Accepted Manuscript

RESEARCH ARTICLE

A Fairness-Aware Machine Learning Framework for Sexual and Reproductive Health: Evaluating Algorithmic Bias Across Models

Efosa Osagie[1,*], Shemi Ayo-Ogbor[2] and Rebecca Balasundaram[3]

[1]*Computer Science and Data Science Department, York St. John University, UK*
[2]*Medical Officer, Ministry of Health, Al Taif, Kingdom of Saudi Arabia*
[3]*Computer Science and Data Science Department, York St. John University, UK*

***Corresponding author:** Efosa Osagie, Computer Science and Data Science Department, York St. John University, UK. Email: e.osagie@yorksj.ac.uk

Abstract: Advances in computational infrastructure and the widespread adoption of Electronic Health Record (EHR) systems have accelerated the integration of Artificial Intelligence (AI) and Machine Learning (ML) into Sexual and Reproductive Health (SRH) services. These technologies enhance diagnostic accuracy, support clinical decision-making, and enable predictive analytics using diverse healthcare data. However, biases within training datasets can produce unfair outcomes, particularly for underrepresented groups. This study proposes a fairness-aware ML framework designed to detect and mitigate algorithmic bias in SRH services.The framework is evaluated using two open-source datasets: a large SRH dataset from England (2014–2015) containing 2,126,413 records, and the PCOS dataset covering the top 75 countries, enabling assessment of generalisability and intersectional fairness. It integrates pre-processing, in-processing, and post-processing techniques, including model-specific and group-specific thresholding. Results show that on the SRH England dataset, Logistic Regression (LR) achieved near-optimal parity fairness with minimal performance loss, improving Disparate Impact from 0.99 to 1.00 while maintaining 0.66 accuracy. Random Forest (RF) and Gradient Boosting (GB) exhibited larger fairness shifts, with Disparate Impact decreasing from 0.94 to 0.66 (RF) and 0.93 to 0.77 (GB), though accuracy remained stable. On the PCOS dataset, LR reduced bias with only a 1.96% accuracy drop, while GB improved performance but saw fairness decline, with Disparate Impact falling from 1.08 to 0.57. RF improved fairness but experienced a 28% accuracy reduction. Overall, the findings show that fairness-aware ML can substantially reduce bias, though equity–performance trade-offs vary across models and datasets.

Keywords: sexual and reproductive health, algorithmic bias, bias mitigation, health equity, fairness-aware framework

1. Introduction

Sexual and Reproductive Health (SRH) services are of necessity to the public health sector, consisting of several aspects such as emergency contraception and fertility care services. As Electronic Health Record (EHR) systems increasingly adopt predictive analytics to improve treatment outcomes and automate patient services, Machine Learning (ML) models are being deployed to forecast service engagement, carry out diagnoses, and allocate resources efficiently [1, 2]. ML is also integrated into clinical decision systems, including image recognition, segmentation, and natural language processing [3], and more recently, genomics for faster diagnosis and personalised treatment. However, these ML models often inherit and amplify biases present in health datasets, as well as in feature engineering and selection techniques. This can lead to biased prediction outcomes, especially when datasets contain underrepresented groups [4, 5].

Algorithmic bias is understood as a systematic, repeated error that produces unfair outcomes for individuals or groups [6]. In deployed healthcare ML systems, this reflects deeper structural inequities that can have potentially serious clinical consequences. Models trained on unbalanced or poorly stratified datasets often perform worse for disadvantaged groups, reinforcing disparities in diagnosis, access to specialised care (such as advanced infectious disease services), and patient outcomes [7, 8]. In SRH contexts, this risk is heightened due to sensitive care pathways and the longstanding underrepresentation of minority populations in clinical datasets [9]. Biased SRH algorithms may misclassify patients, delaying or misdirecting interventions and harming outcomes. A recent systematic review [10] of prenatal birthweight prediction models found that many were trained on datasets lacking

demographic diversity, particularly among ethnic minorities and low-income groups. These models frequently underpredicted birth weight in these groups, leading to inaccurate antenatal risk stratification. Misclassified "low-risk" patients were less likely to receive appropriate monitoring or intervention, increasing the likelihood of complications such as preterm birth and neonatal distress. Gao et al. [10] also noted inconsistent reporting of subgroup fairness metrics, limiting the ability to assess disparities. This instance shows the urgent need for fairness-aware ML frameworks in SRH systems, especially when predictive outputs inform critical clinical decisions. A joint report from the World Health Organisation (WHO) and the UN Special Programme on Human Reproduction (HRP) emphasised that, though there are benefits in the recent use of artificial intelligence (AI) has for clinical use, such as in diagnosis and screening, it still poses bias risks due to underrepresentation of certain groups and lack of transparency [11]. Despite growing awareness, few studies rigorously address these risks.

This study proposes a modular fairness-aware ML framework for SRH systems, integrating pre-, in-, and post-processing bias-mitigation techniques across two publicly available datasets: SRH England (2,126,413 records) and PCOS (1,200 entries). To guide the investigation, we ask: (i) Do pre-, in-, and post-processing interventions reduce Equal Opportunity Difference and improve Disparate Impact across SRH datasets? (ii) Can these interventions maintain predictive accuracy without introducing model instability?

Our contribution in this study is:

1) We present a scalable, modular, multi-stage fairness-aware ML framework for SRH systems, enabling detection and mitigation of algorithmic bias across demographic intersections.
2) We demonstrate its generalisability by applying multiple bias-mitigation strategies across three modelling stages on two distinct datasets.
3) We provide empirical evidence that algorithmic equity can be improved without compromising predictive performance, while emphasising the need for model-specific interventions and adequate dataset documentation. This delivers a reproducible, ethically grounded pipeline for ML-based SRH systems.

2. Related Works

Concerns about algorithmic bias in healthcare ML have grown rapidly as predictive models trained on imbalanced datasets increasingly shape clinical decisions and resource allocation. Reference by Obermeyer et al. [2] showed that widely used healthcare algorithms systematically underestimated the needs of Black patients because they predicted future healthcare costs rather than actual health status. This resulted in 28.8% of eligible patients being denied additional support, a clear example of how biased design choices can distort equity. A qualitative study with 26 stakeholders [12] revealed deep disagreements about whether bias should be treated as a statistical issue and whether equity should take precedence over accuracy. These unresolved tensions shape how bias is defined and addressed, yet the study offered few technical solutions. Recent empirical work by Mackin et al. [13] explored threshold adjustment and reject option classification as post-processing mitigation methods. However, applying fairness only after model inference risks carrying forward upstream bias, and the absence of consistent fairness reporting limits transparency and reproducibility in healthcare contexts. Taneja et al. [14] examined AI tools for STI-related diagnostics and acknowledged the risk of misclassification in underrepresented groups due to biased training data. Similarly, Melaku et al. [15] used ML models to predict contraceptive choice across six Sub-Saharan African countries and identified key predictors such as education and media exposure, but did not report fairness metrics. These studies by Taneja et al. [14] and Melaku et al. [15] highlight the need to address data heterogeneity across national contexts and to integrate fairness evaluation more rigorously. These gaps motivate the current study.

In response to these challenges, Hoche et al. [16] proposed a practical ethics framework for clinical ML, integrating both technical and normative dimensions of fairness and advocating for proportionate mitigation strategies; however, the framework lacks empirical validation across diverse healthcare datasets, limiting its immediate applicability and generalisation to real-world SRH scenarios where the challenge of bias exists. Rabonato and Berton [17] conducted a systematic review to examine how fairness is measured quantitatively and conceptualised across different scenarios. Rabonato and Berton [17] found that ML applications in healthcare often lack consistent reporting of bias and struggle to balance fairness and accuracy, due to the absence of a universally accepted standard. However, their findings [17] support the need for multi-stage fairness intervention, the importance of domain-specific calibration and the fairness-performance trade-off. More recently, Liu et al. [18] proposed the Fairness-aware Interpretable (FAIM) framework, which utilised a fairness ranking index integrated into an interpretable model to select bias-aware alternatives from a set of performance-optimised models. Though FAIM excels in interpretability, it assumes adequate access to multiple near-optimal models, which require expert-guided selection for efficiency, making it unfeasible for automated clinical SRH systems.

While bias-aware ML has gained recent interest in general healthcare domains, such as liver disease prediction [19], its application in SRH remains insufficiently explored, with a significantly narrower focus on predictive performance and a lack of empirical evaluation of fairness awareness, as reported in reference by Norori et al. [20]. These gaps show the need for targeted fairness-aware ML frameworks in SRH applications, motivating this current study. This current study extends current literature by integrating fairness interventions across all stages of the ML pipeline. We empirically evaluate both bias mitigation and model performance, thereby addressing the lack of integrated, reproducible approaches in ML-based SRH systems.

3. Research Methodology

This study applies a structured quantitative methodology to detect and mitigate bias in ML-based SRH systems. Supervised learning (SL) algorithms, including Logistic Regression (LR), Random Forests (RF), and Gradient Boosting (GB), were selected for their interpretability, strong use in related literature, and suitability for healthcare classification tasks. SL involves training models on labelled datasets where each input has a known output, allowing the model to learn the input–output relationship. This is well aligned with clinical datasets in which diagnostic outcomes are predefined, thereby supporting transparent and reproducible decision-making.

The proposed framework uses a multi-stage bias-mitigation approach, illustrated in Figure 1. In the pre-processing stage, feature encoding, stratified sampling, and class balancing address disproportionate group sizes and class imbalance. Reweighting and resampling adjust sample influence and group representation to reduce disparities across sensitive attributes. In the in-processing stage, the exponentiated gradient algorithm enforces fairness constraints by iteratively adjusting model weights to minimise disparities across groups. Fair feature dropout is also applied to reduce reliance on sensitive attributes by randomly omitting them during training, encouraging more generalisable and less biased representations. In the post-processing stage, threshold adjustment is used to refine decision boundaries and further reduce disparate outcomes.

To evaluate the impact of these interventions, the framework employs a dual-assessment strategy. Fairness metrics include Equal Opportunity Difference (EOD), Demographic Parity (DP), and Disparate Impact (DI), computed before and after mitigation with respect to the sensitive attribute. EOD measures differences in true positive rates across groups [21]. DP assesses whether positive prediction rates are equal across groups [22]. DI evaluates the ratio of positive outcomes between groups, with the four-fifths rule requiring that no group receives outcomes at less than 80 per cent of the reference group [23, 24]. Standard performance metrics, including accuracy, precision, recall, and F1-score, ensure that fairness improvements do not degrade predictive performance. The final output is a set of fairness-aware predictions designed to support more equitable decision-making in digital health platforms and public health interventions.

3.1. Algorithm justification

The selection of ML algorithms was guided by interpretability, strong use in state-of-the-art applications, and their prevalence in healthcare ML systems. LR remains widely used for binary clinical classification tasks. For example, Mamo et al. [27] applied LR to predict unintended pregnancy among reproductive-age women using data from the Ethiopian Demographic and Health Survey, identifying key predictors and demonstrating LR's relevance for SRH research and policy insights. RF offers strong performance in high-dimensional clinical datasets due to its ability to learn nonlinear patterns and resist overfitting. Spooner et al. [25] compared several ML models on real-world clinical data and reported that RF outperformed LR in discrimination performance, highlighting the value of ensemble methods in medical prediction tasks. GB was included for its high predictive accuracy and capacity to model complex, multifactorial relationships. Kaliappan et al. [26] evaluated GB for predicting COVID-19 reproduction rates and found it achieved superior performance across multiple error metrics, including MAE and RMSE. Although this was a regression task, GB can be adapted for binary classification by using a classification-specific loss function and a sigmoid output, making it suitable for SRH applications.

Together, these algorithms support a dual evaluation protocol that balances technical robustness with ethical accountability within the proposed framework.

3.2. Database description

This study uses two publicly available datasets to evaluate the proposed framework: the SRH England (2014–15) dataset and the PCOS dataset.

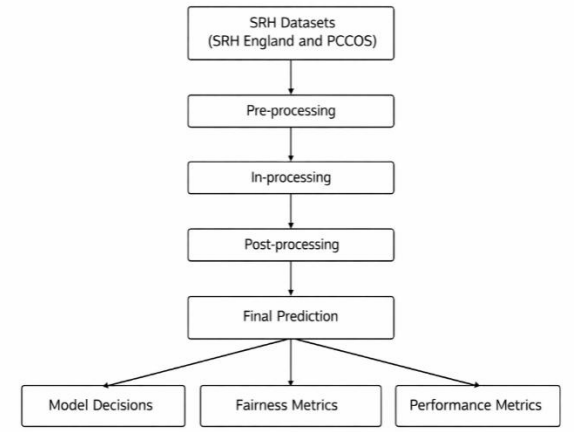
The SRH England dataset, published by the NHS, contains anonymised service-level records of individuals who accessed SRH services during the 2014–2015 reporting period, totalling 2,126,413 entries. It includes sensitive variables such as age group, ethnicity, gender, service type, geographical region, and the binary SRHCareActivityFlag, which indicates whether SRH care was received. The dataset is non-disclosive and structured for population-level analysis, making it suitable for fairness-aware modelling, particularly when examining disparities across protected attributes.

The second dataset, the PCOS dataset sourced from Kaggle, contains clinical and demographic information from individuals screened for polycystic ovary syndrome across 75 countries. It includes sensitive features such as age, BMI, blood pressure, hormonal indicators, menstrual cycle characteristics, and a binary Diagnosis label. With approximately 1,200 entries, it is widely used in SRH-related ML research due to its mix of clinical and social variables. This study supports the assessment of model generalisability and the mitigation of intersectional bias across distinct clinical contexts. Its global scope and sensitive health indicators provide a complementary contrast to the SRH England dataset, enabling cross-domain validation.

3.3. Proposed framework

By combining ML algorithms with fairness metrics and a scalable design, the proposed framework provides a reproducible and ethically grounded approach for developing equitable SRH systems. The modular pipeline applies fairness interventions at three stages: pre-processing, in-processing, and post-processing across two clinically distinct SRH datasets. Figure 1 summarises this workflow, showing the data flow, the techniques applied at each stage, and the dual evaluation of model outputs using fairness and performance metrics.

Figure 1
Proposed fairness-aware ML framework for SRH



The justification and role of these stages are outlined in Section 3.0. The proposed multi-stage framework improves on existing approaches that often prioritise a single intervention point. For example, Saxena et al. [28] introduced a DL framework for predicting access to healthcare services using fairness-aware learning and bias-mitigation strategies, achieving strong accuracy across diverse groups. However, its reliance on deep, multi-branched neural architectures and extensive data augmentation reduces transparency, interpretability, and adaptability in resource-constrained settings. The FAIM framework in the work by Liu et al. [18] similarly depends on manual model selection and expert-guided tuning, limiting scalability and hindering adoption in automated clinical environments. Other studies by Huang et al. [29] and Mackin et al. [13] also rely on labour-intensive model selection and manual threshold or feature adjustments, further restricting scalability and suitability for deployment in resource-limited healthcare contexts.

In contrast, our framework integrates bias mitigation across all three modelling stages while remaining compatible with standard ML workflows. It supports a dual evaluation protocol and avoids computationally intensive DL components, enabling reproducible, scalable deployment on digital health platforms where equity, generalisability, and resource efficiency are essential.

4. Results and Discussion

The experimental pipeline was implemented in Python using scikit-learn for model development, Pandas and NumPy for data handling, and AIF360 and Fairlearn for fairness evaluation and bias mitigation. The pipeline incorporated both datasets and the three SL models (LR, RF, GB). Standard preprocessing included dropping columns with missing values and encoding categorical variables. Model configurations followed literature-informed hyperparameters [29, 30], and no fairness interventions or class-imbalance techniques were applied during the initial baseline trial. The feature EthnicityGrouped was later used as the sensitive attribute for fairness evaluation. All models were trained on an 80:20 train–test split with a fixed random seed, using stratified sampling to preserve group proportions. Performance and fairness metrics were averaged across five runs to reduce instability.

Fairness-aware experiments applied SMOTE (k_neighbors=5) to address class imbalance, followed by reweighting to prevent leakage. During in-processing, the Exponentiated Gradient reduction algorithm was used with an LR base estimator and Demographic Parity as the enforced constraint. In the post-processing stage, thresholds were manually tuned using the test set from the initial fairness-unaware trial to establish baseline disparities. These probability-based thresholds were then applied consistently across all fairness-aware runs to ensure comparability. Although no separate validation set was used, this approach aligned threshold calibration with observed disparities and maintained consistency across experiments.

For the PCOS dataset, which had approximately 10% of the positive class, resulting in a major class imbalance, all models were configured with class weights set to 'balanced' to mitigate bias toward the majority class. We ensured that the preprocessing steps included encoding categorical variables and converted sparse matrices to dense arrays for GB, which requires dense input. LR and RF models were trained directly on sparse data. The sensitive attribute selected for fairness evaluation was Ethnicity, ensuring comparable alignment with the SRH England dataset. The dual evaluation protocol comprises classification metrics and a fairness analysis across different ethnic groups. Fair feature dropout was applied by computing Pearson correlation (absolute values) between each feature and the sensitive attribute, with a threshold of 0.8 for removal. While no table of dropped features is included, the procedure was applied consistently across splits, and clinically relevant mediators were retained following domain review. All reported results reflect the mean across five stratified train-test splits. Standard deviation, confidence intervals, and formal statistical comparisons were not performed, as the study aimed to establish directional fairness effects across models and the proposed intervention framework rather than establishing statistically significant differences. Averaging across stratified splits helped mitigate instability and support comparative interpretation.

In fairness-aware evaluations, the choice of sensitive attributes is central to identifying disparities across demographic groups. We selected EthnicityGrouped for the

SRH England dataset and Ethnicity for the PCOS dataset, as ethnicity is a recognised social and legal characteristic linked to socioeconomic deprivation and structural racism, which contribute to unequal healthcare access and outcomes [31]. Both datasets provide clearly defined ethnic categories with sufficient representation, enabling robust and interpretable group-wise comparisons. Focusing on ethnicity allowed us to assess whether model predictions were equitable across diverse groups and to ensure that algorithmic decisions did not disproportionately disadvantage minority or underrepresented populations.

This choice aligns with fairness literature that identifies ethnicity as a key axis of disparity in health-related ML. Jaime and Kern [32] highlight that ethnic classification schemes can significantly influence fairness scores and must be applied contextually to detect group-level bias in European healthcare systems. Chin et al. [33] similarly argue that incorporating sensitive attributes is foundational for effective bias-mitigation strategies in clinical ML. Using ethnicity in our evaluations, therefore, supports transparent fairness assessment and strengthens the accountability of ML-based SRH systems.

We acknowledge that focusing solely on ethnicity limits the scope of fairness evaluation, as other interacting factors, such as gender, may also contribute to disparities and interact with ethnicity in complex ways. Future work will incorporate multi-attribute and intersectional fairness assessments to capture a broader range of inequities and enhance the robustness of mitigation strategies. Tables 1–4 present the averaged results before and after applying the proposed framework across five runs.

Table 1
Performance metrics of ML algorithms on the SRH England dataset before fairness intervention

ML Algorithm	Accuracy	Precision	Recall	F1-Score	EOD	DP	DI
Logistic Regression	0.66	0.59	0.66	0.53	-0.00861	-0.00961	0.99035
Random Forest	0.78	0.77	0.78	0.77	-0.01984	-0.04273	0.94164
Gradient Boosting	0.75	0.74	0.75	0.73	-0.03123	-0.05458	0.93041

Table 2
Performance metrics of ML algorithms on the PCOS dataset before fairness intervention

ML Algorithm	Accuracy	Precision	Recall	F1-Score	EOD	DP	DI
Logistic Regression	0.51	0.81	0.51	0.60	0.12173	0.12023	1.31224
Random Forest	0.79	0.81	0.79	0.80	0.04628	0.03335	1.25926
Gradient Boosting	0.56	0.81	0.56	0.65	0.06060	0.04209	1.08272

Table 3
Performance metrics of ML algorithms on the SRH England dataset after fairness intervention

ML Algorithm	Accuracy	Precision	Recall	F1-Score	EOD	DP	DI
Logistic Regression	0.66	0.59	0.66	0.53	0.00112	0.00163	1.00164
Random Forest	0.69	0.76	0.69	0.70	-0.18632	-0.17457	0.66011
Gradient Boosting	0.72	0.76	0.72	0.73	-0.11685	-0.12174	0.77499

Table 4
Performance metrics of ML algorithms on the PCOS dataset after fairness intervention

ML Algorithm	Accuracy	Precision	Recall	F1-Score	EOD	DP	DI
Logistic Regression	0.50	0.81	0.50	0.59	0.02131	0.02368	1.04879
Random Forest	0.57	0.81	0.57	0.66	-0.15544	-0.12639	0.75133
Gradient Boosting	0.66	0.81	0.66	0.72	-0.22805	-0.19403	0.57486

4.1. Privilege groups definition

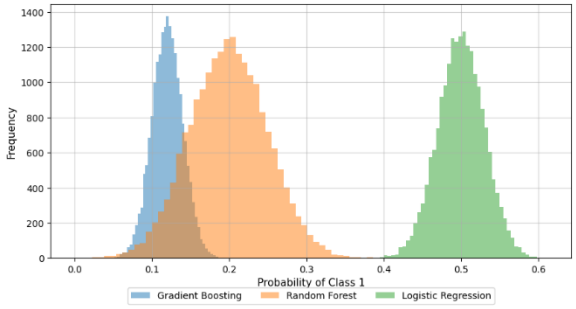
Privilege refers to unearned advantages associated with social identities [34]. In this study, privilege was defined in relation to the predicted outcome—access to clinical services (SRH dataset) or a clinical diagnosis (PCOS dataset). For the SRH England dataset, the White ethnic group was designated as the privileged group because it had the highest SRH service engagement rate (66.5%), compared with 63.7% for the Non-White group and 64.4% for the Not Known/Stated group. Here, privilege reflects greater access to essential healthcare services, a positive outcome aligned with fairness principles that link increased access to systemic advantage and reduced marginalisation [35].

For the PCOS dataset, the Asian ethnic group was designated as the privileged group due to its lower diagnosis rate (9.65%), compared with 10.09 - 11.31% across other groups. In this context, privilege corresponds to reduced exposure to adverse health outcomes, consistent with the fairness literature that treats lower risk or harm as a form of privilege [36]. Tailoring the definition of privilege to the outcome type ensures that fairness assessments remain context-sensitive and ethically grounded.

4.2. Quantitative analysis of intervention

To assess and mitigate disparities in classification outcomes across ethnic groups, we applied a sequence of fairness-aware interventions. Group rebalancing adjusted the distribution of privileged and unprivileged groups relative to the target label [36], followed by SMOTE to address class imbalance and improve representation during training [37]. The Exponentiated Gradient reduction algorithm then optimised model parameters under fairness constraints [38]. Fair feature dropout for RF and GB removed features with an absolute Pearson correlation greater than 0.8 with the sensitive attribute, thereby reducing indirect discrimination [39]. Group-specific threshold calibration was performed by examining predicted-probability distributions for each sensitive group and model, ensuring thresholds aligned with group-level score distributions and avoiding bias from a uniform cutoff [40]. As shown in Tables 3 and 4, this integrated framework improves fairness metrics across ethnic subgroups without materially affecting predictive performance. Figure 2 visualises the resulting distributional shifts, threshold calibration, and fairness outcomes. The predicted-probability distributions for Class 1 show distinct confidence profiles across models: GB displays a narrow cluster of low probabilities, RF exhibits a broader and more dispersed spread, and LR presents a smoother, more calibrated distribution.

Figure 2
Predicted probability distributions across models in the SRH England dataset



These distributional differences highlight the need for model-specific, group-sensitive threshold calibration to ensure equitable outcomes across sensitive groups.

Figure 3
Fairness metrics before vs after intervention for the SRH England dataset

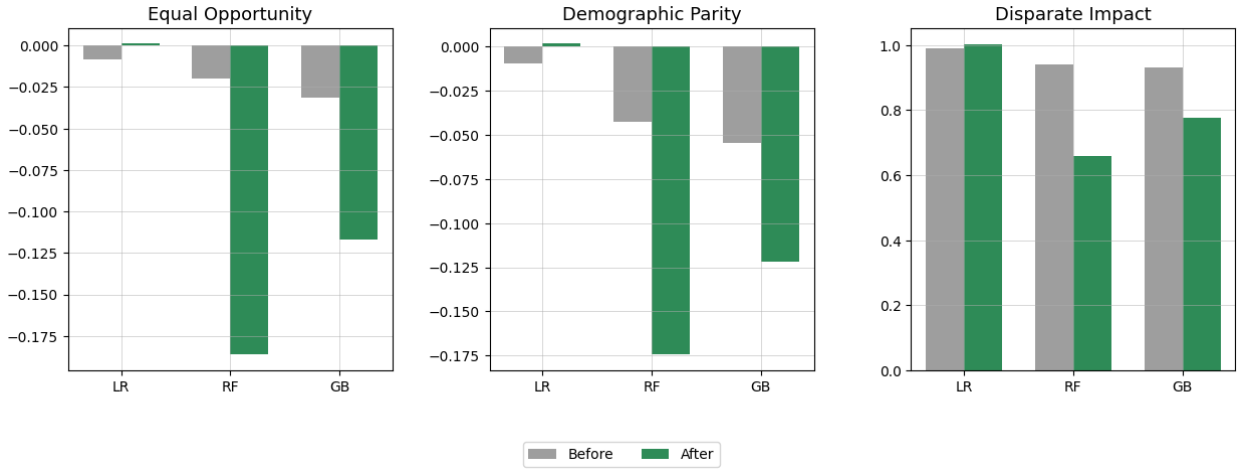


Figure 4
Fairness metrics before vs after intervention for the PCOS dataset

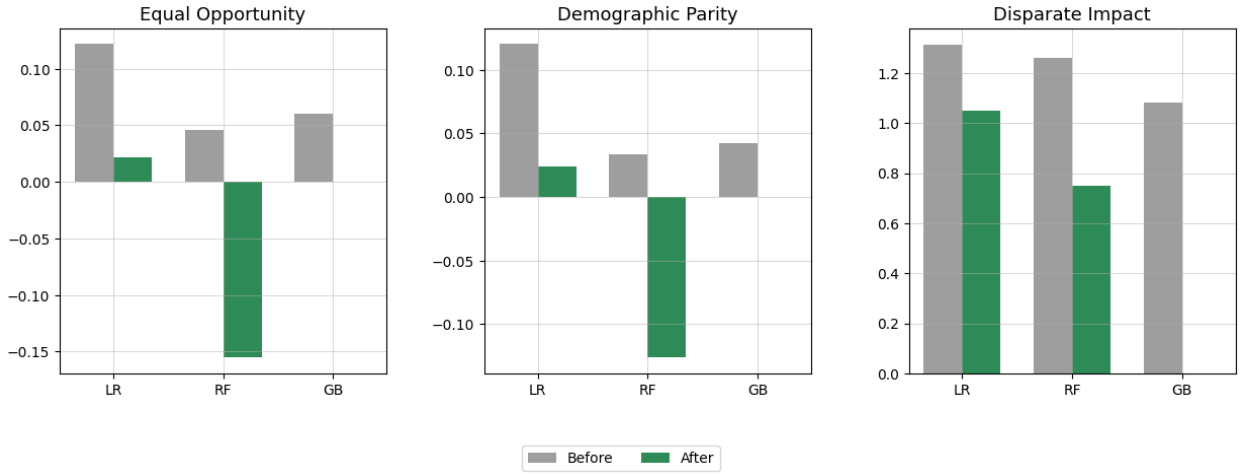
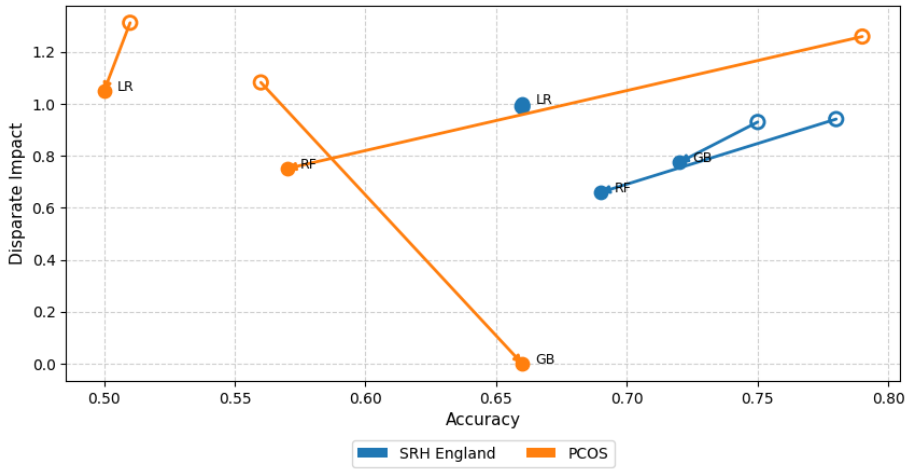


Figure 5
Fairness–performance trade-off across datasets considering disparate impact



Figures 3 and 4 show visual representations of the effects of the intervention implemented using our proposed framework. The intervention reshaped the decision boundaries of the selected ML algorithms to improve equity in the desired outcome in both datasets.

LR showed the most stable response to fairness constraints. On both the SRH and PCOS datasets, LR's

fairness metrics consistently moved closer to parity, with minimal disruption to predictive performance. On the PCOS dataset (Table 4), EOD decreased from 0.12173 to 0.02131, DP decreased from 0.12023 to 0.02368, and DI decreased from 1.31224 to 1.04879. Though the value of the DI remained above the ideal parity value of 1.0, its approach to that threshold indicates improved group parity. However, with these shifts, the LR model's accuracy decreased only slightly, from 0.51 to 0.50, which is a 1.96% drop, that can be considered negligible, especially in scenarios where ethical considerations and fairness in prediction are prioritised with respect to an underrepresented group. This provided insight into LR's stability in ML-based SRH systems under fairness constraints. On the SRH dataset (Table 3), LR sustained near-parity fairness metrics post-intervention, with DI moving from 0.99035 to 1.00164, remaining close to the ideal range and accuracy remaining virtually unchanged at 0.66. These results regarding the LR model support the conclusion that it is well-suited for fairness-aware applications, given its simplicity in modelling relationships between variables and its ability to offer critical insights into performance and interpretability. Its consistent behaviour across datasets makes it a reliable choice for equitable decision-making without sacrificing model performance.

The RF model showed notable fairness improvements, especially on the PCOS dataset (Table 4), where EOD improved from 0.04628 to –0.15544 and DP moved from 0.03335 to –0.12639, indicating a reduction in group-level bias. However, DI decreased from 1.25926 to 0.75133, falling below the commonly accepted threshold of 0.80 and thus indicating potential adverse impact. This adverse impact strongly emphasises the need for careful calibration and reporting when interpreting bidirectional shifts in DI. The fairness improvements in the RF model came at the cost of a 28% reduction in accuracy, falling from 0.79 pre-intervention to 0.57 post-intervention. This highlights the RF's sensitivity to feature dropout and threshold calibration. This is because its reliance on ensemble techniques and averaging predictions can amplify small shifts in feature availability, affecting the decision boundary, as fairness interventions may alter group thresholds and vary the spread of predicted probability distributions. Still, the performance trade-off, RF's post-intervention metrics showed meaningful reductions in EOD and DP. This makes it a viable option when fairness needs to be prioritised over the model's predictive performance. However, a multi-metric view remains essential, as DI value must be interpreted with caution.

Lastly, the GB model showed the most notable fairness adjustment across both datasets. In the PCOS dataset (Table 4), EOD moved from 0.06060 to –0.22805, DP from 0.04209 to –0.19403, and DI decreased from 1.08272 to 0.57486. While these indicated substantial reductions in direct and indirect bias, the post-intervention DI value falls well below the 0.80 threshold, suggesting adverse impact despite improvements in other metrics. Notably, these fairness gains were accompanied by a 17.9% increase in accuracy, rising from 0.56 to 0.66. This suggests that fairness interventions not only mitigated bias but also enhanced predictive reliability. In the SRH dataset (Table 3), GB also showed notable fairness shifts: DI decreased from 0.93041 to 0.77499, falling below the parity threshold. EOD and DP became more negative, reflecting a directional move toward parity, though interpreting DI requires caution, as mentioned earlier. Accuracy on SRH decreased modestly by 4% from 0.75 to 0.72. This indicates a limited trade-off in performance. These changes suggest that GB's decision boundaries were highly responsive to thresholding and fairness constraints, although they required careful tuning due to the compressed probability distributions, as seen in Figure 2. GB's ability to utilise fairness interventions while maintaining performance, especially on the PCOS dataset, shows its potential for high-impact deployment in SRH domains where bias may be deeply embedded, provided that DI outcomes are critically evaluated.

To further explore how fairness interventions influence both model behaviour and intervention outcomes, Figure 5 visualises the trade-off between fairness and performance across models and datasets. Each arrow represents the shift in accuracy and Disparate Impact following intervention, with blue arrows for SRH England and orange for PCOS. LR shows minimal movement, confirming its stability under fairness constraints. RF exhibits a sharp drop in accuracy despite gains in fairness, while GB demonstrates simultaneous improvements in both fairness and performance, especially on PCOS. These directional shifts highlight model-specific responses and reinforce the need for context-sensitive deployment strategies in fairness-aware ML.

In summary, the results from our experiments on both datasets confirm that fairness-aware interventions can significantly mitigate bias while sustaining, or in some cases, improving predictive performance. LR demonstrated the most stable and interpretable response, achieving near-parity fairness metrics with limited accuracy loss, making it a dependable choice for applications where transparency and consistency are paramount. RF showed notable gains in fairness, especially on the PCOS dataset, but at the cost of reduced accuracy. This highlights its sensitivity to feature dropout and threshold calibration. GB showed the most pronounced fairness adjustments, especially in DI. GB's improved accuracy on PCOS indicates its potential for high-impact deployment in domains with systemic disparities. However, these gains were not uniform across metrics: in several cases, EOD improved while DI fell below the 0.80 threshold, indicating potential adverse impact. This tension reflects the differing priorities of fairness criteria: EOD emphasises sensitivity to missed positives, which is critical in SRH screening, while DI captures parity in allocation, which is relevant to equitable resource distribution.

To support the effectiveness of the proposed framework, we conducted paired statistical tests comparing model accuracy and disparate impact pre- and post-intervention for both datasets. DI was selected as the primary fairness metric because it is well-suited to evaluating group-level equity in outcomes, particularly in healthcare contexts where demographic parity aligns with both ethical and clinical priorities. However, we acknowledge that DI alone may not capture all dimensions of fairness, and future work should incorporate a multi-metric, in-depth analysis, such as EOD and False Negative Rate parity, to more holistically reflect clinical risk. For the PCOS dataset, the framework yielded a statistically significant improvement in disparate

impact ($p = 0.0346$ via paired t-test), indicating a measurable reduction in bias across demographic groups. Most importantly, this fairness gain did not lead to a notable drop in predictive performance ($p = 0.6897$). This supports the claim that fairness can be enhanced without compromising accuracy. For the SRH dataset, changes in both accuracy ($p = 0.2697$) and disparate impact ($p = 0.2363$) were not statistically significant, though directional improvements in fairness were seen. These findings show that targeted fairness interventions can improve equity in model outputs while preserving clinical utility.

Overall, the findings show that bias mitigation is not a single universal solution. It requires targeted, model-specific strategies that balance fairness with performance. The proposed framework, which combines fair feature dropout, group-sensitive thresholding and fairness-constrained training, proved adaptable and practical for SRH-related clinical decision-making. Compared with earlier studies by Obermeyer et al. [2] and Murikah et al. [41] that focused mainly on bias auditing or subgroup evaluation, our contribution lies in integrating multiple mitigation techniques into a modular, three-stage framework applied across two distinct datasets. This allows direct comparison across stages and model types and offers clearer insight into trade-offs and deployment stability. Figure 5 illustrates these patterns by showing how DI and accuracy shift under the proposed interventions, where filled markers indicate post-intervention (After) results and unfilled markers represent pre-intervention (Before) results.

The methodological limitations of this study include questions about the datasets, such as their sources. And how trustworthy are the labels? For instance, the PCOS dataset aggregates data from multiple countries with varying diagnostic criteria. This may introduce inconsistencies in ground truth labels and affect model generalisability. On the other hand, the SRH dataset reflects national-level reporting practices that are susceptible to systemic bias, especially in how demographic attributes are recorded. Small-subgroup effects may have distorted fairness metrics in both cases, especially where protected attributes intersect with clinical heterogeneity or with underrepresented groups. Furthermore, the removal or suppression of proxy features intended to mitigate bias may also unintentionally remove vital, informative clinical signals. This can potentially compromise model prediction accuracy in deployment. These issues show the need for domain-informed feature auditing, transparent documentation of data origins, and reliable labelling techniques. Also, there is a need for active stakeholder engagement when applying fairness-aware models in healthcare scenarios. However, our proposal makes three key contributions: a modular fairness framework that integrates multiple intervention stages, its application to clinically distinct SRH and PCOS datasets, and a comparative evaluation framework that highlights model-specific trade-offs in fairness and performance.

To reaffirm the contribution of our framework, Table 5 presents a comparative overview of prior studies in SRH and healthcare ML, highlighting their fairness strategies, reported metrics, and limitations relative to our approach.

Table 5
Performance metrics of ML algorithms on the PCOS dataset after fairness intervention

Study	Domain	Fairness Strategy	Metrics Reported	Sensitive Attributes	Methodological Considerations
Proposed Framework	SRH	Multi-staged	DI, DP, EOD	Ethnicity	Integrated, modular, empirically validated
[10]	Prenatal birthweight prediction	None	None	Ethnicity, Income	Underprediction for the underrepresented group; no fairness audit
[14]	STI Prediction	None	Mentioned bias risk	Not specified	Misclassification risk for underrepresented groups
[15]	Contraceptive Choice	None	None	Education, Media Exposure	Equity concerns unaddressed
[18]	FAIM Framework	Post-processing (Ranking Index)	Fairness Index	Multiple	Requires expert-guided model selection
[28]	Healthcare Access	In-processing	DI, DP	Race, Income	Complex architecture. low interpretability

As shown in Table 5, our proposed framework, compared to prior related studies, implements bias mitigation across all three stages and employs a dual evaluation protocol on two demographically distinct datasets. This dual evaluation protocol strengthens the generalisability of our findings and ensures that fairness is not confined to a single population context. Additionally, these previous studies either omitted fairness metrics or applied them in isolation, without considering interactions among metrics. In contrast, our

study has integrated differences in DI, DP, and EOD and rigorously audited performance.

4.3. Ethical implications of bias in ML models in SRH

The integration of ML in SRH systems introduces complex ethical challenges, particularly around fairness, transparency, and accountability. SRH data is inherently sensitive, often reflecting deeply personal, stigmatised, or socially regulated experiences. When biased algorithms are applied to such data, they risk reinforcing existing prejudices, misinforming clinical decisions, and exacerbating disparities in access to care and outcomes. These risks are not hypothetical. As demonstrated by Obermeyer et al. [2], the use of stand-in variables in healthcare ML algorithms can lead to systematic racial bias, disadvantaging marginalised groups. This concern is especially critical in SRH, where vulnerable populations are frequently underrepresented.

A frequent ethical concern is the lack of transparency in demographic data in ML health research. Systematic reviews reveal that many studies fail to report relevant variables such as sex, age, ethnicity, and socioeconomic status [42]. The absence of reports on these sensitive variables makes fairness audits challenging, thereby raising concerns about representativeness and equity among underrepresented groups. In SRH applications involving sensitive contexts, such as contraceptive recommendations, this lack of transparency can erode clinicians' and stakeholders' trust, leading to unequal outcomes for individuals whose characteristics differ from those in the dominant training subset [43, 44].

To address these ethical gaps, recent research has advocated open-science practices and comprehensive dataset documentation [45, 46]. These approaches support comprehensive reporting on subgroup analysis and fairness evaluation, enabling ML developers and clinicians to identify and correct biases before deployment. However, ethical responsibility extends beyond technical fixes to these issues. This calls for thoughtful design choices, from tuning model settings and choosing the right architecture to understanding the data itself, all with a focus on inclusivity, informed consent, and transparency at every stage, from data collection to model deployment.

Specifically, in SRH contexts, where clinical decisions can influence reproductive autonomy, individuals' privacy, and access to care, the ethical stakes are particularly high. Bias mitigation must be integrated not only into algorithmic design but also into clinical workflows, governance structures, and user engagement strategies. Without these safeguards, ML systems risk amplifying the very inequities they aim to address, compromising both clinical integrity and public trust.

4.4. Clinical implications of bias in ML models in SRH

Clinicians are trained to deliver equitable care across diverse patient populations. For a clinician to accept a model for use, it must earn the trust of unbiased experts across a wide range of demographics and align with the clinician's own clinical reasoning. To foster adoption, a model must provide explainability on how a decision was reached based on data input and demonstrate fairness during validation, extending to subgroups of a population, while showing normal sensitivity or specificity across race, gender and must support an override with justification, which can allow clinicians to refuse an AI recommendation, thereby fostering collaboration rather than authority.

ML Models are trained to learn representation from patterns and correlation, and hence prioritise this attribute over proper clinical judgement [47]. On the other hand, the human aspect of clinical decision-making often involves non-quantifiable factors, such as psychosocial and behavioural factors, that ML models cannot capture. The consequence of this inability of these ML models can amplify existing bias, which can lead to substandard clinical decisions and the worsening of longstanding health care disparities amongst underrepresented groups. An unbiased ML model can significantly assist clinicians by supporting decision-making, reducing error and enabling better patient treatment outcomes. An example is improved diagnostic accuracy, which occurs when a model identifies disease patterns consistently across all patients, regardless of demographics, thereby preventing misdiagnosis or delayed diagnosis in underrepresented groups.

Finally, the continued integration of ML into SRH systems must be guided by human-centred reasoning and a commitment to clinical equity and transparency [46]. Although algorithmic precision offers value for early detection and diagnostic support, its usefulness depends on whether models reflect the diversity of patient experiences and maintain fairness across demographic groups [48]. The consequences of bias are not only technical; they are ethical and systemic, and they directly affect patient outcomes.

While the results show promising improvements in fairness, several limitations must be acknowledged. The datasets are secondary and retrospective, which may restrict generalisability to real clinical settings. The models have also not been validated in operational SRH environments, where user behaviour, data drift and institutional constraints may influence fairness outcomes. Future work will address these issues through prospective validation, stakeholder engagement, and real-time monitoring.

5. Conclusion

This study has demonstrated that using a multi-stage fairness intervention framework can reduce bias in ML-based SRH systems without significantly impacting predictive performance. The study employed a dual evaluation protocol across two distinct datasets, and the findings from the experiments provide adequate insights into the earlier raised research questions:

1) In both datasets, used for the experiments, the intervention showed directional improvements in EOD and DI for some models, especially noticeable in LR. In the SRH England dataset, DI values moved closer to parity for LR, while GB showed partial improvement but remained below the commonly accepted threshold of 0.80. EOD improved only

marginally or worsened for RF. In the PCOS dataset, fairness gains were again mainly noticeable in LR, with RF and GB showing mixed or adverse shifts, including DI values falling below the 0.80 threshold, indicating a potential adverse impact. These findings indicate that while the multi-stage framework can be effective in mitigating group disparities, the outcomes are mainly model-dependent.

2) The predictive performance, as shown in the Accuracy and F1-score results, remained stable or improved for LR across both datasets, and for GB on the PCOS dataset. However, GB showed a slight drop in accuracy on the SRH England dataset, while RF exhibited notable degradation post-intervention, particularly in the PCOS dataset. These findings show that fairness interventions can be integrated without significantly impacting negatively on the predictive ability of the models. However, there are performance trade-offs to consider depending on the model's architecture, configuration, and the nature of the dataset. This further shows the need for model-aware calibration and that fairness–utility balance is not uniformly guaranteed.

Although the proposed framework proved adaptable to the task, several limitations became evident. The availability of sensitive attributes and the reliance on group-level metrics restricted the scope of fairness assessment, particularly in datasets such as PCOS, where demographic information varies across countries and diagnostic practices may differ. Manual calibration of thresholds and dropout parameters also limits scalability in dynamic EHR environments, where fairness constraints must adapt to shifting data distributions. In addition, the study did not explore ensemble or multi-objective optimisation strategies, which may have improved the balance between fairness and performance. The results also revealed a tension between the fairness criteria. EOD often improved, indicating better sensitivity to missed positives, while DI sometimes fell below accepted thresholds, signalling potential adverse impact. This divergence highlights the need for context-specific prioritisation of fairness metrics and careful model-aware calibration, especially in SRH settings where both clinical risk and equitable access matter. Even with these challenges, the framework consistently improved EOD and DP across models, particularly for LR, without compromising predictive performance. Future work will focus on extending the framework to include automated fairness tuning, enhanced statistical reporting, and more comprehensive intersectional analysis, supported by both primary data collection and longitudinal datasets.

Recommendations

Based on the outcomes of this study, we recommend the early integration of fairness-aware interventions into clinical AI pipelines to prevent models from reinforcing existing biases that may restrict treatment access for underrepresented groups. Fairness evaluation should extend beyond technical metrics to include clinical relevance, ethical accountability and human-centred assessments of how fairness interventions influence resource allocation and patient outcomes. In addition, model selection for healthcare applications should prioritise a balance between fairness and predictive reliability, while also improving interpretability.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest in this work.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analysed in this study. The PCOS dataset used in this study is publicly available on Kaggle at https://www.kaggle.com/datasets/ankushpanday1/pcos-prediction-datasettop-75-countries, and the NHS Sexual and Reproductive Health (SRH) Services England 2014–15 dataset is publicly available through NHS Digital at https://digital.nhs.uk/data-and-information/publications/statistical/sexual-and-reproductive-health-services/sexual-and-reproductive-health-services-england-2014-15.

Author Contribution Statement

Efosa Osagie: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing -original draft, Writing - review & editing, Visualization, Supervision, Project administration. **Shemi Ayo-Ogbor:** Investigation, Writing - original draft. **Rebecca Balasundaram:** Investigation, Writing - original draft.

References

[1] Markham, S. (2025). Patient perspective on predictive models in healthcare: Translation into practice, ethical implications and limitations? *BMJ Health & Care Informatics*, *32*(1), e101153. https://doi.org/10.1136/bmjhci-2024-101153

[2] 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. https://doi.org/10.1126/science.aax2342

[3] Hanna, M. G., Pantanowitz, L., Jackson, B., Palmer, O., Visweswaran, S., Pantanowitz, J., ..., & Rashidi, H. H. (2025). Ethical and bias considerations in artificial intelligence/machine learning. *Modern Pathology*, *38*(3), 100686. https://doi.org/10.1016/j.modpat.2024.100686

[4] Chen, I. Y., Szolovits, P., & Ghassemi, M. (2019). Can AI help reduce disparities in general medical and mental health care? *AMA Journal of Ethics, 21*(2), E167-179. https://doi.org/10.1001/amajethics.2019.167

[5] Pierson, E., Cutler, D. M., Leskovec, J., Mullainathan, S., & Obermeyer, Z. (2021). An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nature Medicine*, *27*(1), 136–140. https://doi.org/10.1038/s41591-020-01192-7

[6] Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems*, *31*(3), 388–409. https://doi.org/10.1080/0960085x.2021.1927212

[7] Oyeniyi, J. G. (2025). From Lab to Clinic: Addressing Bias and Generalizability in AI Diagnostic Systems. World Journal of Advanced Research and Reviews, 28(3), 2134–2179. https://doi.org/10.30574/wjarr.2025.28.3.4249

[8] Hasanzadeh, F., Josephson, C. B., Waters, G., Adedinsewo, D., Azizi, Z., & White, J. A. (2025). Bias recognition and mitigation strategies in artificial intelligence healthcare applications. *npj Digital Medicine*, *8*(1), 154. https://doi.org/10.1038/s41746-025-01503-7

[9] Joseph, J. (2025). Algorithmic bias in public health AI: a silent threat to equity in low-resource settings. Frontiers in Public Health, 13. https://doi.org/10.3389/fpubh.2025.1643180

[10] Gao, J., Yao, Y., Xue, J., Chen, R., Yang, X., Xu, J., & Cheng, W. (2025). Methodological conduct and risk of bias in studies on prenatal birthweight prediction models using machine learning techniques: A systematic review. *BMC Pregnancy and Childbirth*, *25*(1), 696. https://doi.org/10.1186/s12884-025-07727-5

[11] World Health Organisation. (2024). *Unpacking artificial intelligence in sexual and reproductive health and rights: Risks and opportunities* (WHO Technical Brief). https://www.who.int/news/item/22-03-2024-unpacking-artifical-intelligence-in-sexual-and-reproductive-health-and-rights

[12] Aquino, Y. S. J., Carter, S. M., Houssami, N., Braunack-Mayer, A., Win, K. T., Degeling, C., ..., & Rogers, W. A. (2025). Practical, epistemic and normative implications of algorithmic bias in healthcare artificial intelligence: A qualitative study of multidisciplinary expert perspectives. *Journal of Medical Ethics*, *51*(6), 420-428. https://doi.org/10.1136/jme-2022-108850

[13] Mackin, S., Major, V. J., Chunara, R., & Newton-Dame, R. (2025). Post-processing methods for mitigating algorithmic bias in healthcare classification models: An extended umbrella review. *BMC Digital Health*, *3*(1), 26. https://doi.org/10.1186/s44247-025-00166-4

[14] Taneja, J., Ghosh, J., Kant, R., & Christodoulides, M. (2025). Artificial intelligence in predicting, diagnosing and preventing sexually transmitted infections (STIs). *Venereology*, *4*(2), 5. https://doi.org/10.3390/venereology4020005

[15] Melaku, M. S., Yohannes, L., Sharew, B., Derseh, M. H., & Taye, E. A. (2025). Application of machine learning algorithms to model predictors of informed contraceptive choice among reproductive age women in six high fertility rate sub Sahara Africa countries. *BMC Public Health*, *25*(1), 1986. https://doi.org/10.1186/s12889-025-23242-w

[16] Hoche, M., Mineeva, O., Rätsch, G., Vayena, E., & Blasimme, A. (2025). What makes clinical machine learning fair? A practical ethics framework. *PLOS Digital Health*, *4*(3), e0000728. https://doi.org/10.1371/journal.pdig.0000728

[17] Rabonato, R. T., & Berton, L. (2025). A systematic review of fairness in machine learning. *AI and Ethics*, *5*(3), 1943-1954. https://doi.org/10.1007/s43681-024-00577-5

[18] Liu, M., Ning, Y., Ke, Y., Shang, Y., Chakraborty, B., Ong, M. E. H., ..., & Liu, N. (2024). FAIM: Fairness-aware interpretable modeling for trustworthy machine learning in healthcare. *Patterns*, *5*(10), 101059. https://doi.org/10.1016/j.patter.2024.101059

[19] Straw, I., & Wu, H. (2022). Investigating for bias in healthcare algorithms: A sex-stratified analysis of supervised machine learning models in liver disease prediction. *BMJ Health & Care Informatics*, *29*(1), e100457. https://doi.org/10.1136/bmjhci-2021-100457

[20] Norori, N., Hu, Q., Aellen, F. M., Faraci, F. D., & Tzovara, A. (2021). Addressing bias in big data and AI for health care: A call for open science. *Patterns*, *2*(10), 100347. https://doi.org/10.1016/j.patter.2021.100347

[21] Li, F., Wu, P., Ong, H. H., Peterson, J. F., Wei, W.-Q., & Zhao, J. (2023). Evaluating and mitigating bias in machine learning models for cardiovascular disease prediction. *Journal of Biomedical Informatics*, *138*, 104294. https://doi.org/10.1016/j.jbi.2023.104294

[22] Xu, J., Xiao, Y., Wang, W. H., Ning, Y., Shenkman, E. A., Bian, J., & Wang, F. (2022). Algorithmic fairness in computational medicine. *eBioMedicine*, *84*, 104250. https://doi.org/10.1016/j.ebiom.2022.104250

[23] Aigner, D. J., del Ángel, M., & Wiles, J. (2024). Statistical approaches for assessing disparate impact in fair housing cases. *Statistics and Public Policy*, *11*(1), 2263038. https://doi.org/10.1080/2330443X.2023.2263038

[24] Chen, R. J., Wang, J. J., Williamson, D. F. K., Chen, T. Y., Lipkova, J., Lu, M. Y., ..., & Mahmood, F. (2023). Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature Biomedical Engineering*, *7*(6), 719–742. https://doi.org/10.1038/s41551-023-01056-8

[25] Spooner, A., Moridani, M. K., Toplis, B., Behary, J., Safarchi, A., Maher, S., ..., & Sowmya, A. (2025). Benchmarking ensemble machine learning algorithms for multi-class, multi-omics data integration in clinical outcome prediction. *Briefings in Bioinformatics*, *26*(2), bbaf116. https://doi.org/10.1093/bib/bbaf116

[26] Kaliappan, J., Srinivasan, K., Mian Qaisar, S., Sundararajan, K., Chang, C.-Y., & C, S. (2021). Performance evaluation of regression models for the prediction of the COVID-19 reproduction rate. *Frontiers in Public Health*, *9*, 729795. https://doi.org/10.3389/fpubh.2021.729795

[27] Mamo, D. N., Gebremariam, Y. H., Adem, J. B., Kebede, S. D., & Walle, A. D. (2024). Machine learning to predict unintended pregnancy among reproductive-age women in Ethiopia: Evidence from

EDHS 2016. *BMC Women's Health*, *24*(1), 57. https://doi.org/10.1186/s12905-024-02893-8

[28] Saxena, A., Sharma, S., Kumar Johari, P., Pandey, A., & Kumar, S. (2025). A fair and interpretable deep learning approach for healthcare access prediction in underserved communities. *Discover Artificial Intelligence*, *5*(1), 185. https://doi.org/10.1007/s44163-025-00425-3

[29] Huang, Y., Guo, J., Chen, W.-H., Lin, H.-Y., Tang, H., Wang, F., ..., & Bian, J. (2024). A scoping review of fair machine learning techniques when using real-world data. *Journal of Biomedical Informatics*, *151*, 104622. https://doi.org/10.1016/j.jbi.2024.104622

[30] Albaroudi, E., Mansouri, T., & Alameer, A. (2024). A Comprehensive Review of AI Techniques for Addressing Algorithmic Bias in Job Hiring. AI, 5(1), 383–404. https://doi.org/10.3390/ai5010019

[31] Macias-Konstantopoulos, W. L., Collins, K. A., Diaz, R., Duber, H. C., Edwards, C. D., Hsu, A. P., ..., & Sachs, C. J. (2023). Race, healthcare, and health disparities: A critical review and recommendations for advancing health equity. *Western Journal of Emergency Medicine*, *24*(5), 906-918. https://doi.org/10.5811/WESTJEM.58408

[32] Jaime, S., & Kern, C. (2024). Ethnic classifications in algorithmic fairness: Concepts, measures and implications in practice. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 237-253. https://doi.org/10.1145/3630106.3658902

[33] Chin, M. H., Afsar-Manesh, N., Bierman, A. S., Chang, C., Colón-Rodríguez, C. J., Dullabh, P., ..., & Ohno-Machado, L. (2023). Guiding principles to address the impact of algorithm bias on racial and ethnic disparities in health and health care. *JAMA Network Open*, *6*(12), e2345050. https://doi.org/10.1001/jamanetworkopen.2023.45050

[34] Abujaber, A. A., & Nashwan, A. J. (2024). Nursing privilege: A concept analysis. *Nursing Open*, *11*(3), e2120. https://doi.org/10.1002/nop2.2120

[35] Jindal, M., Chaiyachati, K. H., Fung, V., Manson, S. M., & Mortensen, K. (2023). Eliminating health care inequities through strengthening access to care. *Health Services Research*, *58*(S3), 300–310. https://doi.org/10.1111/1475-6773.14202

[36] Khakurel, U., Abdelmoumin, G., & Rawat, D. B. (2025). Performance evaluation for detecting and alleviating biases in predictive machine learning models. *ACM Transactions on Probabilistic Machine Learning*, *1*(2), 12. https://doi.org/10.1145/3729432

[37] Chakraborty, J., Majumder, S., & Menzies, T. (2021). Bias in machine learning software: Why? How? What to do? In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 429-440. https://doi.org/10.1145/3468264.3468537

[38] Wang, Y., & Singh, L. (2025). Impact on bias mitigation algorithms to variations in inferred sensitive attribute uncertainty. *Frontiers in Artificial Intelligence*, *8*, 1520330. https://doi.org/10.3389/frai.2025.1520330

[39] Kim, D., Woo, H., & Lee, Y. (2024). Addressing bias and fairness using fair federated learning: A synthetic review. *Electronics*, *13*(23), 4664. https://doi.org/10.3390/electronics13234664

[40] Kpatcha, E. (2025). Balancing fairness and accuracy in machine learning-based probability of default modeling via threshold optimization. *Journal of Risk and Financial Management*, *18*(12), 724. https://doi.org/10.3390/jrfm18120724

[41] Murikah, W., Nthenge, J. K., & Musyoka, F. M. (2024). Bias and ethics of AI systems applied in auditing - A systematic review. *Scientific African*, *25*, e02281. https://doi.org/10.1016/j.sciaf.2024.e02281

[42] Alderman, J. E., Palmer, J., Laws, E., McCradden, M. D., Ordish, J., Ghassemi, M., ..., Liu, X. (2025). Tackling algorithmic bias and promoting transparency in health datasets: The STANDING Together consensus recommendations. *The Lancet Digital Health*, *7*(1), e64-e88. https://doi.org/10.1016/S2589-7500(24)00224-3

[43] Punzi, M. C., & Thuis, T. (2025). Mapping ethical concerns in algorithm-driven period and fertility tracking technologies. *Contraception*. Advance online publication. https://doi.org/10.1016/j.contraception.2025.110837

[44] Döring, N., Le, T. D., Vowels, L. M., Vowels, M. J., & Marcantonio, T. L. (2024). The impact of artificial intelligence on human sexuality: A five-year literature review 2020–2024. *Current Sexual Health Reports*, *17*(1), 4. https://doi.org/10.1007/s11930-024-00397-y

[45] Isaksson, A. (2025). Mitigation measures for addressing gender bias in artificial intelligence within healthcare settings: A critical area of sociological inquiry. *AI & Society*, *40*(4), 3009-3018. https://doi.org/10.1007/s00146-024-02067-y

[46] Rojas, J. C., Fahrenbach, J., Makhni, S., Cook, S. C., Williams, J. S., Umscheid, C. A., & Chin, M. H. (2022). Framework for integrating equity into machine learning models. *Chest*, *161*(6), 1621-1627. https://doi.org/10.1016/j.chest.2022.02.001

[47] Cowley, H. P., Robinette, M. S., Matelsky, J. K., Xenes, D., Kashyap, A., Ibrahim, N. F., ..., & Gray-Roncal, W. (2023). Using machine learning on clinical data to identify unexpected patterns in groups of COVID-19 patients. *Scientific Reports*, *13*(1), 2236. https://doi.org/10.1038/s41598-022-26294-9

[48] Cross, J. L., Choma, M. A., & Onofrey, J. A. (2024). Bias in medical AI: Implications for clinical decision-making. *PLOS Digital Health*, *3*(11), e0000651. https://doi.org/10.1371/journal.pdig.0000651