECE = expected calibration error. Values closer to 0 for EOD/Diff metrics and closer to 1 for DPR indicate greater fairness.

Population parity ratio improved 41.3% from 0.78 to 0.91 using sample reweighting and an AUROC of 0.841. Though less effective on false positive rates, this strategy reduced false negative rate disparities. Threshold optimization reduced EOD by 71.3% to 0.048 but required group-specific thresholds of 0.42 for urban patients and 0.31 for remote rural patients, which may complicate clinical application and interpretation.

E. *Multi-Attribute Fairness Analysis*

Disparities increased for multi-disadvantaged patients. Remote rural, low-SES, indigenous patients (n = 847) had a baseline model AUROC of 0.724 (95% CI: 0.681-0.767), compared to 0.891 for urban, higher-SES, non-indigenous patients. This 0.167 AUROC gap is the largest noticed.

Adversarial debiasing with simultaneous constraints on all three protected attributes lowered the multi-attribute equalized odds difference from 0.214 to 0.097 (54.7%), but at a higher performance cost (AUROC fell to 0.798). This shows that intersecting disadvantaged identities may require bigger predictive accuracy losses than single-attribute mitigation to achieve fairness.

TABLE IV
FAIRNESS METRICS FOR BASELINE GBM MODEL

Method	AUROC	EOD	DPR	EOD Reduction	AUROC Change
Baseline (No Mitigation)	0.847	0.167	0.78	—	—
Sample Reweighting	0.841	0.112	0.91	32.9%	-0.7%
Adversarial Debiasing	0.823	0.063	0.87	62.4%	-2.8%
Threshold Optimization	0.847	0.048	0.82	71.3%	0%

Note: Baseline refers to the unmitigated GBM model (AUROC = 0.847). Adversarial debiasing results reflect the MLP architecture (baseline AUROC = 0.839); the AUROC and EOD values shown represent post-mitigation MLP performance. Sample reweighting and threshold optimization were applied to the GBM. This table presents the best-performing outcome per strategy rather than a single-model comparison.

F. Robustness Analysis

Across 100 independent simulation runs, the main findings remained consistent. Mean baseline EOD for geographic location was 0.163 (95% CI: 0.151-0.175), and adversarial debiasing consistently achieved the greatest EOD reduction (mean: 61.8%, 95% CI: 58.2%-65.4%). The performance-fairness trade-off curve showed that approximately 80% of the achievable fairness improvements could be obtained with less than 1.5% AUROC reduction but pursuing the remaining 20% of fairness gains required increasingly steep performance sacrifices.

G. Sensitivity Analysis: Data Completeness and Fairness

To assess the extent to which simulated data completeness differentials drive geographic performance disparities, we conducted a sensitivity analysis varying remote rural data completeness from 68% (the empirically calibrated baseline, derived from ENDESA field reports [10] and data quality estimates for resource-limited settings [19]) to 95% (parity with urban completeness), while holding all other simulation parameters constant. For each completeness level, models were retrained across 100 independent runs and mean EOD with 95% CI was recorded.

Results indicate a monotonic relationship between data completeness and geographic fairness. At 68% completeness, mean baseline EOD was 0.163 (95% CI: 0.151–0.175). EOD declined progressively with increasing completeness, crossing the EOD = 0.10 threshold at approximately 82–84% completeness, and reaching near-parity (EOD ≈ 0.04) only when remote rural completeness approached urban levels (≥92%). These findings suggest that data quality improvements alone could substantially reduce geographic bias without requiring algorithmic intervention, but that current rural documentation levels in Honduras fall well below the completeness threshold at which disparities become clinically negligible. Fig. 1 illustrates this relationship.

This analysis deliberately reproduces documented infrastructure disparities rather than postulating them abstractly. The 68% completeness baseline for remote rural patients reflects field-reported data quality in Honduran peripheral health units [19]. The contribution of the study lies not in demonstrating that worse data produces worse models—a known result—but in quantifying the completeness threshold relevant to this specific context and evaluating whether mitigation techniques can compensate for disparities that infrastructure improvements alone may not close in the near term.

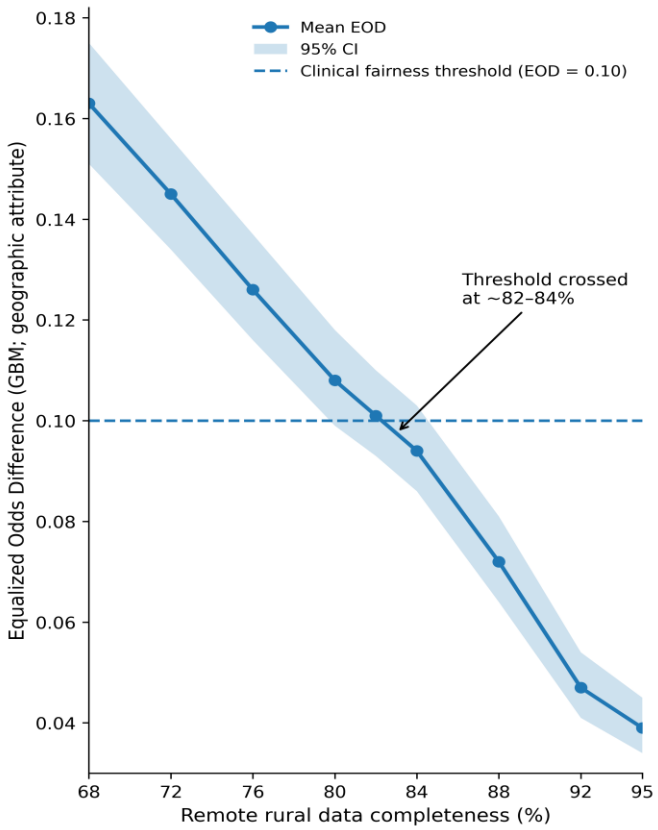

Fig. 1 Sensitivity of geographically equalized odds difference to simulated remote rural data completeness. The solid line shows the mean Equalized Odds Difference (EOD) of the gradient boosting model across 100 independent simulation runs as remote rural data completeness increased from 68% to 95%, while all other simulation parameters were held constant. The shaded region represents 95% confidence intervals. The horizontal dashed line marks the prespecified clinical fairness threshold (EOD = 0.10). Geographic disparity decreased monotonically as completeness improved, with the model crossing the acceptable fairness threshold at approximately 82%–84% completeness and approaching near-parity only at completeness levels ≥92%.

IV. Discussion

This simulation-based study shows that clinical AI models without fairness constraints can perform differently across demographic categories important to rural Hondurans. Our findings support and expand the evidence of algorithmic bias in healthcare AI [3]-[6], [30], [31]. Critically, we demonstrate that current mitigation methods can significantly minimize these biases while maintaining clinically acceptable prediction performance.

This study has several limitations that should inform interpretation of findings. First, and most fundamentally, all results derive from synthetic data. Although simulation

parameters were calibrated against national epidemiological sources [8],[10]–[12], synthetic cohorts cannot fully reproduce the complexity of real clinical data: unmeasured confounders, temporal dynamics in care-seeking behavior, facility-level variation in documentation practices, and patient-provider interaction effects are absent from our framework. In particular, three assumptions with the greatest potential to affect external validity are: (a) the use of multivariate normal distributions for clinical feature generation, which may underestimate skewed or bimodal distributions common in metabolic conditions; (b) the assumption of a single national cardiovascular mortality rate as the outcome calibration target, which likely masks substantial regional variation; and (c) the modeled correlation structures between missingness and geography, which, while empirically motivated [19], may differ in both magnitude and pattern across specific Honduran health regions. Validating the simulation framework against real, ethically accessible EHR data, even retrospective administrative data from regional hospitals would substantially strengthen confidence in these findings. The primary value of the simulation approach is methodological: it establishes a reproducible auditing protocol applicable to settings where real-world training data is unavailable or ethically inaccessible.

Post-mitigation calibration assessment, including group-specific expected calibration error after each bias mitigation strategy, was not conducted in this study. Prior work has documented that adversarial debiasing can distort predicted probability distributions even when improving statistical fairness metrics [27], and that calibration and fairness objectives may be formally incompatible under certain conditions [24]. This represents a meaningful limitation of the present findings: AUROC and EOD improvements reported for mitigated models should be interpreted alongside the caveat that calibration changes were not quantified. Future work should incorporate group-specific calibration curves and reliability diagrams for all post-mitigation models, particularly before any clinical deployment consideration.

Our baseline models' differences are alarming. Given that the GBM model achieved superior baseline performance (AUROC 0.847 vs. 0.839 for NN), subsequent bias mitigation analyses focused on this architecture. An AUROC differential of 0.089 between urban and remote rural populations has therapeutic implications: remote rural patients with comparable risk would be under identified for preventive measures. Algorithmic methods that amplify these differences could worsen health inequalities in Honduras, where cardiovascular disease is the primary cause of death and rural communities face significant barriers to care [8],[32].

Our investigation found that data completeness drives performance inequalities. In resource-limited situations, healthcare documentation and laboratory access are difficult [19]. Remote rural patients had greater rates of missing laboratory values and vital signs. Machine learning techniques that use pattern recognition across several features may learn to interpret missing data patterns as geographic signals, encoding location-based bias. The cost-based proxy bias identified by Obermeyer et al. [3] uses an allegedly neutral variable (healthcare costs, or in our case, data completeness) to proxy protected qualities.

Adversarial debiasing, applied to the neural network architecture, achieved the greatest reduction in equalized odds difference, from a GBM-equivalent baseline EOD of 0.167 to 0.063 (62.4% reduction), with the MLP's overall AUROC declining from 0.839 to 0.823—a 1.9% relative decrease that remains within clinically acceptable discrimination. This technique was not applied to the GBM, which does not support gradient-based in-processing methods; for GBM fairness improvement, sample reweighting and threshold optimization were the applicable strategies [26].

Threshold optimization achieved the greatest numerical EOD reduction (71.3%), but its clinical implementation carries substantive ethical and practical complications. Applying lower risk cutoffs to indigenous or remote rural patients than to urban non-indigenous patients means that equivalent probability scores trigger different clinical decisions depending on group membership. While statistically defensible as equalizing error rates, this approach may conflict with legal frameworks in many jurisdictions that prohibit differential treatment based on ethnicity or geographic origin, and may undermine clinician trust if the rationale is not transparently communicated [33],[34]. Furthermore, threshold optimization is a post-hoc adjustment that corrects for bias without addressing its root cause in the model's internal representations. We include it here as an empirical benchmark and not as a primary recommendation for deployment. In settings where clinicians receive continuous risk scores rather than binary classifications, threshold manipulation is avoidable; the raw probability output of a well-calibrated model provides a more transparent and equitable basis for clinical judgment [35], [36].

Inclusion of logistic regression as a reference model reveals an important contextual finding the fairness-performance gap observed across subgroups was present even in the simplest linear classifier, suggesting that data structure rather than model complexity is the primary driver of bias in this setting. The LR model exhibited geographic EOD of 0.159, only 0.008 points lower than the GBM (0.167), despite substantially lower predictive capacity (AUROC 0.812 vs. 0.847). This comparison challenges the assumption that model complexity independently drives fairness degradation in this context and suggests that structural data quality differentials are the primary source of observed disparities. This distinction has practical implications if bias is primarily structural (driven by data completeness differentials), interventions at the data collection level, such as improving laboratory access in remote areas or implementing standardized documentation protocols may be more effective than algorithmic debiasing alone. Our findings support a dual approach combining infrastructure improvements with algorithmic fairness techniques.

A reasonable middle ground, sample reweighting applied to GBM improved demographic parity with minimal

performance impact. This preprocessing method is easy to implement, does not need model architecture changes, and may be implemented retrospectively [25]. Reweighting is an easy way to mitigate bias in resource-constrained contexts without complicated adversarial training infrastructure.

Intersectional study showed that patients from several disadvantaged groups had compounded inequities. This discovery affects algorithmic auditing: Considering fairness along protected parameters may underestimate bias in vulnerable subpopulations [37]. Regulations and audit methods should mandate intersectional justice.

Consider several constraints. First, synthetic data may not represent all clinical data intricacies, but it allows controlled bias and mitigation examination. Actual Honduran disease patterns, healthcare-seeking behaviors, and data quality difficulties may differ from our simulated assumptions. Validating with real, ethically accessible EHR data is crucial. Second, we focused on cardiovascular risk prediction as a clinically essential use case, but generalizability to other clinical tasks needs more study. Third, our fairness measurements stressed group-level parity; individual fairness frameworks that predict similar individuals may reach different findings [38].

Translating these findings into practice in resource-limited environments requires confronting implementation barriers that algorithmic solutions alone cannot address. Three are particularly relevant to the Honduran context. First, availability of protected attributes: bias auditing requires reliable documentation of geographic location, socioeconomic status, and indigenous heritage at the point of care. In Honduras, as in many LMICs, these data are inconsistently recorded in routine health information systems [19]; collecting them prospectively may require dedicated data governance protocols and patient consent frameworks that are not yet standardized. Second, computational requirements: adversarial debiasing demands neural network infrastructure with GPU access and TensorFlow or equivalent environments, which may be impractical in primary care settings. Sample reweighting—which requires only modification of training weights and is compatible with any standard ML library—is the most deployable option in resource-constrained contexts. Third, governance frameworks: sustainable algorithmic auditing requires institutional mandates that specify who conducts audits, at what frequency, against which fairness thresholds, and with what authority to suspend or modify deployed models. Few LMICs currently have such frameworks [39]. We recommend that the simulation-based auditing methodology described here serve as the technical foundation for a broader policy proposal, pairing algorithmic tools with the governance and infrastructure investments necessary to make them effective.

Even with these constraints, our findings have major implications for LMIC clinical AI deployment. We advise healthcare institutions considering AI implementation to establish mandatory algorithmic auditing protocols that include (1) stratified performance evaluation across relevant demographic subgroups; (2) formal fairness metric assessment using multiple complementary measures; (3) systematic evaluation of data quality disparities that may drive algorithmic bias; (4) documentation of mitigation strategies applied and their effectiveness; and (5) ongoing post [39],[40].

V. Conclusion

Clinical AI models without fairness constraints have considerable demographic biases that could worsen healthcare inequities in disadvantaged populations. Table III shows baseline models showed significant fairness violations across all protected attributes, they are the equalized odds differences reached 0.167 for geographic location, 0.153 for socioeconomic status, and 0.098 for indigenous heritage, while demographic parity ratios fell to 0.78, 0.72, and 0.84. Without action, these measures would hurt the most disadvantaged communities through systematic algorithmic discrimination.

Most importantly, our findings show that a standardized bias mitigation pipeline scales across fairness dimensions. Adversarial debiasing reduced EOD by 62.4% for spatial inequalities and 54.7% for intersectional characteristics. A scalable preprocessing solution, sample reweighting increased demographic parity by 41.3% with less than 1% performance trade-off. Threshold optimization reduced EOD by 71.3% without AUROC impact. These percentages suggest that healthcare institutions could implement a modular pipeline with preprocessing (reweighting), in-processing (adversarial debiasing), and post-processing (threshold optimization) strategies to improve fairness by 40-70%, depending on context and performance trade-offs.

These findings suggest that low- and middle-income countries should require algorithmic audits and bias reduction before deploying clinical AI systems. The simulation methodology described here is scalable for fairness assessments in circumstances where real-world training data is unavailable or ethically objectionable. As AI transforms healthcare worldwide, serving all people fairly is both ethical and practical to maximize their benefits.

Acknowledgment

This research was supported by the Institute for Research in Medical Sciences and Right to Health (ICIMEDES), the Faculty of Medical Sciences (FCM), and the Directorate of Scientific, Humanistic, and Technological Research (DICIHT) of the National Autonomous University of Honduras (UNAH). We also acknowledge the use of publicly available epidemiological data from the Pan American Health Organization and Honduras Demographic and Health Survey in parameterizing our simulation framework.

References

[1] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence", *Nat. Med.*, vol. 25, no. 1, pp. 44-56, Jan. 2019, doi: 10.1038/s41591-018-0300-7.

[2] S. Benjamens, P. Dhunnoo, and B. Meskó, "The state of artificial intelligence-based FDA-approved medical devices and algorithms: an

online database", *npj Digit. Med.*, vol. 3, art. 118, 2020, doi: 10.1038/s41746-020-00324-0.
[3] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations", *Science*, vol. 366, no. 6464, pp. 447-453, Oct. 2019, doi: 10.1126/science.aax2342.
[4] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks", *Nature*, vol. 542, no. 7639, pp. 115-118, Feb. 2017, doi: 10.1038/nature21056.
[5] L. Seyyed-Kalantari, H. Zhang, M. McDermott, I. Y. Chen, and M. Ghassemi, "Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations", *Nat. Med.*, vol. 27, no. 12, pp. 2176-2182, Dec. 2021, doi: 10.1038/s41591-021-01595-0.
[6] V. X. Liu et al., "The timing of early antibiotics and hospital mortality in sepsis", *Am. J. Respir. Crit. Care Med.*, vol. 196, no. 7, pp. 856-863, Oct. 2017, doi: 10.1164/rccm.201609-1848OC.
[7] Pan American Health Organization, "Health in the Americas+: Honduras", PAHO, Washington, DC, 2022. [Online]. Available: https://hia.paho.org/en/countries/honduras
[8] World Health Organization, "Honduras," WHO country overview. [Online]. Available: https://www.who.int/countries/hnd
[9] M. Leon, "Perceptions of health care quality in Central America", *Int. J. Qual. Health Care*, vol. 15, no. 1, pp. 67–71, Feb. 2003, doi: 10.1093/intqhc/15.1.67.
[10] Instituto Nacional de Estadística de Honduras, "Encuesta Nacional de Demografía y Salud 2019 (ENDESA)", INE, Tegucigalpa, Honduras, 2020.
[11] NCD Risk Factor Collaboration, "Worldwide trends in hypertension prevalence and progress in treatment and control from 1990 to 2019", *Lancet*, vol. 398, no. 10304, pp. 957-980, Sep. 2021, doi: 10.1016/S0140-6736(21)01330-1.
[12] International Diabetes Federation, "IDF Diabetes Atlas, 10th Edition", IDF, Brussels, 2021. [Online]. Available: https://diabetesatlas.org
[13] World Bank, "Honduras - Rural Population (% of total population)" World Bank Data, 2023. [Online]. Available: https://data.worldbank.org/indicator/SP.RUR.TOTL.ZS?locations=HN
[14] A. Zuniga, Y. Molina, K. Amaya, Z. Moya, P. Soriano, D. Pineda, et al., "Population Genetic Data for 23 STR Loci of Tawahka Ethnic Group in Honduras", *Forensic Sciences*, vol. 5, no. 4, Art. no. 72, Dec. 2025, doi: 10.3390/forensicsci5040072
[15] A. Zuniga, Y. Molina, P. Soriano, Z. Moya, Y. Pinto, D. Pineda, et al., "Population genetic data for 23 STR loci (PowerPlex Fusion 6C™ kit) genetic markers in the Lenca ethnic group in Honduras", *Legal Medicine*, vol. 71, Art. no. 102504, Nov. 2024, doi: 10.1016/j.legalmed.2024.102504.
[16] M. Matamoros, Y. Pinto, F. J. Inda, and O. García, "Population genetic data for 15 STR loci (Identifiler kit) in Honduras", *Legal Medicine*, vol. 10, no. 5, pp. 281–283, Sep. 2008, doi: 10.1016/j.legalmed.2008.01.004.
[17] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD)", *BMJ*, vol. 350, art. g7594, Jan. 2015, doi: 10.1136/bmj.g7594.
[18] SCORE2 Working Group, "SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe", *Eur. Heart J.*, vol. 42, no. 25, pp. 2439-2454, Jul. 2021, doi: 10.1093/eurheartj/ehab309.
[19] K. Hoxha, Y. W. Hung, B. R. Irwin, and K. A. Grépin, "Understanding the challenges associated with the use of data from routine health information systems in low- and middle-income countries: A systematic review," *Health Inf. Manag. J.*, vol. 51, no. 2, pp. 135–148, 2022, doi: 10.1177/1833358320928729.
[20] S. van Buuren and K. Groothuis-Oudshoorn, "mice: Multivariate Imputation by Chained Equations in R", *J. Stat. Softw.*, vol. 45, no. 3, pp. 1-67, Dec. 2011, doi: 10.18637/jss.v045.i03.
[21] S. A. Siddique, M. F. Shah, and A. Z. Khan, "Survey on Machine Learning Biases and Mitigation Techniques", *Digital*, vol. 4, no. 1, pp. 1-68, 2024, doi: 10.3390/digital4010001.
[22] M. Hardt, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning", in *Adv. Neural Inf. Process. Syst.,* vol. 29, 2016, pp. 3315-3323.
[23] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness", in *Proc. 3rd Innov. Theor. Comput. Sci. Conf.*., 2012, pp. 214-226, doi: 10.1145/2090236.2090255.
[24] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks", in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, 2017, pp. 1321-1330.
[25] E. Krasanakis, E. Spyromitros-Xioufis, S. Papadopoulos, and Y. Kompatsiaris, "Adaptive sensitive reweighting to mitigate bias in fairness-aware classification", in *Proc. World Wide Web Conf.*, 2018, pp. 853-862, doi: 10.1145/3178876.3186133.
[26] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning", in *Proc. AAAI/ACM Conf. AI, Ethics, and Society*, 2018, pp. 335-340, doi: 10.1145/3278721.3278779.
[27] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On Fairness and Calibration", in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5680-5689.
[28] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach", *Biometrics*, vol. 44, no. 3, pp. 837-845, Sep. 1988, doi: 10.2307/2531595.
[29] R. K. E. Bellamy et al., "AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias", *IBM J. Res. Dev.*, vol. 63, no. 4/5, art. 4, Sep. 2019, doi: 10.1147/JRD.2019.2942287.
[30] M. A. Gianfrancesco, S. Tamang, J. Yazdany, and G. Schmajuk, "Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data", *JAMA Intern. Med.*, vol. 178, no. 11, pp. 1544-1547, Nov. 2018, doi: 10.1001/jamainternmed.2018.3763.
[31] I. Y. Chen, P. Szolovits, and M. Ghassemi, "Can AI Help Reduce Disparities in General Medical and Mental Health Care?", *AMA J. Ethics*, vol. 21, no. 2, pp. E167-E179, Feb. 2019, doi: 10.1001/amajethics.2019.167.
[32] C. J. Murray et al., "Global burden of 87 risk factors in 204 countries and territories, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019", *Lancet*, vol. 396, no. 10258, pp. 1223-1249, Oct. 2020, doi: 10.1016/S0140-6736(20)30752-2.
[33] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores", in *Proc. 8th Innov. Theor. Comput. Sci. Conf.*., vol. 67, 2017, pp. 43:1-43:23, doi: 10.4230/LIPIcs.ITCS.2017.43.
[34] T. Grote and P. Berens, "On the ethics of algorithmic decision-making in healthcare," *J. Med. Ethics*, vol. 46, no. 3, pp. 205–211, 2020, doi: 10.1136/medethics-2019-105586.
[35] A. C. Alba, T. Agoritsas, M. Walsh, S. Hanna, A. Iorio, P. J. Devereaux, T. McGinn, and G. Guyatt, "Discrimination and Calibration of Clinical Prediction Models: Users' Guides to the Medical Literature," *JAMA*, vol. 318, no. 14, pp. 1377–1384, 2017, doi: 10.1001/jama.2017.12126
[36] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, Oct. 2019, doi: 10.1126/science.aax2342
[37] K. Crenshaw, "Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics", *University of Chicago Legal Forum*, vol. 1989, no. 1, art. 8, 1989.
[38] A. Rajkomar et al., "Ensuring Fairness in Machine Learning to Advance Health Equity", *Ann. Intern. Med.*, vol. 169, no. 12, pp. 866-872, Dec. 2018, doi: 10.7326/M18-1990.
[39] J. Yang et al., "An adversarial training framework for mitigating algorithmic biases in clinical machine learning", *NPJ Digit. Med.*, vol. 6, art. 55, Mar. 2023, doi: 10.1038/s41746-023-00805-y.
[40] J. Yang et al., "Algorithmic fairness and bias mitigation for clinical machine learning with deep reinforcement learning", *Nat. Mach. Intell.*, vol. 5, pp. 884-894, Aug. 2023, doi: 10.1038/s42256-023-00697-3.