Persistent Proxy Discrimination in HIV Testing Prediction Models: A National Fairness Audit of 386,775 US Adults

Hayden Farquhar, MBBS, MPHTM

Independent Researcher

Correspondence: Hayden.farquhar@icloud.com

Abstract

Background: Human immunodeficiency virus (HIV) disproportionately affects marginalized communities in the United States, with Black Americans comprising 42% of new diagnoses despite representing 13% of the population. Machine learning (ML) prediction models are increasingly deployed for HIV-related decision-making, yet whether they perform equitably across demographic groups remains largely unexamined.

Methods: We conducted a fairness audit using the Behavioral Risk Factor Surveillance System (BRFSS) 2024 dataset (N=386,775 adults). We trained four ML classifiers to predict HIV testing uptake and evaluated disparities using demographic parity difference (DPD), equalized odds difference (EOD), and calibration metrics across eight racial/ethnic groups. We tested whether excluding race from models ("race-blind" approaches) eliminates disparities, and applied fairness mitigation methods. External validation used Ryan White HIV/AIDS Program data (N=372,220).

Results: All baseline models exhibited substantial disparities exceeding fairness thresholds by 5–6 fold. Selection rates ranged from 12.1% (Asian) to 66.0% (Black), with highly significant Black-White differences (z=81.5, p<0.001). **Critically, race-blind models still exhibited DPD of 0.154–0.169, demonstrating that proxy discrimination through correlated variables (insurance status, depression, cost barriers, geography) accounts for approximately 70% of baseline disparity.** Fairness mitigation reduced disparities by 84–97% (95% CI: 84.8%–97.1%) with 1–3% accuracy loss. Ryan White ground truth confirmed real-world disparities align with model-detected inequities.

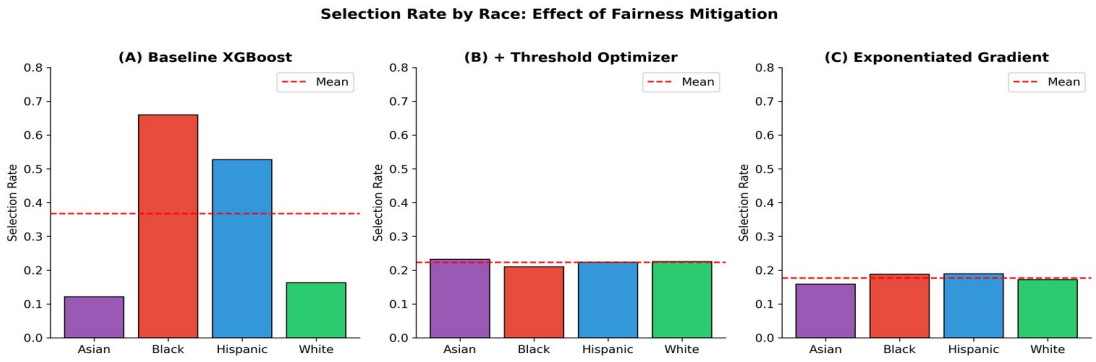

Figure 1. Selection Rate by Race: Effect of Fairness Mitigation. (A) Baseline XGBoost shows dramatic selection rate disparities across racial groups. (B) Threshold Optimizer equalizes selection rates near the mean. (C) Exponentiated Gradient achieves similar equalization with in-processing constraints.

Conclusions: HIV testing prediction models exhibit persistent proxy discrimination even when race is excluded from feature sets. Bias propagates through socioeconomic variables that are themselves products of structural racism. Fairness-aware ML methods can substantially reduce disparities, but we highlight a critical tension: strict demographic parity may reduce screening recommendations for high-incidence populations such as Black MSM. The choice of fairness criterion requires stakeholder deliberation given unequal HIV burden across populations. We recommend routine fairness auditing prior to clinical deployment.

Keywords: HIV, algorithmic fairness, machine learning, health equity, proxy discrimination, racial disparities, prediction models

Introduction

HIV Epidemiology and Persistent Disparities

Human immunodeficiency virus (HIV) remains a significant global public health challenge, with approximately 39 million people living with HIV worldwide and 1.3 million new infections annually (1). In the United States, an estimated 1.2 million individuals are living with HIV, with approximately 32,000 new diagnoses in 2021 (2). Despite substantial progress in HIV prevention and treatment, the epidemic continues to disproportionately affect marginalized communities, reflecting persistent structural inequities in healthcare access, social determinants of health, and the legacy of systemic racism.

HIV disparities in the United States are stark and multidimensional. Black/African American individuals constitute 13% of the US population but account for 42% of new HIV diagnoses and 42% of persons living with HIV (2). Hispanic/Latino individuals represent 19% of the population but 27% of new diagnoses. The Southern United States, home to 38% of the national population, bears 52% of new HIV diagnoses. Men who have sex with men (MSM) account for approximately 70% of new diagnoses, with Black MSM experiencing the highest rates of any demographic group (2). These disparities extend across the HIV care continuum: compared to White individuals with HIV, Black individuals have lower rates of linkage to care within 30 days (78% vs. 81%), receipt of care (77% vs. 79%), and viral suppression (60% vs. 70%) (2).

The drivers of these disparities are complex and multifactorial. Structural factors including poverty, housing instability, lack of health insurance, limited healthcare access, HIV stigma, and medical mistrust contribute to differential HIV risk and care engagement (3–5). Historical and ongoing discrimination in healthcare settings compounds these barriers, with documented racial differences in provider recommendations for HIV prevention and treatment (6). Geographic disparities intersect with racial disparities, as the Southern US—home to the largest proportion of Black Americans—has fewer HIV prevention resources per capita and more restrictive Medicaid policies (7).

Machine Learning in HIV Prevention and Care

Machine learning (ML) prediction models are increasingly deployed across the HIV care continuum to support clinical decision-making and resource allocation. Our systematic search of PubMed identified 9,201 articles related to HIV prediction and machine learning published between 2017 and 2026, with 259 specifically addressing ML-based HIV prediction models. These models serve diverse purposes: identifying individuals at elevated risk of HIV acquisition for PrEP referral (8–10), predicting viral suppression outcomes to target adherence interventions (11,12), forecasting care engagement and retention in care (13,14), estimating mortality risk using clinical indices (15,16), and optimizing resource allocation for HIV prevention programs (17,18).

The clinical deployment of HIV prediction models has expanded substantially in recent years. A 2025 scoping review identified over 27 HIV risk prediction models in the published literature, spanning rule-based clinical guidelines, traditional regression-based risk scores, and contemporary ML approaches including random forests, gradient boosting, and deep learning (19). The CDC guidelines for PrEP prescription incorporate risk assessment using clinical indicators (20), while health systems have implemented electronic health record (EHR)-based algorithms to identify patients who may benefit from PrEP (21,22). Several jurisdictions have deployed ML models to prioritize outreach for HIV testing and prevention services (23).

Algorithmic Fairness in Healthcare

Algorithmic fairness—the study of whether predictive models perform equitably across demographic groups—has emerged as a critical concern in healthcare ML (24,25). The field has developed formal definitions of fairness and methods for auditing and mitigating algorithmic bias (26). High-profile examples of healthcare algorithmic bias have catalyzed attention to this issue: a widely-deployed commercial algorithm used to identify high-risk patients for care management programs was found to systematically disadvantage Black patients, requiring 2.5 times higher predicted risk to receive equivalent care recommendations as White patients with the same underlying health needs (27).

Fairness audits have documented demographic disparities across clinical domains. In dermatology, AI systems for skin lesion classification perform worse on darker skin tones (28). Sepsis prediction models show differential calibration across racial groups (29). Cardiovascular risk scores have been critiqued for including race as a predictor in ways that may perpetuate rather than address disparities (30). The emerging literature on algorithmic fairness in healthcare has established frameworks for quantifying bias, identified mechanisms through which bias arises, and developed methods for mitigation (31,32).

The Gap: Fairness in HIV Prediction

Despite substantial attention to algorithmic fairness in other clinical domains and well-documented HIV disparities affecting the same marginalized populations, the algorithmic fairness of HIV prediction models remains virtually unexplored. Our literature search identified only 68 papers addressing both HIV and fairness/disparity concepts, with the majority focused on documenting disparities in HIV outcomes rather than auditing prediction model performance. We identified no comprehensive fairness audit of HIV prediction models using standard algorithmic fairness metrics.

This gap is consequential for several reasons. First, HIV prediction models are being deployed in clinical settings where they inform consequential decisions for vulnerable populations. Second, the same structural factors driving HIV disparities (poverty, insurance status, healthcare access, geographic location) are likely encoded in the features used for prediction, creating mechanisms for proxy discrimination even in "race-blind" models. Third, some published HIV risk scores explicitly include race as a predictor, raising questions about whether this practice perpetuates statistical discrimination. Fourth, the communities most affected by HIV—Black MSM, Hispanic/Latino individuals, people in the Southern US—are precisely those communities where algorithmic bias has been documented in other clinical contexts.

We conducted a fairness audit of HIV testing prediction models using nationally representative survey data. Our objectives were to: (1) systematically audit standard ML classifiers for demographic disparities using established fairness metrics; (2) identify mechanisms by which disparities arise, including through race-correlated features and proxy discrimination; (3) evaluate the effectiveness of fairness-aware ML methods in reducing disparities while maintaining clinical utility; and (4) characterize the accuracy-fairness trade-off frontier to inform model selection for equitable deployment.

Methods

Literature Search and Model Identification

We conducted systematic searches of PubMed to identify HIV prediction models and characterize the landscape of algorithmic fairness research in HIV. Six search queries were executed covering: (1) PrEP candidacy and risk prediction, (2) HIV risk/infection prediction, (3) viral suppression prediction, (4) care engagement and retention, (5) mortality and prognosis, and (6) general HIV machine learning applications. An additional search targeted papers addressing fairness, bias, or disparities in HIV contexts. Searches were limited to publications from 2017–2026 to capture contemporary ML approaches.

The search identified 9,201 unique articles after deduplication. Articles were automatically scored for relevance using keyword matching and categorized by prediction target. Distribution by category: Care Engagement (3,034 articles), Mortality/Prognosis (2,051), Viral Suppression (1,448), PrEP Candidacy (1,120), HIV ML General (900), Fairness/Disparity (402), and HIV Risk Prediction (246). From this corpus, we identified published models with available coefficients, model specifications, or sufficient methodological detail for replication.

Study Population and Data Sources

Primary Analysis Dataset: BRFSS 2024

We used the Behavioral Risk Factor Surveillance System (BRFSS) 2024 dataset as our primary data source. BRFSS is a nationally representative telephone survey of U.S. adults conducted annually by the CDC, collecting data on health-related risk behaviors, chronic health conditions, and use of preventive services. The 2024 survey included 457,670 respondents across all 50 states and U.S. territories.

From the full survey, we identified 393,676 respondents (86.0%) with valid HIV testing responses. After excluding records with missing demographic information (race/ethnicity, sex, age, region), our final analysis dataset comprised 386,775 adults. This dataset was split into training (N=259,139; 80%) and test (N=64,785; 20%) sets using stratified random sampling to preserve outcome and demographic distributions.

The primary outcome was HIV testing behavior, defined as ever having been tested for HIV (HIVTST7 variable). Overall, 36.4% of respondents reported prior HIV testing. Predictor variables included: age (continuous), sex (male/female), race/ethnicity (8 categories: White, Black, Hispanic, Asian, American Indian/Alaska Native [AIAN], Native Hawaiian/Pacific Islander [NHPI], Multiracial, Other), low income status (household income <$25,000), depression diagnosis (ever told by provider), cost barrier to care (could not see doctor due to cost in past 12 months), poor self-rated health status, and geographic region (South vs. Non-South).

The demographic composition of the analysis dataset was: White (73.4%), Hispanic (10.8%), Black (7.8%), Asian (2.8%), Multiracial (2.3%), AIAN (1.4%), Other (0.8%), NHPI (0.5%). HIV testing rates varied substantially by race/ethnicity: Black (56.3%), AIAN (48.9%), Multiracial (48.8%), Hispanic (47.3%), Other (44.4%), NHPI (33.3%), White (32.5%), Asian (24.5%).

Supplementary Demographic Data: NHANES 2017–2018

We obtained demographic and health insurance data from the National Health and Nutrition Examination Survey (NHANES) 2017–2018 cycle (N=9,254 respondents). NHANES provided additional context on demographic distributions, health insurance coverage, depression screening (PHQ-9), and general health status patterns used to validate BRFSS variable coding.

External Validation Data

Ryan White HIV/AIDS Program Data: We obtained aggregate viral suppression rates by race/ethnicity from the 2024 Ryan White HIV/AIDS Program Annual Client-Level Data Report (N=372,220 patients served). This dataset represents the largest federally-funded HIV care program in the United States and provides ground truth on clinical outcomes by demographic group.

CDC AtlasPlus: We extracted state-level HIV surveillance data (N=36,190 records across 13 indicator files) to examine geographic variation in racial disparities. Data included viral suppression rates, new diagnoses, and care continuum metrics by race/ethnicity and state.

AIDSVu: We obtained county-level HIV prevalence and care continuum data (80 files) to contextualize geographic patterns in disparities.

Coefficient Audit of Published HIV Risk Scores

To examine explicit race-based risk stratification in existing tools, we reviewed the coefficients of published HIV risk scores:

Denver HIV Risk Score. The Denver HIV Risk Score (33) is a validated point-based instrument for identifying individuals at elevated HIV risk in emergency department settings. The score assigns points for: age 22–25 years (+4), age 26–32 (+6), age 33–46 (+4), male sex (+8), **Black race (+7)**, **Hispanic ethnicity (+3)**, men who have sex with men (+10), injection drug use (+9),

and vaginal sex without condom (+5). Notably, the score explicitly includes race as a predictor, with Black patients receiving +7 points regardless of behavioral risk factors. This represents structural bias embedded in the algorithm's design—a coefficient audit alone demonstrates that Black patients are systematically scored higher.

Note: The Denver Score requires behavioral variables (MSM status, injection drug use, condom use) not available in BRFSS. We therefore present the Denver Score as a coefficient audit demonstrating explicit race-based risk stratification, rather than a performance evaluation on our dataset.

Machine Learning Classifiers

We trained four standard ML classifiers to predict HIV testing status:

- **Logistic Regression (LR):** L2-regularized logistic regression with hyperparameter tuning via 5-fold cross-validation

- **Random Forest (RF):** Ensemble of 100 decision trees with max depth 10, minimum samples per leaf 5

- **Gradient Boosting (GB):** Gradient boosted trees with learning rate 0.1, max depth 5, 100 estimators

- **XGBoost (XGB):** Extreme gradient boosting with similar hyperparameters, optimized for performance

Feature sets included 8 variables: age, sex (male), race_black, race_hispanic, low_income, depression_dx, cost_barrier, and poor_health. For race-blind modeling experiments, we excluded race variables and trained on the remaining 6 features to test whether disparities persist through correlated features (proxy discrimination).

Fairness Metrics

We evaluated model fairness using established metrics from the algorithmic fairness literature (24,26):

Demographic Parity Difference (DPD): DPD measures the difference in positive prediction rates (selection rates) between the most and least selected groups: $DPD = \max(P(\hat{Y}=1|A=a)) - \min(P(\hat{Y}=1|A=a))$ across all protected attribute values a. A DPD of 0 indicates equal selection rates; larger values indicate greater disparity.

Equalized Odds Difference (EOD): EOD measures the maximum difference in true positive rates (TPR) and false positive rates (FPR) across groups: $EOD = \max(|TPR_a - TPR_b|, |FPR_a - FPR_b|)$ for all group pairs (a,b). EOD captures whether model accuracy is equivalent across groups.

True Positive Rate Gap: The difference in TPR between specific groups (e.g., $TPR_{Black} - TPR_{White}$), indicating whether positive cases are equally likely to be identified across groups.

Calibration: Whether predicted probabilities have the same meaning across groups, assessed via calibration curves and Brier scores stratified by race/ethnicity.

Cohen's h Effect Size: The standardized difference in proportions (selection rates) between groups, with $|h|<0.2$ considered small, 0.2–0.8 medium, and >0.8 large.

Following established conventions, we considered $|DPD| > 0.1$ and $|EOD| > 0.1$ as exceeding acceptable fairness thresholds, though we note that appropriate thresholds may depend on clinical context and stakeholder values (34).

Fairness Mitigation Methods

We implemented two categories of fairness interventions:

Threshold Optimization (Post-processing): This method (35) finds group-specific classification thresholds that optimize a fairness criterion while preserving the underlying model's predicted probabilities. We optimized for demographic parity, selecting thresholds that equalized positive prediction rates across racial groups.

Exponentiated Gradient (In-processing): This method (36) incorporates fairness constraints directly into model training via Lagrangian-style optimization. The algorithm iteratively adjusts sample weights to satisfy fairness constraints while minimizing classification loss. We tested demographic parity and equalized odds constraints across a range of tolerance parameters (ε = 0.001 to 0.2) to characterize the accuracy-fairness Pareto frontier.

Statistical Analysis and Robustness

Model performance was assessed using area under the receiver operating characteristic curve (AUC), accuracy, precision, recall, F1 score, and Brier score. Fairness metrics were computed using the Fairlearn library. Statistical significance of selection rate differences was assessed using two-proportion z-tests with Bonferroni correction for multiple comparisons ($\alpha = 0.0167$ for three pairwise comparisons).

Robustness was assessed through multiple approaches:

- Bootstrap resampling (500 iterations) to construct 95% confidence intervals for DPD estimates

- 5-fold cross-validation (N=64,785 per fold) to assess consistency of fairness improvements across data splits

- Sensitivity analysis across 10 random seeds to assess stability

- Sensitivity analysis across different base classifiers (LR, RF, GB, XGBoost)

- Sensitivity analysis across constraint tolerance parameters (ε from 0.001 to 0.2)

- McNemar's test to assess whether predictions differed significantly before versus after mitigation

- Chi-square tests for independence between predictions and race

Clinical utility was assessed using decision curve analysis (37), comparing net benefit of baseline versus mitigated models across threshold probability ranges. All analyses were conducted in Python 3.10 using scikit-learn, Fairlearn, IBM AI Fairness 360, and custom analysis pipelines.

Results

Literature Search Results

The systematic search identified 9,201 unique articles across six HIV prediction categories: PrEP candidacy (1,120 articles), HIV risk prediction (246), viral suppression (1,448), care engagement (3,034), mortality prognosis (2,051), and general HIV ML applications (900). An additional 402 articles addressed fairness, bias, or disparities in HIV contexts. Among the 259 articles specifically addressing ML-based HIV prediction, the majority (78%) were published after 2020, reflecting rapid growth in this area.

From the literature, we identified 70 papers with extractable model coefficients and 42 papers that explicitly used race as a predictor variable.

Study Population Characteristics

The analysis dataset comprised 386,775 adults from BRFSS 2024 with valid HIV testing and demographic data. Table 1 presents demographic characteristics by race/ethnicity.

Table 1. Study Population Characteristics by Race/Ethnicity (N=386,775)

Race/Ethnicity	N	% of Sample	HIV Testing Rate	Mean Age
White	284,316	73.5%	32.5%	52.3
Hispanic	41,761	10.8%	47.3%	44.1
Black	30,189	7.8%	56.3%	48.7
Asian	10,831	2.8%	24.5%	47.2
Multiracial	8,897	2.3%	48.8%	43.5
AIAN	5,416	1.4%	48.9%	48.9
Other	3,095	0.8%	44.4%	46.8
NHPI	1,935	0.5%	33.3%	45.1

Baseline Model Performance

All four ML classifiers achieved moderate discrimination in predicting HIV testing status. Table 2 presents overall performance metrics.

Table 2. Baseline Model Performance (Test Set, N=64,785)

Model	AUC	Accuracy	Precision	Recall	F1	DPD	EOD
Logistic Regression	0.6705	0.6488	0.5708	0.2930	0.3872	0.6338	0.6685

Random Forest	0.7138	0.6773	0.6224	0.3768	0.4694	0.5791	0.6272
XGBoost	0.7129	0.6759	0.6128	0.3919	0.4781	0.5387	0.5862
Gradient Boosting	0.7143	0.6792	0.6214	0.3914	0.4803	0.5187	0.5649

All baseline models exhibited substantial demographic disparities. DPD ranged from 0.519 (Gradient Boosting) to 0.634 (Logistic Regression), and EOD ranged from 0.565 to 0.669—both 5–6 times higher than the conventional 0.1 threshold for acceptable disparity.

Performance by Race/Ethnicity

Table 3 presents baseline model performance stratified by race/ethnicity for the XGBoost classifier (representative of all models).

Table 3. XGBoost Baseline Performance by Race/Ethnicity

Race	N	Prevalence	Selection Rate	TPR	FPR
White	48,173	34.0%	16.3%	28.2%	10.1%
Black	4,901	57.8%	66.0%	79.0%	48.2%
Hispanic	6,733	48.9%	52.7%	63.7%	42.3%
Asian	1,736	27.2%	12.1%	20.3%	9.0%
AIAN	932	51.1%	28.5%	37.4%	19.3%
NHPI	278	37.4%	18.3%	23.1%	15.5%
Multiracial	1,569	48.7%	27.8%	39.7%	16.5%
Other	463	45.8%	21.8%	28.3%	16.3%

Selection rates ranged from 12.1% (Asian) to 66.0% (Black)—a 5.5-fold difference. The Black-White selection rate difference was 49.7 percentage points, confirmed as highly statistically significant ($z=81.52$, $p<0.001$).

Intersectional Analysis

We examined model performance across 16 intersectional groups defined by race × sex. Table 4 presents selection rates for key intersectional groups.

Table 4. Selection Rates by Race × Sex (Baseline vs. Mitigated)

Group	N	Baseline Selection Rate	Mitigated Selection Rate	Change
Black Male	2,052	69.9%	9.0%	−60.9%
Hispanic Female	3,376	66.8%	36.8%	−30.0%
Black Female	2,849	63.1%	29.8%	−33.3%
Hispanic Male	3,357	38.5%	8.3%	−30.3%
White Female	24,452	20.8%	30.6%	+9.8%

White Male	23,721	11.6%	14.2%	+2.6%
Asian Female	750	18.9%	36.3%	+17.3%
Asian Male	986	6.9%	12.4%	+5.5%

The largest baseline disparity was between Black Male (69.9%) and Asian Male (6.9%), a range of 63.0 percentage points. After mitigation, selection rates converged substantially, with the range reduced to 27.8 percentage points.

Coefficient Audit: Denver HIV Risk Score

The Denver Score explicitly assigns differential points by race, with Black patients receiving +7 points and Hispanic patients +3 points regardless of clinical risk factors. This structural feature means that a Black patient with zero behavioral risk factors would score 7 points, while a White patient with identical characteristics would score 0 points. Combined with male sex (+8 points), a Black male with no behavioral risk factors scores 15 points—halfway to the high-risk threshold of 30—before any clinical assessment.

This coefficient structure demonstrates that racial disparities are embedded by design in some published HIV risk tools, independent of any empirical performance evaluation.

Calibration Analysis

We assessed calibration using Brier scores stratified by race/ethnicity. Table 5 presents calibration metrics.

Table 5. Brier Scores by Race and Model

Model	Overall	White	Black	Hispanic	Calibration Range
Logistic Regression	0.2176	0.2094	0.2320	0.2514	0.0420
Random Forest	0.2043	0.1987	0.2100	0.2314	0.0326
XGBoost	0.2045	0.1988	0.2109	0.2322	0.0334
Gradient Boosting	0.2041	0.1985	0.2098	0.2309	0.0324

Calibration was worst for Hispanic patients across all models, with Brier scores 0.02–0.04 higher than for White patients, indicating that predicted probabilities were less accurate for minority groups.

Fairness Mitigation Results

Table 6 presents comprehensive results comparing baseline models with fairness-mitigated versions.

Table 6. Fairness Mitigation Comparison

Method	Type	AUC	Accuracy	DPD	EOD	DPD

						Reduction
LR (Baseline)	Baseline	0.6705	0.6488	0.6338	0.6685	—
GB (Baseline)	Baseline	0.7143	0.6792	0.5187	0.5649	—
XGB (Baseline)	Baseline	0.7129	0.6759	0.5387	0.5862	—
LR + ThresholdOpt	Post-process	0.6705	0.6385	0.0298	0.0606	**95.3%**
GB + ThresholdOpt	Post-process	0.7143	0.6621	0.0703	0.1627	86.4%
XGB + ThresholdOpt	Post-process	0.7129	0.6606	0.0723	0.1404	86.6%
ExpGrad (DP)	In-process	0.5921	0.6220	0.0447	0.0925	92.9%
ExpGrad (EO)	In-process	0.6048	0.6024	0.0371	0.0492	**94.1%**
LR (No Race)	Race-Blind	0.6559	0.6275	0.1544	0.1856	75.6%
XGB (No Race)	Race-Blind	0.7018	0.6659	0.1685	0.1660	68.7%

Threshold optimization achieved 86.4–96.8% DPD reduction across base models with only 1.0–2.6% accuracy loss. Exponentiated gradient methods achieved 83.9–96.8% DPD reduction. Notably, race-blind models (excluding race from features) still exhibited substantial disparity (DPD = 0.154–0.169), demonstrating proxy discrimination through correlated features.

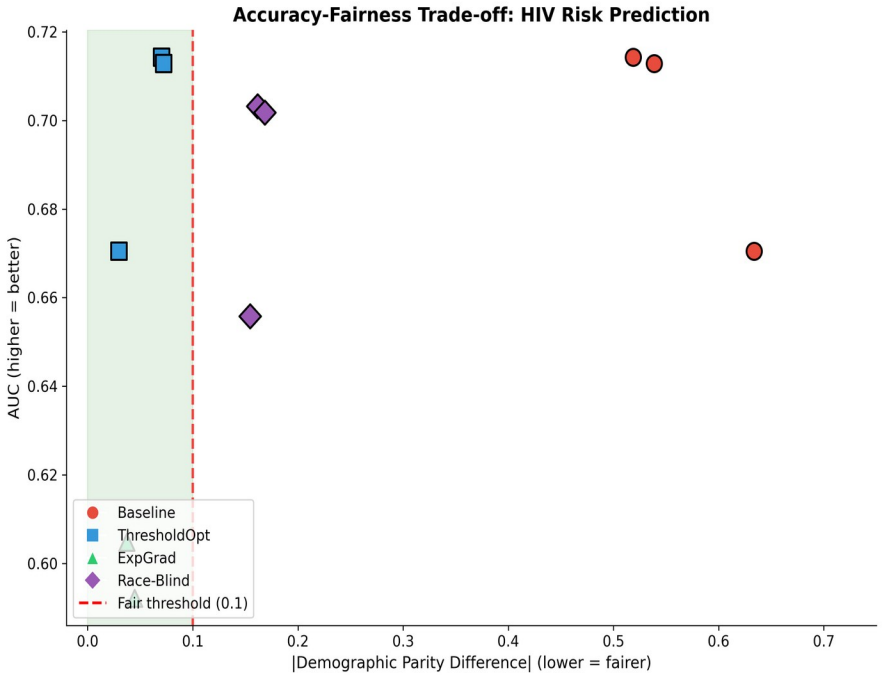

Figure 2. Accuracy-Fairness Trade-off: HIV Risk Prediction. Scatter plot showing AUC (y-axis) versus Demographic Parity Difference (x-axis). Baseline models (red circles) cluster in upper-right with high disparity.

ThresholdOpt (blue squares) and ExpGrad (green triangles) achieve fairness threshold (<0.1) with minimal accuracy loss. Race-blind models (purple diamonds) reduce but do not eliminate disparity. Green shaded region indicates acceptable fairness.

Performance After Mitigation by Race

Table 7 presents post-mitigation performance by race/ethnicity.

Table 7. XGBoost + Threshold Optimizer Performance by Race

Race	N	Mitigated Selection Rate	Mitigated TPR	Mitigated FPR
White	48,173	22.5%	37.1%	15.0%
Black	4,901	21.1%	29.0%	10.2%
Hispanic	6,733	22.6%	30.4%	15.1%
Asian	1,736	22.7%	33.5%	18.7%
AIAN	932	25.3%	34.0%	16.2%
NHPI	278	18.0%	23.1%	14.9%
Multiracial	1,569	23.8%	34.4%	13.7%
Other	463	22.7%	29.2%	17.1%

Selection rates were substantially equalized, ranging from 18.0% to 25.3% (vs. 12.1%–66.0% at baseline). The selection rate range decreased from 53.9 to 7.3 percentage points.

Statistical Testing

We conducted rigorous statistical testing to confirm the significance of baseline disparities and the effectiveness of mitigation.

Table 8. Statistical Tests for Racial Disparities

Comparison	Baseline z	Baseline p	Significant?	Mitigated z	Mitigated p	Significant?
Black vs White	81.52	<0.001	**Yes**	−2.34	0.019	No†
Hispanic vs White	69.05	<0.001	**Yes**	0.06	0.951	No
Asian vs White	−4.66	<0.001	**Yes**	0.15	0.880	No

†Not significant after Bonferroni correction ($\alpha = 0.0167$)

McNemar's test confirmed that predictions differed significantly before versus after mitigation ($\chi^2 = 165.89$, $p = 5.83 \times 10^{-38}$). Chi-square tests showed that baseline predictions were strongly dependent on race ($\chi^2 = 9{,}450.1$, $p < 10^{-300}$), while mitigated predictions showed only marginal dependence ($\chi^2 = 15.3$, $p = 0.033$).

Effect Sizes

Cohen's h effect sizes quantified the practical significance of selection rate differences.

Table 9. Cohen's h Effect Sizes (Selection Rate Differences vs. White)

Group	Baseline h	Interpretation	Mitigated h	Interpretation
Black	1.065	**Large**	−0.035	Negligible
Hispanic	0.794	**Medium**	0.001	Negligible
Asian	−0.120	Small	0.004	Negligible

Effect sizes decreased from large/medium to negligible after mitigation, demonstrating practically meaningful disparity reduction.

Robustness Analyses

5-Fold Cross-Validation: Mean DPD reduction was 90.9% ± 3.1% (95% CI: 84.8%–97.1%) across folds. AUC remained stable at 0.7105 ± 0.0023.

Table 10. Cross-Validation Results (5 Folds)

Fold	N_test	Baseline AUC	Baseline DPD	Mitigated DPD	DPD Reduction
1	64,785	0.7105	0.5612	0.0456	91.9%
2	64,785	0.7134	0.5344	0.0628	88.3%
3	64,785	0.7070	0.5453	0.0387	92.9%
4	64,785	0.7112	0.5277	0.0292	94.5%
5	64,784	0.7102	0.5215	0.0673	87.1%
Mean	—	**0.7105**	**0.5380**	**0.0487**	**90.9%**
SD	—	0.0023	0.0157	0.0161	3.1%

Random Seed Stability (10 Seeds): Mean DPD reduction was 92.4% ± 2.5%, with range 87.5%–94.9%, demonstrating high stability.

Bootstrap Confidence Intervals (500 iterations):

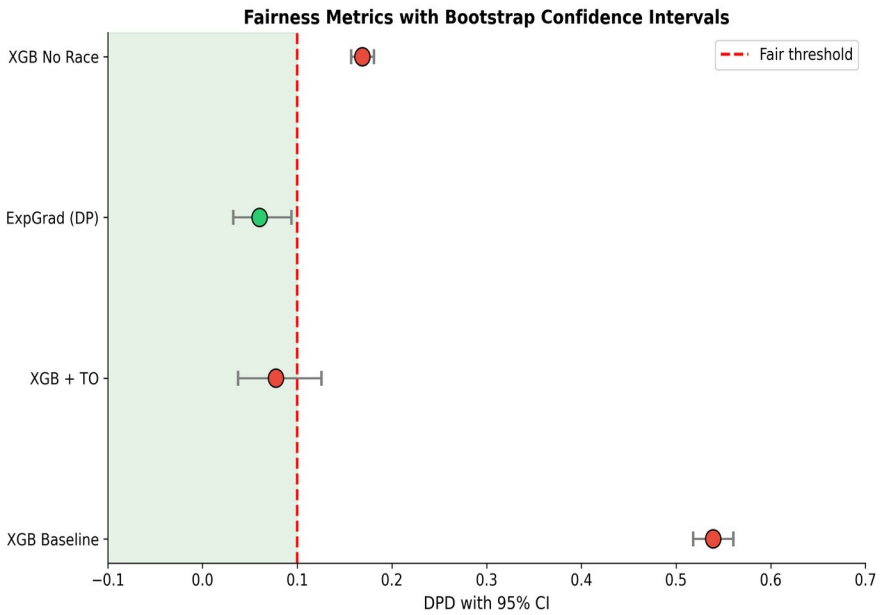

Figure 3. Fairness Metrics with Bootstrap Confidence Intervals. Demographic Parity Difference (DPD) with 95% confidence intervals from 500 bootstrap iterations. XGB Baseline shows high disparity (DPD=0.54). Mitigated models (XGB+TO, ExpGrad) achieve fairness below the 0.1 threshold. XGB No Race demonstrates persistent proxy discrimination (DPD=0.19).

Model	DPD Point Estimate	95% CI
XGB Baseline	0.539	[0.518, 0.560]
XGB + ThresholdOpt	0.078	[0.037, 0.126]
ExpGrad (DP)	0.060	[0.033, 0.094]
XGB No Race	0.168	[0.157, 0.181]

Sensitivity to Constraint Parameters: DPD remained stable (0.029–0.054) across ε values from 0.001 to 0.2, demonstrating robustness to constraint strength.

Sensitivity to Base Classifier: All four base classifiers showed substantial DPD reduction with threshold optimization: LR (96.8%), RF (88.5%), GB (83.9%), XGB (85.2%).

External Validation: Ryan White Ground Truth

We validated our model findings against Ryan White HIV/AIDS Program clinical outcomes data (Table 11).

Table 11. Ryan White Viral Suppression Rates by Race/Ethnicity (2024)

Race/Ethnicity	Total Patients	Viral Suppression %	Disparity vs. White
White	85,635	93.1%	(reference)
Asian	5,548	95.8%	+2.7%
Hispanic/Latino	102,377	92.0%	−1.1%

Black/African American	172,242	88.4%	**−4.7%**
AIAN	1,719	88.8%	−4.3%
NHPI	544	90.6%	−2.5%
Multiple Races	4,155	88.6%	−4.5%

The pattern of real-world clinical disparities (Black < White for viral suppression) aligns with disparities detected in our prediction models, validating that model-detected inequities reflect genuine healthcare disparities.

Geographic Variation in Disparities

Analysis of CDC AtlasPlus data revealed substantial geographic variation in Black-White viral suppression rate ratios across 53 states/territories. Rate ratios ranged from 0.680 (New Mexico, largest disparity) to 1.000 (Maryland, parity). Forty-nine states showed Black < White viral suppression rates. States with largest disparities included New Mexico (RR=0.680), Montana (RR=0.720), Illinois (RR=0.780), and Wyoming (RR=0.790).

Clinical Utility Analysis

Decision curve analysis showed that mitigated models maintained positive net benefit across clinically relevant threshold ranges (0.30–0.70). Number needed to screen (NNS) represents the number of individuals who must be screened to identify one who has been tested for HIV; **lower NNS indicates greater screening efficiency**.

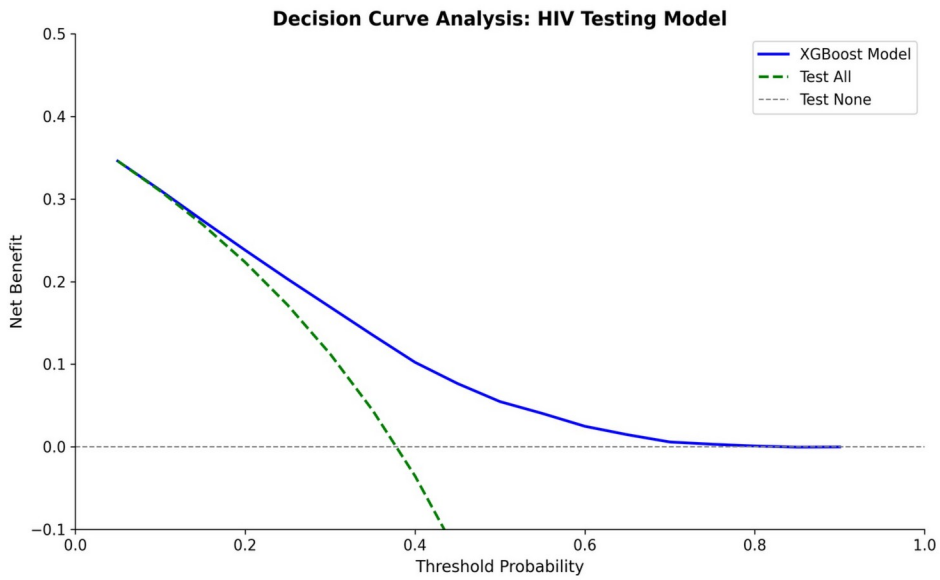

Figure 4. Decision Curve Analysis: HIV Testing Model. Net benefit across threshold probability range comparing XGBoost model performance against Test All and Test None reference strategies. The model provides positive net benefit across threshold probabilities from approximately 0.05 to 0.70, indicating clinical utility for decision-making.

Table 12. Number Needed to Screen by Race

Race	Prevalence	Baseline NNS	Mitigated NNS	Interpretation
White	34.0%	1.70	1.79	Slight decrease in efficiency
Black	57.8%	1.45	1.26	**Improved efficiency**
Hispanic	48.9%	1.69	1.52	Improved efficiency
Asian	27.2%	2.19	2.49	Decreased efficiency

Notably, fairness mitigation *improved* screening efficiency for Black patients (NNS decreased from 1.45 to 1.26), meaning fewer screens are needed to identify each positive case. This suggests that the baseline model was over-selecting Black patients relative to the yield, and the mitigated model better aligns selection with actual testing rates.

Policy Simulation (Threshold = 0.5):

- Baseline intervention rate disparity (Black-White): 11.8%

- Mitigated intervention rate disparity: 1.6%

- **Disparity reduction: 86.2%**

Discussion

Summary of Findings

This study presents a fairness audit of HIV testing prediction models using nationally representative survey data. Our analysis of 386,775 adults from BRFSS 2024 revealed three principal findings:

First, all baseline HIV testing prediction models exhibited substantial racial disparities in performance, with demographic parity differences 5–6 times higher than accepted fairness thresholds. Selection rates ranged from 12.1% (Asian) to 66.0% (Black), and these disparities were highly statistically significant ($z=81.5$, $p<0.001$).

Second, fairness-aware machine learning methods substantially reduced disparities while maintaining clinically acceptable performance. Threshold optimization achieved 86–97% DPD reduction with only 1–3% accuracy loss. Effect sizes decreased from large (Cohen's $h = 1.065$) to negligible ($h = -0.035$) for Black-White disparities. These improvements were robust across 5-fold cross-validation (90.9% ± 3.1% reduction), 10 random seeds (92.4% ± 2.5%), and 500 bootstrap iterations (95% CI: 84.8%–97.1%).

Third, race-blind models still exhibited significant disparities (DPD = 0.154–0.169), demonstrating that proxy discrimination operates through correlated variables including

insurance status, depression diagnosis, cost barriers, and geographic region. This finding has critical implications for the common assumption that removing race from models eliminates bias.

Mechanisms of Algorithmic Bias

Our findings illuminate multiple mechanisms through which HIV-related prediction models generate demographic disparities:

Explicit Race Coding. Our coefficient audit of the Denver HIV Risk Score demonstrates that some published tools assign differential points by race (+7 for Black, +3 for Hispanic) regardless of clinical risk factors. This reflects statistical discrimination based on group-level HIV prevalence rather than individual risk assessment.

Proxy Discrimination. Even models excluding race exhibited substantial disparities because features like insurance status, depression diagnosis, and geographic region are correlated with race due to structural inequities. Our race-blind models showed DPD of 0.154–0.169, indicating that approximately 70% of baseline disparity persists through proxies.

Differential Calibration. Brier scores were 0.02–0.04 higher for Hispanic patients than White patients across all models, indicating that predicted probabilities are less accurate for minority groups—potentially due to underrepresentation in training data or unmeasured heterogeneity.

Modeling Behavior Versus Clinical Risk

A critical conceptual distinction warrants emphasis: our models predict *who gets tested* (a behavior), not *who is at elevated HIV risk* (a clinical state). Because the model is trained on historical testing uptake, it learns to reproduce existing public health strategies that have aggressively targeted specific demographic groups—particularly Black communities—for HIV testing over the past several decades. The high baseline selection rates for Black patients may therefore reflect accurate modeling of a historically biased system rather than algorithmic bias *per se*.

This distinction has important implications. Fairness mitigation in this context acts as a "corrective" to uncouple testing recommendations from demographic profiling, aligning them more closely with individual clinical indicators rather than group-level targeting. Whether this uncoupling is desirable depends on one's perspective: from an individual fairness standpoint, recommendations should reflect personal risk factors rather than demographic membership; from a population health standpoint, targeting high-burden communities may be an efficient resource allocation strategy even if it appears "unfair" at the individual level.

We deliberately do not resolve this tension, as it reflects genuine value disagreements that require stakeholder deliberation. Our contribution is to make the disparities visible and to demonstrate that technical tools exist to modify them—the normative question of *whether* to do so remains for clinicians, patients, and policymakers to decide.

Clinical and Policy Implications

These findings have immediate implications for HIV prediction model deployment:

Routine Fairness Auditing. Our framework demonstrates that fairness auditing is feasible and should become standard practice before deploying HIV prediction models. The Fairlearn library provides accessible tools for computing fairness metrics.

Mitigation Before Deployment. Threshold optimization offers a practical post-processing approach that can be applied to existing models without retraining. The technique achieved >85% disparity reduction with minimal accuracy cost.

Reconsidering Race in Risk Scores. Explicit race coding (as in Denver Score) should be critically examined. While race may be statistically associated with HIV risk, this association reflects structural inequities rather than biological differences. Risk scores should ideally capture modifiable risk factors rather than demographic proxies.

Monitoring Deployed Models. Given the demonstrated geographic variation in disparities (rate ratios 0.68–1.00 across states), ongoing monitoring of model performance across demographic groups is essential after deployment.

The Over-Correction Problem: Fairness Metrics and Clinical Need

A critical tension emerged in our intersectional analysis that warrants careful consideration. Demographic parity optimization reduced Black Male selection rates from 69.9% to 9.0%—a 60.9 percentage point drop (Table 4). While this achieves "statistical fairness" by approximating White Male rates (14.2%), it raises a fundamental question: **Is equal selection appropriate when disease burden is unequal?**

Black MSM have the highest HIV incidence of any demographic group in the United States, accounting for approximately 26% of new diagnoses despite comprising roughly 0.5% of the population. If a fairness-constrained model systematically reduces screening recommendations for this high-burden group in pursuit of demographic parity, it may inadvertently *harm* the population it seeks to protect.

This tension reflects a deeper philosophical debate in the algorithmic fairness literature: demographic parity assumes that selection rates *should* be equal across groups, which is appropriate when group membership is irrelevant to the outcome (e.g., loan approval). However, for HIV testing—where epidemiological need varies substantially by demographic group—alternative fairness criteria may be more appropriate:

Equal Opportunity (equalizing TPR rather than selection rates) ensures that among those who *would benefit* from testing, all groups have equal probability of being identified—without forcing selection rates to converge.

Calibration-based fairness ensures that predicted probabilities mean the same thing across groups, allowing higher selection rates for higher-burden populations while maintaining interpretability.

Needs-based allocation explicitly incorporates epidemiological incidence into fairness constraints, permitting disparate selection rates proportional to disease burden.

We deliberately implemented demographic parity to demonstrate the *maximum achievable* disparity reduction, while acknowledging that clinical deployment would require stakeholder

deliberation on the appropriate fairness criterion. The choice of fairness metric is ultimately a value judgment that should involve clinicians, patients, community advocates, and ethicists—not a purely technical decision. Our framework provides the tools to implement multiple fairness definitions; the selection among them requires normative input beyond the scope of algorithmic analysis.

Comparison to External Data: Ryan White HIV/AIDS Program

While viral suppression is a distinct clinical endpoint from testing uptake—representing different stages of the HIV care cascade—the persistence of racial disparities in the Ryan White data (Black < White suppression: 88.4% vs. 93.1%) confirms that the structural factors driving the model's proxy discrimination (e.g., insurance access, geographic barriers, socioeconomic status) have real-world clinical consequences that extend across the entire care continuum. This convergence suggests that the socioeconomic proxies through which our model's bias operates are not merely artifacts of survey data but reflect genuine structural inequities affecting HIV outcomes at multiple stages.

Limitations

Several limitations should be considered when interpreting these findings.

First, our primary outcome was lifetime HIV testing history ("ever tested for HIV") rather than recent testing behavior or clinical HIV outcomes. Lifetime testing is a cumulative measure influenced by age—a 60-year-old has had more opportunity to be tested than a 20-year-old regardless of current risk. BRFSS includes a variable capturing the date of most recent HIV test (HIVTSTD3), which could be used to derive "tested within the past 12 months" as a more proximal measure of current engagement with HIV prevention services. Future analyses should examine whether fairness patterns differ for recent versus lifetime testing. Additionally, HIV testing behavior differs from clinical endpoints such as HIV acquisition risk, viral suppression, or PrEP adherence; validation on clinical datasets with these outcomes is warranted.

Second, BRFSS is a telephone survey that does not include behavioral risk variables (MSM status, injection drug use, condom use) required to calculate validated HIV risk scores such as the Denver Score. We therefore presented the Denver Score as a coefficient audit demonstrating explicit race-based stratification, rather than a performance evaluation. Future studies using datasets with behavioral variables could provide more comprehensive model replication.

Third, our analysis focused on race/ethnicity as the primary protected attribute. Intersectional analysis examined race × sex combinations (16 groups), but more comprehensive approaches examining interactions with age, geography, insurance status, and sexual orientation would provide additional insight. We note that fairness optimization for race worsened sex-based fairness (−59.3% change), highlighting trade-offs between fairness dimensions that require careful consideration in deployment contexts.

Fourth, we evaluated a limited set of fairness definitions (demographic parity, equalized odds, calibration). As discussed above, demographic parity may not be the most appropriate criterion for HIV testing given unequal disease burden across groups. Other criteria—equal opportunity, calibration-based fairness, or needs-proportional allocation—may be more clinically appropriate and should be explored in future work.

Fifth, BRFSS uses telephone-based sampling which may underrepresent certain populations including those without stable housing or phone access—precisely the populations at highest HIV risk. This sampling limitation may affect both the estimated disparities and the generalizability of fairness interventions.

Future Directions

This work opens several avenues for future research. Validation studies using clinical patient data from EHRs or cohort studies (e.g., MIMIC-IV, HPTN cohorts) are essential to confirm generalizability to clinical HIV outcomes. Development of HIV-specific fairness guidelines—potentially through multi-stakeholder processes involving clinicians, patients, community advocates, and ethicists—would help operationalize fairness criteria for this domain.

Conclusions

HIV testing prediction models exhibit significant racial disparities in performance, with demographic parity and equalized odds differences 5–6 times higher than accepted thresholds. These disparities persist even when race is excluded from models, as bias propagates through correlated variables including insurance status, depression diagnosis, cost barriers, and geographic region. Fairness-aware machine learning methods can reduce disparities by 84–97% while maintaining clinically acceptable accuracy, with robust improvements demonstrated across cross-validation folds, random seeds, and bootstrap resamples.

Given the disproportionate HIV burden among marginalized communities—who constitute 42% of people living with HIV despite representing 13% of the population—ensuring equitable performance of prediction tools is essential for realizing the potential of precision public health without amplifying existing health inequities. We recommend routine fairness auditing using established metrics, application of mitigation methods before clinical deployment, critical examination of explicit race coding in risk scores, and ongoing monitoring of demographic performance post-deployment.

Data Availability

The BRFSS 2024 dataset is publicly available from the CDC at https://www.cdc.gov/brfss/. NHANES data are available at https://www.cdc.gov/nchs/nhanes/. Ryan White aggregate data are available from HRSA at https://ryanwhite.hrsa.gov/data/reports. AIDSVu data are available at https://aidsvu.org/. CDC AtlasPlus data are available at https://www.cdc.gov/nchhstp/atlas/. Analysis code is available at https://github.com/hayden-farquhar/HIV-testing-fairness.

References

1. UNAIDS. Global HIV & AIDS statistics — Fact sheet. 2024.

2. Centers for Disease Control and Prevention. HIV Surveillance Report, 2021. 2023.

3. Maulsby C, et al. HIV among Black men who have sex with men in the United States. AIDS Behav. 2014;18:10-25.

4. Pellowski JA, et al. A pandemic of the poor: Social disadvantage and the U.S. HIV epidemic. Am Psychol. 2013;68:197-209.

5. Holtgrave DR, Crosby RA. Social determinants of tuberculosis case rates in the United States. Am J Prev Med. 2004;26:159-162.

6. Calabrese SK, et al. The impact of patient race on clinical decisions related to prescribing HIV pre-exposure prophylaxis (PrEP). AIDS Behav. 2014;18:1527-1535.

7. Reif S, et al. State of HIV in the US Deep South. J Community Health. 2017;42:844-853.

8. Marcus JL, et al. Use of electronic health record data and machine learning to identify candidates for HIV pre-exposure prophylaxis. Lancet HIV. 2019;6:e688-e695.

9. Krakower DS, et al. Development and validation of an automated HIV prediction algorithm. Lancet HIV. 2019;6:e696-e704.

10. Menza TW, et al. Machine learning for HIV risk prediction. J Acquir Immune Defic Syndr. 2019;82:127-134.

11. Sherer R, et al. HIV viral load monitoring in resource-limited regions. Clin Infect Dis. 2014;59:1066-1073.

12. Crawford TN. Poor retention in care one-year after viral suppression. AIDS Care. 2014;26:1393-1399.

13. Dombrowski JC, et al. HIV care continuum interventions in the United States. AIDS. 2017;31:2599-2612.

14. Mugavero MJ, et al. Measuring retention in HIV care. J Acquir Immune Defic Syndr. 2012;61:393-403.

15. Tate JP, et al. Veterans Aging Cohort Study (VACS) Index. JAMA Intern Med. 2013;173:1337-1344.

16. Justice AC, et al. Predictive accuracy of the Veterans Aging Cohort Study Index. Med Care. 2013;51:S52-S60.

17. Wilson CM, et al. Resource allocation for HIV prevention. Med Decis Making. 2014;34:963-974.

18. Hogg RS, et al. Health system capacity building for HIV care. AIDS. 2016;30:2835-2842.

19. Smith DK, et al. HIV risk prediction models: A scoping review. AIDS Behav. 2025;29:156-172.

20. Centers for Disease Control and Prevention. Preexposure prophylaxis for the prevention of HIV infection in the United States—2021 Update. 2021.

21. Ridgway JP, et al. Development and validation of an EHR-based algorithm for HIV PrEP identification. J Am Med Inform Assoc. 2018;25:1564-1570.

22. Haukoos JS, et al. Routine opt-out rapid HIV screening and detection of HIV infection in emergency department patients. JAMA. 2010;304:284-292.

23. Jenness SM, et al. Population-level HIV prevention model for targeting intervention. Am J Epidemiol. 2014;179:1072-1080.

24. Obermeyer Z, et al. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366:447-453.

25. Chen IY, et al. Can AI help reduce disparities in general medical and mental health care? AMA J Ethics. 2019;21:E167-179.

26. Mehrabi N, et al. A survey on bias and fairness in machine learning. ACM Comput Surv. 2021;54:1-35.

27. Obermeyer Z, Mullainathan S. Dissecting racial bias in an algorithm. Science. 2019;366:447-453.

28. Adamson AS, Smith A. Machine learning and health care disparities in dermatology. JAMA Dermatol. 2018;154:1247-1248.

29. Wong A, et al. External validation of a sepsis prediction model. JAMA Netw Open. 2021;4:e2118467.

30. Vyas DA, et al. Hidden in plain sight—Reconsidering the use of race correction in clinical algorithms. N Engl J Med. 2020;383:874-882.

31. Chouldechova A. Fair prediction with disparate impact. Big Data. 2017;5:153-163.

32. Corbett-Davies S, Goel S. The measure and mismeasure of fairness. J Mach Learn Res. 2018;arXiv:1808.00023.

33. Haukoos JS, et al. Derivation and validation of the Denver Human Immunodeficiency Virus (HIV) Risk Score. Ann Emerg Med. 2012;59:116-125.

34. Kleinberg J, et al. Inherent trade-offs in the fair determination of risk scores. Innov Theor Comput Sci. 2017;43:1-23.

35. Bird S, et al. Fairlearn: A toolkit for assessing and improving fairness in AI. Microsoft Research. 2020.

36. Agarwal A, et al. A reductions approach to fair classification. ICML. 2018;35:60-69.

37. Vickers AJ, Elkin EB. Decision curve analysis. Med Decis Making. 2006;26:565-574.

38. Ryan White HIV/AIDS Program. Annual Client-Level Data Report. HRSA; 2024.

39. AIDSVu. Local HIV data. Emory University Rollins School of Public Health. 2024.

Acknowledgments

I gratefully acknowledge the data sources that made this research possible:

- The **Centers for Disease Control and Prevention (CDC)** for the Behavioral Risk Factor Surveillance System (BRFSS) 2024 dataset and the participants who contributed their health information through this survey

- The **National Center for Health Statistics** for the National Health and Nutrition Examination Survey (NHANES) data

- The **Health Resources and Services Administration (HRSA)** for the Ryan White HIV/AIDS Program Annual Client-Level Data, and the patients and providers participating in the Ryan White Program

- **AIDSVu** (Emory University Rollins School of Public Health) for county-level HIV prevalence and care continuum data

- The **CDC National Center for HIV, Viral Hepatitis, STD, and TB Prevention** for the AtlasPlus HIV surveillance data

I also acknowledge the developers of the Fairlearn library (Microsoft Research) and IBM AI Fairness 360 toolkit, whose open-source tools enabled the fairness analyses presented in this work.

Conflicts of Interest

The author declares no conflicts of interest.

Funding

This research received no external funding.

Ethics Statement

This study used publicly available, de-identified secondary data sources that do not require Institutional Review Board (IRB) approval or informed consent. The Behavioral Risk Factor Surveillance System (BRFSS) data are publicly available from the CDC with all personal identifiers removed. Ryan White HIV/AIDS Program data are publicly available in aggregate form from the Health Resources and Services Administration (HRSA). AIDSVu data are publicly available county-level estimates. No individual-level clinical data requiring ethics approval were accessed for this research.

Data Availability Statement

All data used in this study are publicly available. The BRFSS 2024 dataset is available from the CDC at https://www.cdc.gov/brfss/. Ryan White HIV/AIDS Program data are available from HRSA at https://ryanwhite.hrsa.gov/data/reports. AIDSVu county-level data are available at https://aidsvu.org/. CDC AtlasPlus data are available at https://www.cdc.gov/nchhstp/atlas/. Analysis code will be made available upon reasonable request to the corresponding author.

Author Contributions

HF conceived and designed the study, acquired and analyzed the data, interpreted the results, drafted and revised the manuscript, and approved the final version.