IRSTI 28.23.15 DOI: https://doi.org/10.26577/JMMCS201011426

A. Ramu

Heriot-Watt University International Faculty, K. Zhubanov Aktobe Regional University, Aktobe, Kazakhstan
e-mail: arulmr@gmail.com

A HYBRID BIAS-MITIGATION FRAMEWORK FOR FAIR CLINICAL READMISSION PREDICTION

Machine-learning models increasingly guide the allocation of clinical care, yet models optimised for aggregate accuracy can encode substantial disparities in error rates across demographic groups. This study addresses fair prediction of 30-day hospital readmission for diabetic inpatients under two constraints that are common in practice but rarely modelled together: the deployed model must not use a protected attribute at decision time, and the positive class is rare. We propose FAIR-Health, a hybrid bias-mitigation framework that couples Kamiran–Calders instance reweighing (pre-processing) with a differentiable equalized-odds surrogate penalty embedded in the training objective (in-processing). The result is a single attribute-blind model with one tunable control over the fairness/accuracy trade-off. Using the UCI Diabetes 130-US Hospitals cohort (71,518 patients after leakage-free, patient-level deduplication), we benchmark FAIR-Health against pre-, in-, and post-processing baselines across ten random data splits (seeds) with paired significance testing. Compared to the unmitigated gradient-boosted model, FAIR-Health reduced the equalized-odds difference by 44.7% for gender and 65.6% for race, at a within-model-family AUROC cost of at most 0.004 and with no loss of calibration. Measured against the unmitigated logistic model, the true-positive-rate gap between racial groups fell from 0.111 to 0.019. We further show that hard-constraint methods attain near-zero unfairness only by collapsing to a near-constant classifier, which removes their main practical value. The framework and benchmark give practical guidance for deploying equitable clinical prediction models.

Keywords: algorithmic fairness, responsible artificial intelligence, clinical prediction, bias mitigation, equalized odds.

А. Раму

Хериот-Уатт университетінің халықаралық факультеті, Қ. Жұбанов атындағы Ақтөбе өңірлік университеті, Ақтөбе, Қазақстан
e-mail: arulmr@gmail.com

Әділ клиникалық қайта госпитализацияны болжауға арналған гибридті ауытқуды азайту әдістемесі

Машиналық оқыту модельдері клиникалық көмекті бөлуде барған сайын кеңінен қолданылады, алайда жалпы дәлдікке оңтайландырылған модельдер демографиялық топтар арасында қателік деңгейлерінде елеулі теңсіздіктерді қалыптастыруы мүмкін. Бұл зерттеу қант диабеті бар науқастардың 30 күн ішінде қайта госпитализациялануын әділ болжау мәселесін тәжірибеде жиі кездесетін, бірақ сирек бірге ескерілетін екі шектеу аясында қарастырады: орналастырылған модель шешім қабылдау кезінде қорғалатын белгіні пайдаланбауы тиіс және оң класс сирек кездеседі. Біз FAIR-Health гибридті ауытқуды азайту әдістемесін ұсынамыз, ол Kamiran–Calders салмақтауын (алдын ала өңдеу) оқыту мақсатына енгізілген дифференциалданатын теңестірілген шанстар (equalized odds) айыппұлымен (ішкі өңдеу) біріктіреді. Нәтижесінде әділдік пен дәлдік арасындағы тепе-теңдікті реттейтін бір параметрі бар, атрибутқа тәуелсіз бір ғана модель алынады. UCI Diabetes 130-US Hospitals деректер жиынтығын (ақпарат ағуынсыз (leakage-free), науқас деңгейіндегі дедупликациядан кейін 71 518 науқас) пайдалана отырып, біз FAIR-Health әдісін алдын ала, ішкі және кейінгі өңдеу әдістерімен он кездейсоқ бөліну (seed) бойынша, жұптық маңыздылық тесттерін қолдана отырып салыстырамыз. Алгоритмдік біржақтылықты азайту әдістері қолданылмаған градиенттік бустинг моделімен салыстырғанда FAIR-Health теңестірілген шанстар айырмашылығын (EOD) жынысы бойынша 44,7%-ға және нәсілі бойынша 65,6%-ға азайтты, бұл ретте бір модельдер

© 2026 Al-Farabi Kazakh National University

тобы шеңберінде AUROC шығыны 0,004-тен аспады және калибрлеу нашарламады. Алгоритмдік біржақтылықты азайту әдістері қолданылмаған логистикалық регрессия моделімен салыстырғанда, нәсілдік топтар арасындағы шынайы оң нәтижелер үлесінің айырмашылығы (true-positive-rate gap) 0,111-ден 0,019-ға дейін төмендеді. Сонымен қатар, қатаң шектеулі әдістер әділдіктің нөлге жуық мәндеріне классификаторды тұрақтыға жуық етіп қана жеткізеді, бұл олардың практикалық құндылығын жояды. Ұсынылған әдістеме мен салыстыру әділ клиникалық болжау модельдерін орналастыруға практикалық нұсқаулық береді.
Түйін сөздер: алгоритмдік әділдік, жауапты жасанды интеллект, клиникалық болжау, ауытқуды азайту, теңестірілген шанстар (equalized odds).

А. Раму
Международный факультет Университета Хериот-Ватт, Актюбинский региональный университет им. К. Жубанова, Актобе, Казахстан
e-mail: arulmr@gmail.com

Гибридная методика снижения смещения для справедливого прогнозирования повторной госпитализации

Модели машинного обучения всё чаще определяют распределение клинической помощи, однако модели, оптимизированные по общей точности, могут закреплять существенные различия в уровнях ошибок между демографическими группами. В работе рассматривается задача справедливого прогнозирования повторной госпитализации в течение 30 дней у пациентов с сахарным диабетом при двух практических ограничениях, которые редко моделируются совместно: развёрнутая модель не должна использовать защищённый признак в момент принятия решения, а положительный класс является редким. Мы предлагаем FAIR-Health — гибридную методику снижения смещения, которая объединяет взвешивание Kamiran–Calders (предобработка) с дифференцируемым штрафом уравненных шансов (equalized odds) в целевой функции обучения (внутренняя обработка). В результате формируется единая модель, не использующая защищённый признак, с одним настраиваемым параметром баланса справедливости и точности. На данных UCI Diabetes 130-US Hospitals (71 518 пациентов после дедупликации без утечки (leakage-free) на уровне пациента) FAIR-Health сравнивается с методами предварительной, внутренней и постобработки по десяти случайным разбиениям выборки (seed), с парными тестами значимости. По сравнению с моделью градиентного бустинга без коррекции смещения FAIR-Health снизил разность уравненных шансов на 44,7% по полу и на 65,6% по расе; при этом потеря AUROC в пределах семейства моделей не превысила 0,004, а калибровка не ухудшилась. Относительно логистической модели без коррекции смещения разрыв в доле истинно положительных результатов между расовыми группами сократился с 0,111 до 0,019. Мы также показываем, что методы с жёсткими ограничениями достигают почти нулевой несправедливости только за счёт вырождения в почти постоянный классификатор, что лишает их практической ценности. Предложенная методика и сравнительный анализ дают практические рекомендации по внедрению справедливых моделей клинического прогнозирования.
Ключевые слова: алгоритмическая справедливость, ответственный искусственный интеллект, клиническое прогнозирование, снижение смещения, уравненные шансы.

1 Introduction

Predictive models are increasingly embedded in the operational fabric of health systems, where they triage patients, forecast deterioration, and steer the allocation of follow-up care. Because these decisions bear directly on who receives attention, a model that is accurate on average but systematically less sensitive for one demographic group does not merely misrank patients; it withdraws care from those who need it. The canonical demonstration is Obermeyer et al. [1], who showed that a widely deployed population-health algorithm, by using historical cost as a proxy for need, systematically under-referred Black patients at equal

levels of illness. Similar disparities have since been documented across medical imaging [2, 3] and administrative prediction, establishing that unfairness in clinical artificial intelligence is a measurable and recurring failure mode rather than a hypothetical one [4]. Surveys of the area note that no single mitigation family dominates and that the right choice depends on data access and deployment constraints [5].

Thirty-day hospital readmission is a high-stakes instance of this problem. Readmission-risk scores drive enrolment in transitional-care and case-management programmes; a score that under-identifies at-risk patients in a protected group diverts preventive resources away from them. Yet the readmission setting has two properties that make naive fairness interventions unreliable and that are central to this work. First, the positive class is rare (roughly one in eleven patients in our cohort is readmitted within 30 days), so methods that enforce fairness as a hard constraint can satisfy it degenerately by predicting almost no positives. Second, deployment constraints in care settings often forbid using a patient's race or sex as an explicit input to a decision rule, so post-processing methods that apply group-specific thresholds, while attaining statistical parity, raise a distinct ethical and legal objection (disparate treatment) and require the protected attribute to be collected at inference time.

We therefore ask a deliberately practical question: among mitigation methods that yield a single, attribute-blind model and preserve clinically useful decision utility, which most reduces error-rate disparity on imbalanced clinical prediction, and by what mechanism? Our contributions are fourfold. (i) We cast clinical bias mitigation as reducing error-rate disparity subject to attribute-blind, decision-useful deployment on an imbalanced target. (ii) We propose FAIR-Health, a single-objective hybrid that composes instance reweighing with a differentiable equalized-odds surrogate penalty. (iii) We provide a leakage-free, multi-metric, ten-seed benchmark of six pre-, in-, and post-processing strategies with paired significance testing. (iv) We report four findings, including negative ones: the hidden disparity in unmitigated models; FAIR-Health's significant fairness gain at negligible utility cost; the decision collapse of hard-constraint methods under imbalance; and an ablation isolating the in-processing penalty as the effective component. Group-fairness criteria are provably incompatible when base rates differ [6]; we adopt equalized odds [7] as the primary target because, in a screening context, both missing an at-risk patient and over-flagging a healthy one carry group-specific harms [8–10].

2 Materials and Methods

2.1 Dataset and preprocessing

We use the UCI Diabetes 130-US Hospitals dataset [11], comprising 101,766 inpatient encounters for diabetic patients across 130 US hospitals (1999–2008). To prevent patient-level leakage we retain the first encounter per patient (71,518 patients) and remove encounters whose discharge disposition records death or hospice. The prediction target is readmission within 30 days (positive prevalence 9.0%). After removing identifiers, free-text diagnosis codes, and high-missingness fields, 17 numeric predictors remain (utilisation counts, admission/discharge codes, number of diagnoses, age midpoint, and medication-change and laboratory-measurement indicators). The sensitive attributes are gender (Female vs. Male) and race (African-American vs. Caucasian). Features are standardised on the training fold

of each split.

2.2 Fairness and utility metrics

Let $x \in \mathbb{R}^d$ be patient features, $y \in \{0,1\}$ the readmission label (1 = readmitted within 30 days), and $a \in \{0,1\}$ a binary protected attribute. A model outputs a score $s(x) \in [0,1]$ and a decision $\hat{y} = \mathbb{1}[s(x) \geq t]$. For group a let TPR_a, FPR_a and SR_a denote the true-positive rate, false-positive rate and selection rate. We report the equalized-odds difference, demographic-parity difference and disparate-impact ratio,

$$\mathrm{EOD} = \max\big(|\mathrm{TPR}_1 - \mathrm{TPR}_0|,\, |\mathrm{FPR}_1 - \mathrm{FPR}_0|\big), \mathrm{DPD} = |\mathrm{SR}_1 - \mathrm{SR}_0|, \mathrm{DI} = \frac{\min(\mathrm{SR}_0, \mathrm{SR}_1)}{\max(\mathrm{SR}_0, \mathrm{SR}_1)}, \quad (1)$$

together with utility metrics: AUROC (primary), AUPRC (appropriate under class imbalance), balanced accuracy at the operating threshold, and the Brier score for calibration.

2.3 The FAIR-Health framework

FAIR-Health composes two stages that attack complementary sources of disparity and that both yield a single model which does not consume the protected attribute at inference (Figure 1).

Stage 1: reweighing (pre-processing). Following Kamiran and Calders [12], each training instance receives a weight $w_i = P(A{=}a_i)\,P(Y{=}y_i) \,/\, P(A{=}a_i, Y{=}y_i)$ estimated from empirical frequencies. These weights make A and Y independent in the weighted training distribution, removing the base-rate component of disparity without altering any feature or requiring a at test time.

Stage 2: fairness-regularised objective (in-processing). We train a linear scorer $s(x) = \sigma(\theta^\top x)$ by minimising a weighted log-loss augmented with a differentiable equalized-odds surrogate. Writing $m_{a,c}$ for the mean predicted score in the group with attribute a and true label c, the objective is

$$\mathcal{L}(\theta) = \frac{1}{n}\sum_i w_i\, \ell\big(y_i, s(x_i)\big) + \lambda \sum_{c\in\{0,1\}} \big(m_{1,c} - m_{0,c}\big)^2 + \frac{\mu}{2}\|\theta\|_2^2. \quad (2)$$

The penalty shrinks, within each true-outcome stratum, the gap between the two groups' mean predicted scores. Equal mean scores within a stratum do not force equal TPR or FPR at a given decision threshold, so the penalty is a smooth surrogate for matching TPR and FPR rather than an exact relaxation. It is fully differentiable, with $\nabla_\theta m_{a,c} = \frac{1}{|G_{a,c}|}\sum_{i\in G_{a,c}} s_i(1-s_i)x_i$, so Eq. (2) is minimised directly by L-BFGS-B. The objective is non-convex in θ because of the squared difference of group mean sigmoid scores, so L-BFGS-B returns a local minimiser; we initialise θ at the zero vector and stop when the projected-gradient norm falls below 10^{-5} (the L-BFGS-B default) or after at most 500 iterations. The non-convexity did not cause instability in practice: the per-seed selected λ lay in a narrow range (reported below) and the solutions reproduced across seeds. Alternative in-processing designs replace this penalty with adversarial removal of attribute information [13] or with a constrained reduction to cost-weighted classification [14]; we use a soft penalty because, unlike a hard constraint, it degrades gracefully as λ grows. At $\lambda = 0$ the model reduces to

weighted logistic regression. The resulting model applies one decision rule to every patient and uses the protected attribute only during training, so that fairness through unawareness holds at inference.

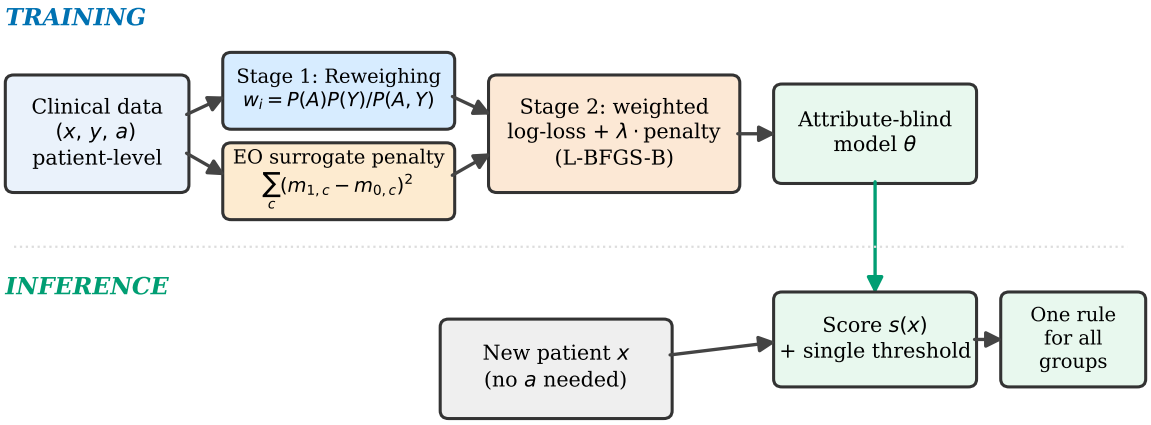


Figure 1: The FAIR-Health framework: two-stage attribute-aware training and attribute-blind inference.

2.4 Baselines, threshold, and protocol

We compare against: unmitigated gradient-boosted trees (utility ceiling); plain logistic regression (linear reference); reweighing only [12]; correlation removal [15]; exponentiated-gradient reduction with an equalized-odds constraint [14]; and equalized-odds threshold optimisation (post-processing) [7]. Implementations use scikit-learn [16] and Fairlearn [17]. We use a 70/30 stratified patient-level split, fixed per seed and reused across all methods, reporting mean over ten seeds with 95% confidence intervals; significance uses the paired Wilcoxon signed-rank test with Holm–Bonferroni correction, following prediction-model reporting guidance [18].

For every single-model method the decision threshold is tuned on the training fold to maximise balanced accuracy and is then applied as one uniform rule to all patients; the threshold is never fitted on the test fold. Only the post-processing baseline departs from this, applying group-specific thresholds by construction. The fairness penalty of FAIR-Health has a single hyperparameter λ, which we select on an internal validation split (a 25% holdout of the training fold, with no access to the test fold): among the grid $\lambda \in \{0, 0.5, 1, 2, 5, 10, 20, 35, 50\}$ we choose the value that minimises validation EOD subject to validation AUROC staying within 0.01 of the unconstrained model. Averaged across seeds the selected value was $\lambda \approx 47$ for gender and $\lambda \approx 41$ for race. The 25% validation holdout is used only to choose λ: once λ is fixed, the final FAIR-Health model is refitted on the entire training fold, and its decision threshold is tuned on that same full training fold (not on the 75% fitting subset), consistently with the other single-model methods. The fairness/accuracy trade-off is smooth and near-monotone in λ (Figure 2): for race, EOD falls from about 0.11 at $\lambda = 0$ to about 0.03 at

$\lambda = 50$ for an AUROC cost of roughly 0.005.

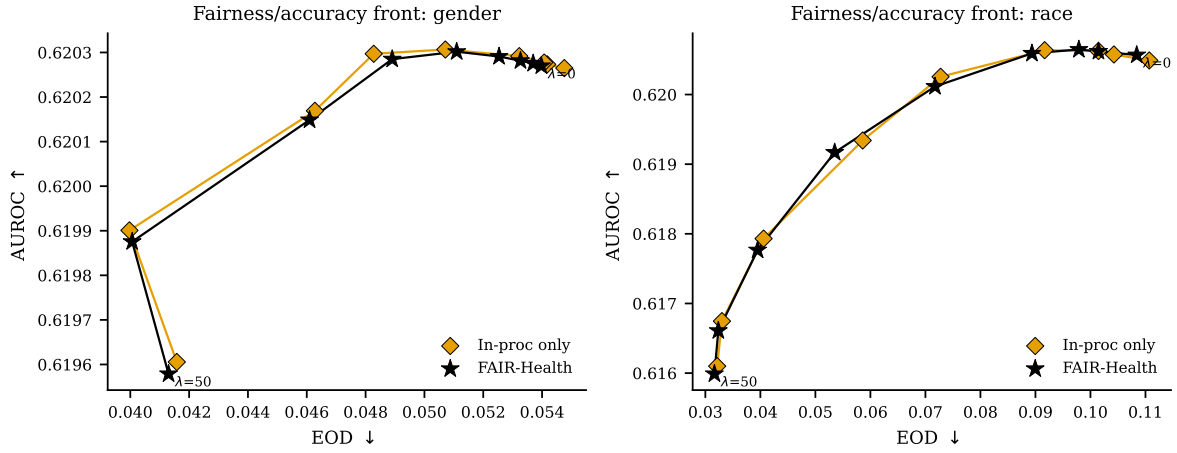


Figure 2: Fairness/accuracy front traced by the penalty strength λ (from 0 to 50) for FAIR-Health and in-processing only. The trade-off is smooth and near-monotone; the two curves nearly coincide, consistent with the ablation in Section 3.

3 Results and Discussion

Table 1 reports all methods on both sensitive attributes, with 95% confidence intervals over ten seeds for AUROC and EOD and the [min, max] range over ten seeds for DI; Figure 4 gives a normalised multi-metric overview.

Finding 1: unmitigated models encode substantial disparity that aggregate metrics hide. The unmitigated gradient-boosted model attains the best AUROC (0.641 gender, 0.640 race) while still showing a large EOD (0.076 for gender, 0.099 for race); the unmitigated logistic model is comparable or worse on fairness (0.055 for gender, 0.114 for race), so moving to a stronger learner does not remove the disparity. A subgroup decomposition for race is stark: the unmitigated logistic model flags African-American patients who are truly readmitted far less often than Caucasian patients (TPR 0.494 vs. 0.605; a gap of 0.111), the precise direction of harm identified by Obermeyer et al. [1]. Reporting AUROC alone would completely conceal this (Figure 3).

Finding 2: FAIR-Health significantly reduces disparity at negligible utility cost. FAIR-Health lowers EOD to 0.042 (gender) and 0.034 (race), reductions of 44.7% and 65.6% versus the unmitigated gradient-boosted model (percentages computed from unrounded means); all comparisons against the unmitigated gradient-boosted, logistic and reweighing baselines are significant after Holm–Bonferroni correction ($p_{\text{adj}} < 0.01$). The within-model-family utility cost (that is, FAIR-Health against the logistic model, both linear) is at most 0.004 AUROC; the roughly 0.02 AUROC gap to the gradient-boosted model is the cost of using a linear model, not of the fairness constraint. AUPRC and Brier are essentially unchanged: Brier is unchanged to three decimals, and AUPRC moves by at most 0.001 (for race it decreases from 0.148 for the logistic model to 0.147 for FAIR-Health). Relative to the unmitigated logistic model, FAIR-Health shrinks the African-American vs. Caucasian gap in

Table 1: Utility and fairness on 30-day readmission (mean over 10 seeds). For AUROC and EOD, ± is the 95% confidence interval computed as a Student-t interval over the ten seeds ($t_{9,0.975} = 2.262$). The disparate-impact ratio DI is bounded in $[0, 1]$, so for DI we report the mean and the [min, max] range across the ten seeds instead. Bal.Acc is balanced accuracy at the operating threshold.

Method	AUROC	AUPRC	Bal.Acc	Brier	EOD	DPD	DI [min,max]
Sensitive attribute: gender							
Unmitigated (GBM)	0.641±0.005	0.165	0.603	0.079	0.076±0.013	0.050	0.867 [0.83,0.90]
Logistic	0.620±0.004	0.144	0.587	0.080	0.055±0.007	0.045	0.905 [0.89,0.93]
Reweighing	0.620±0.004	0.144	0.587	0.080	0.054±0.008	0.044	0.906 [0.89,0.93]
Corr. Remover	0.620±0.004	0.144	0.588	0.080	0.021±0.009	0.012	0.972 [0.95,1.00]
In-proc (Stage 2)	0.620±0.004	0.144	0.586	0.080	0.043±0.008	0.034	0.924 [0.89,0.95]
Exp. Gradient	0.501±0.000	0.091	0.501	0.090	0.001±0.001	0.000	0.716 [0.35,1.00]
Post-proc (EO)	0.620±0.004	0.144	0.500	0.080	0.001±0.001	0.000	0.533 [0.00,1.00]
FAIR-Health	0.620±0.004	0.144	0.586	0.080	**0.042±0.009**	0.032	0.927 [0.89,0.95]
Sensitive attribute: race							
Unmitigated (GBM)	0.640±0.003	0.167	0.601	0.080	0.099±0.016	0.072	0.802 [0.78,0.82]
Logistic	0.620±0.002	0.148	0.585	0.081	0.114±0.017	0.088	0.804 [0.78,0.82]
Reweighing	0.620±0.003	0.148	0.585	0.081	0.107±0.017	0.084	0.807 [0.79,0.83]
Corr. Remover	0.620±0.003	0.147	0.584	0.081	0.039±0.015	0.011	0.974 [0.95,1.00]
In-proc (Stage 2)	0.617±0.004	0.147	0.584	0.081	0.035±0.011	0.020	0.950 [0.90,0.98]
Exp. Gradient	0.501±0.000	0.092	0.501	0.091	0.004±0.002	0.000	0.589 [0.00,0.92]
Post-proc (EO)	0.620±0.002	0.148	0.500	0.081	0.000±0.000	0.000	0.800 [0.00,1.00]
FAIR-Health	0.616±0.004	0.147	0.583	0.081	**0.034±0.013**	0.020	0.947 [0.90,0.98]

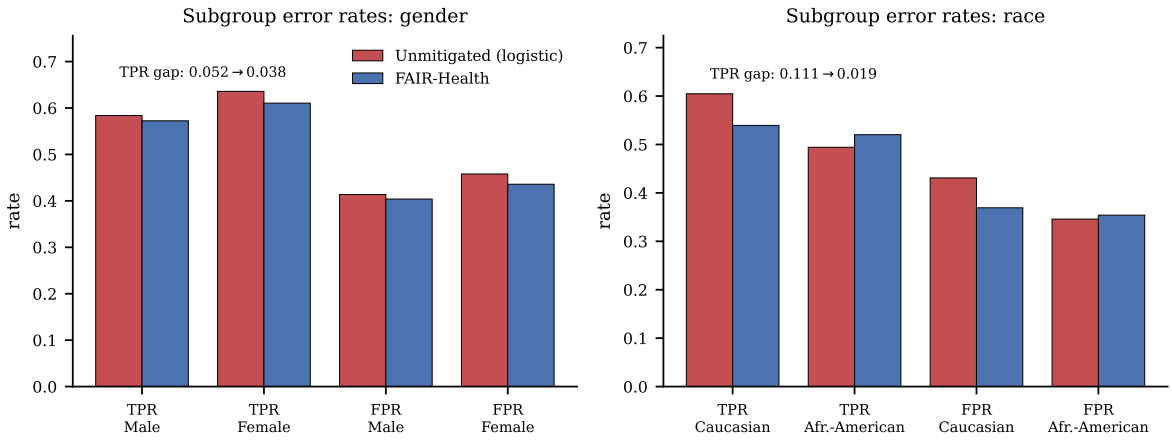


Figure 3: Subgroup true- and false-positive rates before (unmitigated logistic) and after FAIR-Health. For race, the gap in mean true-positive rate between groups falls from 0.111 to about 0.019.

mean TPR from 0.111 to about 0.019 and in mean FPR from 0.085 to about 0.015, without harming per-group calibration.

Finding 3: hard-constraint methods reach parity only by collapsing decision utility. Exponentiated-gradient reduction and equalized-odds post-processing both drive EOD below 0.005, but at their operating points their balanced accuracy falls to ≈ 0.50: they become near-constant classifiers that predict almost no readmissions. Both also make the disparate-impact ratio uninformative: because they predict a positive outcome for almost no patients, the selection rate is near zero in both groups, so DI is computed from very small, near-equal

selection rates and ranges across almost the whole $[0, 1]$ interval across seeds (Table 1); its point value is not meaningful. Post-processing additionally requires the protected attribute at inference. Under the severe class imbalance typical of readmission, "perfect" fairness scores from these methods are therefore of little clinical value. Among methods that retain non-trivial decision utility (balanced accuracy > 0.58), FAIR-Health and correlation removal achieve the lowest EOD (Figure 4).



Figure 4: Normalised multi-metric overview (green best, red worst; raw values overlaid). FAIR-Health and correlation removal are the only methods strong on both utility and fairness.

Finding 4: the in-processing penalty, not reweighing, drives the gain. Adding reweighing to the logistic baseline barely moves EOD (gender $0.055 \rightarrow 0.054$; race $0.114 \rightarrow 0.107$), whereas adding the in-processing penalty produces the large reduction, and the full hybrid shows no statistically significant difference from in-processing alone (with only ten seeds this is a null result, not a demonstration of equivalence). Reweighing is near-inert here because the two groups' readmission base rates are already close, so the weights are near-uniform; its value would be expected to grow on cohorts with larger base-rate gaps. This indicates that where a fairness signal is injected matters more than combining two stages.

Discussion. The results support a nuanced conclusion. On the axis we set out to optimise, namely reducing error-rate disparity while keeping a single, attribute-blind, decision-useful model, FAIR-Health performs among the best available methods, cutting EOD by up to two-thirds at a utility cost that is negligible within its model family. At the same time, the results caution against two oversimplifications. A low fairness number is not self-evidently good (the two methods with the best EOD achieve it by not making useful decisions), and our own hybrid is not uniformly dominant: correlation removal matches it on race and attains a lower EOD on gender (0.021 vs. 0.042) at the same utility. Method selection should therefore be driven by deployment constraints and the operative fairness criterion, not by a single leaderboard number.

Limitations. This study is single-cohort and retrospective; external validation on a second site would strengthen generalisability, as would evaluation on cohorts with larger

between-group base-rate gaps, where the reweighing stage may contribute more. Race and sex are binarised for tractability, so intersectional and multi-group fairness are left for future work, although the framework admits them. The in-processing scorer is linear; combining the surrogate with a non-linear base learner is a natural extension that must be weighed against the interpretability of linear clinical models. Finally, group-fairness metrics are necessary but not sufficient: they do not substitute for addressing the upstream measurement and access inequities that generate biased labels [1, 10].

Conclusion. We presented FAIR-Health, a hybrid pre- and in-processing bias-mitigation framework for equitable clinical readmission prediction, and a leakage-free, multi-metric, ten-seed benchmark against strong baselines. On the UCI Diabetes cohort, FAIR-Health reduced the equalized-odds difference by up to 66% at negligible within-family utility cost while remaining a single, attribute-blind, decision-useful model, closing a real gap in the true-positive rate between racial groups from 0.11 to about 0.02. The benchmark also surfaced two practitioner-relevant cautions: the decision collapse of hard-constraint methods under class imbalance, and the primacy of the in-processing penalty over reweighing. Code and a reproducibility artifact are released to support external scrutiny and extension.

4 Acknowledgments

The dataset analysed is publicly available from the UCI Machine Learning Repository (Diabetes 130-US Hospitals, id 296, `https://doi.org/10.24432/C5230J`) under a CC BY 4.0 licence, and required no additional ethical approval. An artificial intelligence assistant was used to support code implementation and experiment execution; the author verified all methods, results, and claims and takes full responsibility for the content. The author declares no competing interests and received no specific funding for this work.

References

[1] Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science* 366 (6464): 447–453.

[2] Seyyed-Kalantari, Laleh, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen, and Marzyeh Ghassemi. 2021. "Underdiagnosis Bias of Artificial Intelligence Algorithms Applied to Chest Radiographs in Under-Served Patient Populations." *Nature Medicine* 27 (12): 2176–2182.

[3] Gichoya, Judy Wawira, Imon Banerjee, Ananth Reddy Bhimireddy, John L. Burns, Leo Anthony Celi, Li-Ching Chen, et al. 2022. "AI Recognition of Patient Race in Medical Imaging: A Modelling Study." *The Lancet Digital Health* 4 (6): e406–e414.

[4] Rajkomar, Alvin, Michaela Hardt, Michael D. Howell, Greg Corrado, and Marshall H. Chin. 2018. "Ensuring Fairness in Machine Learning to Advance Health Equity." *Annals of Internal Medicine* 169 (12): 866–872.

[5] Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. “A Survey on Bias and Fairness in Machine Learning.” *ACM Computing Surveys* 54 (6): 1–35.

[6] Chouldechova, Alexandra. 2017. “Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments.” *Big Data* 5 (2): 153–163.

[7] Hardt, Moritz, Eric Price, and Nathan Srebro. 2016. “Equality of Opportunity in Supervised Learning.” In *Advances in Neural Information Processing Systems*, 29: 3315–3323.

[8] Verma, Sahil, and Julia Rubin. 2018. “Fairness Definitions Explained.” In *Proceedings of the International Workshop on Software Fairness (FairWare)*, 1–7. New York: ACM.

[9] Barocas, Solon, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. Cambridge, MA: MIT Press.

[10] Corbett-Davies, Sam, Johann D. Gaebler, Hamed Nilforoshan, Ravi Shroff, and Sharad Goel. 2023. “The Measure and Mismeasure of Fairness.” *Journal of Machine Learning Research* 24 (312): 1–117.

[11] Strack, Beata, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore. 2014. “Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records.” *BioMed Research International* 2014: 781670.

[12] Kamiran, Faisal, and Toon Calders. 2012. “Data Preprocessing Techniques for Classification Without Discrimination.” *Knowledge and Information Systems* 33 (1): 1–33.

[13] Zhang, Brian Hu, Blake Lemoine, and Margaret Mitchell. 2018. “Mitigating Unwanted Biases with Adversarial Learning.” In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 335–340. New York: ACM.

[14] Agarwal, Alekh, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. “A Reductions Approach to Fair Classification.” In *Proceedings of the 35th International Conference on Machine Learning*, PMLR 80: 60–69.

[15] Feldman, Michael, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. “Certifying and Removing Disparate Impact.” In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268. New York: ACM.

[16] Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, et al. 2011. “Scikit-learn: Machine Learning in Python.” *Journal of Machine Learning Research* 12: 2825–2830.

[17] Bird, Sarah, Miroslav Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. *Fairlearn: A Toolkit for Assessing and Improving Fairness in AI.* Technical Report MSR-TR-2020-32. Redmond, WA: Microsoft.

[18] Collins, Gary S., Karel G. M. Moons, Paula Dhiman, Richard D. Riley, Andrew L. Beam, Ben Van Calster, et al. 2024. "TRIPOD+AI Statement: Updated Guidance for Reporting Clinical Prediction Models That Use Regression or Machine Learning Methods." *BMJ* 385: e078378.

Information about author:

Arulmurugan Ramu – PhD, Associate Professor, Heriot-Watt University International Faculty, K. Zhubanov Aktobe Regional University (Aktobe, Kazakhstan, email: arulmr@gmail.com).

Автор туралы мәлімет:

Арулмуруган Раму – PhD, қауымдастырылған профессор, Хериот-Уатт университетінің халықаралық факультеті, Қ. Жұбанов атындағы Ақтөбе өңірлік университеті (Ақтөбе, Қазақстан, электрондық пошта: arulmr@gmail.com).

Информация об авторе:

Арулмуруган Раму – PhD, ассоциированный профессор, Международный факультет Университета Хериот-Ватт, Актюбинский региональный университет им. К. Жубанова (Актобе, Казахстан, электронная почта: arulmr@gmail.com).

Revised: September 1, 2026
Accepted: November 20, 2026