Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

fairGATE: traceable least-well-served group fairness correction of prediction models for binary and continuous outcomes

Rhys Holland
rhys.2.holland@kcl.ac.uk
King's College London
Raquel Iniesta
King's College London
Lei Luo
King's College London
Diana Shamsutdinova
King's College London

Method Article

Keywords: Algorithmic fairness, Machine learning, Fairness in prediction models, Clinical prediction models, Health data science, Biostatistics, Mixture of experts, Interpretable and auditable machine learning, Gradient boosting, R software development

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11811415/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

fairGATE: traceable least-well-served group fairness correction of prediction models for binary and continuous outcomes

Rhys Holland*[1,2], Lei Luo[2], Diana Shamsutdinova[2], Raquel Iniesta[2,3]

*Correspondence to Rhys Holland (rhys.2.holland@kcl.ac.uk).

[1] *Pharos AI, London, UK*

[2] *Department of Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK*

[3] *King's Institute for Artificial Intelligence, King's College London, London, UK*

Abstract

Models that inform decisions in healthcare, finance and public services should be accurate and fair across population subgroups, yet those fitted to historical records are often accurate but unfair. We present fairGATE, an in-processing fairness correction for a bias-naive prediction model. Available as an R package, it targets the error-rate gap of the least-well-served group rather than a group average, records every adjustment, and supports binary and continuous outcomes with any number of groups or intersections. Under a pre-specified design, we evaluated it on eleven held-out tasks against two fairness-aware tree ensembles. On all five census tasks with race as the protected attribute, fairGATE reduced worst-group disparity by 33% to 61% relative to the fairer ensemble, achieved higher accuracy and changed about 6% of binary decisions. On five further tasks, the fair methods did not differ significantly; on one, an ensemble was fairer but less accurate.

Main

Prediction models increasingly inform decisions about follow-up after hospital discharge and credit approval, as well as income estimates used to screen people for benefit eligibility [1–3]. In each setting, models must be accurate and treat subgroups defined by a protected attribute, such as race or sex, comparably. Models fitted to historical records can meet the first requirement while failing the second [2,4,5]. We address the problem of correcting such models while retaining their accuracy and allowing users to inspect and justify the correction.

Fairness-aware learning comprises pre-processing, in-processing and post-processing methods [6–10]. Pre-processing and reduction methods reweight or relabel training data and refit the model [7,11,12]. Post-processing adjusts the scores or thresholds of a fitted model [10,13]. Group-specific thresholds require the protected attribute at decision time and, applied to a well-tuned model, can reach the fairness–accuracy frontier of in-processing methods [14]. We focus on in-processing methods, which correct group performance during model fitting without altering the training data or decision rule. For tabular data, gradient-boosted trees such as XGBoost and LightGBM [15,16] are standard accuracy benchmarks. Our comparators, fairness-aware tree ensembles, are the most direct in-processing alternatives: Fair Forests adds a protected-attribute term to each split criterion [17], whereas FairGBM trains a boosted ensemble under a fairness constraint [18]. The closest approaches learn an additive correction to a frozen model under a fairness penalty, framed as post-processing [19], or train a gated mixture of experts with fairness penalties [20]. Like the former, fairGATE corrects a fixed base model. It learns the correction through a gated mixture of experts on cross-fitted base scores and penalises the worst-treated group's gap. Fair regression has a separate literature [21–24], so continuous targets often require a different method and protocol.

Formal definitions of fairness cannot generally be satisfied simultaneously [5,25], so the choice depends on the domain. Throughout this paper, fairness means separation [3,10]: among people with the same observed outcome, predictions should not depend on group. For binary outcomes, this requires equal true- and false-positive rates across groups, summarised by the average odds difference (AOD). For continuous outcomes, its counterpart is the conditional mean prediction difference (CMPD): a group's gap in mean prediction relative to the reference group among people with similar outcomes. Separation is a natural target when the recorded outcome is accepted and the concern is unequal errors, for example when applicants who would repay are refused credit at different rates across

groups. It differs from demographic parity, which requires equal predictions irrespective of outcomes.

A fairness-aware decision model should meet three requirements. First, multi-group coverage: fairness must be assessed at every level of each protected attribute and at their intersections, because these attributes often have several levels and people belong to several groups [26,27]. A pooled summary of all non-reference groups can decrease while the disparity for the worst-treated group, the group least well served by the model, remains unchanged. Minimax and distributionally robust formulations minimise the worst group's loss or error rate [28,29]; here, the concern is its gap in error rates relative to the reference group. Second, traceability: users justifying a correction need to know which decisions changed, by how much, and whether fewer changes would have sufficed; post hoc explanation is a weak substitute [30,31]. Third, protocol-controlled evaluation: methods must be compared using identical partitions and selection rules, against a bias-naive baseline fitted under the same protocol, with design choices tuned on data separate from those used to evaluate success.

This version of fairGATE follows two of our earlier designs: a gated mixture of neural experts [32–34] with a penalty on the variance of outcome-stratified group means, followed by the addition of an independently fitted gradient-boosting base learner, a random-effects (DerSimonian–Laird) penalty [35] and an exported per-person adjustment. In the second design's evaluation, Fair Forests had smaller pooled disparity on census income by race, whereas our design was fairer to the worst-treated group. This finding motivated the present worst-group target. Because those data informed this choice, the method is tested here only on individuals outside the development data.

The present design retains the independently fitted base learner (XGBoost, or a neural network) and the base-plus-adjustment decomposition, with three changes. The objective can directly target the worst-treated group's error-rate gaps at the decision threshold, and a tuned intervention cost balances disparity reduction against the number of decisions changed. Models are selected and evaluated using worst-group disparity: the largest

absolute gap of any sufficiently represented group from the reference, for binary and continuous outcomes and intersectional cells. Each person's adjustment is exported with an uncertainty interval and routing weights; the protected attribute is not an input to any component at prediction time.

We evaluate fairGATE on eleven prediction tasks, each a dataset with one outcome and one protected attribute, drawn from census data, hospital follow-up data and simulated clinical samples not used to develop the method. We call these confirmation tasks. The method, search space, selection rule and success criteria were frozen using development data before any confirmation task was analysed. Every method, including a bias-naive XGBoost baseline, uses the same 50 data partitions. Each task has a single pre-specified primary criterion: fairGATE's worst-group disparity must be significantly smaller than that of the fairer of the two ensembles on that task, without a significant loss of accuracy (AUC or R-squared) beyond a tolerance of 0.005.

Results

fairGATE and the confirmation design

fairGATE (Fig. 1) learns a correction to the score of a cross-fitted base model fitted without fairness input (XGBoost here). For each person, a softmax gate weights the base learner and K experts, using the predictors, the base score and a cross-fitted disparity signal, which estimates the person's contribution to their group's error-rate gap. In the residual coupling, the experts are initialised at zero, so the untrained model is the base learner. The objective combines the task loss with a tuned group penalty, an optional thresholded penalty on the worst group's gap in true- and false-positive rates, and an intervention cost on the mean absolute change to the base score; continuous outcomes use squared-error loss and the CMPD as the penalty.

The eleven confirmation tasks (Extended Data Table 1) use data that informed no design decision: four tasks from a new California sample of the 2018 American Community Survey (50,000 people, disjoint by person and household from the development sample), two from a

Texas sample of the same survey (50,000 people), three from the 18,054 eligible Diabetes 130-US hospitals patients outside the development sample and two from a new draw of the simulated coronary cohort (10,000 people). They cover binary and continuous outcomes, two-group and five-group attributes, intersectional race-by-sex cells and transfer to a second state. The base learner's AUC ranged from 0.65 (hospital readmission) to 0.90 (census income), and its R-squared from 0.07 (length of stay) to 0.61 (log income).

Pre-specified primary criterion

The primary criterion was met on five of the eleven confirmation tasks (Fig. 2): all five census tasks with race as the protected attribute. On California income by race, fairGATE's worst-group |AOD| was 0.082 versus 0.190 for Fair Forests, the fairer comparator; on California log income, |CMPD| was 0.071 versus 0.106 (Fair Forests); on California income by race and sex, disparity was 0.143 versus 0.235 (FairGBM); on Texas income by race, 0.071 versus 0.183 (FairGBM); and on Texas log income, 0.054 versus 0.114 (Fair Forests), with $p < 0.001$ on each task. Across these tasks, fairGATE's worst-group disparity was 33% to 61% lower than the fairer comparator's (median 53%) and 40% to 62% lower than the other comparator's. It was lower in all 50 partitions of each task. Its AUC or R-squared was significantly higher than that of both comparators on every task, by 0.004 to 0.026 over the fairer comparator.

On the remaining six tasks, fairGATE's disparity did not differ significantly from the fairer comparator's on five and was larger on one. On California income by sex, the fair methods achieved the same small disparity (fairGATE 0.0071, FairGBM 0.0071; $p = 0.89$), 60% below XGBoost's 0.0178. On the three hospital tasks and the binary simulated task, no difference between fairGATE and the fairer comparator was detected (p from 0.22 to 0.76). On the simulated coronary risk index, Fair Forests was significantly fairer (|CMPD| 0.066 compared with 0.074, $p < 0.001$), but its R-squared was lower by 0.062 (0.355 compared with 0.417). fairGATE's mean accuracy was never below the fairer comparator's on any of the eleven tasks.

Fairness at little cost in accuracy

Relative to the bias-naive XGBoost baseline, fairGATE significantly reduced worst-group disparity on nine of the eleven tasks; the exceptions were the two readmission tasks (Fig. 3 and Extended Data Table 2). On the census race tasks, disparity decreased by 39% to 60%, with a loss of 0.007 to 0.012 in AUC or R-squared. Across all eleven tasks, fairGATE's accuracy remained within 0.023 of XGBoost's (median cost 0.007), whereas the loss reached 0.062 for Fair Forests (median 0.019) and 0.026 for FairGBM (median 0.015). On readmission by race, the frozen selection effectively returned the base learner: 99.98% of adjustments were below 0.01 on the probability scale, and 0.08% of decisions differed from XGBoost's. In the simulated cohort, where bias was introduced into the outcome, the Bayes-optimal reference had greater worst-group disparity than any fitted model (Extended Data Fig. 1).

Minimal intervention

Across the three binary census race tasks, fairGATE changed 6.7%, 5.5% and 6.1% of decisions relative to XGBoost and removed 1.65 to 1.78 units of worst-group |AOD| per unit share of decisions changed, compared with at most 0.13 for Fair Forests and FairGBM (Fig. 4). It changed significantly fewer decisions than Fair Forests on all five census race tasks and fewer than FairGBM on three. The exceptions were a tie on California income by race ($p = 0.60$) and Texas log income, where FairGBM changed slightly fewer predictions ($p = 0.026$). The pre-specified continuous definition, a change of more than 0.05 outcome standard deviations, classified 51% to 85% of predictions as changed for every fair method on the four continuous tasks, providing little discrimination between methods.

Multi-group and intersectional structure

Under XGBoost, the worst-treated census group was 'Some other race' (California AOD −0.201; Texas −0.176). Fair Forests and FairGBM left that gap almost unchanged while increasing the Asian group's gap from 0.023 to 0.106 and 0.065, respectively. This reduces the pooled contrast by offsetting gaps of opposite sign (Fig. 5). fairGATE reduced the 'Some other race' gap to −0.081 (Texas −0.067) and kept the Asian gap at 0.017. Its mean pooled

disparity was also lower than XGBoost's on every census task. However, fairGATE met the pooled version of the criterion on at most one task, and a comparator had significantly smaller pooled disparity on four census tasks (Extended Data Fig. 2).

On California income by race and sex, fairGATE reduced both the worst-cell gap and the race-marginal gap (0.117 compared with 0.206 for FairGBM), but its sex-marginal gap exceeded those of FairGBM (0.023 compared with 0.006, $p < 0.001$) and XGBoost (0.019). On readmission by race, the minimum-cell rule excluded three of the five race groups (Hispanic, Asian and Other), so the worst-group summary includes only the African American group (per-group gaps in Extended Data Fig. 3).

Group differences and sampling noise

Whether tasks met the criterion may reflect the size of the base learner's group differences relative to sampling noise. We quantified this post hoc using the noise floor: the share of the variance-of-means penalty at the base learner expected from sampling noise alone (Methods). On census tasks, the noise floor was 0.8% to 8.8%, and the criterion was met on five of six; on hospital and simulated tasks, it was 29% to 175%, and none met the criterion (Spearman correlation −0.81 over eleven tasks; Fig. 6b). The association is descriptive and confounded with data source: all tasks meeting the criterion were census tasks. It does not explain California income by sex, where the noise floor was low (4.3%) but the small baseline gap allowed both fair methods to reach the same value. On the two readmission tasks, no fair method significantly reduced the base learner's disparity; on length of stay, fairGATE and Fair Forests both approximately halved it (0.024 and 0.023 compared with 0.051).

Mechanism and traceability

Component ablations were refitted at fairGATE's selected configuration on every confirmation task (Fig. 6a). Removing the fairness penalties significantly increased worst-group disparity on nine of the eleven tasks; these penalties cost at most 0.020 in accuracy (length of stay by race). Replacing or removing the thresholded penalty increased disparity on three of the six tasks where it was selected, and the disparity signal significantly reduced

disparity on one (simulated coronary event by race). Three components increased disparity on some tasks: the intervention cost on four of ten (California income by race 0.082 with the cost compared with 0.063 without, while changing fewer decisions), smoothed early stopping on three of eleven, and residual coupling on one (0.143 compared with 0.129 with expert coupling). Development ablations are in Extended Data Fig. 4.

The routing safeguard held on twenty-four of the twenty-nine tasks: a linear probe recovered the protected attribute less accurately from the routing weights than from the predictors (Fig. 6c and Extended Data Fig. 5). The safeguard failed on the four simulated coronary tasks and GENDEP remission by sex (coronary event: routing AUC 0.610 compared with 0.553 from the predictors).

For every held-out person, the user receives the base prediction, fairGATE's adjustment, its interval, the weight on the base learner and the routing vector (Fig. 1c). One package call writes these to a per-person audit file (four rows in Supplementary Table 1). The exact base-plus-adjustment decomposition identifies which decisions changed, in which direction, for whom and by how much. On California income by race (Fig. 6d), adjustments were centred on zero for four of the five groups and positive for the 'Some other race' group, the worst-treated group under XGBoost. Nearly all changed decisions in this group crossed the threshold upwards. The mean base weight ranged from 0.08 to 0.46 across confirmation tasks, and the interval excluded zero for 30% to 79% of held-out people.

Calibration within groups

In a post hoc analysis of held-out predictions (Methods), the correction traded within-group calibration for separation on census tasks, as expected when base rates differ [36,37]. Relative to XGBoost, every fair method increased the largest within-group gap between observed and mean predicted outcome on all six census tasks. On California income by race, fairGATE's mean predicted risk for 'Some other race' exceeded the observed rate by 0.070, compared with 0.014 for XGBoost and 0.054 for Fair Forests (Supplementary Fig. 1 and Supplementary Table 2). fairGATE's worst-group calibration gap was larger than the

fairer comparator's on four of the five census race tasks and smaller on the other six tasks. Its overall calibration slope was between 0.93 and 1.06 on ten tasks (1.55 on length of stay), whereas both ensembles' slopes exceeded 1 on every task, indicating compressed predictions.

Choice of base learner

In an exploratory analysis, fairGATE was refitted on a cross-fitted neural-network base with the same partitions, search space and selection rule (Methods and Extended Data Fig. 6). The criterion was met on the same five census race tasks, and worst-group disparity after correction did not differ between the two bases on any task (mean differences −0.0052 to 0.0124). Accuracy reflected the choice of base learner: on census tasks, the bias-naive network's AUC or R-squared was 0.004 to 0.017 below that of XGBoost, the package default. The per-person export is identical for either base. A network base could also track learning from each group during training through per-sample gradients [38] or each group's training loss [39].

Discussion

We tested fairGATE under a single frozen protocol across binary and continuous outcomes, two-group, five-group and intersectional attributes, and census, hospital follow-up and simulated clinical data. On every census task with race as the protected attribute, including transfer to a second state and race-by-sex cells, worst-group disparity was 33% to 61% lower than that of the fairer ensemble. Accuracy exceeded that of both ensembles and remained within 0.007 to 0.012 of the bias-naive XGBoost baseline. On these binary census tasks, fairGATE changed about 6% of decisions, removing more than ten times as much worst-group disparity per changed decision as either ensemble.

Disparity did not differ significantly between fair methods on the other six tasks, except on the simulated risk index, where Fair Forests was fairer but had an R-squared lower by 0.062. Group differences in the base learner were small relative to sampling noise on hospital and simulated tasks. The base learner's disparity was already small on the two readmission

tasks, the most difficult binary tasks in the suite, and on California income by sex. fairGATE was never less accurate than the fairer comparator on these tasks and left the base learner almost unchanged on readmission by race. Although the association with the noise floor is descriptive, the noise floor could serve as a preliminary check: it can be computed from the base learner's held-out predictions before fitting a correction.

fairGATE optimises and reports separation for the worst-treated group. A pooled summary can instead be reduced by widening the positive gap of a favourably treated group while the worst negative gap remains nearly unchanged, as the ensembles did on the binary census race tasks. An ensemble achieved significantly smaller pooled disparity on four census tasks (Fair Forests had the smallest mean on three). Users whose obligation is a pooled parity statistic should therefore prefer one of the ensembles on these data. When a hard constraint on a specific error rate is required, constrained boosting, such as FairGBM, imposes it directly [18]. When group-specific thresholds are permissible, post-processing targets the chosen error rates [10,14] and is simpler to audit. None of the methods, as selected here, targets demographic parity.

Gains over the fairer ensemble were demonstrated only on census tasks with race as the protected attribute, not on hospital or simulated data or tasks whose group differences lie within sampling noise. The minimal-intervention claim applies only to binary outcomes. fairGATE is most useful when one method must support binary and continuous outcomes, the protected attribute has several levels or intersections, accuracy close to gradient boosting is required, and every change to a decision must be recorded and explained.

Several limitations remain. Per-person intervals derive from ten warm-started bootstrap refits and are uncalibrated; they rank adjustments by stability but do not provide coverage. The routing safeguard, based on a linear probe, failed on the simulated coronary cohort and GENDEP remission by sex; unawareness of the attribute is not evidence of non-use [40]. The minimum-cell rule excludes groups too small for disparity estimation from the worst-group summary, including three of five groups on readmission by race. Penalising

intersectional cells also did not bound the sex-marginal gap on the race-by-sex task. The California confirmation pool is not a random subset of the survey frame, no correction for multiplicity was pre-specified (although all five successes remain significant after Bonferroni correction over the eleven tasks), and the ensembles as implemented use a binarised attribute. Separation measures unequal errors against the recorded outcome but does not address whether that outcome is itself biased, as the simulated cohort's Bayes-optimal reference illustrates [2].

These findings suggest three directions for further work. First, separation and within-group calibration cannot generally hold simultaneously when base rates differ [36,37], as the census results illustrate. Because fairGATE exports the base prediction and adjustment separately, the correction's effect on each group's calibration can be measured [41], and a calibration term could be added to the objective when calibrated group-level risks are needed. Second, the modular group penalty could accommodate other definitions, such as demographic parity or bounded group loss [21], when the fairness obligation differs. Third, comparisons should extend beyond the two in-processing ensembles to approaches that mitigate bias through group reweighting [7,12], a worst-group weighted loss [29] or group-specific thresholds on the same base model [10,14]. The pre-specified protocol should also be repeated on lower-signal tasks and in additional domains, reporting the noise floor alongside the criterion to establish where the gains persist.

Methods

Study design

The fairGATE method, search space, selection rule for every fairness-aware method, minimum-cell rule and success criteria were fixed in a design document before fitting any confirmation model or computing any confirmation result. This document is deposited with the analysis code. Design decisions used only development data: the inner validation rows of partition 1 of the development tasks and, for one descriptive check, outer partitions 1 to 5 of

those tasks. The confirmatory claim rests on eleven tasks that informed no design decision; all other task results are development results and are reported descriptively.

Datasets, tasks and roles

A task consists of a dataset with one outcome and one protected attribute. The protected attribute was removed from the feature matrix and used only in the fairness objective, the target of the cross-fitted disparity signal and evaluation. The reference is the privileged group in the fairness metrics. Group sizes, outcome rates and roles are reported in Extended Data Table 1. Supplementary Notes 1 to 5 describe the derivation of confirmation samples and development tasks built for this study; derivations of the remaining development samples are deposited with the analysis code.

Confirmation samples. Four samples were built for this study, each disjoint from previously used rows. ACS California: applying the ACSIncome inclusion rule [42] to the 2018 American Community Survey 1-year person file for California [43] yielded 195,665 people. At a fixed seed, 50,000 were drawn from the 106,681 with no person or household overlap with the development sample (an earlier draw, disjoint by person only, was replaced before any analysis). ACS Texas: 50,000 of the 135,924 people meeting the same rule were drawn at the same seed, with columns aligned to California by name. Diabetes: all 18,054 eligible first encounters in the Diabetes 130-US hospitals data [44,45] outside the development sample were included unsampled, with feature encodings frozen at development values. Simulated coronary cohort: 10,000 individuals were drawn from the unchanged development generator at a new seed, with no identifier or predictor record shared with the development cohort. The bias mechanisms (a group offset in outcome log-odds, reduced prescription of preventive drugs and adverse shifts in social and clinical risk factors) follow documented disparities [46–54]; their magnitudes were not calibrated to any published estimate.

Confirmation tasks. California: income above USD 50,000 by race in five groups (White as reference), log income by race, income by sex (men as reference) and income by race-by-sex cell (ten cells, White men as reference). Texas: income and log income by race.

Diabetes: readmission within 30 days and length of stay by race in the five source categories (Caucasian as reference), and readmission by race-by-sex cell. Simulated cohort: a coronary care event and a continuous risk index by race in five groups. Intersectional cells followed a rule fixed solely on cell counts: a cell with fewer than 150 people or fewer than 30 in the smaller outcome class was collapsed over sex and, if still too small, pooled with the next smallest collapsed level. Development tasks and their confirmation counterparts used the same cell structure. All ten census cells were retained; the hospital data yielded five cells, with Hispanic, Asian and Other patients pooled.

Development tasks. Eighteen tasks form the development data: census income by race and by sex, log income by race, credit default by sex and by age band, readmission and length of stay by race, the simulated coronary event and risk index, GENDEP remission by age, sex and marital status [55,56], three Adult tasks [57,58] and three intersectional tasks (census income and readmission by race and sex; credit default by sex and age [59,60]). Credit and GENDEP have no unused rows. Adult is reported solely for comparability (Supplementary Fig. 2), following the recommendation to retire it as a fairness benchmark [42].

Ethics

The authors do not declare any ethical concerns.

Partitions and preprocessing

Each task was evaluated on 50 stratified 80:20 partitions, generated once at a fixed seed and used by every runner. Stratification crossed the outcome (or its equal-frequency quintile) with the group code. Hyperparameters were selected using an inner 75:25 partition of partition 1's training rows, so no outer held-out individual informed selection. Imputation and standardisation were computed from each partition's training rows.

The fairGATE model

Let x_i denote the features of person i, g_i the protected group, y_i the outcome and b_i the link-scale score of the cross-fitted base learner. In this study, the base learner was XGBoost [15], using the configuration selected for the task's XGBoost comparator (the package also offers

a neural-network base learner; see below). Training rows receive 5-fold out-of-fold scores; test rows receive predictions from a fit on all training rows, identical to the comparator's [61,62]. K experts f_k (two hidden layers with batch normalisation, rectified linear units and dropout) receive the standardised features and b_i. A gate network h over the features, b_i and the disparity signal s_i produces $K + 1$ weights through a temperature-scaled softmax,

$$\pi_{ik} = \frac{\exp\{h_k(x_i, b_i, s_i)/\tau\}}{\sum_{j=0}^{K} \exp\{h_j(x_i, b_i, s_i)/\tau\}}, \quad k = 0, \ldots, K, \tag{1}$$

with temperature τ, where component 0 denotes the base learner. The search covers residual and expert couplings:

$$\eta_i = b_i + \sum_{k=1}^{K} \pi_{ik} f_k(x_i, b_i), \tag{2}$$

$$\eta_i = \pi_{i0} b_i + \sum_{k=1}^{K} \pi_{ik} f_k(x_i, b_i). \tag{3}$$

In the residual coupling, the experts' last layers are initialised at zero, so the untrained model is exactly the base learner. The prediction $\hat{y}_i$ is $\sigma(\eta_i)$ for a binary outcome, with σ denoting the logistic function, and the back-transformed η_i for a continuous outcome. The adjustment is $\hat{y}_i$ minus the base learner's prediction. The protected attribute is not an input to the base learner, the gate or the experts.

Disparity signal. A target is computed for each training individual. For a binary outcome, it is the signed gap in positive-decision rate between the person's group and the reference group within their outcome class at the prevalence-matched threshold of the out-of-fold base score and weighted by a Gaussian kernel of bandwidth 1 on the logit scale around the threshold. For a continuous outcome, it is the stratum-conditional mean gap in outcome standard deviations. An XGBoost model (200 trees, depth 4, learning rate 0.05) predicts this target from the features and b_i, using 5-fold out-of-fold predictions for training rows and a fit on all training rows for test rows. Thus, s_i uses the attribute only through a cross-fitted regression, as b_i uses the outcome.

Objective. Parameters are estimated by minimising

$$L = L_{\text{task}} + \lambda P_{\text{type}} + \lambda_{\text{thr}} T + \mu \frac{1}{n} \sum_{i=1}^{n} |\eta_i - b_i|, \quad (4)$$

with L_{task} denoting cross-entropy for a binary outcome and squared error on the standardised outcome for a continuous one. P_{type} is one of three group penalties, selected per task: the DerSimonian and Laird between-group variance component of outcome-stratified mean predictions [35,63], the variance of the cell means, or T. For a binary outcome

$$T = \max_{g \neq r} 1/2 \left(\Delta\text{TPR}_g^2 + \Delta\text{FPR}_g^2 \right), \quad (5)$$

where ΔTPR_g and ΔFPR_g are the differences between group g and the reference group r in true- and false-positive rates, computed using soft decisions $\sigma\{(p_i - t)/0.05\}$, with p_i denoting the predicted probability and t denoting the batch quantile at one minus the prevalence of the fitted rows (detached from the gradient); for a continuous outcome T is the worst group's squared conditional mean prediction difference. The last term of equation (4) is the intervention cost, weighted by μ. Variance-component cells require two rows; for T, a binary group requires two rows in each outcome class.

Training and intervals. Training used full-batch Adam [64], with early stopping on the monitored objective evaluated on a stratified 15% subset of each partition's training rows. The objective was smoothed using a trailing moving average over 5 epochs, with the earliest selectable epoch 10, patience 30 and a cap of 300 epochs; the parameters of the best window were restored. Adjustment intervals are the 5th and 95th percentiles of each held-out person's adjustment over 10 warm-started bootstrap replicates of 10 epochs each on a resample of the fitted rows; point predictions are unchanged. This approximates full bootstrap refits within the time budget; the intervals have not been calibrated.

Search space and selection rule

fairGATE drew 48 distinct random configurations from a generator with a fixed seed and algorithm, over learning rate {0.001, 0.002}, hidden width {32, 64, 128}, dropout {0.1, 0.2}, gate temperature τ {0.5, 1.0}, λ {1, 4, 16, 64}, penalty type {random effects, variance, thresholded}, λ_{thr} {0, 4, 16} (set to 0 when the penalty is thresholded), intervention cost μ {0,

0.02, 0.1} and coupling {residual, expert}. The same rule selects fairGATE, Fair Forests and FairGBM: among configurations with inner-validation AUC or R-squared within 0.01 of that method's best, we select the configuration with the smallest worst-group |AOD| (binary, at the prevalence-matched threshold) or |CMPD| (continuous), subject to the minimum-cell rule. XGBoost uses its best validation performance. Selected confirmation configurations are listed in Extended Data Table 3.

Comparators and references

XGBoost [15], the bias-naive baseline and the base learner, receives no protected-attribute information. **Fair Forests** [17] is a bagged ensemble of 200 trees that scores each split by its normalised label-purity gain minus a weighted normalised protected-attribute-purity gain. **FairGBM** [18] was reimplemented on LightGBM [16] through its custom-objective interface, using the published proxy-Lagrangian game for equality of opportunity. Both were implemented from published descriptions because no reference implementation ran on our platform. Both use a binarised attribute (reference against the rest), and their regression variants are our extensions. Each was tuned over 24 seeded random configurations. **Ablations** were refitted at fairGATE's selected configuration with the same seeds on all 50 partitions, each disabling one component: the fairness penalties ($\lambda = \lambda_{\text{thr}} = 0$), the thresholded penalty (replaced by the random-effects penalty, or the added term removed), the disparity signal, smoothed early stopping, the intervention cost, the tuned penalty type (fixed at random effects) and the residual coupling (replaced by the expert coupling). An ablation identical to fairGATE at the selected configuration was not run and is reported as not applicable. **Bayes-optimal reference.** For the simulated cohort, the true event probability and the rescaled latent predictor were scored by the same code; they quantify the disparity available for removal and do not define a fairness target.

Neural-network base learner (exploratory). The base learner was replaced by a cross-fitted multilayer perceptron, with all other components unchanged. The network was tuned once per task on the inner split of partition 1 by 24 seeded random draws from the package's

default grid (layers, width, dropout, learning rate, weight decay, batch size; epochs by early stopping) and cross-fitted in every partition as for the XGBoost base. The bias-naive network is evaluated using the same base scores. The disparity signal was recomputed from the network scores, and fairGATE was tuned, selected and refitted with the frozen search space, selection rule, partitions and scorer. These fits used fairGATE 1.0.0 with torch 0.17.0 on R 4.5.3. This analysis is exploratory because the confirmation tasks had already been used.

Evaluation

Performance is AUC for binary outcomes or R-squared for continuous outcomes. Binary fairness metrics follow the AI Fairness 360 definitions [65] at the prevalence-matched threshold, which classifies a share of held-out people equal to the training prevalence as positive, using no held-out labels. The average odds difference (AOD) of group g is the mean of its true- and false-positive rate differences from the reference group r,

$$\mathrm{AOD}_g = 1/2\left[(\mathrm{TPR}_g - \mathrm{TPR}_r) + (\mathrm{FPR}_g - \mathrm{FPR}_r)\right]. \quad (6)$$

For continuous outcomes, the conditional mean prediction difference (CMPD) averages the between-group difference in mean prediction across outcome quintiles, expressed in outcome standard deviations,

$$\mathrm{CMPD}_g = \frac{1}{Q}\sum_{q=1}^{Q} \frac{\bar{y}_{gq}^{\mathrm{pred}} - \bar{y}_{rq}^{\mathrm{pred}}}{s_y}, \quad (7)$$

where $\bar{y}_{gq}^{\mathrm{pred}}$ is the mean prediction for group g in band q of the observed outcome, s_y is the outcome standard deviation, and $Q = 5$ equal-frequency bands are used. The worst-group disparity, $\max_g|\mathrm{AOD}_g|$ or $\max_g|\mathrm{CMPD}_g|$, is the largest absolute gap of any non-reference group against the reference among groups that pass the minimum-cell rule: at least 10 expected positives and 10 expected negatives per test set for binary tasks, and at least 50 expected people per test set for continuous tasks. Groups failing this rule are reported individually and excluded from the maximum for every method. The pooled disparity contrasts all non-reference groups with the reference; the joint disparity is the size-weighted mean absolute gap over eligible groups. For intersectional tasks, the worst gap for each

marginal attribute is also reported. A decision is changed when it differs from XGBoost's at each model's own threshold (binary) or when the prediction moves by more than 0.05 outcome standard deviations (continuous); disparity removed per changed decision is the reduction in worst-group disparity relative to XGBoost divided by the proportion changed.

Choice of outcome bands for CMPD. CMPD estimates separation [3,10] for a continuous outcome by comparing groups within bands of the observed outcome. A raw difference in mean predictions (demographic parity) would require equal predictions for groups whose outcomes differ, whereas bounded group loss [21] equalises error magnitude rather than direction. Five equal-frequency bands, with boundaries fixed once per task using the whole cohort, were chosen by convention before any results. These bands removed 88% to 90% of the between-group difference in the observed outcome on the three untied confirmation outcomes, close to the 90% expected for five subclasses of a normally distributed variable [66,67], while retaining about 10 expected people per group-by-band cell for the smallest eligible group. In a post hoc sensitivity analysis, using 2, 3, 4, 6, 8, 10 or 20 bands, or replacing bands with a natural cubic spline of the outcome, did not change any primary-criterion outcome (Extended Data Fig. 7).

Routing safeguard

Within each partition, held-out rows were randomly halved, and a logistic model predicting group membership from the routing weights was fitted on one half and scored on the other (AUC, or mean one-versus-rest AUC). Results are reported alongside the same probe fitted to standardised predictors [68]. An AUC below the predictor probe bounds leakage only against a linear probe.

Noise floor

The noise floor is the part of the variance-of-means penalty expected from sampling noise alone (the mean of the cell sampling variances), divided by that penalty at XGBoost's held-

out predictions; it was computed descriptively and examined against the criterion outcome post hoc.

Calibration (post hoc)

Calibration was assessed post hoc from saved held-out predictions for the reference group and every group passing the minimum-cell rule, without refitting models. Calibration-in-the-large is the observed minus the mean predicted outcome (probability scale for binary outcomes, outcome standard deviations for continuous ones), so negative values indicate over-prediction; the worst-group value is its largest absolute value over groups. The calibration slope is the coefficient of the logit of the predicted probability in a logistic regression of the outcome (binary) or of the prediction in a linear regression (continuous); values above 1 indicate predictions compressed towards the mean. Values are means over the 50 partitions, and fairGATE was compared with XGBoost and the fairer comparator by paired Wilcoxon tests on the worst-group value.

Uncertainty, comparisons and pre-specified criteria

Point estimates are means over the 50 partitions. Within each partition, held-out rows were resampled 200 times, and intervals are the 2.5th and 97.5th percentiles of the pooled values. These are resampling intervals for a single cohort, not population confidence intervals. Comparisons are paired across partitions using two-sided Wilcoxon signed-rank tests [69] on per-partition absolute disparities or signed performance.

For each confirmation task, the primary criterion requires fairGATE to have smaller worst-group disparity than the fairer of Fair Forests and FairGBM (the comparator with the smaller mean worst-group disparity), with paired Wilcoxon $p < 0.05$ and a lower mean for fairGATE. Its performance must either not be significantly lower than that comparator's or be lower by less than 0.005. Secondary quantities in the frozen design document are reported for every model, and every task is assessed against the criterion. No adjustment for multiplicity was pre-specified, and none is applied to the reported p values. Choices not specified in the frozen text were fixed in code before any confirmation outcome was read: a group with an

undefined gap in a partition is excluded from that partition's maximum, and joint disparity weights groups by test-set counts. Two sensitivity analyses are reported: the criterion at a fixed threshold of 0.50 and the pooled-contrast version of the criterion.

Development of the design

The method was developed using development data only. Following the two earlier designs described in the introduction, candidate components were compared in a pilot using the inner validation rows of twelve development tasks, with one component disabled at a time. The retained components and the four excluded components, reported as negative results, are described in Supplementary Note 6 (Supplementary Fig. 3). On the development tasks, the frozen design met the criterion on seven of seventeen testable tasks (Supplementary Table 3).

Software and reproducibility

fairGATE is implemented in the open-source fairGATE R package (version 1.0.0, CRAN [70]), using torch for R 0.17.0 with one thread; its interface, defaults and a worked example are provided in Supplementary Note 7. The main analysis used R 4.4.3 and the development package subsequently released as version 1.0.0. With the same base scores, the released version exactly reproduces the main-analysis fits on six test fixtures (identical predictions, adjustments, routing weights and stopping epochs), and its one-call function reproduces the analysis runner to within 6×10^{-17}; release changes are confined to a result-summary function and a penalty-helper code path unused during training. Comparators, base scores, the disparity signal and scoring use Python 3.11.16 with NumPy 2.4.6, pandas 3.0.6, SciPy 1.17.1, scikit-learn 1.9.1, XGBoost 3.4.2 and LightGBM 4.7.0. The package tests verify exact equality of the untrained residual coupling and the base learner. On a laptop with one torch thread, a fairGATE fit with intervals took about 24 s per partition at the configuration selected on Adult. With its own XGBoost base learner and the folds used here, the package reproduces the Python base scores of this study to within 10^{-15}.

Use of large language models

Claude Opus 5.5 (Anthropic) was used to draft and copy-edit text, generate the LaTeX for equations, debug analysis and figure code, and search for references. The authors fully designed the study and the method, specified every analysis, manually reviewed all code and text, and checked each reported number against the deposited outputs; they take full responsibility for this paper.

Data availability

The American Community Survey 2018 1-year Public Use Microdata Sample person files for California and Texas are publicly available from the US Census Bureau [43] and were read under the ACSIncome task definition [42]. The Diabetes 130-US hospitals dataset [45], the default of credit card clients dataset [60] and the Adult dataset [57] are publicly available from the UCI Machine Learning Repository. The confirmation samples and development samples are regenerated exactly by the build scripts listed under Code availability at their fixed seeds, and the person and household identifiers of every sample, together with the checks of disjointness from the development data, are deposited with the analysis code [71].

The simulated coronary cohorts contain no real individuals and regenerate exactly from the simulation script at their fixed seeds; the generated files are also deposited.

GENDEP data are not publicly available. They are identifiable clinical trial records from participants who did not consent to open redistribution, and they remain under the control of the GENDEP investigators. Access requires a data access agreement with the GENDEP investigators and approval of the proposed analysis; requests should be directed to the corresponding author, who will forward them. The per-partition held-out predictions, group labels and outcomes needed to reproduce every GENDEP result are deposited in a form that contains no individual-level predictor values.

The complete result tables underlying every figure and table (per-partition metrics, pooled metrics with resampling intervals, paired contrasts, criteria, ablation and mechanism tables),

all held-out predictions and per-individual attribution files for every model and task, the frozen design document and the audit table are deposited with the analysis code.

Code availability

fairGATE is implemented in the open-source fairGATE R package (MIT licence). Version 1.0.0, used here, is available from CRAN at https://CRAN.R-project.org/package=fairGATE and is citable as https://doi.org/10.32614/CRAN.package.fairGATE [70]; the source is developed at https://github.com/rhysholland/FairGATE. The analysis code is archived at Zenodo (https://doi.org/10.5281/zenodo.23804615) [71]: the cohort build scripts, the partition files, the runners for every model (including the cross-fitted base learner and the disparity signal), the scoring, criteria, ablation, calibration and mechanism code, the figure code and the test suite. It includes our implementations of Fair Forests and FairGBM, written from the published descriptions because no reference implementation ran on our platform, and their regression variants, which are our extensions and not part of either published method.

References

1. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J. AI in health and medicine. Nat. Med. 28, 31–38 (2022).
2. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 366, 447–453 (2019).
3. Barocas, S., Hardt, M. & Narayanan, A. Fairness and Machine Learning: Limitations and Opportunities (MIT Press, 2023).
4. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. A survey on bias and fairness in machine learning. ACM Comput. Surv. 54, 1–35 (2021).
5. Chouldechova, A. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big Data 5, 153–163 (2017).
6. Pessach, D. & Shmueli, E. A review on fairness in machine learning. ACM Comput. Surv. 55, 1–44 (2023).
7. Kamiran, F. & Calders, T. Data preprocessing techniques for classification without discrimination. Knowl. Inf. Syst. 33, 1–33 (2012).
8. Zafar, M. B., Valera, I., Gomez Rodriguez, M. & Gummadi, K. P. Fairness beyond disparate treatment and disparate impact: learning classification without disparate mistreatment. In Proc. 26th International Conference on World Wide Web 1171–1180 (International World Wide Web Conferences Steering Committee, 2017).
9. Zhang, B. H., Lemoine, B. & Mitchell, M. Mitigating unwanted biases with adversarial learning. In Proc. 2018 AAAI/ACM Conference on AI, Ethics, and Society 335–340 (ACM, 2018).
10. Hardt, M., Price, E. & Srebro, N. Equality of opportunity in supervised learning. Adv. Neural Inf. Process. Syst. 29 (2016).

11. Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J. & Wallach, H. A reductions approach to fair classification. In Proc. 35th International Conference on Machine Learning 80, 60–69 (PMLR, 2018).
12. Jiang, H. & Nachum, O. Identifying and correcting label bias in machine learning. In Proc. 23rd International Conference on Artificial Intelligence and Statistics 108, 702–712 (PMLR, 2020).
13. Hébert-Johnson, U., Kim, M., Reingold, O. & Rothblum, G. Multicalibration: calibration for the (computationally-identifiable) masses. In Proc. 35th International Conference on Machine Learning 80, 1939–1948 (PMLR, 2018).
14. Cruz, A. F. & Hardt, M. Unprocessing seven years of algorithmic fairness. In Proc. 12th International Conference on Learning Representations (ICLR, 2024).
15. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 785–794 (ACM, 2016).
16. Ke, G. et al. LightGBM: a highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. 30 (2017).
17. Raff, E., Sylvester, J. & Mills, S. Fair forests: regularized tree induction to minimize model bias. In Proc. 2018 AAAI/ACM Conference on AI, Ethics, and Society 243–250 (ACM, 2018).
18. Cruz, A. F. et al. FairGBM: gradient boosting with fairness constraints. In Proc. 11th International Conference on Learning Representations (ICLR, 2023).
19. Tifrea, A. et al. FRAPPÉ: a group fairness framework for post-processing everything. In Proc. 41st International Conference on Machine Learning 235, 48321–48343 (PMLR, 2024).
20. Sharma, S., Henderson, J. & Ghosh, J. FEAMOE: fair, explainable and adaptive mixture of experts. In Proc. 32nd International Joint Conference on Artificial Intelligence 492–500 (IJCAI, 2023).
21. Agarwal, A., Dudík, M. & Wu, Z. S. Fair regression: quantitative definitions and reduction-based algorithms. In Proc. 36th International Conference on Machine Learning 97, 120–129 (PMLR, 2019).
22. Berk, R. et al. A convex framework for fair regression. Preprint at https://arxiv.org/abs/1706.02409 (2017).
23. Komiyama, J., Takeda, A., Honda, J. & Shimao, H. Nonconvex optimization for regression with fairness constraints. In Proc. 35th International Conference on Machine Learning 80, 2737–2746 (PMLR, 2018).
24. Chzhen, E., Denis, C., Hebiri, M., Oneto, L. & Pontil, M. Fair regression with Wasserstein barycenters. Adv. Neural Inf. Process. Syst. 33 (2020).
25. Corbett-Davies, S., Gaebler, J. D., Nilforoshan, H., Shroff, R. & Goel, S. The measure and mismeasure of fairness. J. Mach. Learn. Res. 24, 1–117 (2023).
26. Kearns, M., Neel, S., Roth, A. & Wu, Z. S. Preventing fairness gerrymandering: auditing and learning for subgroup fairness. In Proc. 35th International Conference on Machine Learning 80, 2564–2572 (PMLR, 2018).
27. Foulds, J. R., Islam, R., Keya, K. N. & Pan, S. An intersectional definition of fairness. In Proc. 2020 IEEE 36th International Conference on Data Engineering 1918–1921 (IEEE, 2020).
28. Diana, E., Gill, W., Kearns, M., Kenthapadi, K. & Roth, A. Minimax group fairness: algorithms and experiments. In Proc. 2021 AAAI/ACM Conference on AI, Ethics, and Society 66–76 (ACM, 2021).
29. Sagawa, S., Koh, P. W., Hashimoto, T. B. & Liang, P. Distributionally robust neural networks for group shifts: on the importance of regularization for worst-case generalization. In Proc. 8th International Conference on Learning Representations (ICLR, 2020).
30. Ghassemi, M., Oakden-Rayner, L. & Beam, A. L. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit. Health 3, e745–e750 (2021).
31. Wiens, J. et al. Do no harm: a roadmap for responsible machine learning for health care. Nat. Med. 25, 1337–1340 (2019).
32. Jacobs, R. A., Jordan, M. I., Nowlan, S. J. & Hinton, G. E. Adaptive mixtures of local experts. Neural Comput. 3, 79–87 (1991).

33. Jordan, M. I. & Jacobs, R. A. Hierarchical mixtures of experts and the EM algorithm. Neural Comput. 6, 181–214 (1994).
34. Shazeer, N. et al. Outrageously large neural networks: the sparsely-gated mixture-of-experts layer. In Proc. 5th International Conference on Learning Representations (ICLR, 2017).
35. DerSimonian, R. & Laird, N. Meta-analysis in clinical trials. Control. Clin. Trials 7, 177–188 (1986).
36. Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J. & Weinberger, K. Q. On fairness and calibration. Adv. Neural Inf. Process. Syst. 30 (2017).
37. Kleinberg, J., Mullainathan, S. & Raghavan, M. Inherent trade-offs in the fair determination of risk scores. In Proc. 8th Innovations in Theoretical Computer Science Conference 43:1–43:23 (Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2017).
38. Pruthi, G., Liu, F., Kale, S. & Sundararajan, M. Estimating training data influence by tracing gradient descent. Adv. Neural Inf. Process. Syst. 33 (2020).
39. Capstick, A., Palermo, F., Cui, T. & Barnaghi, P. Training neural networks on data sources with unknown reliability. Inf. Fusion 123, 103239 (2025).
40. Dwork, C., Hardt, M., Pitassi, T., Reingold, O. & Zemel, R. Fairness through awareness. In Proc. 3rd Innovations in Theoretical Computer Science Conference 214–226 (ACM, 2012).
41. Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L. & Steyerberg, E. W. Calibration: the Achilles heel of predictive analytics. BMC Med. 17, 230 (2019).
42. Ding, F., Hardt, M., Miller, J. & Schmidt, L. Retiring Adult: new datasets for fair machine learning. Adv. Neural Inf. Process. Syst. 34 (2021).
43. US Census Bureau. American Community Survey 2018 1-year Public Use Microdata Sample (US Census Bureau, 2019). https://www2.census.gov/programs-surveys/acs/data/pums/2018/1-Year/
44. Strack, B. et al. Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. Biomed Res. Int. 2014, 781670 (2014).
45. Clore, J., Cios, K., DeShazo, J. & Strack, B. Diabetes 130-US hospitals for years 1999–2008. UCI Machine Learning Repository (2014). https://doi.org/10.24432/C5230J
46. Smedley, B. D., Stith, A. Y. & Nelson, A. R. (eds) Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care (National Academies Press, 2003). https://doi.org/10.17226/10260
47. National Academies of Sciences, Engineering, and Medicine. Ending Unequal Treatment: Strategies to Achieve Equitable Health Care and Optimal Health for All (National Academies Press, 2024). https://doi.org/10.17226/27820
48. Peterson, E. D. et al. Racial variation in the use of coronary-revascularization procedures: are the differences real? Do they matter? N. Engl. J. Med. 336, 480–486 (1997).
49. Nanna, M. G. et al. Association of patient perceptions of cardiovascular risk and beliefs on statin drugs with racial differences in statin use: insights from the Patient and Provider Assessment of Lipid Management (PALM) Registry. JAMA Cardiol. 3, 739–748 (2018).
50. Eastwood, S. V. et al. Ethnic differences in guideline-indicated statin initiation for people with type 2 diabetes: a cohort study, 2006–2019. PLoS Med. 18, e1003672 (2021).
51. Escobar, G. et al. Racial and ethnic disparities in statin adherence: insights from the All of Us Research Program. Front. Cardiovasc. Med. 12, 1541082 (2025).
52. Kind, A. J. H. & Buckingham, W. R. Making neighborhood-disadvantage metrics accessible: the Neighborhood Atlas. N. Engl. J. Med. 378, 2456–2458 (2018).
53. Churchwell, K. et al. Call to action: structural racism as a fundamental driver of health disparities: a presidential advisory from the American Heart Association. Circulation 142, e454–e468 (2020).
54. Havranek, E. P. et al. Social determinants of risk and outcomes for cardiovascular disease: a scientific statement from the American Heart Association. Circulation 132, 873–898 (2015).

55. Uher, R. et al. Genome-wide pharmacogenetics of antidepressant response in the GENDEP project. Am. J. Psychiatry 167, 555–564 (2010).
56. Fried, E. I. & Nesse, R. M. Depression is not a consistent syndrome: an investigation of unique symptom patterns in the STAR*D study. J. Affect. Disord. 172, 96–102 (2015).
57. Becker, B. & Kohavi, R. Adult. UCI Machine Learning Repository (1996). https://doi.org/10.24432/C5XW20
58. Kohavi, R. Scaling up the accuracy of naive-Bayes classifiers: a decision-tree hybrid. In Proc. 2nd International Conference on Knowledge Discovery and Data Mining 202–207 (AAAI Press, 1996).
59. Yeh, I.-C. & Lien, C.-H. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Syst. Appl. 36, 2473–2480 (2009).
60. Yeh, I.-C. Default of credit card clients. UCI Machine Learning Repository (2009). https://doi.org/10.24432/C55S3H
61. Wolpert, D. H. Stacked generalization. Neural Netw. 5, 241–259 (1992).
62. Chernozhukov, V. et al. Double/debiased machine learning for treatment and structural parameters. Econom. J. 21, C1–C68 (2018).
63. Cochran, W. G. The combination of estimates from different experiments. Biometrics 10, 101–129 (1954).
64. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. In Proc. 3rd International Conference on Learning Representations (ICLR, 2015).
65. Bellamy, R. K. E. et al. AI Fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. IBM J. Res. Dev. 63, 4:1–4:15 (2019).
66. Cochran, W. G. The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics 24, 295–313 (1968).
67. Rosenbaum, P. R. & Rubin, D. B. Reducing bias in observational studies using subclassification on the propensity score. J. Am. Stat. Assoc. 79, 516–524 (1984).
68. Elazar, Y. & Goldberg, Y. Adversarial removal of demographic attributes from text data. In Proc. 2018 Conference on Empirical Methods in Natural Language Processing 11–21 (Association for Computational Linguistics, 2018).
69. Wilcoxon, F. Individual comparisons by ranking methods. Biometrics Bull. 1, 80–83 (1945).
70. Holland, R. & Iniesta, R. fairGATE: Fair Gated Algorithm for Targeted Equity. R package version 1.0.0. CRAN https://doi.org/10.32614/CRAN.package.fairGATE (2026).
71. Holland, R., Luo, L., Shamsutdinova, D. & Iniesta, R. Analysis code for 'fairGATE: traceable least-well-served group fairness correction of prediction models for binary and continuous outcomes'. Zenodo https://doi.org/10.5281/zenodo.23804615 (2026).
72. Nadeau, C. & Bengio, Y. Inference for the generalization error. Mach. Learn. 52, 239–281 (2003).

Acknowledgements

We thank Payam Barnaghi (Department of Brain Sciences, Imperial College London, and UK Dementia Research Institute) and Christopher Cogley for their comments on the manuscript.

This work was supported by the UK Government through the Research Ventures Catalyst award, Department for Science, Innovation, and Technology (R.H. only).

Author contributions

R.H. conceived the method, wrote the software and performed the analyses; R.I. lead the project and performed crucial background work as well as guiding the review process and refining ideas and methods; D.I. guided the initial methodology and provided continual support as the project progressed; L.L. collated figures and code for this paper.

Competing interests

The authors declare no competing interests.

Figure legends

Figure 1 | fairGATE. a, Forward pass. The base learner b_i is a cross-fitted XGBoost model with no fairness input (5-fold out-of-fold training scores, test predictions from a fit on all training rows); the package also offers a cross-fitted neural network. The disparity signal s_i is a cross-fitted XGBoost prediction, from the features and b_i, of each person's contribution to their group's gap from the reference group. A temperature-scaled softmax gate over the features, b_i and s_i weights K experts and the base learner (component 0). In the residual coupling, the experts add to b_i and are initialised at zero, so the untrained model is the base learner; in the expert coupling b_i is one of $K + 1$ weighted components. The coupling is selected by task-specific search. The protected attribute g_i is not an input to the base learner, the signal model, the gate or the experts; it enters the objective and, through cross-fitting, the target from which the disparity signal is learned. **b**, Objective: task loss, a group penalty selected by tuning (random-effects between-group variance component, variance of cell means, or the thresholded worst-group error-rate penalty T), an optional added thresholded term and an intervention cost on the mean absolute departure from the base learner. The shared selection rule for fairGATE, Fair Forests and FairGBM is: the smallest worst-group disparity among configurations (48 for fairGATE, 24 for each ensemble) within 0.01 of the method's best validation performance. **c**, Quantities exported for every held-out person, and the routing safeguard. **d**, Measurement of worst-group disparity: for a binary outcome, the average odds difference of each group from the reference group at the

prevalence-matched threshold; for a continuous outcome, the conditional mean prediction difference, the mean over five equal-frequency bands of the observed outcome of the group's mean prediction minus the reference group's, in outcome standard deviations (schematic; marker positions are illustrative). Worst-group disparity is the largest absolute value among groups passing the minimum-cell rule.

Figure 2 | The pre-specified primary criterion on the eleven confirmation tasks. a, Paired difference in worst-group |AOD| (binary outcomes) or |CMPD| in outcome standard deviations (continuous outcomes), fairGATE minus the fairer of Fair Forests and FairGBM on that task (coloured by that comparator) and minus XGBoost (grey); negative values favour fairGATE. **b**, Paired difference in AUC or R-squared against the same models; the dashed line is the criterion's tolerance of −0.005. Open markers, mean over the 50 partitions; thick bars, 95% percentile-bootstrap interval of that mean (10,000 resamples of partitions); thin lines, 2.5th to 97.5th percentiles of the per-partition differences, which describe partition variability and can include zero despite a consistently non-zero mean difference. Right, two-sided paired Wilcoxon p for disparity and performance against the fairer comparator, and the criterion outcome; 'comparator fairer' marks a significant difference in the unfavourable direction. Five tasks met the criterion and six did not; both sets are displayed identically. See Methods for the minimum-cell rule.

Figure 3 | Worst-group disparity and accuracy on the confirmation tasks. One panel per task: worst-group |AOD| or |CMPD| (x) against AUC or R-squared (y) for fairGATE, XGBoost, Fair Forests and FairGBM, and the Bayes-optimal reference on the simulated cohort (**j**, **k**). Better results lie towards the upper left. Markers, means over the 50 partitions; bars, 2.5th to 97.5th percentiles over partitions. The paired comparisons testing the criterion are in Fig. 2.

Figure 4 | Minimal intervention: disparity removed against decisions changed. a, For each confirmation task, the share of held-out people whose decision differs from XGBoost's (binary; each model at its own prevalence-matched threshold) or whose prediction changes by more than 0.05 outcome standard deviations (continuous), against the worst-group

disparity removed relative to XGBoost (XGBoost's worst-group disparity minus the model's; negative values indicate added disparity). XGBoost is at the origin by construction. The grey step line joins the non-dominated methods. **b**, Disparity removed per changed decision (ratio of the partition means in **a**); on readmission by race, fairGATE changed 0.08% of decisions. Means over the 50 partitions.

Figure 5 | Multi-group and intersectional structure. a–d, Mean over the 50 partitions of each group's absolute gap against the reference group (AOD, or CMPD in outcome standard deviations), with the pooled gap (all non-reference groups against the reference) and the worst-group gap, for California income by race (**a**), Texas income by race (**b**), California log income by race (**c**) and readmission by race (**d**). Open markers in **d**, groups excluded by the minimum-cell rule (fewer than 10 expected positives or negatives per test set). These groups are excluded from all methods' worst-group gaps. **e**, California income by race and sex: the gap of each intersectional cell against White men, the race-marginal and sex-marginal gaps, and the summaries of the pre-specified intersectional analysis. fairGATE has the smallest worst-cell and worst race-marginal gaps and a larger sex-marginal gap than FairGBM and XGBoost. **f**, Readmission by race and sex (Hispanic, Asian and Other pooled across sex under the minimum-cell rule; open markers, race groups excluded by that rule): none of fairGATE's intersectional summaries differs significantly from that of the fairer comparator (worst-cell $p = 0.76$; largest pairwise gap $p = 0.45$; race marginal $p = 0.38$; sex marginal $p = 0.18$).

Figure 6 | Mechanism and traceability. a, Component ablations on the confirmation tasks, refitted at fairGATE's selected configuration with the same seeds: mean over the 50 partitions of fairGATE's worst-group disparity minus that of the ablation (negative values indicate lower disparity with the component); asterisk, paired Wilcoxon $p < 0.05$; n.a., the ablation coincides with the selected configuration. **b**, Noise floor (part of the variance-of-means penalty at the base learner expected from sampling noise alone, as a percentage of that penalty) against the primary-criterion difference; filled markers, criterion met. The

analysis is post hoc and descriptive. **c**, Routing safeguard on every task: median over partitions of the out-of-sample AUC for the protected attribute predicted from fairGATE's routing weights against the AUC from the predictors using the same logistic probe. Points above the diagonal indicate safeguard failures (the five tasks named in the text). **d**, Per-person adjustments on California income by race (fairGATE minus the base learner, probability scale, all 50 partitions): mean, 25th to 75th and 5th to 95th percentiles by group; share of decisions moved above or below the threshold relative to the base learner; share of people whose adjustment interval excludes zero. The intervals are a warm-started bootstrap approximation and have not been calibrated.

Figures

Figure 1

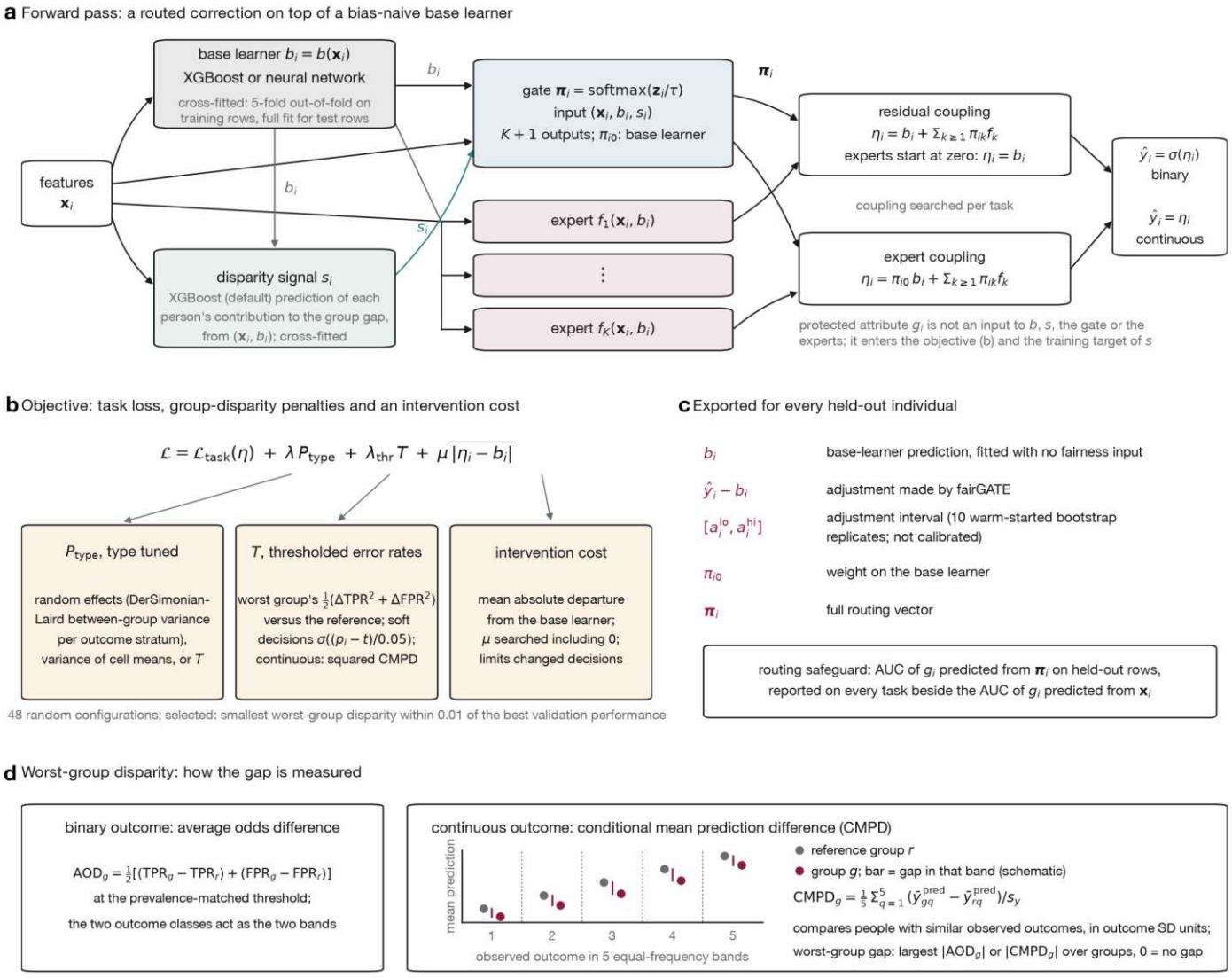


Figure 2

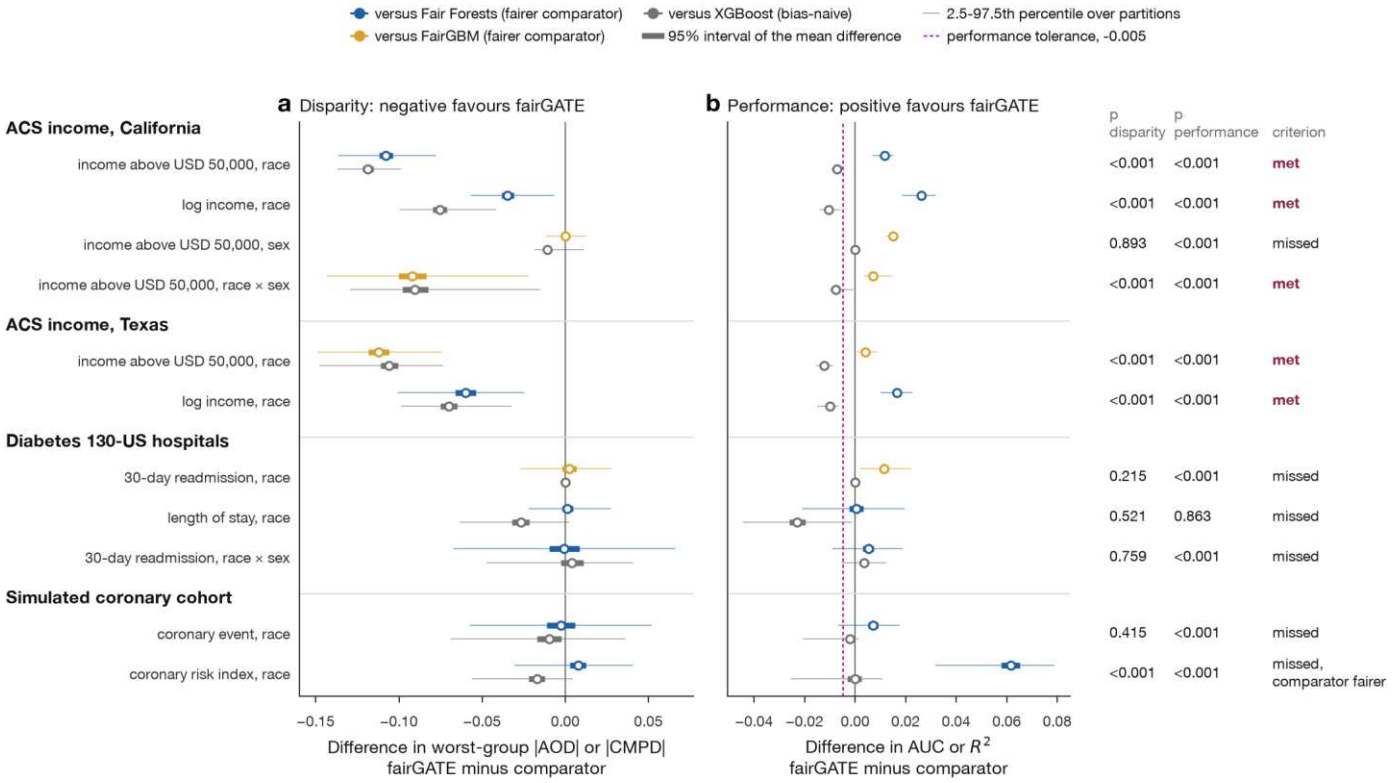
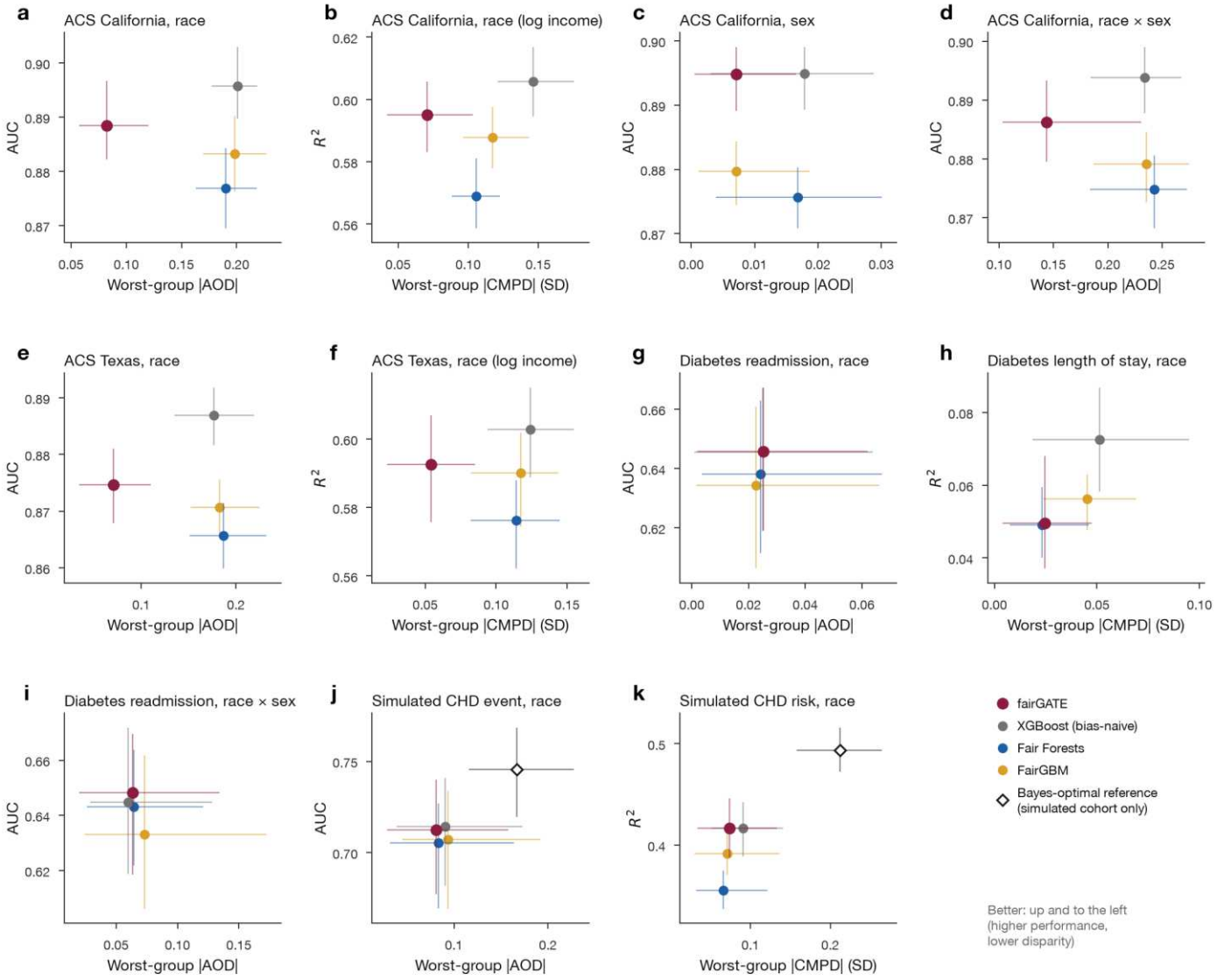

versus Fair Forests (fairer comparator)
versus FairGBM (fairer comparator)
versus XGBoost (bias-naive)
95% interval of the mean difference
2.5-97.5th percentile over partitions
performance tolerance, -0.005
a Disparity: negative favours fairGATE
b Performance: positive favours fairGATE
p disparity
p performance
criterion
ACS income, California
income above USD 50,000, race
<0.001
<0.001
met
log income, race
<0.001
<0.001
met
income above USD 50,000, sex
0.893
<0.001
missed
income above USD 50,000, race × sex
<0.001
<0.001
met
ACS income, Texas
income above USD 50,000, race
<0.001
<0.001
met
log income, race
<0.001
<0.001
met
Diabetes 130-US hospitals
30-day readmission, race
0.215
<0.001
missed
length of stay, race
0.521
0.863
missed
30-day readmission, race × sex
0.759
<0.001
missed
Simulated coronary cohort
coronary event, race
0.415
<0.001
missed
coronary risk index, race
<0.001
<0.001
missed, comparator fairer
−0.15
−0.10
−0.05
0.00
0.05
Difference in worst-group |AOD| or |CMPD|
fairGATE minus comparator
−0.04
−0.02
0.00
0.02
0.04
0.06
0.08
Difference in AUC or R^2
fairGATE minus comparator

Figure 3



841

842 **Figure 4**

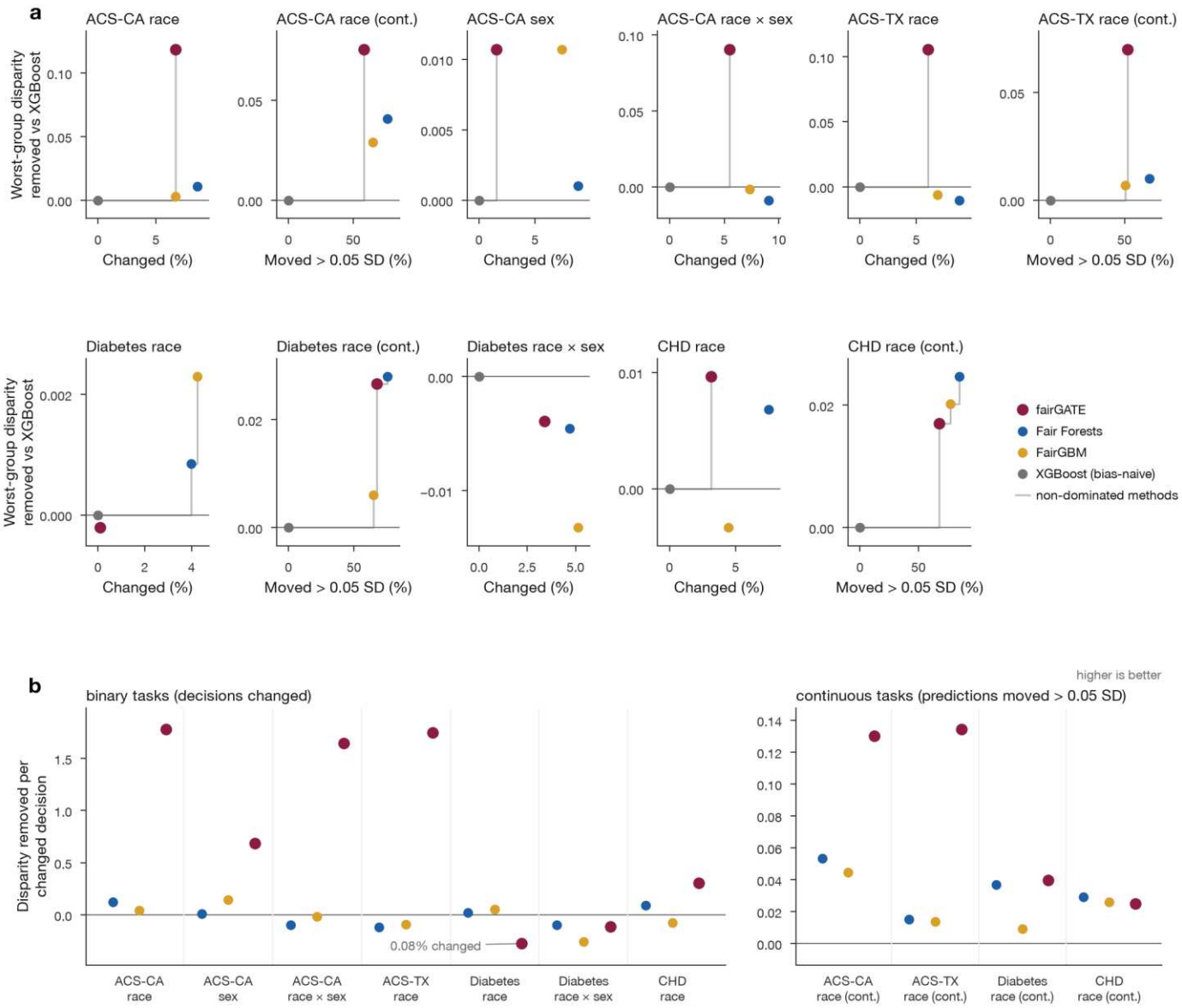

a
ACS-CA race
ACS-CA race (cont.)
ACS-CA sex
ACS-CA race × sex
ACS-TX race
ACS-TX race (cont.)
Diabetes race
Diabetes race (cont.)
Diabetes race × sex
CHD race
CHD race (cont.)
Worst-group disparity removed vs XGBoost
Changed (%)
Moved > 0.05 SD (%)
fairGATE
Fair Forests
FairGBM
XGBoost (bias-naive)
non-dominated methods
b
binary tasks (decisions changed)
continuous tasks (predictions moved > 0.05 SD)
higher is better
Disparity removed per changed decision
0.08% changed
ACS-CA race
ACS-CA sex
ACS-CA race × sex
ACS-TX race
Diabetes race
Diabetes race × sex
CHD race
ACS-CA race (cont.)
ACS-TX race (cont.)
Diabetes race (cont.)
CHD race (cont.)

Figure 5

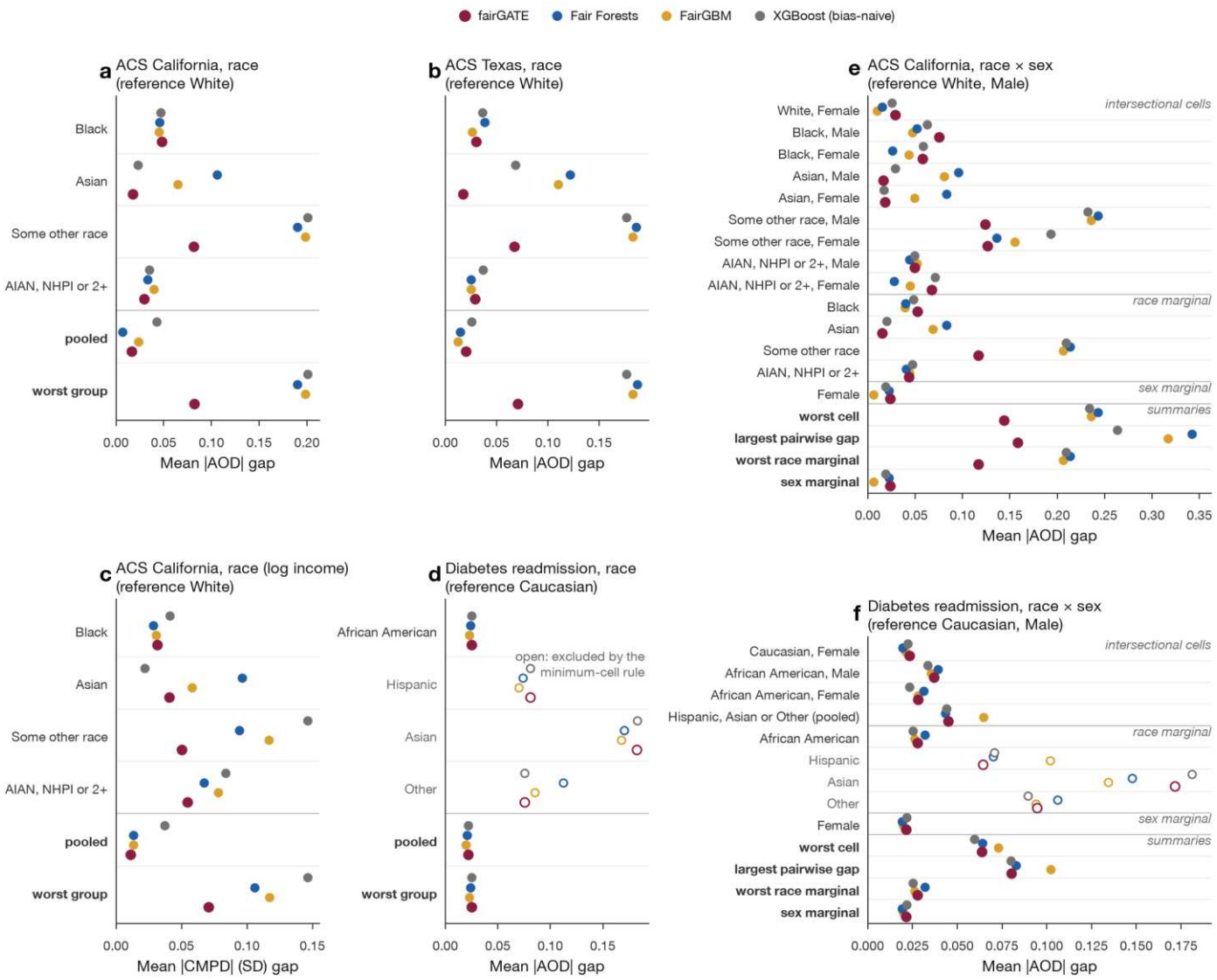
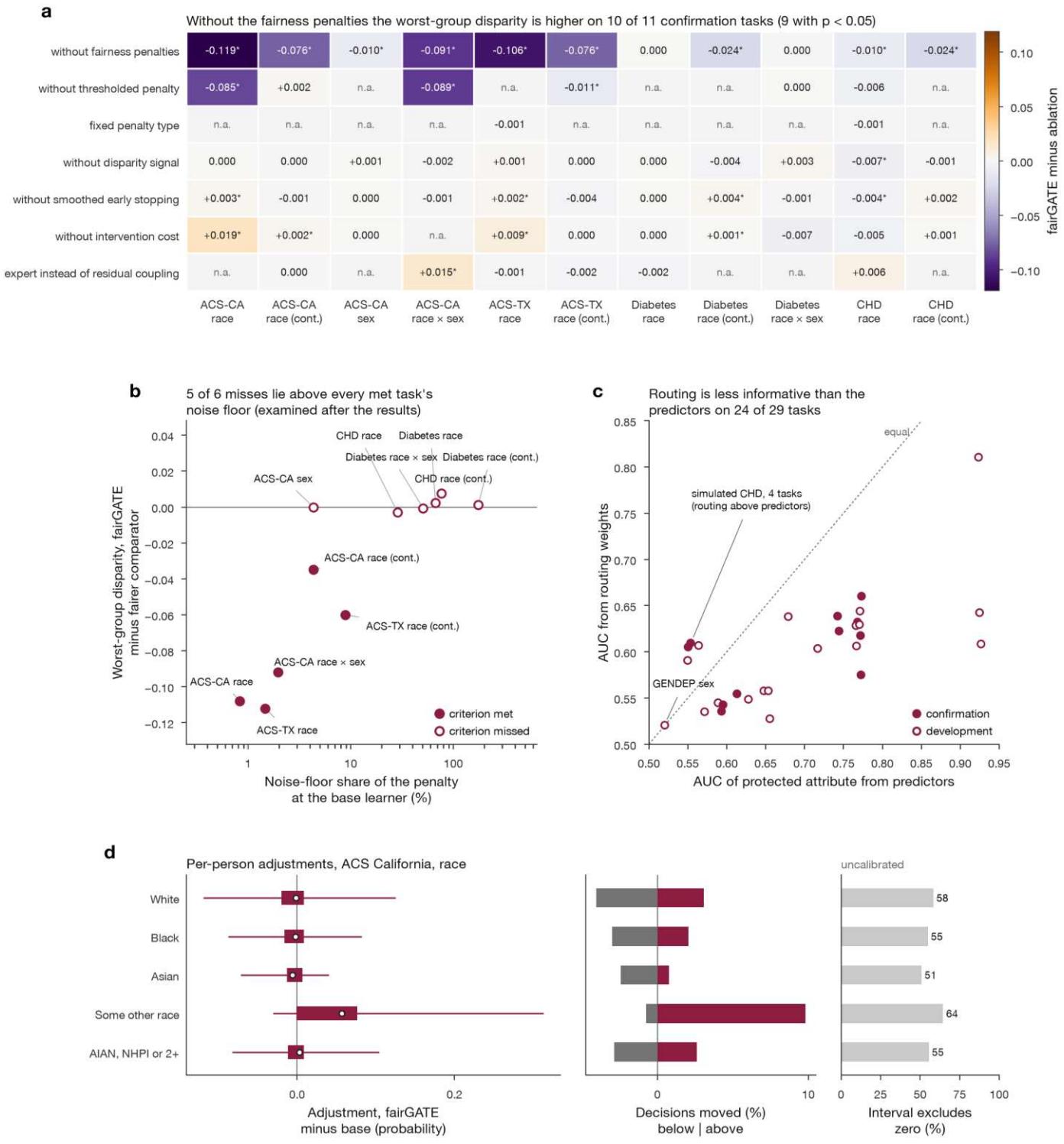

fairGATE
Fair Forests
FairGBM
XGBoost (bias-naive)
a ACS California, race (reference White)
Black
Asian
Some other race
AIAN, NHPI or 2+
pooled
worst group
0.00 0.05 0.10 0.15 0.20
Mean |AOD| gap
b ACS Texas, race (reference White)
0.00 0.05 0.10 0.15
Mean |AOD| gap
e ACS California, race × sex (reference White, Male)
intersectional cells
White, Female
Black, Male
Black, Female
Asian, Male
Asian, Female
Some other race, Male
Some other race, Female
AIAN, NHPI or 2+, Male
AIAN, NHPI or 2+, Female
race marginal
Black
Asian
Some other race
AIAN, NHPI or 2+
sex marginal
Female
summaries
worst cell
largest pairwise gap
worst race marginal
sex marginal
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35
Mean |AOD| gap
c ACS California, race (log income) (reference White)
Black
Asian
Some other race
AIAN, NHPI or 2+
pooled
worst group
0.00 0.05 0.10 0.15
Mean |CMPD| (SD) gap
d Diabetes readmission, race (reference Caucasian)
African American
Hispanic
Asian
Other
pooled
worst group
open: excluded by the minimum-cell rule
0.00 0.05 0.10 0.15
Mean |AOD| gap
f Diabetes readmission, race × sex (reference Caucasian, Male)
intersectional cells
Caucasian, Female
African American, Male
African American, Female
Hispanic, Asian or Other (pooled)
race marginal
African American
Hispanic
Asian
Other
sex marginal
Female
summaries
worst cell
largest pairwise gap
worst race marginal
sex marginal
0.000 0.025 0.050 0.075 0.100 0.125 0.150 0.175
Mean |AOD| gap

Figure 6


a
Without the fairness penalties the worst-group disparity is higher on 10 of 11 confirmation tasks (9 with p < 0.05)
without fairness penalties
without thresholded penalty
fixed penalty type
without disparity signal
without smoothed early stopping
without intervention cost
expert instead of residual coupling
ACS-CA race
ACS-CA race (cont.)
ACS-CA sex
ACS-CA race × sex
ACS-TX race
ACS-TX race (cont.)
Diabetes race
Diabetes race (cont.)
Diabetes race × sex
CHD race
CHD race (cont.)
fairGATE minus ablation
b
5 of 6 misses lie above every met task's noise floor (examined after the results)
Worst-group disparity, fairGATE minus fairer comparator
Noise-floor share of the penalty at the base learner (%)
criterion met
criterion missed
c
Routing is less informative than the predictors on 24 of 29 tasks
AUC from routing weights
AUC of protected attribute from predictors
equal
simulated CHD, 4 tasks (routing above predictors)
GENDEP sex
confirmation
development
d
Per-person adjustments, ACS California, race
White
Black
Asian
Some other race
AIAN, NHPI or 2+
Adjustment, fairGATE minus base (probability)
Decisions moved (%) below | above
uncalibrated
Interval excludes zero (%)

847

Extended Data

Extended Data Figure 1 | Simulated cohort: Bayes-optimal headroom. Worst-group disparity and performance of every model against the Bayes-optimal reference (the true event probability and the rescaled latent predictor) on the simulated confirmation and development tasks. The Bayes-optimal reference has worst-group disparities of 0.167 and 0.212 on the two confirmation tasks, both larger than that of any fitted model, because bias was introduced into the outcome.

Extended Data Figure 2 | Sensitivity analyses. Paired difference, fairGATE minus the fairer comparator of the primary analysis (filled markers; mean over the 50 partitions with 2.5th to 97.5th percentiles), with the primary analysis for reference (open markers), the Wilcoxon p and the criterion outcome. **a**, Worst-group |AOD| at a fixed classification threshold of 0.50 in place of the prevalence-matched threshold (binary tasks only). The criterion is met on three of the seven binary tasks (California income by race, California income by race and sex and Texas income by race), and fairGATE is significantly less fair than the comparator on two (readmission by race and simulated coronary event by race; difference 0.034 on the coronary event). At 0.50, almost no individuals are classified as positive for the low-prevalence outcomes, for which the comparator's worst-group |AOD| is close to zero. These differences are therefore not comparable with those in the primary analysis. **b, The pooled contrast (all non-reference groups against the reference).** Against the primary-analysis comparator, the pooled criterion is met on one task (California log income by race); when the fairer comparator is chosen on the pooled contrast itself, FairGBM replaces Fair Forests on that task (pooled disparities differing by less than 0.0001) and the criterion is met on none. A comparator had a significantly smaller pooled disparity on four tasks (California income by race, California income by race and sex, Texas income by race and Texas log income by race).

Extended Data Figure 3 | Per-group disparities on every multi-group task. Mean signed gap between each group and the reference group over the 50 partitions, for every multi-group and intersectional task (confirmation and development) and every model (columns within each metric: 1 XGBoost, 2 Fair Forests, 3 FairGBM, 4 fairGATE). Binary tasks: SPD, DI (ratio, coloured on a log scale centred at 1), EOD and AOD; continuous tasks: MPD, CMPD, the largest stratum difference (CMPD max), GLD and KS, in outcome standard deviations except KS. On intersectional tasks, the marginal gaps of each attribute are shown below the cells. Grey labels, groups failing the minimum-cell rule, which are excluded from the worst-group summary for every method.

Extended Data Figure 4 | All component ablations, confirmation and development tasks. Left, worst-group disparity of fairGATE minus that of each ablation; right, AUC or R-squared of fairGATE minus that of the ablation; means over the 50 partitions; asterisk, paired Wilcoxon $p < 0.05$; n.a., the ablation coincides with fairGATE's selected configuration. For the confirmation tasks, 57 task-by-ablation pairs were fitted and 20 were not applicable. For the development tasks, 34 pairs on seven tasks were fitted. Hatched cells denote the 55 applicable pairs on the other development tasks not fitted within the computational budget; these are not null effects.

Extended Data Figure 5 | Routing detail. For every task: **a**, normalised gate entropy (1, uniform routing) for fairGATE; **b**, mean routing weight on the base learner; **c**, mean routing weight on the base learner and on each expert for fairGATE; **d**, mean weight on the base learner by protected group for fairGATE. In the residual coupling, the full base score contributes to every prediction, and a larger weight on component 0 leaves less weight for the experts' correction.

Extended Data Figure 6 | fairGATE with a neural-network base learner (exploratory). One panel per confirmation task: worst-group |AOD| or |CMPD| (x) against AUC or R-squared (y) for XGBoost and a cross-fitted neural network (both bias-naive), fairGATE fitted using each base learner, and the fairer comparator of the primary analysis. Markers, mean over the 50 partitions; bars, 2.5th to 97.5th percentiles over partitions; thin lines join each base learner to fairGATE fitted on it (solid, XGBoost; dashed, neural network). The confirmation tasks had already been used for the pre-specified test, so this comparison is exploratory. Better results lie towards the upper left.

Extended Data Figure 7 | Sensitivity of the conditional mean prediction difference to the number of outcome bands. a, Share of the between-group difference in the observed outcome removed by stratifying on equal-frequency bands of that outcome (whole cohort, size-weighted over groups passing the minimum-cell rule), against the number of bands requested; dashed line, the share expected for a normally distributed variable [66]. Length of stay has many tied values, so requesting 8, 10 and 20 bands yields 6, 7 and 9. **b**, Bootstrap standard error of fairGATE's worst-group |CMPD| (solid) and of a single band-specific group difference (dashed), 200 resamples of test individuals in each of partitions 1 to 10; open diamonds, the band-free version. **c**, fairGATE minus the fairer comparator in worst-group |CMPD| (mean over 50 partitions, Nadeau–Bengio corrected 95% interval [72]) for each continuous confirmation task; open diamonds, the band-free version (natural cubic spline of the outcome, 6 df, with group indicators). Grey band, the five bands used throughout. Values

are in outcome standard deviations throughout. The comparator is Fair Forests at every band count and FairGBM for the band-free version on the two census tasks.

Extended Data Table 1 | Datasets, tasks and roles. Confirmation tasks were built for this study from samples that no design decision used and that share no person (and, for the California census sample, no household) with the development data; development tasks informed the design and are reported descriptively. For each task: role, dataset, outcome type and definition, binary prevalence, protected attribute, number of groups or intersectional cells, reference group, group sizes, number of non-reference groups passing the minimum-cell rule, individuals, predictors, partitions, training and test sizes of partition 1, and the disjointness from the development data. Adult tasks are reported in Supplementary Fig. 2 for comparability only.

Extended Data Table 2 | Full results on the confirmation tasks. Every confirmation task and model: performance (AUC or R-squared; RMSE for continuous outcomes), the worst-group disparity on the primary metric (AOD or CMPD) and on two further metrics, the joint and largest pairwise disparity and the pooled gap. The metrics of each task are named beside it. Values are means over the 50 partitions; performance, RMSE and the primary worst-group disparity are shown with 95% resampling intervals (2.5th and 97.5th percentiles of 200 resamples of the held-out rows in each partition, pooled over partitions); these are intervals over one cohort, not population confidence intervals. Shares of decisions changed are shown in Fig. 4; development and Adult tasks are in the deposited result tables [71].

Extended Data Table 3 | Selected configurations. For every method and confirmation task, the configuration chosen by its selection rule (stated in the last column): for fairGATE, coupling, penalty type, λ, λ_{thr}, intervention cost μ, the median (minimum to maximum) selected epoch over partitions, the configuration identifier and the remaining settings; for the comparators, their tuned hyperparameters. XGBoost takes its best validation performance; fairGATE, Fair Forests and FairGBM take the shared selection rule (Methods). Configurations on the development tasks are in the deposited result tables [71].

Extended Data Table 1

Task	Role	Dataset	Outcome type	Outcome	Binary prevalence	Protected attribute	K groups	Reference	Group sizes	Eligible non-reference groups (minimum-cell rule)	Individuals	Predictors	Partitions	Train / test (partition 1)	Disjointness from development data
acsC_race	confirmation	ACS income (California, 2018)	binary	personal income (PINCP) above USD 50,000 in the past 12 months	0.434	race (RAC1P, five groups)	5	White	White 31,398; Black 2,426; Asian 8,195; Some other race 5,315; AIAN, NHPI or 2+ races 2,666	4 of 4	50,000	67	50	40,004 / 9,996	no person or household shared with the development sample (person overlap 0, household overlap 0)
acsC_race_reg	confirmation	ACS income (California, 2018)	continuous	natural logarithm of personal income (PINCP, USD, past 12 months)	n.a.	race (RAC1P, five groups)	5	White	White 31,398; Black 2,426; Asian 8,195; Some other race 5,315; AIAN, NHPI or 2+ races 2,666	4 of 4	50,000	67	50	40,011 / 9,989	no person or household shared with the development sample (person overlap 0, household overlap 0)
acsC_sex	confirmation	ACS income (California, 2018)	binary	personal income (PINCP) above USD 50,000 in the past 12 months	0.434	sex (SEX)	2	Male	Male 26,607; Female 23,393	1 of 1	50,000	71	50	40,001 / 9,999	no person or household shared with the development sample (person overlap 0, household overlap 0)
acsC_race_sex	confirmation	ACS income (California, 2018)	binary	personal income (PINCP) above USD 50,000 in the past 12 months	0.434	race x sex (RAC1P five groups crossed with SEX)	10	White-Male	White-Male 16,974; White-Female 14,424; Black-Male 1,170; Black-Female 1,256; Asian-Male 4,090; Asian-Female 4,105; Some other race-Male 2,996; Some other race-Female 2,319; AIAN, NHPI or 2+ races-Male 1,377; AIAN, NHPI or 2+ races-Female 1,289	9 of 9	50,000	66	50	40,010 / 9,990	no person or household shared with the development sample (person overlap 0, household overlap 0)
acsTX_race	confirmation	ACS income (Texas, 2018)	binary	personal income (PINCP) above USD 50,000 in the past 12 months	0.366	race (RAC1P, five groups)	5	White	White 38,929; Black 4,683; Asian 2,515; Some other race	4 of 4	50,000	67	50	40,005 / 9,995	different state file; no person can be in the development

									2,474; AIAN, NHPI or 2+ races 1,399						sample (overlap 0)
acsTX_race_reg	confirmation	ACS income (Texas, 2018)	continuous	natural logarithm of personal income (PINCP, USD, past 12 months)	n.a.	race (RAC1P, five groups)	5	White	White 38,929; Black 4,683; Asian 2,515; Some other race 2,474; AIAN, NHPI or 2+ races 1,399	4 of 4	50,000	67	50	40,008 / 9,992	different state file; no person can be in the development sample (overlap 0)
diabC_race	confirmation	Diabetes 130-US hospitals, 1999-2008	binary	readmitted within 30 days of discharge (readmitted == '<30')	0.090	race	5	Caucasian	Caucasian 13,824; African American 3,362; Hispanic 431; Asian 124; Other 313	1 of 4	18,054	106	50	14,446 / 3,608	no patient (patient_nbr) shared with the development sample (overlap 0)
diabC_race_reg	confirmation	Diabetes 130-US hospitals, 1999-2008	continuous	length of stay in days (time_in_hospital, integer 1 to 14	n.a.	race	5	Caucasian	Caucasian 13,824; African American 3,362; Hispanic 431; Asian 124; Other 313	3 of 4	18,054	57	50	14,455 / 3,599	no patient (patient_nbr) shared with the development sample (overlap 0)
diabC_race_sex	confirmation	Diabetes 130-US hospitals, 1999-2008	binary	readmitted within 30 days of discharge (readmitted == '<30')	0.090	race x sex	5	Caucasian-Male	Caucasian-Male 6,786; Caucasian-Female 7,038; African American-Male 1,288; African American-Female 2,074; Hispanic, Asian or Other (pooled) 868	4 of 4	18,054	105	50	14,447 / 3,607	no patient (patient_nbr) shared with the development sample (overlap 0)
chdC_race	confirmation	CHD (simulated)	binary	coronary care event	0.191	race	5	White	White 5,499; Black 2,061; Hispanic 1,468; Asian 692; Other 280	4 of 4	10,000	24	50	8,004 / 1,996	fresh simulated cohort; shared patient IDs 0, identical predictor records 0
chdC_race_reg	confirmation	CHD (simulated)	continuous	coronary risk index	n.a.	race	5	White	White 5,499; Black 2,061; Hispanic 1,468; Asian 692; Other 280	4 of 4	10,000	24	50	8,009 / 1,991	fresh simulated cohort; shared patient IDs 0, identical predictor records 0
acs_race	development	ACS income (California, 2018)	binary	personal income (PINCP) above USD 50,000 in the past 12 months	0.412	race (RAC1P, five groups)	5	White	White 31,144; Black 2,061; Asian 8,316; Some other race 5,742; AIAN, NHPI or 2+ races 2,737	4 of 4	50,000	67	50	40,006 / 9,994	development data (used in design decisions)
acs_race_reg	development	ACS income (California,	continuous	natural logarithm of personal	n.a.	race (RAC1P, five	5	White	White 31,144; Black	4 of 4	50,000	67	50	40,007 / 9,993	development data (used in

		2018)		income (PINCP, USD, past 12 months)		groups)			2,061; Asian 8,316; Some other race 5,742; AIAN, NHPI or 2+ races 2,737						design decisions)
acs_sex	development	ACS income (California, 2018)	binary	personal income (PINCP) above USD 50,000 in the past 12 months	0.412	sex (SEX)	2	Male	Male 26,623; Female 23,377	1 of 1	50,000	71	50	40,001 / 9,999	development data (used in design decisions)
credit_sex	development	Credit card default (Taiwan)	binary	no default on the credit card payment due in October 2005 (source column 'default payment next month' = 0)	0.779	sex (SEX)	2	Male	Male 11,888; Female 18,112	1 of 1	30,000	27	50	24,002 / 5,998	development data (used in design decisions)
credit_age	development	Credit card default (Taiwan)	binary	no default on the credit card payment due in October 2005 (source column 'default payment next month' = 0)	0.779	age band (AGE, years: <30, 30-39, 40-49, >=50)	4	30-39	30-39 11,238; <30 9,618; 40-49 6,464; >=50 2,680	3 of 3	30,000	27	50	24,003 / 5,997	development data (used in design decisions)
diab_race	development	Diabetes 130-US hospitals, 1999-2008	binary	readmitted within 30 days of discharge (readmitted == '<30')	0.090	race	5	Caucasian	Caucasian 38,468; African American 9,263; Hispanic 1,069; Asian 364; Other 836	3 of 4	50,000	106	50	40,004 / 9,996	development data (used in design decisions)
diab_race_reg	development	Diabetes 130-US hospitals, 1999-2008	continuous	length of stay in days (time_in_hospital, integer 1 to 14	n.a.	race	5	Caucasian	Caucasian 38,468; African American 9,263; Hispanic 1,069; Asian 364; Other 836	4 of 4	50,000	57	50	40,012 / 9,988	development data (used in design decisions)
chd_race	development	CHD (simulated)	binary	coronary care event	0.203	race	5	White	White 5,470; Black 2,020; Hispanic 1,479; Asian 741; Other 290	4 of 4	10,000	24	50	8,005 / 1,995	development data (used in design decisions)
chd_race_reg	development	CHD (simulated)	continuous	coronary risk index	n.a.	race	5	White	White 5,470; Black 2,020; Hispanic 1,479; Asian 741; Other 290	4 of 4	10,000	24	50	8,009 / 1,991	development data (used in design decisions)
gendep_age	development	GENDEP	binary	remission at 12 weeks (HAMD17 <= 7)	0.386	age group (median split at 42 years)	2	Younger (< 42)	Younger (< 42) 212; Older (>= 42) 218	1 of 1	430	148	50	345 / 85	development data (used in design decisions)
gendep_sex	development	GENDEP	binary	remission at 12 weeks (HAMD17 <= 7)	0.386	sex	2	Male	Male 153; Female 277	1 of 1	430	148	50	346 / 84	development data (used in design decisions)
gendep_marital	development	GENDEP	binary	remission at 12 weeks (HAMD17 <= 7)	0.386	marital status	2	Married	Married 250; Not married 180	1 of 1	430	148	50	345 / 85	development data (used in design decisions)
acs_race_sex	development	ACS income	binary	personal income	0.412	race x sex	10	White-Male	White-Male	9 of 9	50,000	66	50	40,008 /	development

		(California, 2018)		(PINCP) above USD 50,000 in the past 12 months		(RAC1P five groups crossed with SEX)			16,758; White-Female 14,386; Black-Male 1,034; Black-Female 1,027; Asian-Male 4,206; Asian-Female 4,110; Some other race-Male 3,216; Some other race-Female 2,526; AIAN, NHPI or 2+ races-Male 1,409; AIAN, NHPI or 2+ races-Female 1,328					9,992	data (used in design decisions)
credit_sex_age	development	Credit card default (Taiwan)	binary	no default on the credit card payment due in October 2005 (source column 'default payment next month' = 0)	0.779	sex x age band (SEX crossed with AGE bands <30, 30-39, 40-49, >=50)	8	Male-30-39	Male-30-39 4,565; Male-<30 3,281; Male-40-49 2,771; Male->=50 1,271; Female-30-39 6,673; Female-<30 6,337; Female-40-49 3,693; Female->=50 1,409	7 of 7	30,000	26	50	24,007 / 5,993	development data (used in design decisions)
diab_race_sex	development	Diabetes 130-US hospitals, 1999-2008	binary	readmitted within 30 days of discharge (readmitted == '<30')	0.090	race x sex	5	Caucasian-Male	Caucasian-Male 18,556; Caucasian-Female 19,912; African American-Male 3,657; African American-Female 5,606; Hispanic, Asian or Other (pooled) 2,269	4 of 4	50,000	105	50	40,003 / 9,997	development data (used in design decisions)
adult_sex	Adult comparability	Adult	binary	income > USD 50,000	0.239	sex	2	Male	Male 32,650; Female 16,192	1 of 1	48,842	105	50	39,076 / 9,766	pilot data (Extended Data comparability only)
adult_sex_reg	Adult comparability	Adult	continuous	usual hours worked per week	n.a.	sex	2	Male	Male 32,650; Female 16,192	1 of 1	48,842	105	50	39,077 / 9,765	pilot data (Extended Data comparability only)
adult_sex_canon	Adult comparability	Adult	binary	income > USD 50,000	0.239	sex	2	Male	Male 32,650; Female 16,192	1 of 1	48,842	105	1	32,561 / 16,281	pilot data (Extended Data comparability only)

Extended Data Table 2

Task: performance, primary; second/third metric	Model	Performance	RMSE	Worst-group primary	Worst-group second	Worst-group third	Joint primary	Pairwise-max primary	Pooled |gap|
acsC_race: AUC, AOD; EOD/SPD	XGBoost (bias-naive)	0.8957 (0.8868 to 0.9048)	n.a.	0.2008 (0.1572 to 0.2431)	0.2939	0.3022	0.0787	0.2239	0.0425
	Fair Forests	0.8769 (0.8669 to 0.8869)	n.a.	0.1901 (0.1434 to 0.2361)	0.2720	0.2883	0.1117	0.2961	0.0066
	FairGBM	0.8832 (0.8735 to 0.8929)	n.a.	0.1981 (0.1519 to 0.2444)	0.2872	0.2956	0.0967	0.2630	0.0236
	fairGATE	0.8885 (0.8795 to 0.8980)	n.a.	0.0822 (0.0442 to 0.1330)	0.1324	0.2018	0.0415	0.0996	0.0166
acsC_race_reg: R2, CMPD; MPD/GLD	XGBoost (bias-naive)	0.6058 (0.5844 to 0.6264)	0.770 (0.746 to 0.795)	0.1464 (0.1032 to 0.1969)	0.3675	0.0633	0.0688	0.1676	0.037
	Fair Forests	0.5690 (0.5486 to 0.5885)	0.805 (0.781 to 0.830)	0.1056 (0.0815 to 0.1437)	0.2815	0.0617	0.0827	0.1918	0.0132
	FairGBM	0.5878 (0.5669 to 0.6082)	0.787 (0.764 to 0.812)	0.1172 (0.0811 to 0.1645)	0.3181	0.0652	0.0741	0.1753	0.0132
	fairGATE	0.5951 (0.5740 to 0.6158)	0.780 (0.756 to 0.804)	0.0708 (0.0389 to 0.1259)	0.2656	0.0575	0.0441	0.1069	0.0111
acsC_sex: AUC, AOD; EOD/SPD	XGBoost (bias-naive)	0.8949 (0.8866 to 0.9023)	n.a.	0.0178 (0.0013 to 0.0379)	0.0182	0.0873	0.0178	n.a.	0.0178
	Fair Forests	0.8756 (0.8672 to 0.8841)	n.a.	0.0168 (0.0012 to 0.0379)	0.0269	0.0504	0.0168	n.a.	0.0168
	FairGBM	0.8797 (0.8712 to 0.8881)	n.a.	0.0071 (0.0004 to 0.0274)	0.0155	0.0642	0.0071	n.a.	0.0071
	fairGATE	0.8948 (0.8864 to 0.9023)	n.a.	0.0071 (0.0004 to 0.0263)	0.0107	0.0734	0.0071	n.a.	0.0071
acsC_race_sex: AUC, AOD; EOD/SPD	XGBoost (bias-naive)	0.8938 (0.8850 to 0.9020)	n.a.	0.2340 (0.1812 to 0.2932)	0.3509	0.3793	0.0607	0.2631	0.0444
	Fair Forests	0.8748 (0.8654 to 0.8837)	n.a.	0.2428 (0.1758 to 0.2983)	0.3687	0.3392	0.0662	0.3420	0.0066
	FairGBM	0.8791 (0.8697 to 0.8879)	n.a.	0.2354 (0.1736 to 0.2929)	0.3522	0.3505	0.0598	0.3168	0.0186
	fairGATE	0.8862 (0.8767 to 0.8957)	n.a.	0.1435 (0.0859 to 0.2436)	0.2149	0.3317	0.0465	0.1581	0.0365
acsTX_race: AUC, AOD; EOD/SPD	XGBoost (bias-naive)	0.8869 (0.8781 to 0.8952)	n.a.	0.1765 (0.1046 to 0.2522)	0.2588	0.2539	0.0748	0.2446	0.0256
	Fair Forests	0.8656 (0.8563 to 0.8747)	n.a.	0.1869 (0.1263 to 0.2591)	0.2673	0.2544	0.0884	0.3073	0.0143
	FairGBM	0.8706 (0.8614 to 0.8794)	n.a.	0.1827 (0.1206 to 0.2536)	0.2636	0.2539	0.0798	0.2924	0.012
	fairGATE	0.8746 (0.8655 to 0.8841)	n.a.	0.0706 (0.0299 to 0.1445)	0.1212	0.1642	0.0351	0.0827	0.02
acsTX_race_reg: R2, CMPD; MPD/GLD	XGBoost (bias-naive)	0.6028 (0.5808 to 0.6222)	0.762 (0.738 to 0.786)	0.1242 (0.0744 to 0.1841)	0.3108	0.0635	0.0616	0.1911	0.0211
	Fair Forests	0.5762 (0.5554 to 0.5951)	0.787 (0.763 to 0.811)	0.1141 (0.0714 to 0.1749)	0.2599	0.0678	0.0585	0.1992	0.0101
	FairGBM	0.5901 (0.5687 to 0.6091)	0.774 (0.750 to 0.798)	0.1173 (0.0728 to 0.1765)	0.2913	0.0629	0.0583	0.2021	0.0123
	fairGATE	0.5926 (0.5693 to 0.6135)	0.771 (0.746 to 0.796)	0.0541 (0.0250 to 0.1192)	0.2224	0.0623	0.0291	0.0673	0.0166
diabC_race: AUC, AOD; EOD/SPD	XGBoost (bias-naive)	0.6456 (0.6024 to 0.6858)	n.a.	0.0250 (0.0015 to 0.0939)	0.0488	0.0116	0.0250	0.0250	0.0217
	Fair Forests	0.6382 (0.5940 to 0.6794)	n.a.	0.0241 (0.0014 to 0.0953)	0.0462	0.0118	0.0241	0.0241	0.0206

	FairGBM	0.6344 (0.5879 to 0.6767)	n.a.	0.0227 (0.0014 to 0.0919)	0.0426	0.0134	0.0227	0.0227	0.0195
	fairGATE	0.6457 (0.6026 to 0.6858)	n.a.	0.0252 (0.0015 to 0.0931)	0.0490	0.0115	0.0252	0.0252	0.0217
diabC_race_reg: R2, CMPD; MPD/GLD	XGBoost (bias-naive)	0.0725 (0.0476 to 0.0965)	2.844 (2.746 to 2.942)	0.0511 (0.0194 to 0.1466)	0.0473	0.0741	0.0268	0.0620	0.0231
	Fair Forests	0.0491 (0.0326 to 0.0662)	2.880 (2.782 to 2.976)	0.0232 (0.0082 to 0.0801)	0.0202	0.0727	0.0095	0.0306	0.0063
	FairGBM	0.0562 (0.0409 to 0.0704)	2.869 (2.771 to 2.966)	0.0451 (0.0225 to 0.0961)	0.0435	0.0744	0.0314	0.0475	0.03
	fairGATE	0.0495 (0.0322 to 0.0750)	2.879 (2.774 to 2.981)	0.0245 (0.0073 to 0.0830)	0.0218	0.0717	0.0085	0.0289	0.0068
diabC_race_sex: AUC, AOD; EOD/SPD	XGBoost (bias-naive)	0.6447 (0.6025 to 0.6861)	n.a.	0.0596 (0.0299 to 0.1976)	0.1197	0.0243	0.0256	0.0797	0.0174
	Fair Forests	0.6430 (0.6015 to 0.6835)	n.a.	0.0641 (0.0311 to 0.1773)	0.1226	0.0261	0.0258	0.0827	0.017
	FairGBM	0.6331 (0.5906 to 0.6773)	n.a.	0.0729 (0.0314 to 0.2283)	0.1454	0.0252	0.0273	0.1023	0.0161
	fairGATE	0.6482 (0.6065 to 0.6878)	n.a.	0.0635 (0.0308 to 0.2084)	0.1270	0.0258	0.0273	0.0801	0.0195
chdC_race: AUC, AOD; EOD/SPD	XGBoost (bias-naive)	0.7144 (0.6727 to 0.7521)	n.a.	0.0906 (0.0443 to 0.2415)	0.1692	0.0752	0.0479	0.1256	0.0267
	Fair Forests	0.7052 (0.6607 to 0.7427)	n.a.	0.0837 (0.0387 to 0.2417)	0.1653	0.0586	0.0352	0.1074	0.0196
	FairGBM	0.7071 (0.6634 to 0.7446)	n.a.	0.0939 (0.0444 to 0.2482)	0.1671	0.0738	0.0452	0.1278	0.0264
	fairGATE	0.7123 (0.6687 to 0.7508)	n.a.	0.0809 (0.0376 to 0.2425)	0.1589	0.0577	0.0357	0.1026	0.0224
	Bayes-optimal reference	0.7455 (0.7071 to 0.7811)	n.a.	0.1667 (0.0963 to 0.2664)	0.2293	0.1621	0.1022	0.2188	0.0899
chdC_race_reg: R2, CMPD; MPD/GLD	XGBoost (bias-naive)	0.4168 (0.3720 to 0.4592)	10.762 (10.300 to 11.219)	0.0909 (0.0476 to 0.2164)	0.2008	0.0662	0.0567	0.1433	0.0163
	Fair Forests	0.3550 (0.3221 to 0.3874)	11.318 (10.821 to 11.820)	0.0663 (0.0310 to 0.1601)	0.0944	0.0663	0.0339	0.0895	0.0222
	FairGBM	0.3912 (0.3537 to 0.4283)	10.996 (10.510 to 11.477)	0.0707 (0.0329 to 0.1846)	0.1272	0.0647	0.0363	0.0999	0.018
	fairGATE	0.4167 (0.3733 to 0.4583)	10.762 (10.283 to 11.282)	0.0739 (0.0327 to 0.2018)	0.1490	0.0644	0.0386	0.1047	0.019
	Bayes-optimal reference	0.4937 (0.4515 to 0.5318)	10.027 (9.583 to 10.465)	0.2124 (0.1338 to 0.2965)	0.3766	0.0574	0.1308	0.2798	0.109

Extended Data Table 3

Task	fairGATE	XGBoost (bias-naive)	Fair Forests	FairGBM
acsC_race	expert coupling; thresholded penalty; λ 4.0, λ_{thr} 0.0, μ 0.1; epochs 86 (46 to 196); configuration 28.0; lr 0.002; hidden_dim 64; dropout_rate 0.2; temperature 1	n_estimators 900; learning_rate 0.05; max_depth 4; min_child_weight 1; subsample 0.7; colsample_bytree 0.7; reg_lambda 5.0; reg_alpha 0.5	n_estimators 200; max_depth 10; min_samples_leaf 20; min_samples_split 20; max_features 0.3; lambda_fair 4.0	n_estimators 400; learning_rate 0.05; num_leaves 63; max_depth 8; min_child_samples 50; reg_lambda 1.0; lambda_lr 0.2; epsilon 0.02
acsC_race_reg	residual coupling; thresholded penalty; λ 4.0, λ_{thr} 0.0, μ 0.02; epochs 41.5 (14 to 79); configuration 27.0; lr 0.002; hidden_dim 32; dropout_rate 0.1; temperature 1	n_estimators 900; learning_rate 0.02; max_depth 6; min_child_weight 10; subsample 0.7; colsample_bytree 0.7; reg_lambda 0.5; reg_alpha 0.5	n_estimators 200; max_depth 10; min_samples_leaf 20; min_samples_split 20; max_features 0.3; lambda_fair 4.0	n_estimators 400; learning_rate 0.1; num_leaves 31; max_depth 6; min_child_samples 20; reg_lambda 1.0; lambda_lr 0.05; epsilon 0.02
acsC_sex	expert coupling; random_effects penalty; λ 16.0, λ_{thr} 0.0, μ 0.1; epochs 75.5 (60 to 96); configuration 47.0; lr 0.002; hidden_dim 128; dropout_rate 0.1; temperature 0.5	n_estimators 900; learning_rate 0.05; max_depth 4; min_child_weight 1; subsample 0.7; colsample_bytree 0.7; reg_lambda 5.0; reg_alpha 0.5	n_estimators 200; max_depth 8; min_samples_leaf 20; min_samples_split 40; max_features 0.3; lambda_fair 0.25	n_estimators 200; learning_rate 0.1; num_leaves 31; max_depth 8; min_child_samples 50; reg_lambda 1.0; lambda_lr 0.05; epsilon 0.0
acsC_race_sex	residual coupling; thresholded penalty; λ 1.0, λ_{thr} 0.0, μ 0.0; epochs 77 (11 to 174); configuration 41.0; lr 0.001; hidden_dim 32; dropout_rate 0.2; temperature 1	n_estimators 900; learning_rate 0.05; max_depth 4; min_child_weight 1; subsample 0.7; colsample_bytree 0.7; reg_lambda 5.0; reg_alpha 0.5	n_estimators 200; max_depth 8; min_samples_leaf 20; min_samples_split 40; max_features 0.3; lambda_fair 0.5	n_estimators 200; learning_rate 0.1; num_leaves 31; max_depth 8; min_child_samples 50; reg_lambda 1.0; lambda_lr 0.05; epsilon 0.0
acsTX_race	residual coupling; variance penalty; λ 64.0, λ_{thr} 0.0, μ 0.1; epochs 59.5 (25 to 138); configuration 17.0; lr 0.002; hidden_dim 64; dropout_rate 0.1; temperature 0.5	n_estimators 600; learning_rate 0.02; max_depth 6; min_child_weight 10; subsample 0.9; colsample_bytree 0.7; reg_lambda 0.5; reg_alpha 0.5	n_estimators 200; max_depth 6; min_samples_leaf 5; min_samples_split 20; max_features sqrt; lambda_fair 4.0	n_estimators 200; learning_rate 0.1; num_leaves 15; max_depth 8; min_child_samples 20; reg_lambda 0.0; lambda_lr 0.5; epsilon 0.0
acsTX_race_reg	residual coupling; thresholded penalty; λ 16.0, λ_{thr} 0.0, μ 0.02; epochs 33.5 (10 to 138); configuration 38.0; lr 0.002; hidden_dim 128; dropout_rate 0.1; temperature 0.5	n_estimators 300; learning_rate 0.02; max_depth 8; min_child_weight 5; subsample 0.7; colsample_bytree 0.7; reg_lambda 1.0; reg_alpha 0.0	n_estimators 200; max_depth 10; min_samples_leaf 10; min_samples_split 40; max_features 0.3; lambda_fair 0.25	n_estimators 400; learning_rate 0.1; num_leaves 31; max_depth 8; min_child_samples 20; reg_lambda 1.0; lambda_lr 0.02; epsilon 0.02
diabC_race	residual coupling; random_effects penalty; λ 1.0, λ_{thr} 0.0, μ 0.02; epochs 10 (10 to 21); configuration 25.0; lr 0.001; hidden_dim 32; dropout_rate 0.2; temperature 0.5	n_estimators 300; learning_rate 0.02; max_depth 4; min_child_weight 1; subsample 0.9; colsample_bytree 0.7; reg_lambda 0.5; reg_alpha 0.5	n_estimators 200; max_depth 4; min_samples_leaf 5; min_samples_split 20; max_features 0.3; lambda_fair 0.25	n_estimators 200; learning_rate 0.05; num_leaves 63; max_depth 4; min_child_samples 50; reg_lambda 0.0; lambda_lr 0.2; epsilon 0.02
diabC_race_reg	expert coupling; random_effects penalty; λ 64.0, λ_{thr} 0.0, μ 0.1; epochs 14 (10 to 58); configuration 7.0; lr 0.002; hidden_dim 32; dropout_rate 0.2; temperature 0.5	n_estimators 600; learning_rate 0.02; max_depth 6; min_child_weight 10; subsample 0.9; colsample_bytree 0.7; reg_lambda 5.0; reg_alpha 0.0	n_estimators 200; max_depth 8; min_samples_leaf 5; min_samples_split 20; max_features 0.3; lambda_fair 2.0	n_estimators 200; learning_rate 0.1; num_leaves 15; max_depth 8; min_child_samples 20; reg_lambda 0.0; lambda_lr 0.5; epsilon 0.0
diabC_race_sex	expert coupling; thresholded penalty; λ 1.0, λ_{thr} 0.0, μ 0.1; epochs 78 (38 to 111); configuration 1.0; lr 0.002; hidden_dim 128; dropout_rate 0.1; temperature 1	n_estimators 300; learning_rate 0.02; max_depth 6; min_child_weight 5; subsample 0.7; colsample_bytree 0.9; reg_lambda 5.0; reg_alpha 0.5	n_estimators 200; max_depth 8; min_samples_leaf 20; min_samples_split 20; max_features 0.3; lambda_fair 0.5	n_estimators 100; learning_rate 0.05; num_leaves 15; max_depth 8; min_child_samples 50; reg_lambda 0.0; lambda_lr 0.02; epsilon 0.0
chdC_race	residual coupling; variance penalty; λ 16.0, λ_{thr} 4.0, μ 0.1; epochs 19.5 (10 to 109); configuration 9.0; lr 0.002; hidden_dim 128; dropout_rate 0.1; temperature 1	n_estimators 300; learning_rate 0.02; max_depth 4; min_child_weight 1; subsample 0.9; colsample_bytree 0.7; reg_lambda 0.5; reg_alpha 0.5	n_estimators 200; max_depth 8; min_samples_leaf 20; min_samples_split 20; max_features sqrt; lambda_fair 4.0	n_estimators 400; learning_rate 0.1; num_leaves 31; max_depth 4; min_child_samples 20; reg_lambda 1.0; lambda_lr 0.2; epsilon 0.0
chdC_race_reg	expert coupling; random_effects penalty; λ 4.0, λ_{thr} 0.0, μ 0.1; epochs 42.5 (18 to 100); configuration 26.0; lr 0.001; hidden_dim 64; dropout_rate 0.1; temperature 1	n_estimators 600; learning_rate 0.05; max_depth 3; min_child_weight 5; subsample 0.7; colsample_bytree 0.7; reg_lambda 5.0; reg_alpha 0.5	n_estimators 200; max_depth 10; min_samples_leaf 10; min_samples_split 40; max_features 0.3; lambda_fair 0.25	n_estimators 400; learning_rate 0.1; num_leaves 31; max_depth 8; min_child_samples 20; reg_lambda 1.0; lambda_lr 0.02; epsilon 0.02

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFig1.png
- SupplementaryFig1.png
- SupplementaryFig3.png
- ExtendedDataFig1.png
- ExtendedDataFig1.png
- ExtendedDataFig3.png
- ExtendedDataFig4.png
- ExtendedDataFig5.png
- ExtendedDataFig6.png
- ExtendedDataFig7.png