BEYOND SINGLE-THRESHOLD FAIRNESS: AUDITING GROUP ERROR DISPARITIES WITH RESIDUAL DISTRIBUTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Fairness audits often compare groups at a single decision threshold, even though deployed thresholds can vary across settings and over time. We study how false-negative and false-positive differences between groups change over the full threshold range. Using the signed residual, we show that the Wasserstein distance between two groups' residual distributions equals the total area between their group FN and FP curves as the threshold varies. The two components show which error type drives the difference and how prevalence connects these group rates to FNR and FPR. We construct exactly calibrated populations that agree on the examined accuracy, ranking, calibration, average-error, and single-threshold fairness summaries while having different residual distributions and error gaps at other thresholds. Experiments on Adult, ACSIncome, HateXplain, and Measuring Hate Speech show that calibration can move group mean residuals closer to zero while residual distance increases. They also show that different prevalences can enlarge or reduce group FN/FP gaps. Across ACSIncome model configurations, validation FN and FP components preserve rankings of related test-time error differences more closely than fixed-threshold equalized odds or a mean-residual summary for the evaluated targets. Together, residual distributions provide a direct way to audit group error differences across decision thresholds.

1 INTRODUCTION

Fairness audits are often reported at one decision threshold, even though the threshold used in practice can change with application goals, capacity constraints, and time. As the threshold moves, the fractions of each group receiving false negatives and false positives can change in different ways. Demographic parity (DP) and equalized odds (EO) describe group differences at a selected threshold, while calibration and average-error statistics summarize other properties of probabilistic predictions (Hardt et al., 2016; Guo et al., 2017; Pleiss et al., 2017; Chouldechova, 2017). A useful complementary view is therefore to examine group error differences over the full threshold range.

We obtain this view from the signed residual $D = \hat{P} - Y$. Positive examples appear on the negative side of the residual axis and negative examples on the positive side. Comparing two groups' residual distributions therefore separates naturally into a false-negative part and a false-positive part. The Wasserstein distance between the two residual distributions is exactly the sum of these two areas across thresholds (§ 3.2). This gives the distance between residual distributions a direct interpretation in terms of group FN and FP differences.

The same result explains why group FN/FP rates and conditional FNR/FPR can differ. Our FN and FP components measure errors as a fraction of the full group, whereas EO conditions on the true label. Prevalence connects the two: changing the label proportions can enlarge or reduce a group FN or FP gap even when the score distributions among positive and negative examples stay fixed (§ 3.3). The components also relate to costs that place more weight on false negatives or false positives (§ 3.4).

We further show that several familiar summaries can agree while error behavior across thresholds remains different. Two exactly calibrated populations in our construction have the same examined accuracy, AUC, directional xAUC, Brier score, mean residual, and DP and EO gaps at threshold 0.5.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

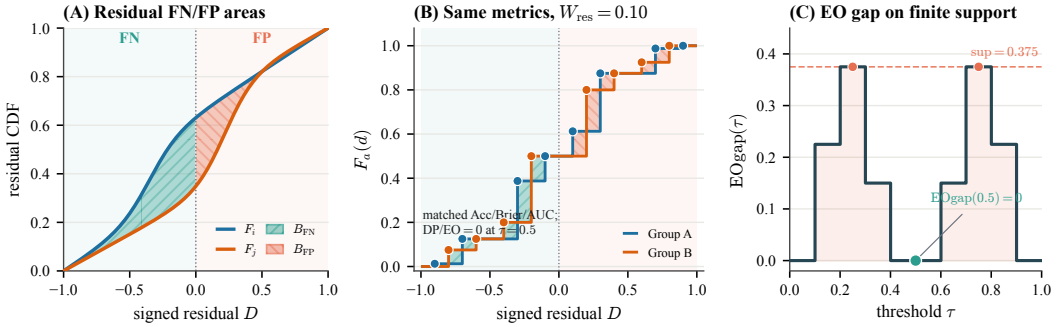

Figure 1: Residual distributions and group errors across thresholds. **(A)** The negative and positive halves of the residual CDF give the FN and FP components B_{FN} and B_{FP} (Equation (3)). **(B)** Exactly calibrated populations can match the listed accuracy, ranking, average-error, and threshold metrics while retaining residual distance $W_{\mathrm{res}} = 0.10$ (§ A). **(C)** On the same construction, $\mathrm{EOgap}(0.5) = 0$ while $\sup_\tau \mathrm{EOgap}(\tau) = 0.375$, showing how the EO gap changes away from the reported threshold.

Their residual distance is 0.10, and their error gaps differ at other thresholds (§ 3.5). The example isolates information that becomes visible only when group errors are followed beyond one threshold.

The experiments test these implications in tabular and text classification. On Adult and ACSIncome, groupwise calibration moves group mean residuals closer to zero while residual distance increases, and the FN/FP split shows which error type changes. A prevalence analysis separates score differences among positive and negative examples from differences caused by the group label proportions. Across 432 ACSIncome model configurations, validation FN and FP components preserve rankings of related test-time error differences more closely than fixed-threshold EO or a mean-residual summary for the evaluated targets. HateXplain and Measuring Hate Speech provide an additional check on text classification.

We make three contributions.

- *An exact error interpretation across thresholds.* We show that residual Wasserstein distance equals the sum of the group FN and FP gap areas over decision thresholds, with separate components for the two error types and connections to corresponding error costs.
- *A separation from common summary metrics.* We construct exactly calibrated populations that agree on the examined threshold-based fairness, ranking, calibration, and average-error summaries while retaining different residual distributions and error gaps across thresholds.
- *Empirical evidence and model comparison.* Experiments show how calibration, prevalence, and threshold choice change the measured group differences, and the 432-model study tests whether validation FN/FP components preserve rankings of related test-time error differences.

2 Related Work

Threshold-based fairness criteria include demographic parity and equalized odds, and widely used auditing toolkits report their group rates and aggregate gaps (Hardt et al., 2016; Bird et al., 2020; Bellamy et al., 2018; Saleiro et al., 2018). Prior work also studies how fairness conclusions vary with the decision threshold (Corbett-Davies et al., 2017). Our analysis focuses on a different quantity: the total difference in group FN and FP rates as the threshold varies.

Ranking- and ROC-based approaches also examine behavior across thresholds. xAUC measures cross-group ranking between positive and negative examples (Kallus & Zhou, 2019); ABROCA integrates gaps between group ROC curves (Gardner et al., 2019); and subgroup AUC summaries capture related ranking differences (Borkan et al., 2019). Other work constrains ROC/AUC fairness over threshold ranges or adjusts scores to improve parity across thresholds (Vogel et al., 2021; Kwegyir-Aggrey et al., 2022). These methods focus on ranking or ROC behavior, while residual distance measures differences in group FN and FP rates over the threshold range.

Other fairness methods compare score distributions with Wasserstein and related distances (Jiang et al., 2019; Chzhen et al., 2020; Gordaliza et al., 2019; Becker et al., 2024). Jiang et al. compare

marginal score distributions, while Becker et al. compare score distributions conditional on the true label (Jiang et al., 2019; Becker et al., 2024). Related work also compares areas between curves or decomposes transport distances (Han et al., 2023; Miroshnikov et al., 2022). Our residual comparison combines positive and negative examples according to group prevalence; § 3.3 gives the relationship used in the experiments. Finite-sample Wasserstein estimates also require uncertainty analysis (Lum et al., 2022), which motivates the permutation tests used here.

Calibration and error-rate parity impose different constraints on probabilistic predictions (Chouldechova, 2017; Kleinberg et al., 2017; Pleiss et al., 2017). Kleinberg et al.'s balance conditions constrain mean scores within each label, while multicalibration and multiaccuracy constrain residual means over selected groups or score regions (Kim et al., 2019; Hebert-Johnson et al., 2018). Our analysis compares the full residual distributions of a fixed predictor and separates the FN and FP parts.

3 Residual Distributions and Error Gaps Across Thresholds

This section defines the quantities used in the experiments and explains what each one measures. Proofs and more general statements are collected in § B.

3.1 Residual Distributions

Let $Y \in \{0, 1\}$ be the observed label, A a categorical protected attribute, and $\hat{P} \in [0, 1]$ a predicted probability. We use the signed residual $D = \hat{P} - Y$. For positive-label examples, $D = \hat{P} - 1 \in [-1, 0]$; for negative-label examples, $D = \hat{P} \in [0, 1]$. The sign therefore identifies the error side: large negative residuals correspond to positive examples assigned low scores, while large positive residuals correspond to negative examples assigned high scores. Values near zero indicate scores closer to the observed label. Thus the two sides of the residual distribution correspond directly to positive- and negative-label errors.

For group a, let F_a be the CDF of D, let $\pi_a = \mathbb{P}(Y = 1 \mid A = a)$ be the positive-label prevalence, and let $\mu_a = \mathbb{E}[D \mid A = a]$ be the group mean residual. For two groups i and j, we compare their residual distributions with Wasserstein-1 distance,

$$W_{\text{res}}(i, j) = W_1(\mathcal{L}(D \mid A = i), \mathcal{L}(D \mid A = j)) = \int_{-1}^{1} |F_i(t) - F_j(t)| \, dt. \tag{1}$$

Geometrically, W_{res} is the total area between the two residual CDFs. It compares the full residual distributions rather than reducing each group to one mean or one decision threshold.

Residual magnitude also records how far a score is from the observed label. For positive examples, lower predicted probabilities move farther along the negative side; for negative examples, higher predicted probabilities move farther along the positive side. Averaging residuals into one number can hide these separate changes.

A simple way to read the residual axis is to separate *which label an example has* from *how far its score is from that label*. The sign identifies the label side, while the magnitude gives the distance from the observed outcome. This is why a single residual distribution can keep FN-related and FP-related behavior together without mixing them into the same average.

The calibration experiment reports two mean summaries alongside W_{res}. The between-group mean residual gap is $|\mu_i - \mu_j|$, while $M_{ij} = |\mu_i| + |\mu_j|$ measures how far the two group means are from zero. Because $|\mu_i - \mu_j| \leq W_{\text{res}}(i, j)$, similar residual distributions also have similar means. FN and FP differences can offset in the mean, however, while remaining visible in the full distributions. § 4.2 tests this distinction, and § B.2 gives the formal result.

3.2 Residual Distance and FN/FP Gaps

A decision threshold τ gives $\hat{Y} = \mathbb{I}[\hat{P} \geq \tau]$. FNR and FPR are conditional on the true label. To measure false negatives and false positives as fractions of the full group, define

$$\text{GFNR}_a(\tau) = \pi_a \text{FNR}_a(\tau), \qquad \text{GFPR}_a(\tau) = (1 - \pi_a)\text{FPR}_a(\tau). \tag{2}$$

Here FNR_a is computed among positive examples and FPR_a among negative examples. After multiplying by the corresponding label proportion, GFNR_a and GFPR_a give the fractions of the full group receiving false negatives and false positives. We call them the group FN and FP rates.

As τ moves from 0 to 1, the negative residual half traces the group FN rate and the positive half traces the group FP rate. We summarize the absolute gaps between two groups by

$$\begin{aligned} B_{\mathrm{FN}}(i,j) &= \int_0^1 |\mathrm{GFNR}_i(\tau) - \mathrm{GFNR}_j(\tau)|\, d\tau, \\ B_{\mathrm{FP}}(i,j) &= \int_0^1 |\mathrm{GFPR}_i(\tau) - \mathrm{GFPR}_j(\tau)|\, d\tau, \\ W_{\mathrm{res}}(i,j) &= B_{\mathrm{FN}}(i,j) + B_{\mathrm{FP}}(i,j). \end{aligned} \tag{3}$$

The last line is the paper's central result. B_{FN} is the area between the two groups' FN-rate curves, and B_{FP} is the corresponding FP area. Their sum is exactly the residual Wasserstein distance, and the identity applies to any probabilistic classifier regardless of calibration. Figure 1(A) shows the same relation geometrically.

This decomposition makes the total distance easier to interpret. A large W_{res} can come mainly from false negatives, mainly from false positives, or from both. The total distance says how far the two residual distributions are apart; the two components show where that difference comes from in terms of prediction errors.

Equation (3) also gives a simple meaning to "across thresholds." Under a uniform sweep from 0 to 1, B_{FN} is the average absolute group FN gap. B_{FP} is the corresponding average group FP gap. § 4.3 also examines selected operating points to show how a single threshold samples only one part of these curves.

The split is useful even when two models have similar total residual distance. One model can differ mainly on false negatives, while another can differ mainly on false positives. Reporting B_{FN} and B_{FP} alongside W_{res} shows which error type accounts for the difference. The calibration experiment in § 4.2, for example, uses the two components to locate the change.

For model comparison we use the same idea over a selected threshold range. Let Π describe how thresholds are weighted. We write

$$B_{\Pi}^{\mathrm{FN}}(i,j) = \mathbb{E}_{\tau\sim\Pi}|\mathrm{GFNR}_i(\tau) - \mathrm{GFNR}_j(\tau)|, \qquad B_{\Pi}^{\mathrm{FP}}(i,j) = \mathbb{E}_{\tau\sim\Pi}|\mathrm{GFPR}_i(\tau) - \mathrm{GFPR}_j(\tau)|. \tag{4}$$

In § 4.4, "Avg. FN" and "Avg. FP" average uniformly over the validation-selected threshold interval corresponding to selection rates from 5% to 50%. Corollary B.9 gives the general weighted form.

3.3 Prevalence and Conditional Error Rates

Residual distance depends on both the score distributions among positive and negative examples and the proportion of each label in each group. Let $W_{Y=1}(i,j)$ be the Wasserstein distance between the groups' scores among positive examples, and let $W_{Y=0}(i,j)$ be the corresponding distance among negative examples. These distances equal the average FNR and FPR differences across thresholds:

$$W_{Y=1} = \int_0^1 |\mathrm{FNR}_i(\tau) - \mathrm{FNR}_j(\tau)|\, d\tau, \qquad W_{Y=0} = \int_0^1 |\mathrm{FPR}_i(\tau) - \mathrm{FPR}_j(\tau)|\, d\tau. \tag{5}$$

Thus $W_{Y=1}$ and $W_{Y=0}$ compare scores among positive and negative examples, while B_{FN} and B_{FP} measure FN and FP differences as fractions of the full groups.

Prevalence connects the two views. When two groups have the same positive-label prevalence, residual distance is the prevalence-weighted sum of the positive- and negative-label distances. When their prevalences differ, the group FN and FP gaps also reflect the different label proportions. Proposition B.12 gives the formal bound, and Equation (18) gives the corresponding identity at one threshold.

The intuition is simple. FNR asks what fraction of the positive examples receive a false negative. The group FN rate asks what fraction of the entire group receives a false negative. Even with the same FNR, the group FN rate is larger when positive examples are more common. The same reasoning

applies to false positives. The two quantities therefore answer different questions: FNR/FPR describe errors conditional on the label, while the group FN/FP rates describe how frequently each error occurs in the full group.

This distinction is useful when reading the experiments. If the conditional rates stay similar but the label proportions differ, the group FN/FP gaps can still be large. If the label proportions are similar, the group gaps more directly reflect the score differences among positive or negative examples. The prevalence analysis in § 4.3 separates these two sources.

The experiments make this distinction explicit by setting both groups to the average prevalence $\bar{\pi} = (\pi_i + \pi_j)/2$ while keeping the score distributions among positive and negative examples fixed. After equalizing prevalence, the residual distance is

$$\bar{\pi}\, W_{Y=1}(i,j) + (1-\bar{\pi})\, W_{Y=0}(i,j). \tag{6}$$

We compare this value with the observed $W_{\mathrm{res}}(i,j)$. Table 2 reports observed minus equalized distance: a positive value means the observed label proportions increase W_{res}, while a negative value means they reduce it. This comparison simply measures how the distance changes after equalizing the label proportions while keeping the score distributions within each label fixed.

This also clarifies the relationship with equalized odds. EO compares FNR and FPR at one chosen threshold. $W_{Y=1}$ and $W_{Y=0}$ compare the corresponding score distributions over all thresholds. The residual components B_{FN} and B_{FP} instead measure FN and FP differences as fractions of the full groups over the threshold range.

3.4 Error Costs and Model Comparison

The FN and FP components also relate to simple error costs. Consider the false-negative cost $c_{\mathrm{FN}}(D) = \max(0, -D)$, which increases as a positive example receives a lower predicted probability. Its group gap satisfies

$$|\mathbb{E}[c_{\mathrm{FN}}(D) \mid A = i] - \mathbb{E}[c_{\mathrm{FN}}(D) \mid A = j]| \le B_{\mathrm{FN}}(i,j). \tag{7}$$

The false-positive side has the same form. For a positive example, $c_{\mathrm{FN}} = 1 - \hat{P}$, so the cost is larger when the model assigns a lower score. The bound therefore links the FN component to both the frequency and size of false-negative errors. Corollary B.8 gives the more general result.

The model-comparison experiment uses this separation directly. FN summaries are compared with FN-based test quantities, and FP summaries are compared with FP-based test quantities. This keeps the comparison tied to the same error type instead of asking one combined statistic to represent both sides at once.

This motivates the validation–test analysis in § 4.4. We compare the validation FN component with test quantities based on false negatives, including the average FN gap and the FN cost gap; the FP component is compared with false-positive quantities. The experiment asks whether a validation FN or FP summary preserves model ordering for a matching test quantity. The mean residual gap serves as a baseline because it combines the FN and FP sides into one signed average.

High-confidence errors look at specific parts of the residual range. At a fixed confidence cutoff, we count the share of each group receiving a confident error. These cutoffs correspond to particular regions of the negative and positive residual halves.

3.5 Residual Distributions Beyond Common Summaries

The residual view can separate populations that look the same under several common summaries. Figure 1(B) constructs two exactly calibrated populations with the same accuracy, AUC, directional xAUC values, DP and EO gaps at threshold 0.5, Brier score, and mean residual, while their residual distance is 0.10. The predictor is Bayes-optimal for squared loss on the information available in the construction. § A gives the score supports, masses, and exact calculations.

The two groups have different residual CDFs, so their FN and FP gaps change differently as the threshold or confidence cutoff moves. Marginal score distances and score distances computed separately within each label can also separate this construction. Our contribution is the direct

Table 1: Effect of groupwise isotonic calibration. Gray columns highlight the main contrast between residual distance and the summed absolute mean residual M_{ij}. FN and FP identify the error side driving the change.

Data	Residual distance	FN component†	FP component	Avg. FN	Avg. FP	M_{ij}
Adult	0.133→0.147	0.074→0.073	0.059→0.074	0.068→0.065	0.062→0.072	0.017→0.008
ACS	0.050→0.062	0.025→0.032	0.025→0.030	0.036→0.039	0.026→0.039	0.024→0.008

interpretation of residual distance as group FN and FP differences across thresholds. The construction keeps the familiar summaries fixed while changing information that becomes visible when the threshold moves.

The purpose of the construction is therefore diagnostic rather than competitive. It isolates a case in which familiar summaries agree and then shows exactly which additional behavior is exposed by following FN and FP gaps across thresholds. This provides a controlled setting for interpreting the empirical differences observed later.

The experiments use W_{res} for the total residual distance, B_{FN} and B_{FP} for its FN and FP parts, and M_{ij} for the group mean residuals. The prevalence analysis also uses $W_{Y=1}$ and $W_{Y=0}$, while the validation–test analysis averages FN and FP gaps over the selected threshold range.

4 Experiments

4.1 Experimental Setup

Adult Female vs. Male is the running pair unless stated otherwise. The experiments address three questions suggested by the result above. First, we ask whether improved mean calibration implies closer residual distributions. Second, we examine why group FN/FP rates differ from conditional FNR/FPR by separating prevalence and threshold effects. Third, we test whether validation FN/FP components preserve rankings of related test-time error differences across model configurations. Adult and ACSIncome provide the primary evidence, with fixed text classifiers as an additional check (§ 4.5).

Adult compares Female and Male groups, and ACSIncome averages ten race-group pairs with sufficient test support (Ding et al., 2021). Logistic regression is the primary model for the calibration and threshold studies; the validation–test study also includes XGBoost and MLP models. All tabular experiments use five group-and-label-stratified 60/20/20 train/validation/test splits. Calibration maps and thresholds are chosen on validation data and then evaluated on the unchanged test predictions. Full preprocessing, model settings, and aggregation details appear in § D and § E.4.

4.2 Calibration and Residual Distance

Groupwise isotonic calibration moves the group mean residuals closer to zero while residual distance can increase, consistent with known calibration–parity trade-offs (Kleinberg et al., 2017; Pleiss et al., 2017). On Adult, the summed absolute mean residual decreases from 0.017 to 0.008 while residual distance increases from 0.133 to 0.147; ACSIncome shows the same overall pattern (Table 1).

Permutation tests separate fixed-score sampling variation from calibration-fitting variation (§ E.5). Every Adult and ACS White–Black split exceeds the fixed-score null. After refitting calibration under shuffled validation labels, three of five Adult splits remain significant, while one ACS White–Black split reaches significance. The Adult result therefore provides clearer evidence of an increase after calibration, while the ACS result is more sensitive to how the calibration map is fitted.

The FN/FP split shows that the Adult increase comes from false positives: the FP component rises from 0.059 to 0.074, while the FN component is essentially unchanged. Figure 2(A) shows the corresponding high-confidence error gaps.

Takeaway. Groupwise calibration can move group mean residuals closer to zero while residual distance rises; on Adult, the increase comes from false positives.

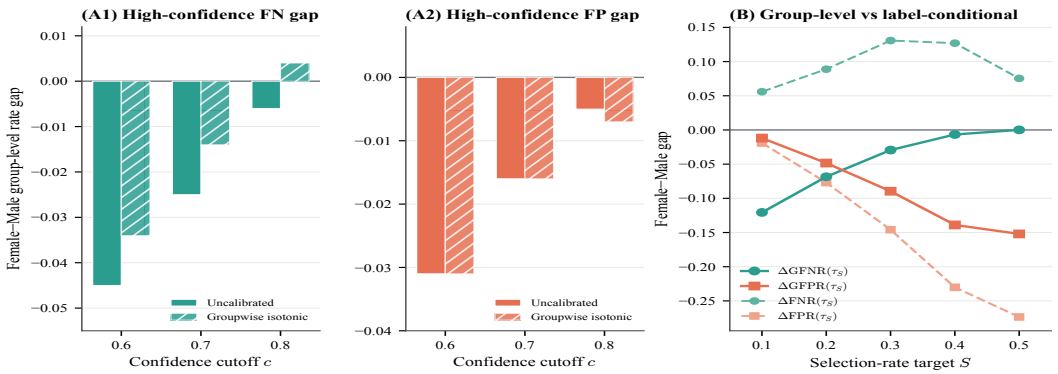

Figure 2: Adult Female vs. Male error gaps. (A) Calibration: high-confidence FN and FP gaps, measured as fractions of the full group, at cutoffs 0.6, 0.7, and 0.8 before and after groupwise isotonic calibration. (B) Thresholds: group FN/FP gaps (solid) and conditional FNR/FPR gaps (dashed) as the target selection rate S increases (Table 6).

Table 2: Residual distance before and after equalizing group prevalence (§ 3.3). "Observed − equalized" is positive when the observed label proportions increase the distance and negative when they reduce it. Adult has a large prevalence contribution on the FN side; for MHS, most of the observed distance comes from the prevalence difference.

Dataset	W_{res}	Distance with same prevalence	Observed − equalized	$W_{Y=1}$ / $W_{Y=0}$
Adult (F vs. M)	0.133	0.097	+0.036	0.070 / 0.104
ACSIncome (10-pair)	0.050	0.068	−0.018	0.101 / 0.052
MHS (RoBERTa)	0.098	0.009	+0.089	0.013 / 0.006
HateXplain (RoBERTa)	0.138	0.237	−0.099	0.208 / 0.299

4.3 Prevalence and Threshold Choice

Residual distance depends on both the score distributions among positive and negative examples and the label proportions in each group. We separate these factors by assigning both groups the average prevalence while keeping their score distributions within each label fixed. Table 2 compares the observed distance with the distance after giving the two groups the same prevalence (§ 3.3).

On Adult, equalizing prevalence reduces the FN component sharply, showing that much of the observed group FN gap comes from unequal label proportions. ACSIncome shows the reverse pattern: equalizing prevalence increases residual distance, so the observed label proportions reduce part of the difference between the score distributions. For Measuring Hate Speech, most of the observed residual distance disappears after prevalence is equalized, while HateXplain retains much larger score differences among positive and negative examples (Table 2). The reweighting procedure used for this comparison is described in § E.6.

We next vary the operating threshold while keeping the fitted score distributions fixed. For each target selection rate, the threshold is selected on validation data and then held fixed for test evaluation (§ E). The threshold changes the reported EO gap in different directions across datasets. At a 50% selection rate, Adult increases relative to threshold 0.5, while the ACSIncome ten-pair mean decreases (Figure 3).

Group FN/FP gaps and conditional FNR/FPR can also differ at the same threshold because the label proportions differ across groups. At the 50% selection point on Adult, the group FN gap is approximately zero while the FNR gap is 0.075. ACSIncome shows the same separation in magnitude across its ten group pairs (Table 3). The exact prevalence decomposition and the full threshold grids are given in Proposition B.11 and § E.

Takeaway. Prevalence and threshold choice can substantially change the relation between group FN/FP gaps and conditional FNR/FPR.

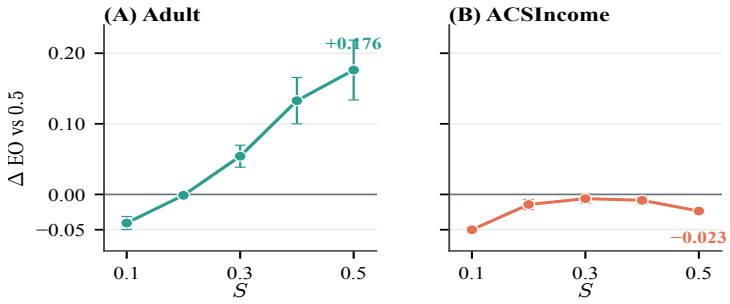

Figure 3: Change in the test-set equalized-odds gap relative to threshold 0.5, at validation-selected thresholds for each target S (connected markers with seed-wise error bars). (A) Adult; (B) ACSIncome. At $S = 0.50$ the changes are $+0.176 \pm 0.042$ and -0.023 ± 0.004. Full grids are in § E.

Table 3: False-negative gaps at selected thresholds. Adult reports Female minus Male; ACS reports the mean absolute gap across ten race-group pairs. Group FN is measured over the full group, while FNR is measured among positive examples. Thresholds are chosen on validation data. The gray row highlights the 50% selection point.

Threshold	Adult group FN	Adult FNR	ACS abs. group FN	ACS abs. FNR
Threshold 0.5	−0.067	+0.090	0.008	0.141
10% selection	−0.121	+0.056	0.080	0.099
30% selection	−0.029	+0.131	0.021	0.141
50% selection	+0.000	+0.075	0.009	0.102

4.4 Validation Metrics for Test-Time Error Gaps

We evaluate 432 ACSIncome White–Black model configurations spanning 12 states, three model families, four score variants, and three random seeds (§ D.3). For each model, validation metrics are compared with related test-set error quantities using Spearman correlation; uncertainty comes from state-clustered bootstrap resampling.

Spearman correlation is used because the question is about model ordering. A validation metric is useful for this experiment when models that rank high or low on validation receive a similar ordering on the related test quantity. The analysis therefore evaluates whether the FN and FP summaries preserve relative model rankings, rather than whether their numerical scales match the test values.

The validation FN and FP components preserve model ordering most strongly for the matching test quantities. For the average FN gap, the within-state mean correlation is 0.578 for the matching validation FN component and −0.146 for EO. After accounting for state prevalence differences, the corresponding partial correlations are 0.795 and 0.182 (Table 8). The pooled results show the same pattern. The validation FN and FP components correlate at 0.941 and 0.797 with their test counterparts, and the validation FN component correlates at 0.845 with the test FN cost gap (Table 4). The corresponding EO correlations are 0.354, 0.588, and 0.209.

The mean residual gap illustrates why a single mean can give a different ranking. It has high validation–test self-correlation (0.913) while correlating negatively with the test FP gap (−0.294). A single signed mean combines FN and FP contributions, so movement on one side can offset movement on the other; separate FN and FP components keep the two error types distinct. Within-state and prevalence-adjusted results are reported in Table 8.

Takeaway. For the evaluated targets, the matching validation residual component preserves test-time model rankings more closely than fixed-threshold EO or a mean-residual summary.

4.5 Text Classification

Toxicity classifiers on HateXplain (Mathew et al., 2021) and Measuring Hate Speech (Sachdeva et al., 2022) provide an additional test on text classification. On Measuring Hate Speech, the RoBERTa residual distance is driven primarily by prevalence, while HateXplain shows much larger score

Table 4: Pooled Spearman correlations across 432 ACSIncome models. Avg. FN/FP gaps are averaged over the selected threshold range; FN cost uses the false-negative cost defined in § 3.4. Shaded cells mark matching FN/FP comparisons, and brackets show state-clustered 95% bootstrap intervals.

Validation predictor	Self	Avg. FN gap	Avg. FP gap	FN cost gap	High confidence FP (0.7)
Avg. FN gap	0.941	0.941 [0.83, 0.96]	0.223 [−0.17, 0.45]	0.845 [0.60, 0.91]	0.090 [−0.18, 0.27]
Avg. FP gap	0.797	0.164 [−0.20, 0.41]	0.797 [0.67, 0.86]	−0.002 [−0.44, 0.30]	0.627 [0.43, 0.73]
EO gap (0.5)	0.725	0.354 [0.02, 0.54]	0.588 [0.39, 0.74]	0.209 [−0.13, 0.41]	0.449 [0.29, 0.61]
Mean residual gap ($\lvert s_{ij} \rvert$)	0.913	0.480 [0.11, 0.68]	−0.294 [−0.58, −0.07]	0.687 [0.54, 0.80]	−0.340 [−0.53, −0.18]
Score distance	0.882	0.448 [−0.02, 0.70]	0.538 [0.21, 0.76]	0.213 [−0.20, 0.44]	0.415 [0.18, 0.62]
Positive-label distance	0.737	0.360 [−0.03, 0.58]	0.591 [0.39, 0.75]	0.144 [−0.19, 0.34]	0.497 [0.33, 0.66]
Negative-label distance	0.864	0.304 [−0.15, 0.63]	0.674 [0.39, 0.84]	0.060 [−0.33, 0.40]	0.535 [0.30, 0.70]

Table 5: Residual results for fixed RoBERTa toxicity scores. Gray columns show residual distance and its FN/FP components; brackets are 95% bootstrap intervals.

Data	Model	Sample sizes	π_i	π_j	Residual distance	FN component	FP component	Positive label distance	Negative label distance	EO gap (0.5)
HateXplain	RoBERTa	439/1576	0.841	0.514	0.138 [0.10, 0.18]	0.054 [0.03, 0.09]	0.084 [0.06, 0.11]	0.208	0.299	0.299 [0.20, 0.43]
MHS	RoBERTa	4888/8668	0.522	0.333	0.098 [0.09, 0.11]	0.015 [0.01, 0.02]	0.083 [0.07, 0.09]	0.013	0.006	0.015 [0.01, 0.03]

differences among positive and negative examples (Tables 2 and 5). The threshold results also show that the group FN gap can differ from the FNR gap. Models and text-classification details are provided in § E.7.

5 Discussion

The central result gives residual distance a direct error interpretation across thresholds rather than replacing existing fairness criteria. Residual FN/FP components measure differences in the fraction of each group receiving each error type as the threshold varies. EO measures conditional rates at a selected threshold, while score-distribution measures capture other properties of the scores. These views answer different questions and can therefore move in different directions. The useful measure depends on whether the application emphasizes conditional performance, group FN/FP rates, ranking, calibration, or an application-specific cost.

Residual distance combines score differences within each label with prevalence, so it is most informative alongside EO and score-distribution measures. A practical report can therefore show W_{res} together with B_{FN} and B_{FP}, group prevalences, and the usual conditional error rates at the selected deployment threshold. W_{res} summarizes the overall difference across thresholds, the FN/FP components identify the error type, and prevalence helps explain why the group gaps can differ from FNR/FPR. These quantities answer related but different questions, so presenting them together gives a clearer picture than relying on one number alone.

6 Conclusion

We establish an exact relation between group residual distance and errors across thresholds: the Wasserstein distance is the sum of the FN and FP group-gap areas. The two components identify which error type drives the difference, connect to conditional score behavior through prevalence, and relate to corresponding error costs. An exactly calibrated construction shows that the examined threshold-based fairness, ranking, calibration, and average-error summaries can agree while residual distance remains. Empirically, group mean residuals can move closer to zero while residual distance grows, and prevalence can either enlarge or reduce the observed group error difference. Across 432 model configurations, validation FN/FP components preserve rankings of related test-time error differences more closely than fixed-threshold EO or a mean-residual summary for the evaluated targets. Together, these results make residual distributions a direct way to examine group error differences across decision thresholds.

Ethics Statement

Our analysis is based on observed labels, which may themselves reflect bias or measurement error. The reported residual differences therefore describe behavior relative to those labels; claims about real-world harm require additional application-specific evidence. Indigenous and NHPI groups in ACSIncome are excluded from the main comparisons because their sample sizes are substantially smaller. We use residual distance as an evaluation measure in this work; optimizing it directly could involve trade-offs with predictive performance and would require application-specific consideration of who benefits from the resulting changes.

Reproducibility Statement

Appendix protocols specify preprocessing, fixed split seeds, hyperparameters, calibration fitting, threshold selection, bootstrap procedures, and permutation tests with scores held fixed and with calibration refitted for ΔW_{res}. We will release anonymized code and scripts to regenerate all tables upon acceptance (or via an anonymous repository link at submission if the venue permits).

AI Use Statement

A language-model assistant was used to help check consistency across reported numerical results and to improve wording and readability. All research ideas, methodological choices, experiments, analyses, and conclusions are the authors' own, and the authors take full responsibility for the paper.

References

Ann-Kristin Becker, Oana Dumitrasc, and Klaus Broelemann. Standardized interpretable fairness measures for continuous risk scores. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 3327–3346. PMLR, 2024. URL https://proceedings.mlr.press/v235/becker24a.html.

Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018. URL https://arxiv.org/abs/1810.01943.

Sarah Bird, Miroslav Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32*, 2020. URL https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *The Web Conference (WWW) Companion*, pp. 491–500, 2019. URL https://doi.org/10.1145/3308560.3317593.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017. URL https://doi.org/10.1089/big.2016.0047.

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Fair regression with wasserstein barycenters. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 7321–7331, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/51cdbd2611e844ece5d80878eb770436-Abstract.html.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 797–806, 2017. URL https://doi.org/10.1145/3097983.3098095.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 6478–6490, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/32e54441e6382a7fbacbbbaf3c450059-Abstract.html.

Josh Gardner, Christopher Brooks, and Ryan Baker. Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge (LAK)*, pp. 225–234, 2019. doi: 10.1145/3303772.3303791. URL https://doi.org/10.1145/3303772.3303791.

Paula Gordaliza, Eustasio Del Barrio, Fabrice Gamboa, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning (ICML)*, pp. 2357–2365, 2019. URL https://proceedings.mlr.press/v97/gordaliza19a.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Xiaotian Han, Zhimeng Jiang, Hongye Jin, Zirui Liu, Na Zou, Qifan Wang, and Xia Hu. Retiring δDP: New distribution-level metrics for demographic parity. *arXiv preprint arXiv:2301.13443*, 2023. URL https://arxiv.org/abs/2301.13443.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html.

Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning (ICML)*, pp. 1939–1948, 2018. URL https://proceedings.mlr.press/v80/hebert-johnson18a.html.

Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. Wasserstein fair classification. In *Uncertainty in Artificial Intelligence (UAI)*, 2019. URL https://proceedings.mlr.press/v115/jiang20a.html.

Nathan Kallus and Angela Zhou. The fairness of risk scores beyond classification: Bipartite ranking and the xAUC metric. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/73e0f7487b8e5297182c5a711d20bf26-Abstract.html.

Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *AAAI/ACM Conference on AI, Ethics, and Society*, pp. 247–254, 2019. URL https://doi.org/10.1145/3306618.3314287.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Innovations in Theoretical Computer Science (ITCS)*, 2017. URL https://doi.org/10.4230/LIPIcs.ITCS.2017.43.

Kweku Kwegyir-Aggrey, Jessica Dai, A. Feder Cooper, John Dickerson, Keegan Hines, and Suresh Venkatasubramanian. Repairing regressors for fair binary classification at any decision threshold. *arXiv preprint arXiv:2203.07490*, 2022. URL https://arxiv.org/abs/2203.07490.

Kristian Lum, Yunfeng Zhang, and Amanda Bower. De-biasing "bias" measurement. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 379–389, 2022. doi: 10.1145/3531146.3533105. URL https://doi.org/10.1145/3531146.3533105.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. HateXplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 14867–14875, 2021. URL https://doi.org/10.1609/aaai.v35i17.17745.

Alexey Miroshnikov, Konstandinos Kotsiopoulos, Ryan Franks, and Arjun Ravi Kannan. Wasserstein-based fairness interpretability framework for machine learning models. *Machine Learning*, 111: 3307–3357, 2022. URL https://doi.org/10.1007/s10994-022-06213-9.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/b8b9c74ac526fffbeb2d39ab038d1cd7-Abstract.html.

Pratik S. Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pp. 83–94. European Language Resources Association, 2022. URL https://aclanthology.org/2022.nlperspectives-1.11/.

Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T. Rodolfa, and Rayid Ghani. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*, 2018. URL https://arxiv.org/abs/1811.05577.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009. URL https://doi.org/10.1007/978-3-540-71050-9.

Robin Vogel, Aurélien Bellet, and Stephan Clémençon. Learning fair scoring functions: Bipartite ranking under ROC-based fairness constraints. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 784–792. PMLR, 2021. URL https://proceedings.mlr.press/v130/vogel21a.html.

A Construction with Equal Summary Metrics

This is an analytical population construction. Let A take values $0, 1$ with positive probabilities. Set $X = Z$, $f(X) = Z$, and $Y \mid Z, A \sim \text{Bernoulli}(Z)$. Conditional on $A = 0$, the scores $(0.1, 0.3, 0.7, 0.9)$ have masses $(1, 3, 3, 1)/8$. Conditional on $A = 1$, the scores $(0.2, 0.4, 0.6, 0.8)$ have masses $(3, 1, 1, 3)/8$. This fully specifies the joint distribution.

A.1 Calibration and Predictive Performance

By construction, $\mathbb{E}[Y \mid Z, A] = Z$. Since $X = Z$ and $f(X) = X$, it follows that $\mathbb{E}[Y \mid X] = X$, so the predictor is Bayes-optimal for squared loss on the σ-algebra generated by X. Both score distributions are symmetric about $1/2$, so $\mathbb{E}[Z \mid A] = \pi_a = 1/2$ and $\mu_a = 0$. At threshold $1/2$, both select half the group and have

$$\mathbb{E}[\max(Z, 1 - Z) \mid A] = \tfrac{3}{4}, \qquad \mathbb{E}[Z\mathbb{I}\{Z \geq 1/2\} \mid A] = \tfrac{3}{8}.$$

Accuracy is $3/4$, TPR is $(3/8)/(1/2) = 3/4$, and FPR is $(1/2 - 3/8)/(1/2) = 1/4$ in each group. The Brier score is

$$\mathbb{E}[(Z - Y)^2 \mid A] = \mathbb{E}[Z(1 - Z) \mid A] = \tfrac{9}{50} = 0.18.$$

The mean absolute residual and one-sided magnitudes also match: $\mathbb{E}[|D| \mid A] = 9/25 = 0.36$ and $\mathbb{E}[\max(0, -D) \mid A] = \mathbb{E}[\max(0, D) \mid A] = 9/50 = 0.18$ in each group. Population calibration error is zero both pointwise on the support and in every score bin of positive probability. The equality of EO stated here holds at threshold $1/2$.

A.2 Residual Distributions

At score s with mass w, the residual has mass $w(1 - s)$ at s and mass ws at $s - 1$. The positive residual masses are $(0.1, 0.3, 0.7, 0.9)$ with weights $(9, 21, 9, 1)/80$ in group 0 and $(0.2, 0.4, 0.6, 0.8)$ with weights $(12, 3, 2, 3)/40$ in group 1. By symmetry of the score distributions, the negative side is the reflection of the positive side, with the same masses at reflected residual values; each side

accounts for half of the group's probability. Both distributions have zero mean and second moment 0.18, but their supports differ. For the sorted union of supports $x_1 < \cdots < x_{16}$, exact summation gives

$$W_1(F_0, F_1) = \sum_{k=1}^{15} (x_{k+1} - x_k)|F_0(x_k) - F_1(x_k)| = \tfrac{1}{10}.$$

By symmetry together with Proposition B.4, each signed area is $1/20$. The quantile signed areas A^+, A^- from Proposition B.4 and the half-axis components $B_{\mathrm{FN}}, B_{\mathrm{FP}}$ are different splits in general; they coincide here only by the construction's symmetry. For example, $Q_0(0.05) = -0.7 > Q_1(0.05) = -0.8$, and $Q_0(0.10) = -0.7 < Q_1(0.10) = -0.6$. These orderings persist on intervals between the cumulative masses.

A.3 High-Confidence FN/FP Rates

Under $\hat{Y} = \mathbb{I}[\hat{P} \geq 1/2]$, a residual with $|D| \geq c$ for $c \geq 1/2$ is a wrong prediction with predicted-class confidence at least c. At cutoffs 0.6, 0.7, and 0.8, the two groups have rates $(1/4, 1/4)$, $(1/4, 3/20)$, and $(1/40, 3/20)$, respectively. The gap is zero at 0.6, positive at 0.7, and negative at 0.8, showing how the full residual distribution captures variation across confidence levels.

A.4 Error Rates Across Thresholds

Because $\pi_0 = \pi_1 = 1/2$, $B_{\mathrm{FN}} = \frac{1}{2}\int_0^1 |\Delta\mathrm{FNR}|\, d\tau$ and $B_{\mathrm{FP}} = \frac{1}{2}\int_0^1 |\Delta\mathrm{FPR}|\, d\tau$, so the sum of the integrated absolute FNR and FPR differences is $2W_{\mathrm{res}} = 1/5$. The score distances within each label are $W_1(\hat{P} \mid Y{=}1) = W_1(\hat{P} \mid Y{=}0) = 1/10$, which gives the same integrated differences. At $\tau = 1/2$ one has $\mathrm{EOgap} = 0$, and $\mathrm{EOgap}(1/4) = \mathrm{EOgap}(3/4) = 3/8$; these are the maximizing thresholds on the finite support. The signed integrals satisfy $\int_0^1 \Delta\mathrm{GFNR}(\tau)\, d\tau = \int_0^1 \Delta\mathrm{GFPR}(\tau)\, d\tau = 0$, and $B_{\mathrm{FN}} = B_{\mathrm{FP}} = 1/20$. Group FN/FP gaps with opposite signs on different threshold intervals cancel in the signed mean but remain in the absolute components. The construction yields $W_{\mathrm{res}} = B_{\mathrm{FN}} + B_{\mathrm{FP}} = 0.10$ with equal blocks 0.05, score distance 0.10 within each label, integrated absolute FNR and FPR differences summing to 0.20, and $\sup_\tau \mathrm{EOgap} = 0.375$, matching Figure 1.

The listed single-threshold, ranking, calibration, and average-error metrics can agree even when high-confidence FN/FP gaps differ across confidence levels. The construction therefore separates agreement in those metrics from agreement of the full residual distributions. A score-distribution audit can also distinguish these groups; the construction concerns the listed single-threshold, ranking, calibration, and average error metrics.

B Proofs and Additional Results

B.1 Error-Rate Representation and Decomposition

The main text states the compact decomposition in Equation (3). This section gives the full integral representation for the group FN and FP rates. For the proofs we write $\pi_a = \mathbb{P}(Y = 1 \mid A = a)$, $F_a(d) = \mathbb{P}(D \leq d \mid A = a)$, $\mathrm{FNR}_a(\tau) = \mathbb{P}(\hat{Y} = 0 \mid Y = 1, A = a)$, $\mathrm{FPR}_a(\tau) = \mathbb{P}(\hat{Y} = 1 \mid Y = 0, A = a)$, and

$$\begin{aligned} \mathrm{GFNR}_a(\tau) &= \mathbb{P}(Y{=}1, \hat{Y}{=}0 \mid A = a) = \pi_a \mathrm{FNR}_a(\tau), \\ \mathrm{GFPR}_a(\tau) &= \mathbb{P}(Y{=}0, \hat{Y}{=}1 \mid A = a) = (1 - \pi_a)\mathrm{FPR}_a(\tau), \end{aligned} \tag{8}$$

the group FN/FP rates (joint label-and-decision probabilities within group a; the fraction of group a receiving an FN or FP), with gaps $\Delta\mathrm{GFNR} = \mathrm{GFNR}_i - \mathrm{GFNR}_j$ and $\Delta\mathrm{GFPR} = \mathrm{GFPR}_i - \mathrm{GFPR}_j$. The EO gap is $\mathrm{EOgap}(\tau) = \max\{|\Delta\mathrm{FNR}(\tau)|, |\Delta\mathrm{FPR}(\tau)|\}$.

Theorem B.1 (Error-curve representation)**.** *Fix a group a with $\pi_a \in (0, 1)$. Under $\hat{Y} = \mathbb{I}[\hat{P} \geq \tau]$ for $\tau \in (0, 1)$,*

$$F_a\big((\tau - 1)^-\big) = \mathrm{GFNR}_a(\tau), \qquad 1 - F_a(\tau^-) = \mathrm{GFPR}_a(\tau). \tag{9}$$

Corollary B.2 (Block decomposition)**.**

$$W_{\mathrm{res}}(i,j) = B_{\mathrm{FN}}(i,j) + B_{\mathrm{FP}}(i,j) = \int_0^1 |\Delta\mathrm{GFNR}(\tau)|\, d\tau + \int_0^1 |\Delta\mathrm{GFPR}(\tau)|\, d\tau, \tag{10}$$

which is Equation (3) *written with the abbreviations above.*

One-sided bounds for costs defined on prediction errors and Π-weighted threshold components are stated and proved in § B.3.

Proof of Theorem B.1

Proof. Since $\hat{P} = D+Y$, the event $\{\hat{Y}=1\}$ is $\{D \geq \tau - Y\}$. On $Y=0$, $D=\hat{P}\in[0,1]$; on $Y=1$, $D=\hat{P}-1\in[-1,0]$. For $\tau\in(0,1)$ the point $\tau-1$ lies in $(-1,0)$, so $\mathbb{P}(D<\tau-1 \mid Y=0, A=a)=0$ and the mixture $F_a=(1-\pi_a)F_a^{Y=0}+\pi_a F_a^{Y=1}$ gives $F_a((\tau-1)^-)=\pi_a\mathbb{P}(D<\tau-1 \mid Y=1, A=a)=\pi_a\mathrm{FNR}_a(\tau)=\mathrm{GFNR}_a(\tau)$. Symmetrically $\mathbb{P}(D\geq\tau \mid Y=1, A=a)=0$ for $\tau>0$, so $1-F_a(\tau^-)=(1-\pi_a)\mathbb{P}(D\geq\tau \mid Y=0, A=a)=\mathrm{GFPR}_a(\tau)$. The left limits make the two displays exact under score atoms; for continuous score distributions they may be dropped. □

Proof of Corollary B.2

Proof. W_1 between distributions on $[-1,1]$ is $\int_{-1}^1 |F_i-F_j|\, dt$ for any CDF versions that differ at most on a countable set. Split the integral at $t=0$ and substitute Equation (9) with $\tau=t+1$ on $[-1,0)$ and $\tau=t$ on $(0,1]$. Each half-axis has Lebesgue length 1, so $B_{\mathrm{FN}}(i,j)=\int_0^1|\Delta\mathrm{GFNR}|\,d\tau$ and $B_{\mathrm{FP}}(i,j)=\int_0^1|\Delta\mathrm{GFPR}|\,d\tau$. Endpoint mass has Lebesgue measure zero in the integral. $F_a((-1)^-)=0$ and $1-F_a(1^-)=0$. □

B.2 Mean and Score-Distance Bounds

Differences at residual quantiles can cancel in a difference of means. Writing $\mu_a=\mathbb{E}[D \mid A=a]$ and $s_{ij}=\mu_i-\mu_j$,

$$W_{\mathrm{res}}(i,j) \geq |s_{ij}| = |\mu_i-\mu_j|, \qquad |\mu_i-\mu_j| \leq |\mu_i|+|\mu_j|. \tag{11}$$

Thus a small mean gap need not imply a small residual distance: opposite-signed quantile differences can cancel in s_{ij} while remaining in W_{res}. A second bound relates residual distance to label proportions and marginal score distance $W_{\mathrm{score}}(i,j)=W_1(\mathcal{L}(\hat{P} \mid A=i), \mathcal{L}(\hat{P} \mid A=j))$.

Proposition B.3 (Lower bound from label proportions and score distributions)**.** *For groups i and j,*

$$W_{\mathrm{res}}(i,j) \;\geq\; \max\{0,\; |\pi_i-\pi_j| \;-\; W_{\mathrm{score}}(i,j)\}. \tag{12}$$

Write $L=\max\{0, |\Delta\pi| - W_{\mathrm{score}}\}$ with $\Delta\pi=\pi_i-\pi_j$ and $\Delta\bar{P}=\mathbb{E}[\hat{P} \mid A=i]-\mathbb{E}[\hat{P} \mid A=j]$. Then

$$L \leq [|\Delta\pi|-|\Delta\bar{P}|]_+ \leq |\Delta\bar{P}-\Delta\pi| = |s_{ij}| \leq W_{\mathrm{res}}(i,j). \tag{13}$$

Under exact population calibration, both group means are zero, so $L=0$ and $W_{\mathrm{score}}\geq|\Delta\pi|$. A zero lower bound leaves residual distance free to stay positive (§ 3.5). The middle expression in Equation (13) uses the mean-score gap and can be tighter than the score-distance bound in Equation (12).

Proposition B.4 (Mean–Wasserstein identity)**.** *Let $\Delta_{ij}(u)=Q_i(u)-Q_j(u)$, $A^+(i,j)=\int_0^1[\Delta_{ij}(u)]_+\,du$ and $A^-(i,j)=\int_0^1[-\Delta_{ij}(u)]_+\,du$ with $[x]_+=\max(x,0)$. Then $W_{\mathrm{res}}=A^++A^-$, $s_{ij}=A^+-A^-=\mu_i-\mu_j$, and $W_{\mathrm{res}}\geq|s_{ij}|$, with equality exactly when Δ_{ij} has one sign almost everywhere.*

Proof. Summing Δ_{ij} over $[0,1]$ gives the difference of means. Writing $\Delta_{ij}=[\Delta_{ij}]_+-[-\Delta_{ij}]_+$ and adding or subtracting the signed areas gives the rest. □

Proposition B.5 (Integrated error-rate differences). *Let* $\bar{q}_a = \mathbb{E}[1 - \hat{P} \mid Y = 1, A = a]$ *and* $\bar{p}_a = \mathbb{E}[\hat{P} \mid Y = 0, A = a]$. *Then* $\int_0^1 \mathrm{FNR}_a \, d\tau = \bar{q}_a$ *and* $\int_0^1 \mathrm{FPR}_a \, d\tau = \bar{p}_a$, *so when each signed gap* $\Delta\mathrm{FNR}$ *and* $\Delta\mathrm{FPR}$ *has one sign almost everywhere on* $[0, 1]$,

$$\int_0^1 |\Delta\mathrm{FNR}| \, d\tau + \int_0^1 |\Delta\mathrm{FPR}| \, d\tau = \left|\bar{q}_i - \bar{q}_j\right| + \left|\bar{p}_i - \bar{p}_j\right|. \tag{14}$$

In general the sum on the left is at least the expression on the right, and it is zero if and only if equalized odds holds at every threshold.

In particular, $\int_0^1 \Delta\mathrm{FNR} \, d\tau = \bar{q}_i - \bar{q}_j$, so balance for the positive class holds iff the signed integrated FNR gap is zero.

Remark B.6 (Uninformative scores and prevalence). If $\hat{P}$ has the same marginal distribution in both groups and is independent of Y given A, then the groups share label-conditional score laws and $\bar{q} + \bar{p} = 1$, so the last clause of Proposition B.12 gives $W_{\mathrm{res}} = |\Delta\pi|$. For Adult-like prevalences near 0.11 and 0.31, this yields $W_{\mathrm{res}} \approx 0.20$, larger than the trained logistic model's 0.133. Because residual distance uses FN/FP rates over the full group, a constant predictor can still have a large value when group prevalences differ. Residual distance should therefore be reported together with score distances computed within each label.

Proof of Proposition B.3

Proof. By Proposition B.4, $W_{\mathrm{res}}(i, j) \geq |s_{ij}| = |\mu_i - \mu_j|$. Writing $\mu_a = \mathbb{E}[\hat{P} \mid A = a] - \pi_a$ gives $\mu_i - \mu_j = \Delta\bar{P} - (\pi_i - \pi_j)$, so

$$W_{\mathrm{res}}(i, j) \;\geq\; |\Delta\bar{P} - (\pi_i - \pi_j)| \;\geq\; |\pi_i - \pi_j| - |\Delta\bar{P}| \;\geq\; |\pi_i - \pi_j| - W_{\mathrm{score}}(i, j),$$

where the last step uses that Wasserstein-1 is at least the absolute mean difference (Villani, 2009). Taking the positive part yields Equation (12). Equal score distributions have $W_{\mathrm{score}} = 0$, giving the stated special case. The independent-score example attains equality and is already detected by signed mean residuals. □

B.3 Error Costs and Weighted Thresholds

This subsection gives the full dual characterization and proofs for the cost bounds summarized in the main text. By Kantorovich–Rubinstein duality, residual distance equals the largest absolute group gap over 1-Lipschitz functions of the residual; the FN and FP components admit analogous dual characterizations on functions restricted to the negative and positive residual ranges. Those components also bound one-sided residual costs. For a general threshold-weight distribution Π with density ρ_Π, the corresponding weighted gaps are dual to absolutely continuous functions with $|f'| \leq \rho_\Pi$ (Corollary B.9), which reduces to the 1-Lipschitz case when Π is Lebesgue on $[0, 1]$.

Theorem B.7 (Dual characterization). $W_{\mathrm{res}}(i, j)$ *equals the largest group gap* $|\mathbb{E}_i f(D) - \mathbb{E}_j f(D)|$ *over* 1*-Lipschitz functions* f *on* $[-1, 1]$. *Restricting* f *to vanish on the false-positive side (functions depending only on negative residuals) gives exactly* $B_{\mathrm{FN}}(i, j)$*; restricting it to vanish on the false-negative side (functions depending only on positive residuals) gives exactly* $B_{\mathrm{FP}}(i, j)$. *Each supremum is attained.*

Write $\mathcal{F}_1$ for the 1-Lipschitz class on $[-1, 1]$, $\mathcal{F}_1^-$ for functions depending only on negative residuals (vanish on $[0, 1]$), and $\mathcal{F}_1^+$ for functions depending only on positive residuals (vanish on $[-1, 0]$). The formal dualities stated in Theorem B.7 are

$$W_{\mathrm{res}}(i, j) = \sup_{f \in \mathcal{F}_1} \left|\mathbb{E}_i f(D) - \mathbb{E}_j f(D)\right|, \tag{15}$$

$$B_{\mathrm{FN}}(i, j) = \sup_{f \in \mathcal{F}_1^-} \left|\mathbb{E}_i f(D) - \mathbb{E}_j f(D)\right|, \qquad B_{\mathrm{FP}}(i, j) = \sup_{f \in \mathcal{F}_1^+} \left|\mathbb{E}_i f(D) - \mathbb{E}_j f(D)\right|. \tag{16}$$

Corollary B.8 (One-sided residual-cost bounds). *For every* $f \in \mathcal{F}_1^-$, *the absolute group gap* $|\mathbb{E}_i f(D) - \mathbb{E}_j f(D)|$ *is at most* $B_{\mathrm{FN}}(i, j)$. *Nonnegativity and monotonicity of a miss cost are only an example of this class: for instance,* $c(D) = \max(0, -D)$ *belongs to* $\mathcal{F}_1^-$ *and yields* $|\pi_i \bar{q}_i - \pi_j \bar{q}_j| \leq B_{\mathrm{FN}}(i, j)$. *The analogous bound for every* $f \in \mathcal{F}_1^+$ *is* $B_{\mathrm{FP}}(i, j)$.

An ε-smoothed high-confidence indicator is $(1/\varepsilon)$-Lipschitz, so its absolute group gap is at most $B_{\mathrm{FN}}/\varepsilon$ (and likewise $B_{\mathrm{FP}}/\varepsilon$ on the FP side).

Corollary B.9 (Weighted operating points)**.** *For a probability measure Π on deployment thresholds in $[0,1]$, the Π-weighted FN component*

$$B_{\Pi}^{\mathrm{FN}}(i,j) := \int_{[0,1]} \big|\Delta\mathrm{GFNR}(\tau)\big|\, \Pi(\mathrm{d}\tau)$$

is the expected absolute group FN gap when thresholds are drawn from Π; B_{Π}^{FP} is defined analogously. When Π is Lebesgue on $[0,1]$, these reduce to B_{FN} and B_{FP} in Equation (3). In the experiments we call the case with Π uniform on the validation threshold interval for selection rates $S \in [0.05, 0.50]$ the average FN/FP gap over that deployment threshold range. If Π has density ρ_Π, then B_{Π}^{FN} equals $\sup|\mathbb{E}_i f - \mathbb{E}_j f|$ over absolutely continuous f with $f(0) = 0$, $f \equiv 0$ on $[0,1]$, and $|f'| \le \rho_\Pi(\cdot + 1)$ a.e. on $[-1,0]$—not, in general, over the 1-Lipschitz class (the bound $|f'| \le \rho_\Pi$ may exceed 1 when Π concentrates on a short band). Accordingly, $B_{\Pi}^{\mathrm{FN}} \le (\sup \rho_\Pi)\, B_{\mathrm{FN}}$, while a 1-Lipschitz cost gap remains bounded by B_{FN} (not by B_{Π}^{FN}).

We write F_a for the right-continuous CDF $\mathbb{P}(D \le \cdot \mid A = a)$ and use left limits $F_a(t^-) = \mathbb{P}(D < t \mid A = a)$ where the decision rule requires them. The maps $t \mapsto F_a(t)$ and $t \mapsto F_a(t^-)$ differ only on the at most countable atom set, so $\int |F_i - F_j|$ and the dual identities below are unaffected by this choice. Every $f \in \mathcal{F}_1$ is absolutely continuous, hence differentiable almost everywhere with $|f'| \le 1$ a.e., and $f(b) - f(a) = \int_a^b f'$ (Villani, 2009).

Lemma B.10 (Half-axis representation)**.** *If $f \in \mathcal{F}_1^-$ (so $f \equiv 0$ on $[0,1]$ and thus $f(0) = 0$), then for every group a,*

$$\mathbb{E}_a f(D) = -\int_{-1}^{0} f'(t)\, F_a(t)\, dt. \tag{17}$$

If $f \in \mathcal{F}_1^+$ (so $f \equiv 0$ on $[-1,0]$ and $f(0) = 0$), then

$$\mathbb{E}_a f(D) = \int_{0}^{1} f'(t)\, \big(1 - F_a(t)\big)\, dt.$$

Proof. On $[-1,0]$, $f(0) = 0$ yields $f(d) = -\int_d^0 f'(u)\, du$. Hence

$$f(D) = -\int_{-1}^{0} f'(u)\, \mathbf{1}\{D \le u\}\, du$$

almost surely. The identity holds on $\{D \le 0\}$, and both sides vanish on $\{D > 0\}$ because f is zero there and $\mathbf{1}\{D \le u\} = 0$ for all $u \le 0$. Taking expectations and applying Fubini (justified by $|f'| \le 1$ and finite measure) gives (17). The positive half-axis argument is symmetric. For $d \in [0,1]$, $f(d) = \int_0^d f'(u)\, du$, so $f(D) = \int_0^1 f'(u)\, \mathbf{1}\{D > u\}\, du$, and $\mathbb{E}_a \mathbf{1}\{D > u\} = 1 - F_a(u)$. Atoms are included through the CDF values; no continuity of F_a is required. □

Proof of Theorem B.7

Proof. The identity (15) is Kantorovich–Rubinstein duality for W_1 on the compact metric space $[-1,1]$; the supremum is attained (Villani, 2009).

For the FN half-axis, Lemma B.10 gives

$$\mathbb{E}_i f(D) - \mathbb{E}_j f(D) = -\int_{-1}^{0} f'(t)\big(F_i(t) - F_j(t)\big)\, dt$$

for every $f \in \mathcal{F}_1^-$.

$$\big|\mathbb{E}_i f(D) - \mathbb{E}_j f(D)\big| \le \int_{-1}^{0} |F_i(t) - F_j(t)|\, dt = B_{\mathrm{FN}}(i,j),$$

since $|f'| \le 1$ a.e. For attainability, define f by $f(0) = 0$, $f'(t) = -\operatorname{sign}\big(F_i(t) - F_j(t)\big)$ for $t \in [-1, 0]$ (with the convention $\operatorname{sign}(0) = 0$), and $f \equiv 0$ on $[0, 1]$. Then $f \in \mathcal{F}_1^-$ and

$$\mathbb{E}_i f(D) - \mathbb{E}_j f(D) = \int_{-1}^{0} |F_i(t) - F_j(t)|\, dt = B_{\mathrm{FN}}(i, j).$$

(The indefinite integral of a bounded measurable function is Lipschitz, so f is well-defined everywhere even though f' is specified only a.e.) The FP half-axis is identical after exchanging F_a for $1 - F_a$ on $[0, 1]$ and using $\mathcal{F}_1^+$. □

Proof of Corollary B.8

Proof. The bound for every $f \in \mathcal{F}_1^-$ (resp. $\mathcal{F}_1^+$) is immediate from (16). For the ramp $c(D) = \max(0, -D)$, one has $\mathbb{E}_a[c(D)] = \pi_a \bar{q}_a$, so $|\pi_i \bar{q}_i - \pi_j \bar{q}_j| \le B_{\mathrm{FN}}(i, j)$. □

Proof of Corollary B.9

Proof. The claim $B_\Pi^{\mathrm{FN}} = \mathbb{E}_{\tau\sim\Pi}|\Delta\mathrm{GFNR}(\tau)|$ is the definition of expectation under Π, with $\Delta\mathrm{GFNR}$ as in Corollary B.2. When Π is Lebesgue measure on $[0, 1]$, which is a probability measure of total mass one, $B_\Pi^{\mathrm{FN}}(i, j) = \int_0^1 |\Delta\mathrm{GFNR}|\, d\tau = B_{\mathrm{FN}}(i, j)$, and likewise for B_{FP}. The open or closed status of the endpoints $\{\tau = 0\}$ and $\{\tau = 1\}$ is immaterial under Lebesgue measure. If Π has density ρ_Π, the change of variables $\tau = t + 1$ on the FN half-axis shows that

$$B_\Pi^{\mathrm{FN}}(i, j) = \int_{-1}^{0} |F_i(t) - F_j(t)|\, \rho_\Pi(t + 1)\, dt,$$

which equals $\sup |\mathbb{E}_i f - \mathbb{E}_j f|$ over absolutely continuous f with $f(0) = 0$, $f \equiv 0$ on $[0, 1]$, and $|f'| \le \rho_\Pi(\cdot + 1)$ a.e. on $[-1, 0]$, by the same integration-by-parts argument as in the proof of Theorem B.7. This class is 1-Lipschitz only when $\rho_\Pi \le 1$ a.e.; for a uniform band of width $w < 1$ one has $\rho_\Pi = 1/w > 1$. Comparing to the unweighted dual, $B_\Pi^{\mathrm{FN}} \le (\sup \rho_\Pi)\, B_{\mathrm{FN}}$, while any 1-Lipschitz gap remains at most B_{FN}. For discrete or mixed Π, keep the integral form. □

B.4 Additional Decompositions

For $b \in \{\mathrm{FN}, \mathrm{FP}\}$, write Δ^b for $\Delta\mathrm{GFNR}$ or $\Delta\mathrm{GFPR}$ as appropriate, and define the signed integral $m^b = \int_0^1 \Delta^b(\tau)\, d\tau$ and the component distance $B^b = \int_0^1 |\Delta^b(\tau)|\, d\tau$. The mean difference is $\mu_i - \mu_j = -m^{\mathrm{FN}} + m^{\mathrm{FP}}$.

For a fixed threshold, the group FN gap can be written as

$$\pi_i \mathrm{FNR}_i(\tau) - \pi_j \mathrm{FNR}_j(\tau) = \bar{\pi}\, \Delta\mathrm{FNR}(\tau) + \Delta\pi\, \frac{\mathrm{FNR}_i(\tau) + \mathrm{FNR}_j(\tau)}{2}, \tag{18}$$

with the analogous identity for false positives.

Proposition B.11 (Component means). *Write $\bar{\pi} = (\pi_i + \pi_j)/2$, $\bar{q}_{\mathrm{mid}} = (\bar{q}_i + \bar{q}_j)/2$, $\bar{\nu} = 1 - \bar{\pi}$, and $\bar{p}_{\mathrm{mid}} = (\bar{p}_i + \bar{p}_j)/2$. Each signed component integral admits the decomposition*

$$\begin{aligned} m^{\mathrm{FN}} &= (\pi_i - \pi_j)\bar{q}_{\mathrm{mid}} + \bar{\pi}(\bar{q}_i - \bar{q}_j), \\ m^{\mathrm{FP}} &= (\pi_j - \pi_i)\bar{p}_{\mathrm{mid}} + \bar{\nu}(\bar{p}_i - \bar{p}_j). \end{aligned} \tag{19}$$

The first term on each line contains the difference in positive-label proportions; the second contains the difference in conditional mean prediction errors.

Proof of Proposition B.11

Proof. The identities $m^{\mathrm{FN}} = \pi_i \bar{q}_i - \pi_j \bar{q}_j$ and $m^{\mathrm{FP}} = (1 - \pi_i)\bar{p}_i - (1 - \pi_j)\bar{p}_j$ follow from Proposition B.5. Averaging the two one-sided references yields Equation (19). □

Proposition B.12 (Residual distance, label-conditional distances, and prevalence)**.** *Let* $\bar{\pi} = (\pi_i + \pi_j)/2$*,* $\Delta\pi = \pi_i - \pi_j$*,* $\bar{q}_{\mathrm{mid}} = (\bar{q}_i + \bar{q}_j)/2$*, and* $\bar{p}_{\mathrm{mid}} = (\bar{p}_i + \bar{p}_j)/2$*. Then*

$$\left|B_{\mathrm{FN}} - \bar{\pi}\, W_{Y=1}\right| \le |\Delta\pi|\, \bar{q}_{\mathrm{mid}}, \qquad \left|B_{\mathrm{FP}} - (1-\bar{\pi})\, W_{Y=0}\right| \le |\Delta\pi|\, \bar{p}_{\mathrm{mid}}. \tag{20}$$

When $\pi_i = \pi_j = \pi$*,*

$$W_{\mathrm{res}}(i,j) = \pi\, W_{Y=1}(i,j) + (1-\pi)\, W_{Y=0}(i,j). \tag{21}$$

If the groups share the same label-conditional score laws, $W_{\mathrm{res}}(i,j) = |\Delta\pi|(\bar{q} + \bar{p})$*.*

Proof of Proposition B.12

Proof. Pointwise, $\Delta\mathrm{GFNR}(\tau) = \bar{\pi}\,\Delta\mathrm{FNR}(\tau) + \Delta\pi\,(\mathrm{FNR}_i(\tau) + \mathrm{FNR}_j(\tau))/2$. Taking absolute values and integrating, the triangle inequality yields

$$\left|B_{\mathrm{FN}} - \bar{\pi}\int_0^1 |\Delta\mathrm{FNR}|\, d\tau\right| \le |\Delta\pi| \int_0^1 \tfrac{\mathrm{FNR}_i + \mathrm{FNR}_j}{2}\, d\tau = |\Delta\pi|\, \bar{q}_{\mathrm{mid}},$$

where the last step uses $\int \mathrm{FNR}_a\, d\tau = \bar{q}_a$ from Proposition B.5. The identity $W_{Y=1} = \int |\Delta\mathrm{FNR}|\, d\tau$ is the label-conditional case of Theorem B.1 (equivalently Becker et al. (2024), Thm. 3.2). The FP half is identical. When $\Delta\pi = 0$, the inequalities collapse to equalities and sum to Equation (21). When the label-conditional score laws agree, $W_{Y=1} = W_{Y=0} = 0$ and $B_{\mathrm{FN}} = |\Delta\pi|\bar{q}$, $B_{\mathrm{FP}} = |\Delta\pi|\bar{p}$. □

$A^-(i,j) = 0$ if and only if $Q_i(u) \ge Q_j(u)$ almost everywhere (first-order stochastic dominance of group i over j); the reverse holds when $A^+(i,j) = 0$. The identity $\mu_a = \mathbb{E}[\hat{P} \mid A = a] - \pi_a$ is linearity of expectation. Group calibration is equivalent to $\mathbb{E}[D \mid \hat{P} = p, A = a] = 0$, which implies $\mu_a = 0$.

B.5 Label-Conditional Parity and Decision Rates

Assume that both outcome strata have positive probability in each compared group. Write $\bar{F}(t) = 1 - F(t^-) = \mathbb{P}(D \ge t)$; the left limit handles atoms consistently with the rule $\hat{Y} = \mathbb{I}[\hat{P} \ge \tau]$. Then

$$\mathbb{P}(\hat{Y} = 1 \mid A = a) = (1-\pi_a)\bar{F}_{D|A=a,Y=0}(\tau) + \pi_a \bar{F}_{D|A=a,Y=1}(\tau - 1).$$

Proposition B.13 (EO at all thresholds)**.** *EO between groups* i, j *holds at every* $\tau \in (0,1)$ *if and only if* $\mathcal{L}(D \mid Y = y, A = i) = \mathcal{L}(D \mid Y = y, A = j)$ *for* $y = 0, 1$*.*

Proof. Within stratum $Y = y$, $D = \hat{P} - y$ and $\mathbb{P}(\hat{Y} = 1 \mid Y = y, A = a) = \mathbb{P}(D \ge \tau - y \mid Y = y, A = a)$. Equality for every score threshold identifies the conditional score distribution on $[0,1]$, including endpoint masses by limits. Translation by $-y$ identifies the corresponding residual distribution. The converse follows from the same identity. □

Corollary B.14 (DP under equal base rates)**.** *If both label-conditional residual distributions agree across groups and* $\pi_i = \pi_j$*, DP holds at every threshold.*

Proof. The conditional tail probabilities in the selection-rate formula agree, as do their positive- and negative-label proportions. □

Corollary B.15 (Label-conditional residual distances)**.** *With* $W_{Y=y}$ *as defined on scores in Proposition B.12, if* $W_{Y=0} = 0$*, FPRs agree at every threshold; if* $W_{Y=1} = 0$*, TPRs agree at every threshold. Both distances vanish if and only if EO holds at all thresholds.*

Proof. A Wasserstein distance is zero precisely when its two distributions agree. Within stratum $Y = y$, $D = \hat{P} - y$ is a translation of the score, so $W_{Y=y} = 0$ if and only if the label-conditional residual laws agree. Apply Proposition B.13 within each outcome stratum. □

The marginal residual distribution is a mixture with weights π_a. These statements use the two label-conditional distributions.

C Computing the Empirical Distances

Sorting each group's n_a residuals takes $O(n_a \log n_a)$. For exact empirical distances between unequal-sized groups, merge the union of empirical quantile breakpoints and sum each constant quantile difference over its interval (linear in $n_i + n_j$ after sorting). Let K denote the number of groups and $n = \sum_a n_a$ the total sample size. For all pairs the sorting-plus-merging cost is $O(n \log n + Kn)$. A fixed G-point grid instead costs $O(n \log n + K^2 G)$ under constant-time quantile lookup after sorting, but adds quadrature error.

D Additional Experimental Details

The primary protocol is described in § 4.1. This section records dataset construction, features, and model hyperparameters. Threshold sensitivity analyses appear in § E.

D.1 Adult Experimental Protocol

OpenML Adult v2 (id 1590) merges the UCI training and test files ($n = 48{,}842$). Complete-case deletion of missing features leaves $n = 45{,}222$. The label is $Y = 1$ iff income is > 50K. Sex is the protected attribute, with groups Female and Male, and is excluded from model inputs. Dropped columns are the sampling weight and the categorical education field (the numeric education code is kept). Numeric features are age, years of education, capital gain, capital loss, and hours per week; categorical features are work class, marital status, occupation, relationship, race, and native country (one-hot encoded).

A train-only standard scaler and one-hot encoder are fit on the training split. Each seed $0, 1, 2, 3, 4$ draws a 60/20/20 split stratified on (sex, Y), yielding train/validation/test sizes 27,132/9,045/9,045. Logistic regression uses ℓ_2 regularization with $C = 1$ and the L-BFGS solver. Uncalibrated scores are the model's predicted class probabilities. Platt scaling is a one-dimensional logistic fit on the uncalibrated decision scores from the validation split only. A global isotonic map is also fit on validation probabilities only. Train, validation, and test sample identifiers are disjoint. Both labels occur in both groups in every split. The Adult calibration, decomposition, and threshold tables use these five test splits.

The selection-rate targets and paired bootstrap procedure are specified in § E.

D.2 ACSIncome California 2018 Protocol

The ACSIncome experiment uses Folktables version 0.0.12 with ACS PUMS 2018 one-year person records for California. The extract contains 378,817 person-level rows. The official ACSIncome filter retains 195,665 rows; no further complete-case deletions were required. The label is personal income above \$50,000 (prevalence 0.411). Model features are the official ACSIncome covariates except race. Continuous features are age and hours worked (train-only standardization). Categorical features are class of worker, education, marital status, occupation, place of birth, relationship, and sex (train-only one-hot encoding). Race and the semantic audit-group label are excluded from X. Splits are 60/20/20, stratified on (audit group, Y), seeds 0–4; each test split has $n = 39{,}133$. Logistic regression uses ℓ_2 regularization with $C = 1$ and L-BFGS; all five fits converged. Group construction follows the 2018 ACS PUMS race codebook and was fixed before residual W_1 was computed. Codes 3–5 are pooled as Indigenous/Alaska Native by the codebook. Asian alone stays separate from Some Other Race. No group is dropped. The main ACSIncome comparisons use the ten pairs among the five large groups with mean test $n \geq 1000$ (White, Black, Asian, Some Other Race, Two or More Races). Indigenous (test $n = 352$) and Native Hawaiian/Pacific Islander (test $n = 128$) appear only in exploratory checks and are excluded from the main comparisons.

D.3 ACSIncome 432-Model Ranking Protocol

The validation–test ranking study in § 4.4 uses $12 \times 3 \times 4 \times 3 = 432$ White–Black model configurations on ACSIncome 2018 one-year extracts.

- *States.* CA, FL, GA, IL, LA, MD, NC, NY, OH, SC, TX, VA (chosen for White and Black test support under the Folktables ACSIncome filter).
- *Model families.* Logistic regression (ℓ_2, C=1, L-BFGS), XGBoost (200 trees, depth 4, learning rate 0.1, row/column subsample 0.8, ℓ_2=1), and MLP (one hidden layer of width 32, Adam, batch size 2048, early stopping with max 25 epochs, ℓ_2=10^{-4}).
- *Score versions.* Uncalibrated probabilities; global isotonic calibration fit on validation; groupwise isotonic calibration (separate map per race group on validation); and prevalence reweighting of residuals to a common $\bar{\pi}$ with the score distributions within each label held fixed.
- *Seeds.* $0, 1, 2$, each drawing a group-and-label-stratified 60/20/20 train/validation/test split within the state.
- *Threshold band.* On each model's validation split, τ_{lo} and τ_{hi} are the validation-selected thresholds for selection rates S=0.05 and S=0.50 (same selection rule as § E). Avg. FN/FP gaps average the absolute group FN/FP gaps uniformly over $\tau \in [\tau_{\text{lo}}, \tau_{\text{hi}}]$. Ramp FN/FP costs use $c_{\text{FN}}(D) = \max(-D, 0)$ and $c_{\text{FP}}(D) = \max(D, 0)$ on the frozen test residuals.

Race and the audit-group label are excluded from X in every fit.

E Additional Sensitivity Analyses

E.1 Selection-Rate Targets

Before inspecting disparity results we fixed the selection-rate targets $S \in \{0.10, 0.20, 0.30, 0.40, 0.50\}$. On the validation split of the audited pair, candidate thresholds are the unique predicted scores together with $\{0, 1\}$. For each target S we set τ_S to the candidate minimizing $|r_{\text{val}}(\tau) - S|$, breaking ties toward the larger τ (more selective rule). We freeze τ_S, evaluate EOgap on test, and keep the threshold fixed on the evaluation sample. Test selection rates track the targets (Table 6). Comparisons of group FN/FP gaps with conditional FNR/FPR gaps use the decomposition in Equation (18). The prevalence term reverses the sign of the FN gap relative to ΔFNR in 54% of ACSIncome pair×selection cells and in 92% of Adult selection×split cells. Across the 250 ACSIncome pair×selection cells, the prevalence term reverses the sign of either the FN or the FP gap relative to its conditional counterpart in 57% of cells, and at $S = 0.50$ it shrinks the absolute FN gap below half the conditional gap in 98% of cells. We report these frequencies rather than a composite "divergence" rate.

E.2 Adult Selection-Rate Bootstrap

For each seed and each $S \in \{0.30, 0.40, 0.50\}$, with the fitted predictor and validation-selected τ_S held fixed, we draw $B = 2000$ stratified resamples of the test sample within each (sex, Y) cell and recompute $\Delta\text{EOgap}(S) = \text{EOgap}(\tau_S) - \text{EOgap}(0.5)$. We report split-wise percentile 95% intervals. A pooled interval across the five overlapping partitions would mix dependent splits. At $S = 0.50$ the five intervals are all strictly positive; the main-text result is the mean $\pm$ standard deviation over five data splits 0.176 ± 0.042.

E.3 ACSIncome Observation Bootstrap

The ACSIncome threshold tables average the ten pairs within each seed, then across seeds. At $S = 0.50$, the seed-wise mean of pairwise $\Delta\text{EOgap}(S) = \text{EOgap}(\tau_S) - \text{EOgap}(0.5)$ is -0.023 ± 0.004, matching the difference of the reported ten-pair mean EO gaps $0.127 - 0.150$ in Table 6.

E.4 Paired Calibration Results

Primary contrast uncalibrated→groupwise isotonic on frozen logistic predictions. Paired changes are computed within each split before aggregation.

Table 6: Full threshold grids on uncalibrated scores. Adult is Female vs. Male (means over five splits). EO gap is the mean of per-split EOgap; EOΔ vs 0.5 is the mean of per-split $\mathrm{EOgap}(\tau_S) - \mathrm{EOgap}(0.5)$, which equals the difference of the EO-gap means before rounding. ACS is the ten-pair mean of absolute gaps. HateXplain is African-targeted minus other on frozen RoBERTa scores. Main-text Table 3 reports the FN columns at τ=0.5 and $S \in \{0.10, 0.30, 0.50\}$.

Adult Female vs. Male							
Operating point	Group-lev. FN	Group-lev. FP	Cond. FNR	Cond. FPR	EO gap	EOΔ vs 0.5	Test sel. rate
Threshold 0.5	-0.067	-0.050	$+0.090$	-0.080	0.097	0	0.205 ± 0.002
10% selection	-0.121	-0.012	$+0.056$	-0.019	0.057	-0.040 ± 0.009	0.099 ± 0.004
20% selection	-0.068	-0.048	$+0.089$	-0.077	0.096	-0.001 ± 0.003	0.201 ± 0.008
30% selection	-0.029	-0.090	$+0.131$	-0.146	0.151	$+0.054 \pm 0.016$	0.298 ± 0.007
40% selection	-0.007	-0.139	$+0.127$	-0.230	0.230	$+0.133 \pm 0.033$	0.395 ± 0.004
50% selection	$+0.000$	-0.152	$+0.075$	-0.273	0.273	$+0.176 \pm 0.042$	0.499 ± 0.004
ACSIncome ten-pair absolute means							
Operating point	Abs. group-lev. FN	Abs. group-lev. FP	Abs. cond. FNR	Abs. cond. FPR	EO gap		
Threshold 0.5	0.008	0.034	0.141	0.075	0.150		
10% selection	0.080	0.011	0.099	0.021	0.100		
20% selection	0.043	0.020	0.135	0.039	0.136		
30% selection	0.021	0.031	0.141	0.065	0.144		
40% selection	0.009	0.038	0.129	0.090	0.142		
50% selection	0.009	0.039	0.102	0.106	0.127		
HateXplain (RoBERTa)							
Operating point	Group-lev. FN	Group-lev. FP	Cond. FNR	Cond. FPR			
10% selection	$+0.264$	-0.009	-0.021	-0.000			
20% selection	$+0.185$	-0.019	-0.061	$+0.066$			
30% selection	$+0.108$	-0.027	-0.108	$+0.174$			
40% selection	-0.009	-0.042	-0.208	$+0.230$			
50% selection	-0.043	-0.064	-0.219	$+0.300$			

Table 7: Paired calibration changes (uncalibrated→groupwise isotonic). Adult rows average within-split changes over five data splits. ACS rows average the ten pairs within each seed, then report the mean and standard deviation (SD) over seeds. Levels are means of the same aggregation. DP and EO gaps use threshold 0.5. Adult EO varies across splits; ACS rows summarize the ten race-group pairs before aggregation across splits. For each of the ten ACSIncome pairs, residual distance increases on average across the five data splits.

Data	Quantity	Uncal.	Groupwise	Change mean	Change SD
Adult	Residual distance	0.133	0.147	$+0.014$	0.006
	FN component	0.074	0.073	-0.001	0.003
	FP component	0.059	0.074	$+0.015$	0.003
	Pos. label score dist.	0.070	0.096	$+0.025$	0.010
	Neg. label score dist.	0.104	0.126	$+0.022$	0.002
	DP (0.5)	0.182	0.209	$+0.027$	0.013
	EO gap (0.5)	0.097	0.105	$+0.008$	0.049
ACS (10-pair)	Residual distance	0.050	0.062	$+0.011$	0.003
	FN component	0.025	0.032	$+0.007$	0.001
	FP component	0.025	0.030	$+0.005$	0.002
	Pos. label score dist.	0.101	0.132	$+0.030$	0.003
	Neg. label score dist.	0.052	0.065	$+0.012$	0.003
	DP (0.5)	0.157	0.182	$+0.025$	0.012
	EO gap (0.5)	0.148	0.195	$+0.047$	0.016

E.5 Adult EO Per-Split Values and Permutation Nulls

Empirical W_1 is upward-biased in finite samples, so we accompany calibration-induced ΔW_{res} with two group-label permutation tests: scores held fixed, and calibration refitted.

Permutation test with scores held fixed. Residuals and sample sizes are held fixed while test group assignments are shuffled, and ΔW_{res} is recomputed under the same already-fitted uncalibrated and groupwise-isotonic scores. This targets empirical W_1 group-association noise conditional on fitted scores. Every Adult and ACS White–Black split exceeds this null's 95% upper quantile (B=500; $p \leq 0.002$).

Permutation test with calibration refitted. Validation group labels are shuffled (preserving sizes), groupwise isotonic maps are re-fit on the shuffled validation set, maps are applied to test using true test labels, and ΔW_{res} is evaluated on test with true labels (B=200). Under re-fitting, three Adult splits remain significant at $p \leq 0.005$ and two are marginal ($p = 0.055$ and 0.060). On ACS White–Black, one split reaches $p = 0.025$, one reaches $p = 0.060$, and the remaining three fall inside the calibration-refit permutation null.

E.6 Reweighting to a Common Positive-Label Proportion

For groups i, j with prevalences π_i, π_j, set $\bar{\pi} = (\pi_i + \pi_j)/2$. Score distributions within each label are held fixed; only the marginal residual mixture $\bar{F}_a = \bar{\pi}\, F_{D|A=a,Y=1} + (1 - \bar{\pi})\, F_{D|A=a,Y=0}$ is rebuilt on the merged residual knot grid (no resampling). Reported means average Adult over five seeds and ACSIncome by first averaging the ten pairs within each seed.

E.7 Text Classification Details

The NLP audits use the Hugging Face releases `s-nlp/roberta_toxicity_classifier` and `unitary/toxic-bert`. HateXplain threshold gaps appear in Table 6.

E.8 Additional Validation–Test Correlations

The bootstrap intervals in Table 4 use state-clustered resampling (B=2000). Table 8 reports within-state and prevalence-adjusted correlations.

Table 8: Within-state and partial Spearman correlations for the ACSIncome White–Black zoo in Table 4. Within-state: mean±SD of Spearman ρ over the 12 states (36 models each). Partial: Spearman after residualizing ranks on state $|\pi_{\text{White}} - \pi_{\text{Black}}|$. Pooled values repeat the main-table point estimates.

Validation → test target	Pooled ρ	Within-state mean±SD	Partial ρ
Avg. FN gap → Avg. FN gap	0.941	0.578±0.258	0.795
Avg. FP gap → Avg. FP gap	0.797	0.565±0.222	0.794
Avg. FN gap → FN cost gap	0.845	0.400±0.383	0.622
Avg. FP gap → High-conf. FP (0.7)	0.627	0.355±0.314	0.614
EO → Avg. FN gap	0.354	−0.146±0.309	0.182
EO → Avg. FP gap	0.588	0.493±0.254	0.562
EO → FN cost gap	0.209	−0.383±0.165	0.004
Score W_1 → Avg. FN gap	0.448	−0.332±0.481	−0.227
Score W_1 → Avg. FP gap	0.538	0.490±0.262	0.479
Mean-res. gap → Avg. FP gap	−0.294	−0.488±0.258	−0.408