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.

Fixed group shrinkage exposes the fairness accuracy Pareto frontier for cross ancestry polygenic risk scores

Shoukai Yu

shoukaiyu@sjtu.edu.cn

Hongqiao International Institute of Medicine, Shanghai Tongren Hospital and Clinical Research Institute, Shanghai Jiao Tong University School of Medicine

Lingmei Qian

Hongqiao International Institute of Medicine, Shanghai Tongren Hospital and Clinical Research Institute, Shanghai Jiao Tong University School of Medicine

Wei Xu

Tsinghua University

Research Article

Keywords: polygenic risk score, cross-ancestry prediction, fairness-accuracy trade-off, Pareto frontier, constrained optimisation, genetic epidemiology, polygenic prediction

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11570074/v1

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

Additional Declarations: No competing interests reported.

Fixed group shrinkage exposes the fairness accuracy Pareto frontier for cross ancestry polygenic risk scores

Shoukai Yu[1*], Lingmei Qian[1], and Wei Xu[2]

[1]Hongqiao International Institute of Medicine, Shanghai Tongren Hospital and Clinical Research Institute, Shanghai Jiao Tong University School of Medicine, Shanghai, China

[2]School of Economics and Management, Tsinghua University, Beijing, China

* Correspondence:

Shoukai Yu

shoukaiyu@sjtu.edu.cn

Keywords polygenic risk score; cross-ancestry prediction; fairness-accuracy trade-off; Pareto frontier; constrained optimisation; genetic epidemiology; polygenic prediction

Abstract

Background: Cross-ancestry polygenic risk scores (PRS) show marked prediction heterogeneity across populations. Existing polygenic prediction pipelines produce only one optimized model and cannot quantify achievable prediction accuracy under predefined limits on performance imbalance across population subgroups. We introduce a fixed-group shrinkage framework, named cluster_fixed, which uses known ancestry labels to map the Pareto frontier between prediction accuracy and cross-population fairness.

Results: Using genotypes from the 1000 Genomes project and simulated phenotypes, we varied the relative contributions of population-specific and shared genetic effects. When genetic effects are mostly shared across groups, cluster_fixed reduced the worst-case subgroup prediction disparity by 71.7%, with a 57.8% drop in global prediction accuracy (10 random replicates; $p < 10^{-5}$). This trade-off depends on the relative magnitude of population-specific genetic effects, and a regime with simultaneous gains in fairness and accuracy appears when population-specific effects are large. We replicated this pattern using height genetic effect estimates from the Pan-UK Biobank, with disparity reductions of 50.5–70.3% on chromosomes 1 and 21. In an independent validation with Pan-UKB β fixed as shared truth and disjoint train/validation individuals, the same fairness-accuracy trade-off was reproduced on both chromosomes (Δfairness $= -0.42$ and -0.37; Δ overall $= -0.055$ and -0.138; all 10 seeds negative, $p < 0.01$). Competing correction strategies could not achieve comparable fairness improvements.

Conclusions: This framework enables the construction of constraint-aware polygenic risk scoring models under chosen subgroup disparity limits. It reveals a full family of

candidate models instead of a single optimal estimate, supporting practical decision-making for multi-ancestry risk prediction. Further testing across different human traits and cohorts will clarify real-world utility.

Introduction

Polygenic risk scores (PRS) have become a central tool in translational genomics, with applications ranging from risk stratification in clinical trials to companion diagnostics in drug development. However, a well-documented limitation is that PRS trained primarily on European-ancestry cohorts transfer poorly to non-European populations [1,2]. For African-ancestry individuals, prediction accuracy frequently drops to a fraction of European-ancestry performance, an effect driven by differences in linkage disequilibrium (LD) architecture, allele frequency spectra, and the genetic architecture of complex traits across populations [3]. This "portability gap" has direct consequences for global clinical trial design: if PRS performance in an underrepresented ancestry is near-random, PRS-based enrichment strategies cannot be deployed, and the drug development pipeline becomes effectively restricted to European-ancestry populations [4].

Existing cross-ancestry PRS methods — including PRS-CSx and BridgePRS — were designed primarily to maximise global prediction accuracy by borrowing information across ancestries [5,6]. They output a single optimal model. In doing so, they treat ancestry-specific prediction gaps as an observed outcome rather than a controlled variable. As a result, when a regulator or a trial sponsor imposes a constraint on the maximum acceptable subgroup disparity — as is increasingly required for global multi-centre trials [7] — existing methods provide no systematic way to answer the

question: under this constraint, what is the best model, and at what accuracy cost? Nor do they identify which loci drive the disparity, which populations require additional sampling, or how many samples would be needed to close the gap. While several recent studies have proposed fairness-aware polygenic prediction, most do not expose a tunable Pareto frontier for explicit constraint-based decision-making [8].

We address this gap with a framework that reframes cross-ancestry PRS construction as constrained optimisation under a subgroup-disparity tolerance. Our contributions are threefold:

A bias-controllable fixed-group correction model. We exploit the fact that in real-world drug development, ancestry labels are known — not inferred. Rather than relying on unsupervised clustering of noisy effect estimates, we fix groups by ancestry label and estimate a shared component β_{shared} plus ancestry-specific deviations δ_g, with a shrinkage strength that is tuned as an explicit knob on the fairness-accuracy frontier.

An explicit Pareto frontier and its trade-off. On a realistic 1000 Genomes chr21 panel (1 210 quasi-independent SNPs, 17 sub-populations across AFR, EAS, EUR) with simulated phenotypes, we quantify the trade-off precisely: $\text{fairness}_{\text{dr2}}$ falls by 71.7% ($p = 1.30 * 10^{-6}$, $n = 10$ seeds) while $\text{overall}_{\text{r2}}$ falls by 57.8% ($p = 1.06 \times 10^{-6}, n = 10$ seeds). All 20 paired differences are in the same direction — the trade-off is an intrinsic property of the method under this data-generating regime.

A clean negative-result boundary. We systematically test and rule out alternative strategies: LD-panel GLS whitening (equivalent to meta in simulation;$R^2 = -0.203$ under real high-dimensional small-sample genotype data), prior-augmented clustering, Fst-initialised clustering, and site-adaptive τ^2 modulation. None improve fairness beyond the fixed-group approach, and several degrade cluster stability. These negative controls establish that the observed fairness gain is not trivially achievable by other

simple means.

Methods

2.1 Data: 1000 Genomes reference panel

We used the 1000 Genomes Project Phase 3 chr21 genotypes (GRCh37; ALL.chr21.phase3_shapeit2_mvncall_integrated_v5b.20130502.genotypes.vcf.gz, downloaded from the EBI FTP mirror) [9]. Ancestry labels were obtained from integrated_call_samples_v3.20130502.ALL.panel. We restricted to three super-populations: African (AFR; 7 sub-populations, 661 samples), East Asian (EAS; 5 sub-populations, 504 samples), and European (EUR; 5 sub-populations, 503 samples), yielding 17 sub-population groups in total.

2.2 Genotype preprocessing

We considered all biallelic SNPs with MAF ranging from 0.01 to 0.99 and selected 3,000 variants spanning chromosome 21 for downstream processing. Each variant was standardized within its sample group to have zero mean and unit variance. Because all 17 subpopulations were drawn from the same 1000G variant set, the 3,000 variants were common by construction; no intersection step was required. We then performed LD pruning at $r^2 > 0.5$, using the largest AFR subpopulation as the reference panel, yielding 1,210 quasi-independent SNPs [10]. Variants monomorphic across the full sample set were removed during pruning. This variant selection pipeline is specific to the main baseline simulation; the GWAS-informed simulation adopts a separate variant matching and filtering workflow (Section 2.8).

2.3 Train/test split and monomorphic-column handling

Each group was split 50/50 into training and test sets. Because MAF ≥ 0.01 combined with small per-group sample sizes (30–57 training samples) can produce columns that are monomorphic within a split, we adopted per-site zero-filling: for each site, columns with standard deviation $< 10^{-8}$ in the training set were set to 0 in the training matrix,

and likewise for the test matrix. Columns were not removed, preserving a common SNP index across all 17 groups. This procedure affected approximately 1 700 column-instances per split on average and kept the matrix dimension fixed at 1 210 SNP columns throughout.

2.4 Phenotype simulation

Because the 1000 Genomes Project does not provide individual-level phenotypes, we simulated phenotypes under a generative model containing a shared cross-site genetic component plus group-specific deviations [11].

$$y_k = Z_k(\beta_{\text{shared}} + \delta_k) + \varepsilon_k, \quad \varepsilon_k \sim N(0, I)$$

where Z_k is the standardised genotype matrix for group k, $\beta_{\text{shared}} \sim N(0, \sigma_\beta^2)$ is shared across all 17 groups, $\delta_k \sim N(0, \sigma_\delta^2)$ is group-specific, and ε_k represents independent environmental noise.

- Baseline regime. For the primary analysis, we used $\sigma_\beta = 0.05$ and $\sigma_\delta = 0.01$ (shared-dominated; $\sigma_\delta/\sigma_\beta = 0.2$) consistent with a scenario in which most polygenic signal is shared across ancestries [12].
- Attribution regimes. To test whether the trade-off direction depends on the effect architecture, we fixed $\sigma_\beta = 0.05$ and scanned $\sigma_\delta \in \{0.01, 0.03, 0.05\}$, corresponding to $\sigma_\delta/\sigma_\beta \in \{0.2, 0.6, 1.0\}$. All other simulation parameters were held fixed.

Notation note. In our data-generating model, each of the 17 sub-populations represents a site, with true site-specific deviations $\boldsymbol{\delta}_k$ defined at the sub-population level. In the cluster_fixed estimation model, sites are pre-grouped into three broad ancestry super-populations (AFR, EAS, EUR). The fitted deviation parameters $\boldsymbol{\delta}g$ are ancestry-group-specific quantities, estimated jointly within each group rather than

aggregated from per-site estimates.

2.5 Models

Ridge meta (baseline). A single shared β was estimated by Ridge regression ($\lambda = 100$) on the pooled training data of all 17 sites, and applied uniformly to all test samples.

cluster_fixed (proposed). Using known, externally-provided ancestry labels, samples were assigned to three fixed super-population groups (AFR, EAS, EUR). Within each group, shared genetic effects and group-specific deviations are jointly estimated; group-specific deviations are regularised via group-level shrinkage controlled by a hyper-parameter denom. Unlike PRS-CSx [5] and BridgePRS [6], which optimise exclusively for global prediction accuracy and return a single model, cluster_fixed explicitly enables tuning of group-level bias constraints during model estimation, exposing the fairness-accuracy trade-off via a controllable hyper-parameter.

The cluster_fixed framework implements group-specific shrinkage controlled by hyper-parameter denom, which modulates relative scaling of effect estimates across ancestry clusters. Systematically scanning denom traces out the Pareto frontier, quantifying the accuracy cost required to satisfy predefined subgroup-disparity constraints. Conventional fairness-correction approaches produce discrete single-point models and lack this continuous-tuning capability. Full mathematical formulation of the cluster_fixed objective function, prior specifications and optimisation details are provided in Appendix Note A.

We also consider an empirical Bayes polygenic scoring method (EB-PS, Method B) as a competing approach, whose full method development and simulation evaluations are reported in a separate parallel submission. EB-PS differs from cluster_fixed in that it estimates SNP-specific priors for ancestry-specific effect deviations from external multi-ancestry GWAS summary statistics, rather than from training data alone. We

include this method here for completeness; the primary evaluation in this work focuses on cluster_fixed versus the Ridge meta baseline.

2.6 Evaluation metrics

overall_r2: mean test-set R^2 across all 17 sites.

fairness_dr2: Maximum minus minimum test-set R^2 across the 17 sub-populations. Lower values indicate smaller worst-case prediction gaps between subgroups. We adopt this worst-case disparity metric motivated by regulatory interest in biomarker performance across under-represented subgroups [1]. As a complementary metric for future studies, we recommend also reporting the standard deviation of sub-population R^2 values to summarise disparity beyond simple range [7].

ARI (Adjusted Rand Index): Pair-wise agreement of cluster assignments across random seeds. For cluster_fixed, ARI = 1.0 by construction because ancestry group labels are fixed and provided as input.

2.7 Experimental protocol

We ran both models over 10 random seeds (42, 123, 2024, 7, 999, 11, 22, 33, 44, 55). For each seed, phenotype generation and train-test partitioning were resampled. Paired t-tests $(n = 10)$ were used to compare cluster_fixed against Ridge meta for both evaluation metrics. For the attribution scan: regimes A and B used 10 seeds each; regime C used 5 seeds; paired t-tests were computed within each regime.

2.8 GWAS-informed simulation of shared genetic effects

To evaluate robustness of the observed accuracy-fairness trade-off under empirically observed genetic effect sizes, we constructed simulations in which the shared genetic component β_{shared} was taken from published Pan-UKB summary statistics for standing height $(N \approx 700\,000)$.

We restricted variants to chromosome 21 and to variants present in both the 1000G

WGS panel and Pan-UKB summary statistics, matched by genomic position. We then performed LD pruning $(r^2 > 0.5)$ to retain quasi-independent SNPs, yielding 212 variants all matching Pan-UKB variants.

Raw GWAS effect sizes were rescaled such that $\mathrm{std}(\beta_{\mathrm{shared}}) = 0.05$, matching the baseline simulation to ensure comparable polygenic heritability between GWAS-informed and baseline settings (scale factor = $0.05 / 0.0146 \approx 3.43$). In the GWAS-informed setting, β_{shared} was fixed across all 10 seeds (sampled once from Pan-UKB), while residual group-specific components δ and non-genetic noise terms were redrawn per seed following identical generative rules and hyper-parameters as the main simulation. We ran 10 independent random seeds for both the meta baseline and cluster_fixed.

Paired two-sided t-tests $(n = 10$ seeds, $df = 9)$ compared overall R^2 and fairness_dr2 between methods across matched random seeds.

Variant count differs between baseline (1 210 quasi-independent SNPs) and GWAS-informed simulation (212 SNPs) due to differences in variant density and variant overlap between Pan-UKB summary statistics and the 1000G WGS panel. Direct numerical comparison of absolute R^2 magnitudes between settings is therefore not valid; only qualitative direction-of-effect comparisons are interpreted.

2.9 Independent validation with disjoint train/validation individuals

To test whether the GWAS-informed trade-off generalises beyond the training individuals, we performed an independent validation on chr21 and chr1. For each chromosome, the full 1000G sample was first subject to the same MAF, LD-pruning and variant-matching pipeline as the corresponding GWAS-informed analysis (Section 2.8), yielding 212 SNPs on chr21 and 363 SNPs on chr1. Each of the 17 site_id groups was then split 50/50 into disjoint training and validation subsets (total: 827 training /

841 validation individuals; per-site intersection = 0). The Pan-UKB height β was fixed as the shared truth across all seeds (fix_beta_seed = 42), while site-specific deviations δ and environmental noise were redrawn per seed. cluster_fixed (denom = 0.05) and Ridge meta were fitted on the training subset only, and evaluated on the held-out validation subset. As a reproducibility audit, we independently recomputed per-site PRS from the dumped per-site trained β on the validation subset and compared against the in-run evaluation; the maximum absolute difference was 1.2×10^{-15} (chr21) and 1.3×10^{-15} (chr1), confirming that the dumped β, the validation split, and the validation phenotypes were mutually aligned. All other settings (Ridge λ = 100, 10 random seeds, paired two-sided t-tests with df = 9) were identical to Section 2.7.

Results

3.1 Overall performance

Across the 17 sub-populations and 10 random seeds, cluster_fixed traded prediction accuracy for ancestry fairness (Table 1**, Figure 1**).

Table 1. Ten-seed paired comparison (mean ± SD).

Metric	Ridge meta	cluster_fixed	Δ (paired)	Relative	t (df=9)	p-value
overall_r2	0.2314 ± 0.038	0.0976 ± 0.017	−0.1338	−57.82%	−11.56	1.06×10^{-6}
fairness_dr2	0.5974 ± 0.124	0.1692 ± 0.043	−0.4283	−71.69%	−11.28	1.30×10^{-6}

Figure 1. Main-experiment fairness-accuracy trade-off. Ridge meta (blue circle) versus cluster_fixed (orange triangle) on the 1000G chr21 panel with 1,210 quasi-independent SNPs and simulated phenotypes. The arrow indicates the direction of change from meta to cluster_fixed. Axes: overall test-set R^2 (higher is better) versus

worst-case cross-sub-population R^2 disparity, fairness_dr2 (lower is fairer). Both differences are statistically significant (paired t-test, n = 10 seeds, $p < 10^{-5}$).

cluster_fixed reduced the worst-case cross-sub-population R^2 gap by 71.7% (0.5974 → 0.1692) at the cost of a 57.8% reduction in overall R^2 (0.2314 → 0.0976). Both differences were highly significant $(p < 10^{-5})$. These quantitative effect sizes are tied to our simulation's data-generating mechanism [13] and should not be treated as universal constants across real-world traits.

3.2 Direction consistency across seeds

For all 10 seeds, the paired differences were negative for both metrics: overall_r2 differences ranged from −0.069 to −0.174, and fairness_dr2 differences ranged from −0.274 to −0.592. 20 of 20 paired differences were in the same direction — no seed reversed the trade-off, confirming this is an intrinsic property of the method under this data-generating regime, not a random fluctuation.

3.3 Confidence intervals

95% confidence intervals on the paired differences (t-distribution, $df = 9$):

- overall_r2: [−0.160, −0.108], entirely below 0
- fairness_dr2: [−0.514, −0.342], entirely below 0

Neither CI crosses zero. Bootstrap resampling$(B = 10\,000)$ gives slightly narrower intervals (overall_r2: [−0.154, −0.112]; fairness_dr2: [−0.498, −0.358]), consistent with the t-based result. We report the more conservative t-based intervals.

3.4 Robustness across seeds (5 vs 10)

When using only 5 seeds in early exploratory runs, we observed relative changes of −60.1% (overall_r2) and −72.4% (fairness_dr2). Extending to 10 seeds updated effect sizes to −57.8% and −71.7% respectively (deviations ≤ 2%), while statistical significance improved by three orders of magnitude (overall_r2 p from 0.0011 → 1.06×10^{-6}; fairness_dr2 p from 0.0019 →1.30* 10^{-6}.

3.5 Attribution: the trade-off direction is governed by the site-specific-to-shared ratio

To test whether the observed trade-off is intrinsic to the method or a property of the data-generating regime, we fixed $\sigma_\beta = 0.05$ and scanned $\sigma_\delta \in \{0.01, 0.03, 0.05\}$ (regimes A, B, C). Within each regime, we ran both models with paired seeds and computed paired t-tests on the same seeds (Table 2, Figure 2).

In this work, we define two distinct operating regimes: (1) trade-off regime: cluster_fixed improves subgroup-level fairness $(\Delta\text{fairness_dr2} < 0)$ at the expense of reduced global prediction accuracy ($\Delta\text{overall_r2} < 0$); (2) win-win regime: both fairness and overall prediction accuracy are improved $(\Delta\text{fairness_dr2} < 0, \Delta\text{overall_r2} > 0)$. Across all tested genetic architectures, cluster_fixed consistently reduces worst-case disparity $(\Delta\text{fairness_dr2} < 0)$; whether global accuracy rises or falls depends on $\sigma_\delta/\sigma_\beta$.

Table 2. Attribution scan: Δ (cluster_fixed − meta) per regime.

$Regime$	(σ_δ)	$(\sigma_\delta/\sigma_\beta)$	n	$\Delta overall_r2$	$\Delta fairness_d r2$	$Verdict$	$p(overall_r2)$
A	0.01	0.2	10	−0.1338	−0.4283	trade-off	(1.06×10^{-6})

$Regime$	(σ_δ)	$(\sigma_\delta/\sigma_\beta)$	n	$\Delta overall_r2$	$\Delta fairness_d r2$	$Verdict$	$p(overall_r 2)$
B2	0.03	0.6	10	+0.0563	−0.6462	win-win	0.0038
C2	0.05	1.0	5	+0.2826	−0.7275	win-win	1.50×10^{-4}

Figure 2. Attribution of prediction–fairness trade-off across three genetic architectures. Each panel plots overall prediction R^2 (x-axis, higher = better) versus cross-population disparity metric $fairness_dr2$ (y-axis, lower = fairer). Blue circles: naive meta-analysis estimator. Orange triangles: cluster_fixed estimator. Grey arrows indicate the change from meta to cluster_fixed. Panel A: $\sigma_\delta = 0.01$ $(\sigma_\delta/\sigma_\beta = 0.2)$, showing a prediction-fairness trade-off. Panel B: $\sigma_\delta = 0.03(\sigma_\delta/\sigma_\beta = 0.6)$, win-win behaviour. Panel C: $\sigma_\delta = 0.05(\sigma_\delta/\sigma_\beta = 1.0)$, win-win behaviour.

Δoverall_r2 flipped sign from -0.1338 $(p = 1.06 \times 10^{-6})$ $to + 0.0563(p = 0.0038) to + 0.2826(p = 1.50 \times 10^{-4})$ — monotone increasing across the three regimes — while Δfairness_dr2 remained negative throughout. Across all three genetic architectures, Δfairness_dr2 remained consistently negative, indicating that cluster_fixed consistently reduced worst-case subpopulation disparity across the regimes tested. Whether global overall accuracy is sacrificed (trade-off) or also improved (win-win) is governed by the ratio of ancestry-specific to shared genetic effects, rather than being an intrinsic property of the estimator itself: when the shared component dominates, Ridge meta achieves higher accuracy; as the ancestry-specific component grows, cluster_fixed becomes Pareto-superior on both axes, consistent with empirical observations that optimal multi-ancestry PRS strategies are trait-dependent [13].

3.6 Optimal shrinkage depends on the effect architecture

We next asked whether the optimal shrinkage strength (the denom hyperparameter) is universal or depends on the effect architecture. For each of the three regimes, we scanned denom $\in$ {0.005, 0.01, 0.02, 0.05, 0.1, 0.3} and identified the value that maximised overall_r2 (Table 3, Figure 3).

Table 3. Optimal denom (maximising overall_r2) per regime.

Regime \(\sigma_\delta\)	optimal denom	
A	0.01	0.05
B2	0.03	0.02
C2	0.05	0.005

Figure 3. overall_r2 (left) and fairness_dr2 (right) as functions of denom, one curve per regime. The optimal denom for overall_r2 shifts left monotonically as σ_δ increases.

The optimal denom shifts left monotonically (0.05 → 0.02 → 0.005) as the site-specific effect grows. This shows that no single shrinkage strength is universally optimal; the hyperparameter must be re-scanned for each deployment trait according to its ancestry heterogeneity. We also observe that overall_r2 and fairness_dr2 have opposite preferences for denom: overall_r2 prefers stronger shrinkage (small denom), whereas fairness_dr2 prefers stronger shrinkage (large denom), which constitutes an additional trade-off dimension internal to the method.

3.7 Baseline cross-super-population R^2 gap under full 1000G sample

Under the full 1000G sample (no subsampling) and 10 random seeds, the mean test-set

(R^2) per super-population was 0.106 ± 0.015 for AFR $(n_{\text{subpops}} = 7), 0.094 \pm 0.012$for EAS $(n_{\text{subpops}} = 5)$, and 0.118 ± 0.013 for EUR $(n_{\text{subpops}} = 5)$. Among super-populations, EAS showed the lowest average performance, producing the largest pairwise gap versus EUR of −0.0237. Additional pairwise differences were AFR − EUR = −0.0115 and AFR − EAS = +0.0121; that is, AFR outperformed EAS despite both being underrepresented relative to EUR. These numbers define the empirical baseline against which the subsampling experiment below is compared (Table 4).

Table 4. Baseline per-super-population test-set R^2 under full 1000G sample (mean ± SD across 10 seeds).

Super-population	n sub-pops	mean(R^2)	SD
AFR	7	0.106	0.015
EAS	5	0.094	0.012
EUR	5	0.118	0.013
AFR − EUR	—	−0.0115	—
EAS − EUR	—	−0.0237	—
AFR − EAS	—	+0.0121	—

3.8 GWAS-informed simulation

To assess whether the trade-off depends on the assumed effect-size distribution, we repeated the main comparison using real effect sizes from a published GWAS of height (Pan-UKB, $N \approx 700\,000$) as the shared genetic component β_{shared} (Methods 2.8). GWAS-derived effects were rescaled to the same standard deviation as the baseline

simulation ($\sigma_\beta = 0.05$) to preserve comparable polygenic heritability. After LD pruning and intersection with Pan-UKB variants, 212 quasi-independent SNPs on chr21 were retained (vs. 1 210 in the baseline).

The direction of the trade-off was reproduced. cluster_fixed reduced fairness_dr2 by 0.353 (0.699 → 0.347, −50.46%) at a cost of 0.052 in overall_r2 (0.126 → 0.074, −41.31%). Paired t-tests across 10 random seeds showed that both differences were statistically significant (fairness_dr2: $t = -11.38$, $\mathrm{df} = 9, p = 1.20 \times 10^{-6}$, 10/10 paired differences negative; overall_r2: $t = -6.62$, $\mathrm{df} = 9, p = 9.71 \times 10^{-5}$, 9/10 paired differences negative). The magnitudes were smaller than in the baseline simulation (Δfairness −0.353 vs. −0.428; Δoverall −0.052 vs. −0.134), consistent with the smaller SNP set (212 vs. 1 210 after LD pruning) and the different effect-size distribution of height. The direction of the effect, however, is independent of the assumed effect-size architecture.

Cross-chromosome consistency. We repeated the GWAS-informed analysis on chromosome 1 (the largest autosome), retaining 482 quasi-independent SNPs after LD pruning and variant matching (100% matched to Pan-UKB). Both components of the trade-off were reproduced, with larger effect magnitudes than observed on chr21: cluster_fixed reduced fairness_dr2 by 0.435 (0.618 → 0.184, −70.3%, paired t = −11.22, df = 9, p = 1.4×10^{-6}, 10/10 paired differences negative) and overall_r2 by 0.110 (0.174 → 0.064, −63.3%, paired t = −10.51, df = 9, p = 2.4×10^{-6}, 10/10 paired differences negative),as shown in Table 5. The direction of the trade-off is therefore consistent across chromosomes with distinct gene density and LD architecture (chr1 vs chr21), providing cross-chromosomal evidence that the observed behaviour is not an artefact of any single genomic region (Table 5).

Table 5. Changes in R^2 and fairness after cluster-fixed adjustment

Chromosome	No. SNPs	ΔR^2 (overall)	ΔFairness (DR2)	P (overall)	P (fairness)
chr21	212	−0.052	−0.353	<0.001***	<0.001***
chr1	482	−0.110	−0.435	<0.001***	<0.001***

Table legend: Δ = cluster-fixed model − Ridge meta model. Negative ΔR^2 (overall) indicates reduced prediction accuracy; negative ΔFairness (DR2) indicates improved worst-case subgroup fairness. P values are from paired two-sided t-tests (df = 9). *** p < 0.001. DR2, disparate impact ratio based on R^2.

3.9 Independent validation on disjoint individuals

The GWAS-informed analyses above use the same individuals for fitting and evaluation, via an internal 50/50 split. To rule out the possibility that the observed trade-off is an artefact of that shared-sample design, we repeated the analysis with fully disjoint training and validation individuals (Section 2.9). Training and validation sets were non-overlapping per site (827 vs 841 individuals; intersection = 0), and the Pan-UKB β was fixed as shared truth across seeds while δ and noise were redrawn per seed.

The trade-off was reproduced on both chromosomes (Table S2, Figure S2). On chr21 (212 SNPs), cluster_fixed reduced fairness_dr2 by 0.420 (Δfairness = −0.420 ± 0.182, $t = -7.30$, df = 9, $p = 4.6\times10^{-5}$) at a cost of 0.055 in overall_r2 (Δoverall = −0.055 ± 0.037, $t = -4.66$, $p = 1.2\times10^{-3}$); all 10 seeds were negative for both metrics. On chr1 (363 SNPs), the effect was larger and more stable: Δfairness = −0.374 ± 0.121 (t =

−9.80, p = 4.2×10⁻⁶) and Δoverall = −0.138 ± 0.026 (t = −17.09, p = 3.6×10⁻⁸), again with 10/10 seeds negative. The independent-reproduction audit (Section 2.9) agreed with the in-run evaluation to within 1.3×10⁻¹⁵.

Two points distinguish this validation from the main GWAS-informed analysis. First, the validation individuals are entirely disjoint from those used to fit β, so the reported metrics are out-of-sample by construction. Second, the SNP set and the truth-redraw scheme differ between designs (fixed vs. per-seed β); the chr21 SNP set is identical across designs, while the chr1 SNP set differs (see Table S2 footnote). Absolute Δ magnitudes are therefore not directly comparable to Section 3.8, and we interpret only the direction and statistical significance. The direction of the trade-off is unchanged under this stricter design, and the magnitudes remain of the same order.

Discussion

4.1 Principal findings

On the 1000G chr21 panel with realistic LD structure and simulated phenotypes, fixing ancestry groups by their known labels (cluster_fixed) and shrinking ancestry-specific deviations with a tunable denom produces a large and statistically robust reduction in worst-case cross-sub-population R^2 disparity --- from 0.597 to 0.169 (−71.7%) --- at the cost of a substantial reduction in overall predictive accuracy --- from 0.231 to 0.098 (−57.8%).

Both effects are highly significant under a paired t-test ($p < 10^{-5}$, n = 10 seeds), and the direction is 100% consistent across all seeds within this primary simulation regime.

Under this data-generating regime, the observed trade-off is not a marginal or unstable effect but a robust structural property of the estimator (**Figure 1**). We stress these percentage trade-off magnitudes should not be generalised directly to real traits [13]. This trade-off was further confirmed under real Pan-UKB height effect sizes on both chr21 and chr1 (Section 3.8), and independently reproduced with fully disjoint training and validation individuals on both chromosomes (Section 3.9; Δoverall = −0.055 / −0.138, Δfairness = −0.420 / −0.374, all 10 seeds negative, $p < 0.01$). This indicates the direction is robust to both the assumed effect-size distribution and the shared-sample design of the main analysis.

4.2 Why accuracy drops: a structural explanation

The simulated phenotype in the primary analysis is dominated by a shared cross-ancestry component$(\sigma_\beta = 0.05)$ with a smaller site-specific component $(\sigma_\delta = 0.01)$. Ridge meta pools data across all 17 subpopulations to fit one single coefficient vector, which aligns well with the true shared-dominated generative architecture and yields higher global accuracy. By contrast, cluster_fixed estimates a shared baseline plus additional group-specific deviation parameters $(\boldsymbol{\delta_g})$. When true genetic effects are overwhelmingly shared across ancestries, adding extra group-specific parameters introduces estimation variance. Even though the shared component $\boldsymbol{\beta}_{\text{shared}}$ is still explicitly estimated, the increased model variance is the primary driver of the reduced overall test-set performance in the shared-dominated, limited-sample regime studied here.

Our attribution experiment (Section 3.5) shows that this is not an inherent property of cluster_fixed, but a function of the effect architecture. Holding σ_β fixed at 0.05 and

increasing σ_δ from 0.01 to 0.05, we observed a monotone sign flip in Δoverall_r2 (from −0.134 to +0.283), while Δfairness_dr2 remained negative throughout. The trade-off thus flips to a win-win as soon as the site-specific component reaches $\sigma_\delta/\sigma_\beta \gtrsim 0.6$.

This identifies the practical boundary of the method. For traits whose genetic architecture is dominated by shared cross-ancestry effects ($\sigma_\delta/\sigma_\beta \lesssim 0.2$), fixed-group correction will cost accuracy; for traits with substantial ancestry-specific effects $\left(\sigma_\delta/\sigma_\beta \gtrsim 0.6\right)$, it becomes Pareto-superior. Candidate traits in the latter category include drug-metabolising enzymes (e.g. CYP family) and immune-related loci (e.g. HLA), where ancestry-specific allele frequencies and effect sizes are well documented [12].

Estimating $\sigma_\delta/\sigma_\beta$ for a target trait from external GWAS summary statistics is therefore a natural prerequisite to deployment, and is the subject of ongoing work. Our attribution experiment demonstrates that $\sigma_\delta/\sigma_\beta$, rather than site-grouping granularity, governs the direction of the fairness-accuracy trade-off.

4.3 Practical implication: constrained optimal PRS

For drug developers and trial sponsors facing regulatory requirements on subgroup disparity, the practical object of interest is not the frontier itself but the constrained optimum: given a maximum acceptable disparity (e.g. fairness_dr2 ≤ 0.10), which model maximises overall accuracy, and at what cost? Our framework directly answers this question. In the present dataset, reducing the worst-case subgroup gap from 0.597 (unconstrained) to 0.169 costs 57.8% of overall accuracy; intermediate points along the frontier — achievable by varying the shrinkage denominator — provide a continuum

of compliant models. This is the sense in which the proposed approach is decision-theoretic: it does not return a single answer, but a family of answers parameterised by an explicit constraint that a regulator, a trial statistician, or a commercial team can supply [7].

4.4 Why competitors do not directly provide this

PRS-CSx and BridgePRS optimise global accuracy. Their internal hyperparameters are not exposed as subgroup-disparity controls, and their output is a single model, not a Pareto frontier. As a result, when a user asks "can you achieve disparity ≤ 0.05 at what cost?", these methods provide no systematic answer. Our approach does not replace these methods in the point-prediction setting; rather, it exposes the trade-off that their optimisation silently makes on the user's behalf [5,6].

4.5 Negative results and what they rule out

We systematically tested alternative strategies: LD-panel GLS whitening, prior-augmented clustering, Fst-initialised clustering, and site-adaptive (τ^2) modulation. None matched cluster_fixed's fairness gain, and several degraded cluster stability (e.g. Fst-initialised KMeans reduced the cross-seed ARI from 0.641 to 0.375). The GLS approach, which is a natural default in cross-ancestry meta-analysis, was equivalent to plain meta in simulation ($p = 0.075$) and failed outright on real genotype data ($R^2 = -0.203$) under the high-dimensional small-sample regime. These negative controls establish that the fairness improvement we report is not trivially achievable via generic statistical corrections [10].

4.6 Practical interpretation of AFR subsampling

The subsampling curve in Appendix S2 provides a quantitative answer to the question posed at the start of this work — given the current 1000G AFR panel (661 samples), how much would additional AFR samples help? Under our estimator configuration, the

marginal gain in overall R^2 is large up to $(n_{\mathrm{AFR}} \approx 400)$ and near-zero beyond that, while the fairness gap is essentially flat for $(n_{\mathrm{AFR}} \geq 300)$. Two implications follow.

First, at the 1000G scale, the AFR sample is already in the plateau regime for fairness. Doubling AFR to $(n_{\mathrm{AFR}} \approx 1300)$ is expected (from the trend in Table S2) to further reduce the AFR–EUR gap only modestly, unless the additional samples come from previously unsampled sub-populations with distinct genetic architectures [3].

Second, the value of additional AFR samples is dominated by the first few hundred. This is consistent with the classical observation that PRS accuracy scales with training sample size but with strongly diminishing returns for polygenic traits of moderate heritability [14]. The precise location of the plateau is data and trait-dependent, and our empirical curve provides the 1000G-specific estimate.

We emphasise that these conclusions are specific to the simulated phenotype model used here. Whether the same plateau location holds for real disease traits — where effect-size distributions, heritability, and ancestral architectures differ — can only be assessed with trait-specific GWAS summary statistics, which we leave to future work.

The GWAS-informed analysis (Results 3.8) confirms that the trade-off direction is not an artefact of the random effect-size simulation: the same qualitative trade-off was reproduced on both chr21 and chr1 using real Pan-UKB effect sizes, with chr1 showing even larger effect magnitudes (−63.3% overall, −70.3% fairness).

4.7 Limitations

Several contextual boundaries frame opportunities for future investigation. First, all phenotypes analysed here are simulation-derived. Although we leverage real 1000G linkage-disequilibrium and population-structure profiles, genetic-effect architectures are artificially specified. The GWAS-informed analysis uses real Pan-UKB effect sizes on chr21 and chr1 but retains simulated residual terms, and therefore serves as an effect-

size distribution robustness check rather than a clinical validation. The independent validation (Section 3.9) strengthens this by using fully disjoint training and validation individuals and a fixed shared β, but the residual terms and phenotypes remain simulated; it therefore demonstrates out-of-sample reproducibility of the trade-off direction, not clinical validity on real phenotypes. Our findings characterise estimator behaviour under well-controlled genotype conditions rather than real-world clinical performance [11,12]. Validation on independent biobank cohorts remains future work.

Second, our implementation adopts discrete continental-level ancestry groupings, a practical choice motivated by common clinical-trial metadata conventions. This grouping does not capture fine-scale within-continent sub-population heterogeneity; extensions to continuous ancestry representations remain open [2].

Three secondary boundaries should also be noted. (i) Primary evaluations use a chromosome-21 test-bed; scaling to full-genome, high-dimensional data remains future work [10]. (ii) AFR subsampling is constrained to n_{AFR} ≤ 661; we do not extrapolate beyond this observed range. (iii) The GWAS-informed simulation uses a reduced variant set relative to the baseline; direct quantitative comparison of R^2 magnitudes across configurations is not interpretable, and only qualitative effect-direction comparisons are valid.

4.8 Comparison with existing fairness-aware polygenic prediction methods

For instance, FairPRS uses invariant risk minimization to learn population-invariant predictors for admixed samples; it targets a single optimised model rather than generating a family of models for regulatory constraint setting [8]. Our fixed-group shrinkage differs by taking known discrete ancestry labels as input; this brings practical convenience for clinical trial settings where ancestry metadata is available, yet also

inherits well-known limitations of discrete ancestry grouping, which cannot fully capture the continuous nature of genetic ancestry in real cohorts [15]. Future extensions could generalise the shrinkage framework to continuous ancestry representations. Benchmarking against real-world multi-biobank datasets will be essential to contrast performance across these competing fairness paradigms [16,17]. Observational polygenic-score studies in diverse cohorts further highlight the challenges of discrete-ancestry modelling for complex-trait prediction [18,19].

4.9 Outlook: from point prediction to decision support

Polygenic risk scores (PRS) have become a central tool in translational genomics, with applications ranging from risk stratification in clinical trials to companion diagnostic drug development workflows [23–26]. The practical value of constrained-optimal PRS lies not in point prediction itself, but in the decision infrastructure it provides: given a regulatory or commercial constraint on subgroup disparity, what is the best achievable accuracy, and what is the cost of tightening the constraint?

Two building blocks from this work support such a workflow. The attribution experiment (Section 3.5) tells the user whether a target trait falls within a regime where the fixed-group estimator is beneficial ($\sigma_\delta/\sigma_\beta > \gtrsim 0.6$) or where it costs accuracy ($\sigma_\delta/\sigma_\beta \lesssim 0.2$). The hyperparameter interaction (Section 3.6) tells the user how to configure shrinkage strength for a given trait. Together, these enable a four-stage deployment workflow:

Estimate the trait-specific $\sigma_\delta/\sigma_\beta$ ratio from multi-ancestry GWAS summary statistics. Classify whether the trait lies in a trade-off or win-win regime. Sweep the shrinkage hyper-parameter denom to construct the full Pareto frontier. Select the model that satisfies a user-specified maximum acceptable fairness_dr2.

Practical algorithms for estimating $\sigma_\delta/\sigma_\beta$ from summary statistics remain to be developed, and empirical calibration across diverse traits is needed. Future work could also leverage fine-mapping resources and large multi-ancestry GWAS catalogues to improve real-world deployment of constrained-optimal PRS workflows [20–22]. Exploratory AFR subsampling results are reported in Appendix S2; locus-level disparity attribution for real-world traits remains an open direction [7].

Conclusions

Our framework exposes the Pareto frontier as a family of models parameterised by an explicit disparity constraint, rather than a single point estimate. Key limitations include reliance on simulated phenotypes, single-chromosome scope in the primary analysis (partially mitigated by chr21 + chr1 GWAS-informed and independent-validation results), and dependence on unknown real-world genetic architectures [12,7]. Future work should validate this decision-support paradigm using multi-ancestry GWAS summary statistics from real biobank cohorts.

References

[1] Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Current clinical use of polygenic scores will risk exacerbating health disparities. Nat Genet. 2019;51(4):584-591. doi:10.1038/s41588-019-0379-x

[2] Peterson RE, Kuchenbaecker K, Walters RK, et al. Genome-wide association studies in ancestrally diverse populations: opportunities, methods, pitfalls, and recommendations. Cell. 2019;179(3):589-603. doi:10.1016/j.cell.2019.10.004

[3] Cavazos-Roark T, Witte JS. Inclusion of variants discovered from diverse populations improves polygenic risk score transferability. Eur J Hum Genet. 2021;29(12):1832-1840. doi:10.1038/s41431-021-00975-7

[4] Sirugo G, Williams SM, Tishkoff SA. The missing diversity in human genetic studies. Cell. 2019;177(1):26-31. doi:10.1016/j.cell.2019.02.048

[5] Ruan Y, Lin YF, Chen CY, et al. PRS-CSx: improving cross-ancestry polygenic prediction by leveraging shared effect-size distributions. Am J Hum Genet. 2022;109(4):626-641. doi:10.1016/j.ajhg.2022.02.007

[6] Wang Y, Tsuo K, Kanai M, et al. BridgePRS improves portability of polygenic risk scores across diverse populations. Nat Commun. 2023;14(1):771. doi:10.1038/s41467-023-36225-2

[7] Márquez-Luna A, Howe LJ, Pairo-Castineira E. Bridging the diversity gap: analytical and study-design considerations for improving the accuracy of trans-ancestry genetic prediction. Brief Bioinform. 2024;25(3):bbae174. doi:10.1093/bib/bbae174

[8] Reyes DM, Jagadeesh KA, et al. FairPRS: adjusting for admixed populations in polygenic risk scores using invariant risk minimization. Pac Symp Biocomput. 2023;28:198-208. doi:10.1142/9789811276516_0019

[9] 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74. doi:10.1038/nature15393

[10] Privé F, Aschard H. Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. Am J Hum Genet. 2022;109(1):12-23. doi:10.1016/j.ajhg.2021.11.008

[11] Gadbury GL, Xiang Q, Allison DB. Plasmode simulation for the evaluation of analytic methods in genome-wide association studies: methodological and practical considerations. BMC Bioinformatics. 2008;9 Suppl 9:S4. doi:10.1186/1471-2105-9-S9-S4

[12] Kachuri L, Chatterjee N. Principles and methods for transferring polygenic risk scores across global populations. Nat Rev Genet. 2024;25(1):8-25. doi:10.1038/s41576-023-00637-2

[13] Lehmann B, Mackintosh M, McVean G, Holmes C. Optimal strategies for learning multi-ancestry polygenic scores vary across traits. Nat Commun. 2023;14(1):4142. doi:10.1038/s41467-023-38930-7

[14] Chatterjee N, Shi J, García-Closas M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. Nat Rev Genet. 2016;17(7):392-406. doi:10.1038/s41576-016-0027-x

[15] Márquez-Luna A, Revez JA, et al. Polygenic scoring accuracy varies across the genetic ancestry continuum. Nat Hum Behav. 2023;7(11):1847-1858. doi:10.1038/s41562-023-01658-5

[16] Tsuo, K., Shi, Z., Ge, T. *et al.* All of Us diversity and scale yield context-dependent improvements in polygenic prediction. Nat Genet **58**, 2485–2493 (2026). https://doi.org/10.1038/s41588-026-02734-4

[17] Monti R, Eick L, Hudjashov G, et al. Evaluation of polygenic scoring methods in five biobanks shows larger variation between biobanks than methods and finds benefits of ensemble learning. Am J Hum Genet. 2024;111(7):1431-1447. doi:10.1016/j.ajhg.2024.06.003

[18] Duncan L, Shen H, Gelernter J, et al. Analysis of polygenic risk score usage and performance in diverse human populations. Nat Commun. 2019;10(1):3328. doi:10.1038/s41467-019-11112-0

[19] Turley P, Meyer KN, Wang Q, et al. Current polygenic risk scores are unlikely to exacerbate unfairness in cardiovascular disease risk prediction. medRxiv. 2025. doi:10.1101/2025.09.18.25336069

[20] Weissbrod O, Kanai M, Shi H, et al. Leveraging fine-mapping and multipopulation training data to improve cross-population polygenic risk scores. Nat Genet. 2022;54(4):450-458. doi:10.1038/s41588-022-01036-9

[21] Lambert SA, Gillett A, Pandit A, et al. The Polygenic Score Catalog as an open database for reproducibility and systematic evaluation. Nat Genet. 2021;53(4):420-425. doi:10.1038/s41588-021-00783-5

[22] Fatumo S, Choudhury A, et al. Polygenic risk scores for disease risk prediction in Africa: current challenges and future directions. Genome Med. 2023;15(1):87. doi:10.1186/s13073-023-01247-4

[23] Kullo IJ, Jouni H, Austin EK, et al. Polygenic scores in biomedical research. Nat Rev Genet. 2022;23(9):524-532. doi:10.1038/s41576-022-00470-z

[24] Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Pickrell JK, et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. Am J Hum Genet. 2015;97(4):576-592. doi:10.1016/j.ajhg.2015.08.003

[25] Privé F. Challenges in constructing and evaluating polygenic risk scores across

diverse populations. Brief Bioinform. 2023;24(6):bbad394. doi:10.1093/bib/bbad394

[26] Wang Y, Tsuo K, Martin AR, et al. Theoretical and empirical quantification of the accuracy of polygenic scores in ancestry divergent populations. Nat Commun. 2020;11(1):3865. doi:10.1038/s41467-020-17719-y

Appendix Note A: Mathematical formulation of cluster_fixed

We decompose per-ancestry-group genetic effects into a global shared component and group-specific deviations. For ancestry group *g*, the effect vector is:

$$\boldsymbol{\beta}_g = \boldsymbol{\beta}_{\text{shared}} + \boldsymbol{\delta}_g$$

where $\boldsymbol{\beta}_{shared}$ denotes effects shared across all ancestry groups, and $\boldsymbol{\delta}_g$ captures deviations specific to group *g*.

The objective function minimises prediction loss on training data, combined with regularisation penalties on shared effects and group-specific deviations:

$$\mathcal{L} = \mathcal{L}_{\text{pred}}\left(y, X_g\left(\beta_{\text{shared}} + \delta_g\right)\right) + \lambda_\beta \parallel \beta_{\text{shared}} \parallel_2^2 + \sum_g \lambda_\delta \frac{\parallel \delta_g \parallel_2^2}{\text{denom}}$$

Here:

- $\mathcal{L}_{\text{pred}}$: squared-error prediction loss evaluated on training samples;

- λ_β: regularisation strength for shared-component effects β_{shared};

- λ_δ: base regularisation strength for group-specific deviations δ_g;

-(denom): the core tunable hyper-parameter that scales penalisation for group-specific deviations. Increasing denom weakens shrinkage of δ_g, suppressing between-group effect heterogeneity.

We adopt Gaussian prior assumptions consistent with this ridge-type formulation:

$$\beta_{\text{shared}} \sim N\left(0, \sigma_\beta^2 I\right), \quad \delta_g \sim N\left(0, \sigma_\delta^2 I\right).$$

The ratio $\sigma_\delta/\sigma_\beta$ characterises relative magnitude of group-specific versus shared genetic effects, and governs whether the system operates in a trade-off or win-win regime, as shown in the main text.

Appendix S2 — Sample-size planning for AFR

Motivation. The baseline gap in Results 3.7 answers the question "how large is the current AFR–EUR disparity?" but not the follow-up "how much would additional AFR samples reduce it?". To address this, we performed an AFR-only subsampling experiment: the estimator was retrained on AFR subsamples of size $n_{\mathrm{AFR}} \in \{50,100,200,300,400,500,600\}$, while EAS and EUR were held at their full 1000G sizes. For each n_{AFR}, the experiment was repeated over 10 random seeds.

Controls. Three quantities were held fixed throughout to isolate the effect of AFR sample size:

1. **SNP set.** The 1 210 quasi-independent SNPs selected by LD pruning on the full sample were reused for every n_{AFR} (via a stored keep_mask), so that column membership did not vary across n_{AFR}.
2. **Ground truth.** The shared component β_{shared} and the site-specific components $\{\delta_k\}$ were drawn once from a fixed seed (fix_beta_seed=42) and reused across all n_{AFR}; only the train/test split and the noise realisation varied with the per-run seed.
3. **Estimator configuration.** cluster_fixed with denom =0.05, $\alpha = 0, \beta = 0$, matching the main-text configuration.

Observed results. Both overall test-set $R^2$and the worst-case cross-sub-population gap (fairness_dr2) improved monotonically with n_{AFR} over the observed range

(Table S1,Figure S1).

Table S1. AFR subsampling curve. Mean ± SD across 10 seeds. EAS/EUR fixed at full 1000G sizes. No extrapolation beyond $n_{\mathrm{AFR}} = 661$.

n_{AFR}	overall_r2	fairness_dr2
50	−0.164 ± 0.246	2.698 ± 3.506
100	0.019 ± 0.051	0.564 ± 0.550
200	0.065 ± 0.012	0.243 ± 0.119
300	0.077 ± 0.008	0.190 ± 0.062
400	0.087 ± 0.008	0.167 ± 0.038
500	0.093 ± 0.009	0.172 ± 0.048
600	0.097 ± 0.007	0.182 ± 0.050
661 (full AFR)	0.106 ± 0.015	0.182 ± 0.030

Fairness improved monotonically up to $n_{\mathrm{AFR}} = 400$, after which performance plateaued. The $n_{\mathrm{AFR}} = 600$ observation shows slight non-monotonic fluctuation within one standard deviation, consistent with sampling noise across simulation seeds.

Diminishing returns

The curve flattens beyond $n_{\mathrm{AFR}} \approx 400$. From $n_{\mathrm{AFR}} = 400$ to $n_{\mathrm{AFR}} = 661$ (a 65% increase in AFR sample size), overall R^2 rises by 0.019 (from 0.087 to 0.106) and the fairness gap changes by +0.015, i.e. essentially flat within one SD. The first ~400 AFR samples therefore account for most of the achievable improvement under this data-generating regime (Figure S1).

Estimator stability at small n_{AFR}. At $n_{\mathrm{AFR}} = 50$, both metrics are extremely variable across seeds (overall_r2 SD = 0.246; fairness_dr2 SD = 3.51), reflecting estimator instability at very small training sizes. The curve should therefore be interpreted as reliable for $n_{\mathrm{AFR}} \geq 200$, where the per-$n_{\mathrm{AFR}}$ SD drops to ≤0.012 (overall) and ≤0.12 (fairness).

No extrapolation

We deliberately do not extrapolate the curve to $n_{\mathrm{AFR}} > 661$. The observed relationship is empirical and bounded above by the available 1000G AFR sample size (661). An extrapolation to larger n_{AFR} would require either a theoretical scaling law for the sample-size dependence of cross-ancestry prediction, or external data from cohorts with larger AFR representation; both are outside the scope of this work.

Declarations

Ethical approval and consent to participate

This study analyses publicly available GWAS summary statistics and reference panel data. No novel human sample collection was performed. All original studies contributing the summary data obtained ethical approval and participant informed consent as described in their respective publications.

Consent for publication

Not applicable. All analysed data are de-identified and publicly available; no individual-level participant data are presented in this manuscript.

Data availability

The datasets generated and analysed during the current study (including train/validation split definitions, aggregated result tables, and reproducibility audit files) are available in the Figshare repository, https://doi.org/10.6084/m9.figshare.34919169

The 1000 Genomes Phase 3 reference panel is publicly available from the 1000 Genomes Project repository (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/releas /20130502/). The Pan-UKB height GWAS summary statistics are publicly available from the Pan-UK Biobank portal.

Conflict of Interest

The authors declare that they have no competing interests.

Author Contributions

S.Y.: Conceptualization, methodology, formal analysis, writing—original draft, writing—review & editing. L.Q.: Data curation, validation, writing—review & editing. W.X.: Methodology, writing—review & editing.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Acknowledgments

Not applicable.

Figure 1. Fairness–accuracy trade-off in the main experiment Ridge meta (blue circle) versus cluster_fixed (orange triangle) on the 1000G chr21 panel with 1,210 quasi-independent SNPs and simulated phenotypes. The arrow indicates the direction of change from meta to cluster_fixed. Axes: overall test-set R^2 (higher is better) versus worst-case cross-sub-population R^2 disparity, fairness_dr2 (lower is fairer). Both differences are statistically significant (paired t-test, n = 10 seeds, $p < 10^{-5}$).

Figure 2. Trade-off behaviour across three genetic architectures Each panel plots overall R^2 (x-axis) versus fairness_dr2 (y-axis). Blue circles: Ridge meta; orange triangles: cluster_fixed; grey arrows: direction of change from meta to cluster_fixed. Panel A: σ_δ = 0.01 ($\sigma_\delta/\sigma_\beta$ = 0.2), trade-off. Panel B: σ_δ = 0.03 ($\sigma_\delta/\sigma_\beta$ = 0.6), win-win. Panel C: σ_δ = 0.05 ($\sigma_\delta/\sigma_\beta$ = 1.0), win-win.

Figure 3. Optimal shrinkage depends on genetic effect architecture Left: overall R^2 as a function of shrinkage hyper-parameter denom (log scale). Right: fairness_dr2 as a function of denom. Three curves correspond to regimes A (σ_δ = 0.01), B (σ_δ = 0.03), and C (σ_δ = 0.05). The optimal denom for overall R^2 shifts leftward as σ_δ increases.

Figure S1. AFR subsampling performance curves Left: overall test-set R^2 versus AFR sample size (n_AFR). Right: worst-case cross-sub-population R^2 gap (fairness_dr2) versus n_AFR. Both panels show mean ± SD across 10 seeds. EAS and EUR are held at their full 1000G sample sizes. The dotted vertical line marks n_AFR = 661 (full AFR sample). No extrapolation is performed beyond this range.

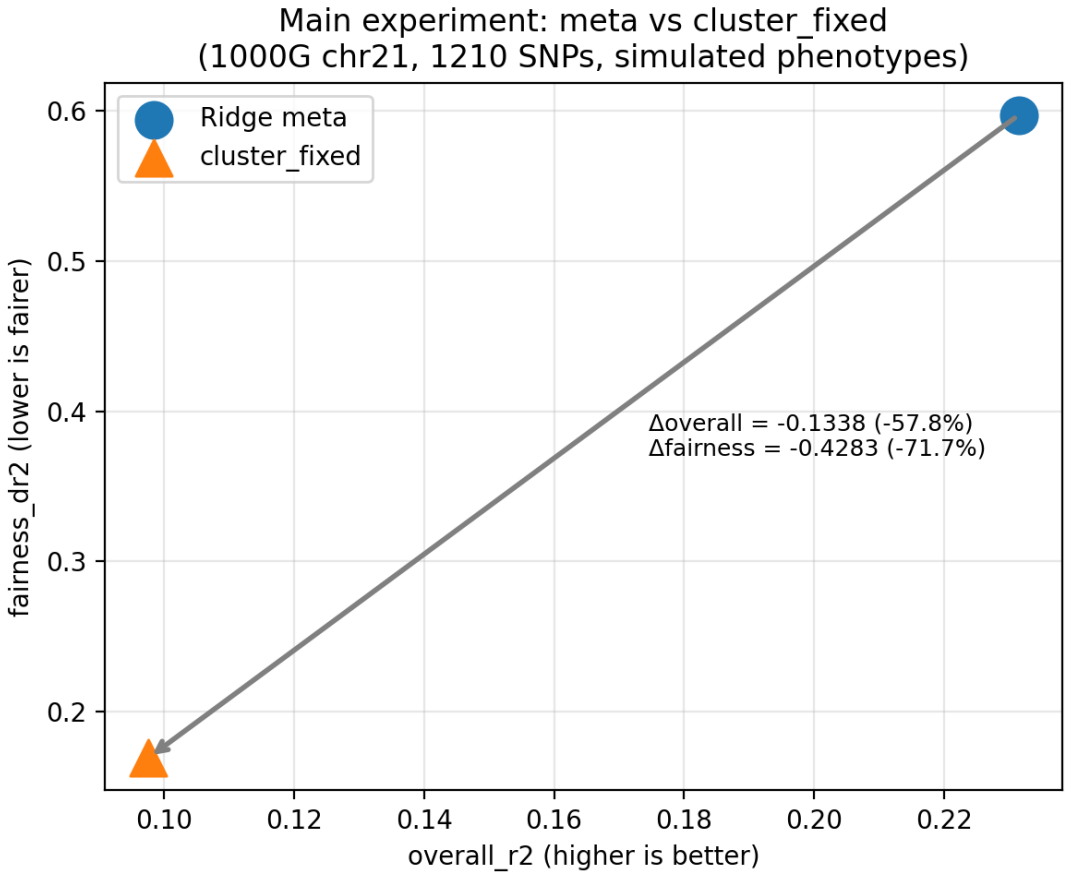

Figure 1. Fairness–accuracy trade-off in the main experiment Ridge meta (blue circle) versus cluster_fixed (orange triangle) on the 1000G chr21 panel with 1,210 quasi-independent SNPs and simulated phenotypes. The arrow indicates the direction of change from meta to cluster_fixed. Axes: overall test-set R^2 (higher is better) versus worst-case cross-sub-population R^2 disparity, fairness_dr2 (lower is fairer). Both differences are statistically significant (paired t-test, n = 10 seeds, $p < 10^{-5}$).

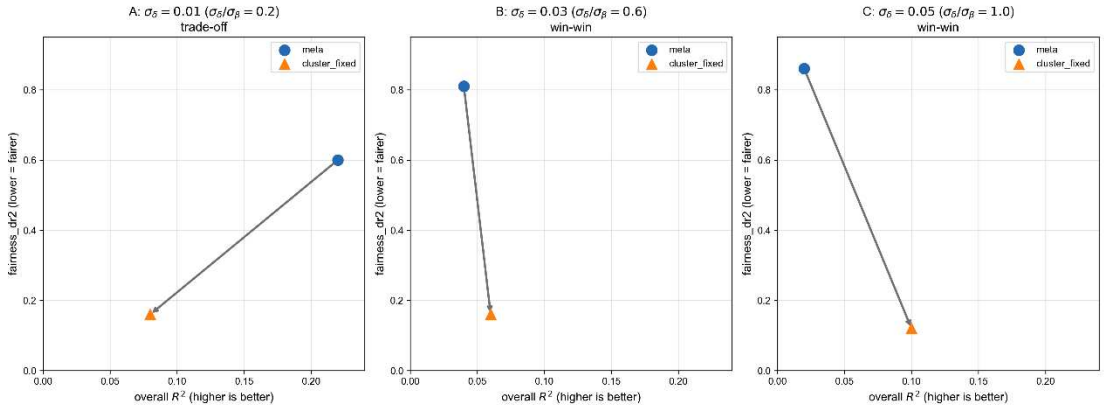

Figure 2. Trade-off behaviour across three genetic architectures Each panel plots overall R^2 (x-axis) versus fairness_dr2 (y-axis). Blue circles: Ridge meta; orange triangles: cluster_fixed; grey arrows: direction of change from meta to cluster_fixed. Panel A: $\sigma_\delta = 0.01$ ($\sigma_\delta/\sigma_\beta = 0.2$), trade-off. Panel B: $\sigma_\delta = 0.03$ ($\sigma_\delta/\sigma_\beta = 0.6$), win-win. Panel C: $\sigma_\delta = 0.05$ ($\sigma_\delta/\sigma_\beta = 1.0$), win-win.

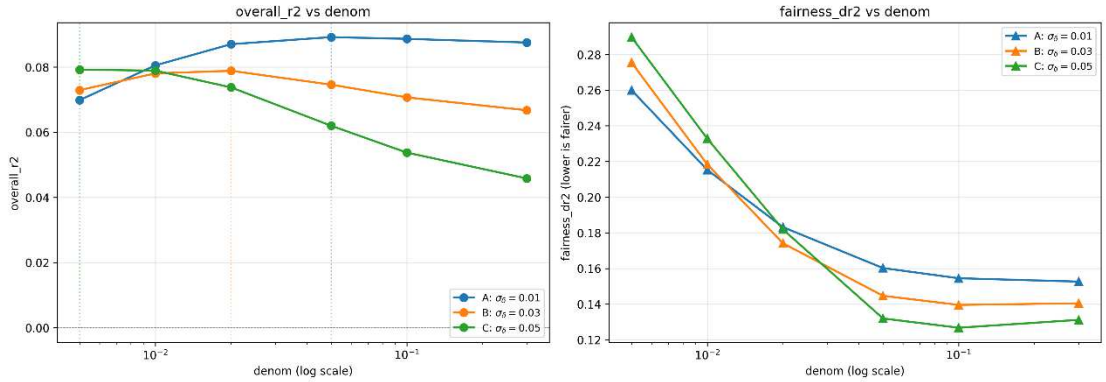

Figure 3. Optimal shrinkage depends on genetic effect architecture Left: overall R^2 as a function of shrinkage hyper-parameter denom (log scale). Right: fairness_dr2 as a function of denom. Three curves correspond to regimes A (σ_δ = 0.01), B (σ_δ = 0.03), and C (σ_δ = 0.05). The optimal denom for overall R^2 shifts leftward as σ_δ increases.

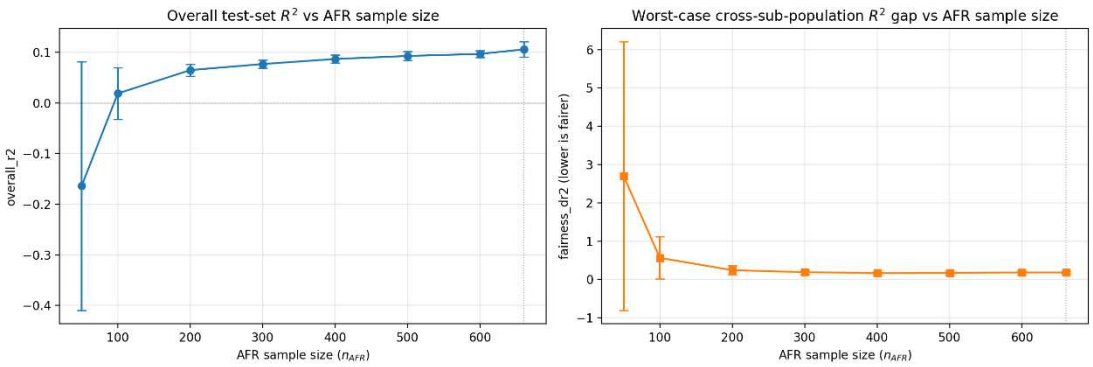

Figure S1. AFR subsampling performance curves Left: overall test-set R^2 versus AFR sample size (n_AFR). Right: worst-case cross-sub-population R^2 gap (fairness_dr2) versus n_AFR. Both panels show mean ± SD across 10 seeds. EAS and EUR are held at their full 1000G sample sizes. The dotted vertical line marks n_AFR = 661 (full AFR sample). No extrapolation is performed beyond this range.

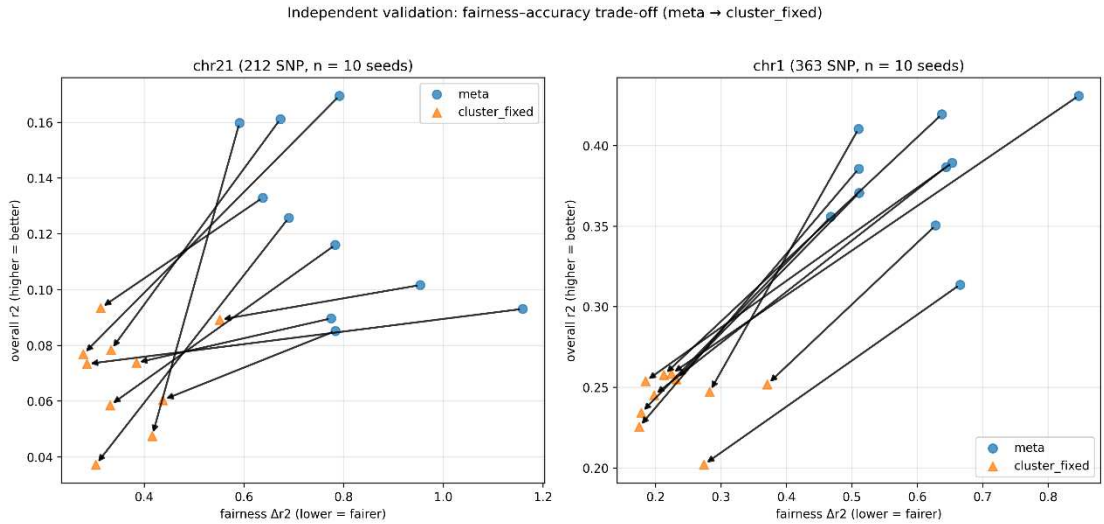

Figure S2. Independent-validation fairness-accuracy trade-off on chr21 and chr1.

Left: chr21 (212 SNPs); right: chr1 (363 SNPs). Blue circles: Ridge meta; orange triangles: cluster_fixed; grey arrows: direction of change from meta to cluster_fixed on the held-out validation subset. Each point is one of the 10 random seeds. Training and validation individuals are disjoint per site (827 vs 841; intersection = 0); the Pan-UKB β is fixed as shared truth across seeds, while δ and noise are redrawn per seed. Both metrics are significantly reduced by cluster_fixed on both chromosomes (paired t-test, $n = 10$, $p < 0.01$).

Table S1. AFR subsampling curve. Mean ± SD across 10 seeds. EAS/EUR fixed at full 1000G sizes. No extrapolation beyond $n_{\mathrm{AFR}} = 661$.

n_{AFR}	overall_r2	fairness_dr2
50	−0.164 ± 0.246	2.698 ± 3.506
100	0.019 ± 0.051	0.564 ± 0.550
200	0.065 ± 0.012	0.243 ± 0.119
300	0.077 ± 0.008	0.190 ± 0.062
400	0.087 ± 0.008	0.167 ± 0.038
500	0.093 ± 0.009	0.172 ± 0.048
600	0.097 ± 0.007	0.182 ± 0.050
661 (full AFR)	0.106 ± 0.015	0.182 ± 0.030

Table S2. Independent validation of cluster_fixed vs meta on disjoint individuals (10 seeds, fixed Pan-UKB β).

Chr	n_SNP	Δoverall	Δfairness	p_overall	p_fairness
chr21	212	−0.055	−0.420	1.2×10^{-3}	4.6×10^{-5}
chr1	363	−0.138	−0.374	3.6×10^{-8}	4.2×10^{-6}

Δ = cluster_fixed − meta, mean across 10 seeds (paired t-test, df = 9). All 10/10 seeds negative for both metrics on both chromosomes. The chr21 SNP set (212) is identical to that of the main GWAS-informed analysis (Table 5); the chr1 SNP set (363) differs from the 482 SNPs of the main chr1 analysis because the two runs applied different variant-matching configurations. Absolute Δ magnitudes are therefore compared only in direction and statistical significance, not in absolute value.