# A-IEOF: ANCHORED INTERVENTIONAL EQUALIZED-ODDS FAIRNESS FOR FEATURE ADDITION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Machine learning systems in high-stakes domains must balance utility with group fairness. Conventional debiasing optimizes fairness on the training distribution and does not address disparities that appear under distributional shift. We introduce Anchored Interventional Equalized-Odds Fairness (A-IEOF), an in-processing framework for *feature addition*, where a newly available feature block $B$ can improve utility but may cause fairness regressions. A-IEOF trains a student model with a minimax objective that (i) enforces worst-case multi-group equalized odds under adversarially generated counterfactual replacements of $B$, (ii) anchors predictions to a no-$B$ teacher to preserve behavior on shared inputs, and (iii) stabilizes calibrated scores against variation in $B$. On a multi-institutional clinical corpus, A-IEOF attains the highest AUC (0.976) and lowest equalized-odds gap (0.085–0.126) across global, per-group, and Hardt-EO post-processing policies, with better calibration (Brier 0.0495–0.0507; ECE 0.0262–0.0343) than adversarial gradient-reversal (GRL) debiasing. To test the robustness claim directly we introduce a controlled tabular feature-addition benchmark in which $P(B \mid S)$ can be shifted by a prescribed total-variation amount: A-IEOF holds its equalized-odds gap essentially flat as the shift grows (degradation $< 0.01$ up to $\Delta_{TV} = 0.5$), consistent with our transfer bound, while an empirical-risk baseline is four times more sensitive to $B$. Over ten seeds A-IEOF matches a constrained-optimization (reductions) baseline in distribution and surpasses invariant-learning baselines (IRM, CORAL) and GRL. On three real tabular datasets (UCI Adult, COMPAS, CT metadata) it attains the lowest equalized-odds gap on Adult and COMPAS and the lowest sensitivity to $B$ on all three. A-IEOF thus offers a principled, auditable route to feature addition that aligns utility, calibration, and fairness under realistic deployment conditions. Our implementation and benchmark are available at `https://anonymous.4open.science/r/AIEOF-778E`.

## 1 INTRODUCTION

Deployed classifiers must deliver both accuracy and fairness across predefined groups. Group fairness is formalized through constraints on error or selection rates and enforced by post-processing or in-processing methods such as reductions, constrained optimization, adversarial debiasing, and representation sanitization Hardt et al. (2016); Agarwal et al. (2018); Zhang et al. (2018); Ravfogel et al. (2020). Practical deployment additionally requires calibrated probabilities, motivating Brier score, Expected Calibration Error (ECE), and AUC as reporting standards Glenn et al. (1950); Guo et al. (2017).

A common real-world event is *feature addition*: a model trained on base inputs acquires a new feature block $B$. Standard observational debiasing optimizes fairness on the training distribution and does not control disparities that arise when $B$'s distribution shifts relative to the remaining inputs and to group membership. Such conditional shifts degrade both calibration and fairness at deployment LEARNING (2009); Ovadia et al. (2019), and causal perspectives argue for reasoning about counterfactual feature changes when disparities can be induced by manipulating a specific block Kusner et al. (2017); Nabi & Shpitser (2018).

We introduce **Anchored Interventional Equalized-Odds Fairness (A-IEOF)** for feature addition, an in-processing framework that aligns training with deployment by controlling interventional



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.

fairness for a designated block. A-IEOF trains a student that uses $B$ while (i) enforcing worst-case multi-group equalized odds under counterfactual replacements of $B$ drawn from a group- and context-conditional generator, (ii) *anchoring* behavior to a no-$B$ teacher to preserve legacy decisions, and (iii) *stabilizing* predictions so calibrated scores change minimally when only $B$ changes. Our evaluation protocol calibrates on validation, fixes global or per-group thresholds, and reports AUC, equalized-odds (EO) gap, demographic-parity (DP) gap, Brier, and ECE on held-out test data.

**Contributions.** (1) A minimax objective targeting worst-case *interventional* multi-group equalized odds on a designated block, in contrast to observational methods that operate on the training distribution Zhang et al. (2018); Mehrabi et al. (2021); Ganin & Lempitsky (2015); (2) a *teacher anchoring* mechanism that reduces regressions when the new block is introduced; (3) a *stability* regularizer on counterfactual replacements, with a transfer bound linking interventional control to fairness robustness under shift; (4) a controlled feature-addition benchmark with prescribed shift magnitude, plus multi-seed evaluations on Adult, COMPAS, and clinical CT metadata against ERM, GRL, IRM, CORAL, and a reductions learner.

## 2 Related Work

**Counterfactual and group fairness.** Counterfactual fairness requires a decision to be unchanged under intervention on the sensitive attribute Kusner et al. (2017); extensions provide post-hoc adjustments that enforce counterfactual constraints without retraining Wang et al. (2023), address attribute–label intersections Wastvedt et al. (2024), and characterize the utility–fairness trade-off for compliant predictors Zhou et al. (2024). Group criteria such as equalized odds are commonly enforced adversarially by suppressing group information in learned representations Zhang et al. (2018), which is effective observationally but brittle under deployment shift; robust-optimization variants therefore target worst-case disparities directly Grari et al. (2024), and causal concept bottlenecks address demographic bias in medical imaging through institutional pathway interventions Qingge et al. (2025). These approaches act at the objective or representation level; A-IEOF instead reasons through explicit interventions on a *designated* feature block and fixes operating points by validation-time post-processing.

**Invariant learning and domain generalization.** Invariant Risk Minimization seeks predictors whose optimal classifier is simultaneously optimal across environments, down-weighting spurious correlations Arjovsky et al. (2019); Deep CORAL aligns second-order feature statistics across domains Sun & Saenko (2016); and classical domain-adaptation theory bounds target error by a source–target divergence Ben-David et al. (2010). These target *predictive* invariance or overall error transfer. A-IEOF differs in two respects: it targets *group-fairness* transfer (equalized-odds gaps) rather than accuracy, and it treats a single named block $B$ as the locus of shift, using counterfactual replacements of $B$ instead of pre-defined environments. We adopt IRM, CORAL, gradient-reversal debiasing, and a constrained-optimization (reductions) fair learner Agarwal et al. (2018) as baselines in Section 5.3.

**Post-hoc adjustment.** Post-processing remains attractive in high-stakes deployments for its auditability and model-agnosticism Wang et al. (2023), and adversarial training reduces across-group error disparities in clinical risk prediction Yang et al. (2023). Because calibration quality and thresholding policy largely determine realized EO/DP gaps, our evaluation separates probability quality (AUC, Brier, ECE) from thresholded fairness (EO/DP) across *global*, *per-group*, and *Hardt-EO* policies.

## 3 Methodology

### 3.1 Problem Setting

Consider binary classification with data $\mathcal{D} = \{(x_i, y_i, s_i)\}_{i=1}^n$ from distribution $P$, where $x_i \in \mathcal{X}$, $y_i \in \{0, 1\}$, and $s_i \in \{0, \dots, G-1\}$ is protected-group membership. In feature addition a **new block** $B$ becomes available that may improve utility but risks fairness regressions, so we decompose $X = (X_{-B}, X_B)$ with $X_B$ (written $B$) the designated block and $X_{-B}$ all remaining features. For

instance, a clinical model using demographics and vitals ($X_{-B}$) gains an institution identifier ($B$) that improves discrimination but may encode bias if institution correlates with protected attributes. We seek $f_\theta$ using $B$ that (i) maintains or improves accuracy, (ii) controls disparities across $S$, and (iii) preserves behavior on $X_{-B}$ relative to a no-$B$ model $f_0$.

Standard debiasing optimizes fairness on $P(X, Y, S)$ Grari et al. (2024). Post-deployment, however, $P(B \mid X_{-B}, S)$ may shift (e.g. patient mix changes across institutions) while $P(X_{-B}, S, Y)$ remains stable. We therefore enforce fairness under **counterfactual replacements** of $B$: for context $(x_{-B}, s)$ we evaluate disparities when $B$ takes an alternative value $b$, writing $x^{(b)} = (x_{-B}, b)$. Throughout, $f_0$ is the frozen teacher, $f_\theta$ the student, $g_\phi$ the adversarial generator over $B$, $\sigma(\cdot)$ the sigmoid, and $p(x) = \sigma(f(x))$ the predicted probability; uppercase letters denote random variables and lowercase their values.

**On the term "interventional."** We use "intervention" in a restricted, operational sense: replacing the value of block $B$ while holding $X_{-B}$ fixed, with replacements drawn from a learned conditional generator $g_\phi(b \mid x_{-B}, s)$. This is a conditional counterfactual replacement, not a $do(\cdot)$ operation identified from a structural causal model, and we do not claim identifiability of institution effects: replacing an institutional identifier need not be clinically realizable, since institution can proxy acquisition pipeline, case mix, and severity. Our guarantees should be read as robustness to a stated family of conditional shifts in $B$ (Section 3.3), not as causal-effect estimates.

### 3.2 The A-IEOF Objective

A-IEOF combines three mechanisms: **teacher anchoring** (a no-$B$ teacher $f_0$ is trained first and frozen; the student matches it on $X_{-B}$), **interventional EO control** (penalizing worst-case multi-group equalized odds under adversarial counterfactual $B$-replacements), and **counterfactual stability** (regularizing the student toward similar scores when only $B$ changes).

**Soft fairness metrics.** For gradient-based optimization we use *soft* group error rates computed from probabilities rather than hard decisions: $\mathrm{TPR}_g^{\mathrm{soft}}(f) = \mathbb{E}[\sigma(f(x)) \mid y{=}1, s{=}g]$ and $\mathrm{FPR}_g^{\mathrm{soft}}(f) = \mathbb{E}[\sigma(f(x)) \mid y{=}0, s{=}g]$, estimated as minibatch means. The weighted soft EO gap is

$$\begin{aligned}\Delta_{\mathrm{EO}}^{\mathrm{soft}}(f) = w_+\big(\max_g \mathrm{TPR}_g^{\mathrm{soft}} - \min_g \mathrm{TPR}_g^{\mathrm{soft}}\big) \\ + w_-\big(\max_g \mathrm{FPR}_g^{\mathrm{soft}} - \min_g \mathrm{FPR}_g^{\mathrm{soft}}\big),\end{aligned} \tag{1}$$

with $w_+{=}1$, $w_-{=}2$ prioritizing FPR parity. At test time, thresholds yield hard decisions and hard rates (Eq. 6).

**Adversarial generator.** The generator $g_\phi(\cdot \mid x_{-B}, s)$ produces counterfactual values of $B$ conditioned on context. For continuous $B$ it is $\mathcal{N}(\mu_\phi(x_{-B}, s), \sigma_\phi^2(x_{-B}, s))$ with $\mu_\phi, \sigma_\phi^2$ neural networks; for categorical $B$ (e.g. institution identifiers) it is $\mathrm{Categorical}(\mathrm{softmax}(\mathrm{logits}_\phi(x_{-B}, s)))$. Each step draws two replacements $b', b'' \sim g_\phi(\cdot \mid x_{-B}, s)$, and $g_\phi$ is updated to *maximize* the resulting soft EO gap, exposing worst-case violations that the student must then withstand.

**Training objective.** Let $b_\odot$ denote a neutralized $B$ (zero, or the mean after standardization), and let $\Delta_{\mathrm{EO}}^{\mathrm{soft}}(f_\theta; x^{(b')}, x^{(b'')})$ be the soft EO gap on counterfactual pairs. A-IEOF minimizes

$$\begin{aligned}\min_\theta\ \mathbb{E}\Big[&\ell(\sigma(f_\theta(x)), y) + \lambda_{\mathrm{dist}} D\big(\sigma(f_\theta(x^{(b_\odot)})), \sigma(f_0(x_{-B}))\big) \\ &+ \lambda_{\mathrm{interv}} \sup_{b', b''} \Delta_{\mathrm{EO}}^{\mathrm{soft}}(f_\theta; x^{(b')}, x^{(b'')}) + \lambda_{\mathrm{full}} \Delta_{\mathrm{EO}}^{\mathrm{soft}}(f_\theta) \\ &+ \lambda_{\mathrm{stab}} \big|\sigma(f_\theta(x^{(b')})) - \sigma(f_\theta(x^{(b'')}))\big|\Big],\end{aligned} \tag{2}$$

where $\ell$ is binary cross-entropy and $D$ is mean-squared error on probabilities. The five terms are, in order: task loss; *teacher anchoring*, matching the student to the teacher when $B$ is masked so that $X_{-B}$ behavior is preserved; *interventional EO*, penalizing worst-case soft EO under the adversary's $B$-values; *observational EO* on factual $B$; and *stability*, damping prediction change when only $B$ varies.

**Algorithm 1** Anchored Interventional Equalized-Odds Fairness (A-IEOF)

**Require:** dataset $\mathcal{D} = \{(x_{-B}, x_B, y, s)\}$, student $f_\theta$, generator $g_\phi$, learning rates $\eta_\theta, \eta_\phi$, generator period $K$, total steps $T$
**Ensure:** debiased student $f_\theta$

**Stage 1: anchor teacher**
Train $f_0$ on $(x_{-B}, y)$ with cross-entropy; freeze $f_0$
**Stage 2: student and generator**
**for** $t \leftarrow 1$ **to** $T$ **do**
  Sample minibatch $\mathcal{B} \sim \mathcal{D}$; write $x = (x_{-B}, x_B)$
  Draw $b', b'' \sim g_\phi(\cdot \mid x_{-B}, s)$; form interventions $x^{(b')}, x^{(b'')}$
  $z \leftarrow f_\theta(x),\ z' \leftarrow f_\theta(x^{(b')}),\ z'' \leftarrow f_\theta(x^{(b'')}),\ z_0 \leftarrow f_0(x_{-B})$
  $\theta \leftarrow \theta - \eta_\theta \nabla_\theta \mathcal{L}_{\text{A-IEOF}}(\theta; z, z', z'', z_0, y)$ ▷ Eq. equation 2
  **if** $t \bmod K = 0$ **then**
    $\phi \leftarrow \phi + \eta_\phi \nabla_\phi \Delta_{\text{EO}}^{\text{soft}}\left(f_\theta; x^{(b')}, x^{(b'')}\right)$ ▷ ascent: adversarial
  **end if**
**end for**
**return** $f_\theta$

**Implementation notes.** $\Delta_{\text{EO}}^{\text{soft}}$ requires at least 30 samples per (group, label) stratum; batches that fail this skip the fairness terms. The generator updates every 4 student steps after a 100-step warm-up. Hyperparameters are selected on validation.

### 3.3 Robustness Under Distribution Shift

Two mechanisms provide robustness when $B$'s distribution changes post-deployment while other variables remain stable: interventional control ($\lambda_{\text{interv}}$) penalizes worst-case soft EO under counterfactual $B$-replacements, and stability regularization ($\lambda_{\text{stab}}$) reduces score oscillation when only $B$ varies. Let $\epsilon_{\text{stab}} = \sup_{x_{-B},s,b,b'} |\sigma(f_\theta(x^{(b)})) - \sigma(f_\theta(x^{(b')}))|$ be the worst-case sensitivity to $B$.

[Fairness transfer] Let $P'$ differ from $P$ only in $P(B \mid X_{-B}, S)$, with $P(X_{-B}, Y, S)$ stable, and assume $B \perp Y \mid (X_{-B}, S)$ and $\epsilon_{\text{stab}} < \infty$. Then

$$\Delta_{\text{EO}}^{P'} \ \le\ \Delta_{\text{EO}}^{P} + C\,\epsilon_{\text{stab}}\,\Delta_{TV}(P, P'), \tag{3}$$

where $\Delta_{TV}$ is total-variation distance and $C = 4(w_+ + w_-)$, so $C = 12$ for $w_+$=1, $w_-$=2.

[Proof sketch] Because $P(X_{-B}, Y, S)$ is stable, $P'(x_{-B} \mid Y{=}1, S{=}g) = P(x_{-B} \mid Y{=}1, S{=}g)$, so the change in a group's true-positive rate isolates to the shift in $B$:

$$\begin{aligned}\text{TPR}_g^{P'} - \text{TPR}_g^{P} = &\sum_{x_{-B}} P(x_{-B} \mid Y{=}1, g) \\ &\times \sum_{b} \big[P'(b \mid \cdot) - P(b \mid \cdot)\big]\,\sigma\big(f_\theta(x^{(b)})\big).\end{aligned}$$

For a fixed context the inner sum is an expectation difference of a function whose oscillation over $b$ is at most $\epsilon_{\text{stab}}$, so it is bounded by $2\epsilon_{\text{stab}}\|P' - P\|_{TV}$. Averaging over $x_{-B}$ (the weights sum to one) and taking the supremum over contexts gives $|\text{TPR}_g^{P'} - \text{TPR}_g^{P}| \le 2\epsilon_{\text{stab}}\Delta_{TV}$, and the same argument with $Y{=}0$ gives the identical bound for $\text{FPR}_g$, since stability holds regardless of the label and $B \perp Y \mid (X_{-B}, S)$. Each max–min spread in Eq. 7 therefore moves by at most $4\epsilon_{\text{stab}}\Delta_{TV}$, and weighting the two spreads by $w_+$ and $w_-$ yields the additive term $4(w_+ + w_-)\epsilon_{\text{stab}}\Delta_{TV}$. The full derivation is in the released supplementary material.

**Scope.** The bound covers shifts confined to $P(B \mid X_{-B}, S)$, not general covariate or label shift. Its practical content is that a model made insensitive to $B$ (small $\epsilon_{\text{stab}}$, as enforced by $\lambda_{\text{stab}}$ and $\lambda_{\text{interv}}$) degrades gracefully and linearly in shift magnitude, whereas a $B$-sensitive model has a loose, possibly vacuous bound. Section 5.4 verifies this: A-IEOF attains $\epsilon_{\text{stab}} \approx 0.04$ with near-zero EO degradation up to $\Delta_{TV} = 0.5$, while an empirical-risk baseline attains $\epsilon_{\text{stab}} \approx 0.15$.

### 3.4 POST-PROCESSING AND METRICS

Post-processing acts on model outputs only; it is fitted on validation and then frozen for test. Given validation logits $z = f(x)$ we fit a temperature $T > 0$ by minimizing negative log-likelihood,

$$p(x) = \sigma(z/T), \tag{4}$$

either globally ($T_{\text{glob}}$) or per group ($\{T_g\}$, regularized toward $T_{\text{glob}}$). Decisions use one of two threshold policies. The **global** policy picks $t_{\text{ref}}$ by Youden index over $[0.01, 0.99]$. The **per-group** policy optimizes $\{t_g\}$ with guarded smoothing,

$$t_g^{\text{guard}} = a_g \,\text{clamp}(t_g, q_{g,0.10}, q_{g,0.90}) + (1 - a_g)\, t_{\text{ref}}, \tag{5}$$

where $a_g = n_g/(n_g + k_0)$ shrinks toward the global threshold, $n_g$ is the validation size of group $g$, $k_0 = 100$, and $q_{g,\alpha}$ is the $\alpha$-quantile of that group's probabilities. Test decisions are $\hat{y} = \mathbf{1}[p(x) \geq t_{\text{ref}}]$ under the global policy and $\hat{y} = \mathbf{1}[p(x) \geq t_s^{\text{guard}}]$ under the per-group policy.

We report AUC for utility. Fairness uses hard per-group rates

$$\begin{aligned} \text{TPR}_g &= \Pr(\hat{Y}{=}1 \mid Y{=}1, S{=}g), \\ \text{FPR}_g &= \Pr(\hat{Y}{=}1 \mid Y{=}0, S{=}g), \end{aligned} \tag{6}$$

aggregated into the equalized-odds gap

$$\begin{aligned} \Delta_{\text{EO}} = w_+ \big( \max_g \text{TPR}_g - \min_g \text{TPR}_g \big) \\ + w_- \big( \max_g \text{FPR}_g - \min_g \text{FPR}_g \big), \end{aligned} \tag{7}$$

with $w_+{=}1$, $w_-{=}2$, and the demographic-parity gap $\Delta_{\text{DP}} = \max_g \Pr(\hat{y}{=}1 \mid s{=}g) - \min_g \Pr(\hat{y}{=}1 \mid s{=}g)$. Calibration uses the Brier score $\frac{1}{n}\sum_i (p(x_i) - y_i)^2$ and the expected calibration error $\text{ECE} = \sum_{k=1}^{M} \frac{n_k}{n} |\text{freq}_k - \text{conf}_k|$ over $M{=}15$ equal-width confidence bins. Lower Brier and ECE indicate better calibration, which matters for threshold transfer from validation to test.

## 4 EXPERIMENTAL SETUP

**Clinical corpus.** We use a multi-institution COVID-19 CT metadata corpus with binary labels (COVID-19, Normal), demographics (age, gender), country, and a site identifier `Institution` Rahimzadeh et al. (2021). We designate **Institution as block** $B$, while fairness is *measured* over intersectional groups `AgeBin` · `Gender` (e.g. f·A<40 denotes females under 40). Preprocessing harmonizes categorical strings, standardizes continuous fields, and bins age for group construction. The evaluation set ($n{=}12{,}281$) is imbalanced across groups: m·A40+ is 41.7% of samples and f·A<40 only 12.4%. Treating `Institution` as $B$ matches the A-IEOF objective: the student uses $B$ for utility while anchoring to a no-$B$ teacher and penalizing worst-case interventional EO, giving robustness to site-mix changes without conflating $B$ with the protected groups.

**Implementation.** Data are split 60/25/15 by stratified sampling (seed 42), images resized to $224{\times}224$, and $B$ is institution ($K{=}10$ sites). Following Algorithm 1, the teacher $f_0(X_{-B})$ trains for 10 epochs on a frozen MobileNetV2 backbone, then the student $f_\theta(X_{-B}, B)$ trains for 30 epochs under Eq. 2 with AdamW and an unfrozen backbone. Loss weights are $\lambda_{\text{dist}}{=}0.05$, $\lambda_{\text{interv}}{=}0.55$, $\lambda_{\text{full}}{=}0.5$, $\lambda_{\text{stab}}{=}0.05$; the generator uses Gumbel-Softmax with temperature annealing. Model selection maximizes $\text{AUC} - \alpha\Delta_{\text{EO}}$ ($\alpha{=}1$) on validation, and all reported test metrics use validation-fixed temperatures and thresholds. Full configurations are in the released code.

## 5 RESULTS

### 5.1 OVERALL PERFORMANCE AND FAIRNESS

Table 1 reports test metrics for four model families under four policies: (A) global threshold with groupwise temperature, (B) per-group thresholds with groupwise temperature, (C) Hardt-EO post-processing, and (D) pure in-processing (global temperature and threshold). A-IEOF achieves the

Table 1: Overall test metrics by post-processing policy on the clinical corpus. Best per policy in **bold**.

| Model | AUC ↑ | EO ↓ | DP ↓ | Brier ↓ | ECE ↓ |
|---|---|---|---|---|---|
| *(A) Global threshold* | | | | | |
| Teacher (no $B$) | 0.928 | 0.215 | 0.181 | 0.0909 | 0.0373 |
| Teacher (with $B$) | 0.923 | 0.289 | 0.166 | 0.0935 | 0.0602 |
| Adv. debias (GRL, $B$) | 0.892 | 0.135 | 0.192 | 0.0868 | 0.0868 |
| **A-IEOF** ($B$) | **0.976** | **0.126** | 0.195 | **0.0495** | **0.0280** |
| *(B) Per-group thresholds* | | | | | |
| Teacher (no $B$) | 0.932 | 0.395 | 0.256 | 0.0905 | 0.0538 |
| Teacher (with $B$) | 0.926 | 0.191 | 0.193 | 0.0947 | 0.0625 |
| Adv. debias (GRL, $B$) | 0.886 | 0.206 | 0.198 | 0.0918 | 0.0918 |
| **A-IEOF** ($B$) | 0.975 | **0.085** | 0.190 | **0.0498** | 0.0343 |
| *(C) Hardt-EO post-processing* | | | | | |
| Teacher (no $B$) | 0.919 | 0.130 | 0.020 | 0.0935 | 0.0416 |
| Teacher (with $B$) | 0.925 | 0.132 | 0.055 | 0.0981 | 0.0527 |
| **A-IEOF** ($B$) | **0.975** | **0.094** | **0.048** | **0.0507** | **0.0262** |
| *(D) Pure in-processing* | | | | | |
| Teacher (no $B$) | 0.926 | 0.236 | 0.145 | 0.0949 | 0.0488 |
| Teacher (with $B$) | 0.925 | 0.220 | 0.188 | 0.0932 | 0.0502 |
| Adv. debias (GRL, $B$) | 0.886 | 0.181 | 0.198 | 0.0918 | 0.0918 |
| A-IEOF ($B$) | 0.899 | 0.135 | 0.198 | 0.0819 | 0.0819 |

Table 2: Per-group TPR/FPR on the clinical test split. Group supports (total/pos/neg): f·A40+ = 106/43/63, f·A<40 = 43/10/33, m·A40+ = 164/73/91, m·A<40 = 90/35/55.

| Policy | Model | f·A40+ | f·A<40 | m·A40+ | m·A<40 |
|---|---|---|---|---|---|
| Global | Teacher (no $B$) | 0.651/0.063 | 0.800/0.030 | 0.836/0.033 | 0.800/0.036 |
| | GRL ($B$) | 0.744/0.016 | 0.800/0.000 | 0.849/0.000 | 0.714/0.000 |
| | **A-IEOF** | 0.884/0.095 | 0.900/0.061 | 0.959/0.044 | 0.886/0.055 |
| Per-gr. | Teacher (no $B$) | 0.860/0.222 | 0.800/0.061 | 0.726/0.011 | 0.914/0.218 |
| | **A-IEOF** | 0.884/0.111 | 0.900/0.091 | 0.945/0.088 | 0.886/0.091 |
| Hardt-EO | Teacher (no $B$) | 0.698/0.127 | 0.700/0.061 | 0.767/0.066 | 0.800/0.109 |
| | **A-IEOF** | 0.767/0.032 | 0.900/0.061 | 0.836/0.011 | 0.857/0.036 |

highest AUC (0.976) and lowest EO gap (0.085–0.126) across policies (A)–(C), with better calibration (Brier 0.0495–0.0507; ECE 0.0262–0.0343) than adversarial debiasing. Per-group thresholds give the smallest EO gap (0.085) while retaining utility, and under Hardt-EO correction A-IEOF preserves the best discrimination and probability quality with minimal EO/DP drift.

### 5.2 Per-Group Analysis

Table 2 gives per-group TPR/FPR at the validation-fixed thresholds. A-IEOF improves consistently across all four groups rather than through a single dominant stratum: under the global policy its TPRs span 0.884–0.959 at moderate FPRs (0.044–0.095), whereas GRL is uneven (TPR 0.714–0.849) and produces zero FPR in several groups, indicating overly conservative thresholds that sacrifice sensitivity. Per-group thresholds balance the rates further (TPR 0.884–0.945, FPR 0.088–0.111, EO $= 0.085$), and Hardt-EO keeps TPRs high (0.767–0.900) while constraining FPRs (0.011–0.061).

These clinical numbers come from a single stratified test split and are point estimates. To quantify variability, and to obtain a design in which the shift magnitude can be controlled as the transfer bound requires, we complement them with the multi-seed benchmark below; the released code also provides group-stratified bootstrap intervals for the clinical per-group rates.

### 5.3 A Controlled Feature-Addition Benchmark

A single clinical corpus cannot establish generality, and off-the-shelf real data does not permit controlling the shift magnitude our theory concerns. We therefore add a structurally different tabular benchmark: eight continuous base features $X_{-B}$, four intersectional groups with the clinical class imbalance ($0.417/0.255/0.204/0.124$), a binary label generated from $X_{-B}$ plus a group base-rate offset, and a categorical block $B$ with $K{=}10$ sites drawn from group-dependent site distributions.

Table 3: Tabular benchmark, global-threshold policy, mean $\pm$ sd over ten seeds. ERM, IRM and CORAL exploit the proxy block and leave large equalized-odds gaps; only the reductions learner and A-IEOF control EO. In distribution A-IEOF does not dominate reductions; its advantage is robustness to shift (Table 4).

| Model | AUC ↑ | EO ↓ | DP ↓ |
|---|---|---|---|
| Teacher (no $B$) | $0.892 \pm 0.064$ | $0.871 \pm 0.303$ | 0.506 |
| ERM ($B$) | $0.893 \pm 0.063$ | $0.902 \pm 0.324$ | 0.512 |
| GRL ($B$) | $0.860 \pm 0.064$ | $0.699 \pm 0.376$ | 0.390 |
| IRM ($B$) | $0.880 \pm 0.060$ | $0.848 \pm 0.284$ | 0.487 |
| CORAL ($B$) | $0.887 \pm 0.068$ | $0.858 \pm 0.352$ | 0.502 |
| Reductions-EO ($B$) | $0.841 \pm 0.076$ | **0.244** $\pm 0.064$ | **0.246** |
| **A-IEOF ($B$)** | $0.825 \pm 0.084$ | $0.305 \pm 0.107$ | 0.261 |

Table 4: Test $\Delta_{\mathrm{EO}}$ versus shift magnitude $\Delta_{TV}$ (mean over ten seeds), with empirical worst-case sensitivity $\epsilon_{\text{stab}}$. A-IEOF keeps its gap essentially constant and is roughly four times less sensitive to $B$ than ERM.

| Method | $\epsilon_{\text{stab}}$ | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 |
|---|---|---|---|---|---|---|---|
| ERM | 0.154 | 0.932 | 0.925 | 0.912 | 0.891 | 0.872 | 0.879 |
| **A-IEOF** | **0.040** | 0.292 | 0.292 | 0.297 | 0.287 | 0.293 | 0.294 |

By construction $B$ is a group-correlated *proxy*: marginally associated with $Y$ through $S$ but conditionally uninformative given $(X_{-B}, S)$, so a model that leans on $B$ looks stronger in training yet becomes unfair once $P(B \mid \cdot)$ shifts. This instantiates exactly the hazard A-IEOF targets and satisfies the stable-base assumption of Theorem B.2 in the Appendix B. All methods use a linear (logistic) backbone, making the study deterministic and reproducible on CPU; the same method core drives the clinical pipeline with a MobileNetV2 backbone.

Table 3 compares A-IEOF against ERM, GRL, IRM, Deep CORAL, and a Lagrangian equalized-odds reductions learner, all using $B$, plus the no-$B$ teacher. Two honest observations follow. First, methods that optimize predictive invariance or accuracy alone (ERM, IRM, CORAL) do not reduce the EO gap here, because the proxy $B$ predicts the group base rate; only methods that explicitly constrain group error rates shrink EO. Second, in distribution A-IEOF is competitive with but does not beat the constrained-optimization baseline. Its distinguishing advantage is stability under shifts in $B$, isolated next.

### 5.4 Distribution-Shift Stress Test

We test the paper's central claim directly. Holding the calibrated probabilities and threshold fixed at their unshifted validation values, we resample $B$ on the test set from a deployment distribution $P'(B \mid S)$ at a prescribed total-variation distance $\Delta_{TV}$ from training, keeping $(X_{-B}, S, Y)$ fixed. Table 4 and Figure 3 in Appendix F report the resulting test EO gap as $\Delta_{TV}$ grows. A-IEOF's degradation up to $\Delta_{TV} = 0.5$ is below $0.01$, comfortably inside the bound $12\,\epsilon_{\text{stab}}\Delta_{TV} \leq 12 \times 0.040 \times 0.5 = 0.24$. For ERM the same bound is loose ($\geq 0.9$) because its sensitivity is large, illustrating that Eq. 3 is informative precisely when the training objective controls $\epsilon_{\text{stab}}$.

### 5.5 Ablation and the Robustness–Fairness Dial

Table 8 ablates each term of Eq. 2 and reports the sensitivity $\epsilon_{\text{stab}}$ that governs the transfer bound. The observational EO term ($\lambda_{\text{full}}$) is the dominant driver of in-distribution fairness: removing it raises EO from $0.31$ to $0.40$. The interventional and stability terms behave differently, having limited effect on the stationary gap but driving $\epsilon_{\text{stab}}$: removing the interventional term more than triples it ($0.066 \to 0.224$), and removing stability raises it to $0.085$. This clarifies the division of labor among terms and corrects the original claim that every term improves every fairness metric.

Accordingly, $\lambda_{\text{interv}}$ is best read as a principled dial between stationary fairness and shift robustness rather than as a fixed hyperparameter. Sweeping it from $0$ to $1$ on the two real datasets where $B$

Table 5: Real datasets, pure in-processing policy, mean over ten seeds. A-IEOF attains the lowest equalized-odds gap on Adult and COMPAS; on the CT metadata, where institution is a strong label proxy, it does not dominate. Best AUC and best EO per dataset in bold.

| | Adult | | COMPAS | | CT metadata | |
|---|---|---|---|---|---|---|
| Model | AUC | EO | AUC | EO | AUC | EO |
| Teacher (no $B$) | 0.796 | 0.187 | 0.728 | 1.163 | 0.692 | 0.374 |
| ERM ($B$) | 0.819 | 0.278 | 0.728 | 1.117 | 0.867 | 0.301 |
| GRL ($B$) | 0.802 | 0.196 | 0.727 | 1.125 | 0.879 | 0.334 |
| IRM ($B$) | 0.729 | 0.284 | 0.730 | 1.084 | 0.741 | 0.378 |
| CORAL ($B$) | 0.824 | 0.309 | **0.733** | 1.149 | **0.894** | **0.269** |
| Reductions-EO ($B$) | 0.807 | 0.151 | 0.633 | 0.865 | 0.795 | 0.383 |
| **A-IEOF** ($B$) | 0.792 | **0.149** | 0.625 | **0.311** | 0.807 | 0.389 |

meaningfully influences the predictor lowers $\epsilon_{\text{stab}}$ monotonically (Adult $0.23 \to 0.10$; CT metadata $0.64 \to 0.35$) at a small, graceful cost in stationary equalized odds ($+0.02$ to $+0.06$) and AUC ($-0.02$ to $-0.05$); on COMPAS the charge-degree block barely affects the model, so $\epsilon_{\text{stab}} \approx 0$ regardless and the dial is moot. Since degradation under $B$-shift is governed by $\epsilon_{\text{stab}}$ (Theorem B.2), a practitioner can trade a little stationary fairness and utility for a quantified reduction in worst-case degradation. We also tested whether a higher-capacity MLP backbone or a stronger fairness penalty Pareto-improves the stationary frontier; over ten seeds it did not, and an over-strong penalty degenerates to a near-constant predictor on low-signal data, so we retain the reported configuration.

### 5.6 Real Datasets: Adult, COMPAS, CT Metadata

We evaluate all seven methods on three real tabular datasets: UCI Adult ($B$=occupation; groups sex×age; $n$=30,718), COMPAS recidivism ($B$=charge degree; groups race×sex; $n$=6,172), and the CT metadata as a tabular task ($B$=institution; groups gender×age; $n$=12,652) Rahimzadeh et al. (2021). We report the pure in-processing policy, which isolates the training objective from post-processing and so addresses the concern that per-group post-processing can confound method comparisons; group-wise temperature scaling proved unstable on severely imbalanced strata, a property of the calibrator rather than of any training method. Values are means over ten seeds, with standard deviations in the released summaries.

Two patterns emerge (Table 5). First, A-IEOF attains the lowest EO gap on Adult ($0.149$) and COMPAS ($0.311$), matching or beating the reductions baseline at a modest utility cost: its AUC is within $0.03$ of the best method on Adult and drops more on COMPAS, where a large across-group base-rate disparity makes fairness expensive for every constrained method. The same ranking holds under Equal Opportunity, the false-negative-sensitive criterion appropriate to disease detection, where A-IEOF is dramatically the fairest on COMPAS ($0.158$ versus $0.40$–$0.50$ for all other methods) and competitive on Adult ($0.100$) and CT ($0.092$). Second, A-IEOF is the least $B$-sensitive learned model on all three datasets with $\epsilon_{\text{stab}}$ of $0.125$ on Adult, $0.001$ on COMPAS and $0.368$ on CT, against $0.328$, $0.048$ and $0.734$ for ERM; and shows the flattest EO gap under the empirical $B$-shift. Because $\epsilon_{\text{stab}}$ is the worst-case change in an individual's prediction when only $B$ is replaced, it is an individual-level, block-$B$ counterfactual-fairness measure; A-IEOF attains the lowest value everywhere, so by that criterion it is the fairest method even where its group-level gaps are not the smallest. The CT metadata is an honest boundary case: institution is strongly predictive of the label, so in distribution A-IEOF does not dominate (CORAL is best), and because the conditional-independence premise of Theorem B.2 then holds only approximately, all methods degrade under shift.

## 6 Discussion

The observed patterns trace to the three components of Eq. 2. *Interventional control* penalizes worst-case soft equalized odds under counterfactual replacements drawn from $g_\phi$, directly targeting disparities that emerge when $P(B \mid X_{-B}, S)$ shifts. For example when patient mix changes after site closures or openings and so reduces reliance on $B$ in ways that would amplify group disparities.

*Counterfactual stability* complements it by damping score changes when only $B$ varies, which the ablation confirms is what actually drives low $\epsilon_{\text{stab}}$. *Teacher anchoring* preserves the utility already captured by the base features (the no-$B$ teacher reaches AUC 0.928) and prevents regressions when $B$ is introduced, which matters when backward compatibility with a legacy system is required. Anchoring, however, inherits whatever the no-$B$ teacher encodes, including its biases; the ablation shows the distillation term contributes little to the fairness gains or to $\epsilon_{\text{stab}}$, so it is best understood as a tunable backward-compatibility mechanism rather than a fairness mechanism. When the legacy teacher is itself biased, its anchoring weight should be reduced or the teacher de-biased first, since the observational and interventional EO terms carry the fairness objective.

The three post-processing policies reflect different deployment constraints. When group-conditional decisions are prohibited, the global policy still delivers the best combined AUC, EO and calibration without needing group labels at test time. When policy permits group-specific thresholds with safeguards, the guarded shrinkage of Eq. 5 gives the smallest EO gap (0.085) while preventing overfitting on small strata. Necessary here, since the largest group holds $3.4\times$ the samples of the smallest. Hardt-EO correction reduces EO for every model, and A-IEOF retains the strongest discrimination and calibration afterwards, showing that interventional training is a good foundation for post-hoc correction; the residual DP gap (0.048) reflects the known tension between equalized odds and selection-rate parity.

A-IEOF outperforms gradient reversal because the two target different objectives. GRL enforces group-invariant representations by training an adversary to predict $S$ from features, an observational penalty on the training distribution, whereas A-IEOF penalizes multi-group soft equalized odds under counterfactual $B$-replacements and therefore constrains disparities that appear when $P(B \mid X_{-B}, S)$ shifts. GRL also has no mechanism for preserving no-$B$ behavior, no way to control sensitivity to a specific block, and only an indirect connection between its adversarial loss and equalized odds. The consequences are visible in probability quality (Brier 0.0495–0.0507 versus 0.0868–0.0918; ECE 0.0262–0.0343 versus 0.0868–0.0918) and in the uneven per-group behavior of GRL noted in Section 5.

## 7 Conclusion

We introduced A-IEOF, a training framework that aligns utility, calibration, and group fairness when a feature block $B$ is added or retained, combining interventional control via counterfactual $B$-replacements, anchoring to a no-$B$ teacher, and stability regularization. On the clinical corpus A-IEOF outperformed adversarial debiasing and the teacher baselines under all three post-processing policies, reaching AUC 0.976 with EO 0.126 under a global threshold, EO 0.085 under guarded per-group thresholds, and AUC 0.975 with EO 0.094 under Hardt-EO correction, with consistent gains across all four protected groups. In practice, use the global policy when group labels are unavailable, guarded per-group thresholds when validation support is adequate, and Hardt-EO for post-hoc correction. The shift stress tests (Section 5.4) and component ablations (Section 5.5) confirm that the interventional and stability terms deliver robustness to shifts in $P(B \mid X_{-B}, S)$ rather than stationary-distribution fairness alone. Future work includes further real-world corpora and simultaneous shifts in multiple feature blocks.

### AI use statement

In this work, generative AI tools were used solely to improve the readability and grammatical correctness of the manuscript. All AI-assisted text was reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

## Ethical Statement

This work studies fairness in clinical decision support, a high-stakes setting. Four considerations warrant note. First, A-IEOF supports group-specific thresholds (Policy B); although these reduce equalized-odds gaps, differential treatment by protected group raises legal and normative concerns, may be prohibited in some jurisdictions, and is unstable when small groups have limited validation support, so we also report the group-blind global policy and recommend it as the default unless

group-specific treatment is clearly justified and auditable. Second, institutional identifiers can encode socioeconomic, geographic and case-mix information correlated with protected attributes; using them for prediction, even with the safeguards proposed here, risks encoding structural inequities, and deployments should audit whether $B$ proxies a sensitive attribute. Third, our guarantees hold under a stated shift model (shifts confined to $P(B \mid X_{-B}, S)$) and do not cover general covariate or label shift, so clinical deployment should include out-of-distribution monitoring rather than relying on the transfer bound alone. Finally, the demographic strata analyzed here are coarse, derived from age and gender; fairness on them does not guarantee fairness at finer granularity. The dataset is a publicly available, de-identified research corpus, and no new human-subjects data were collected.

## Reproducibility and Transparency

All post-processing parameters are selected once on validation and held fixed for test evaluation, preventing overfitting to test data. Temperature parameters (Eq.4) are fit on validation with at least 40 samples per group and ridge regularization (weight 100.0) toward the global $T_{\text{glob}}$. Threshold (Eq.5 ) are optimized on validation via grid search (401 points over [0.02, 0.98]) with constraints on group size ($\geq 50$), positive/negative counts ($\geq 30$ each), and positive rate bounds ([0.15, 0.85]). Tables 1 and 2 report results under this protocol. Figure 2 documents the operating points used for per-group analysis, with markers on ROC and precision-recall curves corresponding to validation-fixed thresholds. The implementation is publicly available at the repository cited in the abstract, enabling independent verification and adaptation to new domains.

## References


Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International conference on machine learning*, pp. 60–69. PMLR, 2018.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27(3):642–669, 1956.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pp. 1180–1189. PMLR, 2015.

W Brier Glenn et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3, 1950.

Vincent Grari, Thibault Laugel, Tatsunori Hashimoto, Marcin Detyniecki, et al. On the fairness road: Robust optimization for adversarial debiasing. In *International Conference on Learning Representations*, volume 2024, pp. 36608–36638, 2024.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL `http://proceedings.mlr.press/v70/guo17a.html`.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016. URL `https://arxiv.org/abs/1610.02413`.

Matt J. Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

TASET SHIFT IN MACHINE LEARNING. Dataset shift in machine learning, 2009.

Pascal Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283, 1990.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 32, 2018. URL https://ojs.aaai.org/index.php/AAAI/article/view/12168.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

Letu Qingge, Hailemicael Lulseged Yimer, Maxwell Sam, Richard Annan, Robert Newman, and Hong Qin. C2bm: Causal concept disentanglement for fair multimodal covid-19 detection. In *Proceedings of the AAAI Symposium Series*, volume 7, pp. 567–575, 2025.

Mohammad Rahimzadeh, Abolfazl Attar, and Seyed Mohammad Sakhaei. A fully automated deep learning-based network for detecting covid-19 from a new and large lung ct scan dataset. *Biomedical Signal Processing and Control*, pp. 102588, 2021. ISSN 1746-8094. doi: https://doi.org/10.1016/j.bspc.2021.102588. URL https://www.sciencedirect.com/science/article/pii/S1746809421001853.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7237–7256. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.647.

Baochen Sun and Kate Saenko. Deep CORAL: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision Workshops (ECCVW)*, pp. 443–450. Springer, 2016.

Yixin Wang, Dhanya Sridhar, and David M. Blei. Adjusting machine learning decisions for equal opportunity and counterfactual fairness. *Transactions on Machine Learning Research*, 2023. Published 07/2023.

Solvejg Wastvedt, Jared D Huling, and Julian Wolfson. An intersectional framework for counterfactual fairness in risk prediction. *Biostatistics*, 25(3):702–717, 2024.

Jenny Yang, Andrew AS Soltan, David W Eyre, Yang Yang, and David A Clifton. An adversarial training framework for mitigating algorithmic biases in clinical machine learning. *NPJ digital medicine*, 6(1):55, 2023.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

Zeyu Zhou, Tianci Liu, Ruqi Bai, Jing Gao, Murat Kocaoglu, and David I Inouye. Counterfactual fairness by combining factual and counterfactual predictions. *Advances in Neural Information Processing Systems*, 37:47876–47907, 2024.

# A Appendix

## A Ablation Studies

This appendix provides a detailed ablation study for the A-IEOF framework, systematically evaluating the contribution of each component to fairness and utility performance.

### A.1 Experimental Setup

We evaluate 11 configurations on the COVID-19 detection dataset, progressively adding components to isolate their individual and combined effects. To keep the large configuration sweep tractable, this appendix ablation uses a lighter ResNet-18 backbone trained for 20 epochs, rather than the MobileNetV2 backbone trained for 30 epochs used for the headline results in the main text (Tables 1 and 2); all other hyperparameters are held consistent except for the ablated terms. Because the backbone and epoch budget differ, the absolute numbers in this appendix are not directly comparable to the main tables and should be read only as within-table comparisons across configurations. A backbone-consistent ablation that uses exactly the main pipeline is reported on the tabular benchmark in Section 5.5. The test set here contains 418 samples across 4 demographic groups (G0: $n = 115$, G1: $n = 52$, G2: $n = 162$, G3: $n = 89$).

### A.2 Configurations Evaluated

We test the following configurations, where $\lambda_f$ denotes $\lambda_{interv}$ (interventional fairness), $\lambda_d$ denotes $\lambda_{dist}$ (distillation), $\lambda_{full}$ denotes $\lambda_{full_fair}$ (full EO regularization), and $\lambda_s$ denotes $\lambda_{stab}$ (stability):

1. **ERM (baseline):** Standard empirical risk minimization with no fairness intervention ($\lambda_f = 0, \lambda_d = 0, \lambda_{full} = 0, \lambda_s = 0$)
2. **+ Full EO only:** Direct EO regularization without other components ($\lambda_{full} = 1.0$)
3. **+ Distillation only:** Teacher anchoring without fairness terms ($\lambda_d = 0.6$)
4. **+ Interventional only:** Representation-level interventional fairness alone ($\lambda_f = 0.05$)
5. **Distill + Full EO:** Core combination without interventional components ($\lambda_d = 0.6, \lambda_{full} = 1.0$)
6. **+ Small interventional:** Adding weak interventional control ($\lambda_f = 0.05, \lambda_d = 0.6, \lambda_{full} = 1.0$)
7. **Full A-IEOF:** Complete framework with all terms ($\lambda_f = 0.1, \lambda_d = 0.6, \lambda_{full} = 1.0, \lambda_s = 0.05$)
8. **Full-Distillation:** Ablating teacher anchoring ($\lambda_d = 0$)
9. **Full-Full EO:** Ablating direct EO regularization ($\lambda_{full} = 0$)
10. **Distill + Stability only:** Post-processing approach without in-training fairness ($\lambda_d = 0.6, \lambda_s = 0.05$)
11. **Best alternative:** Strongest configuration without interventional term ($\lambda_d = 0.7, \lambda_{full} = 1.2$)

### A.3 Results

Table 6 presents comprehensive results for all configurations. We report AUC (discrimination), weighted EO gap (fairness), demographic parity gap (DP), and Brier score (calibration). The EO gap is computed as $\Delta_{EO} = \text{TPR_spread} + 2 \times \text{FPR_spread}$ following Equation equation 13, prioritizing false positive rate parity.

### A.4 Key Findings

#### A.4.1 Individual Component Effects

**Distillation (Teacher Anchoring):** Comparing configurations 1 vs. 3, distillation alone improves EO gap from 0.280 to 0.172 (38% reduction) while boosting AUC from 0.911 to 0.922. This validates our hypothesis that anchoring to a no-institution (*no_B*) teacher provides a fairness-aware initialization. The dramatic DP improvement (0.181 $\rightarrow$ 0.077) suggests distillation reduces institution-correlated biases.

Table 6: Ablation study results. Best values in bold, second-best underlined.

| Configuration | AUC ↑ | EO ↓ | DP ↓ | Brier ↓ |
|---|---|---|---|---|
| 1. ERM (baseline) | 0.911 | 0.280 | 0.181 | 0.101 |
| 2. + Full EO only | 0.908 | 0.190 | 0.122 | 0.109 |
| 3. + Distillation only | 0.922 | 0.172 | <u>0.077</u> | 0.105 |
| 4. + Interventional only | 0.911 | <u>0.158</u> | 0.128 | 0.105 |
| 5. Distill + Full EO | 0.922 | 0.234 | 0.129 | 0.102 |
| 6. + Small interventional | 0.919 | 0.465 | 0.251 | 0.101 |
| 7. **Full A-IEOF** | **0.926** | **0.160** | **0.079** | **0.094** |
| 8. Full-Distillation | 0.921 | 0.370 | 0.207 | 0.098 |
| 9. Full-Full EO | 0.915 | 0.219 | 0.105 | 0.101 |
| 10. Distill + Stability only | 0.922 | 0.339 | 0.207 | 0.103 |
| 11. Best alternative | 0.916 | 0.192 | 0.128 | 0.114 |

**Full EO Regularization:** Configuration 2 shows that direct EO control reduces the gap from 0.280 to 0.190 (32% reduction) but at a slight AUC cost (0.911 $\rightarrow$ 0.908) and calibration degradation (0.101 $\rightarrow$ 0.109). This demonstrates the utility-fairness trade-off inherent in direct fairness constraints.

**Interventional Fairness:** Configuration 4 achieves the second-best single-component EO gap (0.158), validating that representation-level interventional control effectively reduces sensitivity to institution. However, this comes with increased FPR spread across groups, suggesting the need for complementary mechanisms.

### A.4.2 Component Interactions

**Distillation + Full EO (Configuration 5):** This core combination achieves strong utility (AUC = 0.922) but surprisingly high EO gap (0.234), worse than either component alone. This suggests that naive combination can lead to conflicting optimization pressures, motivating the need for careful balancing via interventional control.

**Effect of Adding Interventional Control:** Configuration 6 (adding small $\lambda_f = 0.05$ to configuration 5) dramatically *worsens* fairness (EO: 0.234 $\rightarrow$ 0.465). Examination of per-group metrics reveals extreme suppression in group G1 (TPR drops from 0.833 to 0.556), indicating instability. This highlights the importance of proper hyperparameter tuning and the need for stability regularization.

**Critical Role of Stability:** The Full A-IEOF (configuration 7) adds $\lambda_{stab} = 0.05$ to prevent the instability observed in configuration 6. This recovers strong performance (EO = 0.160) while maintaining the best utility (AUC = 0.926) and calibration (Brier = 0.094). The stability term acts as a crucial stabilizer for the interventional control.

### A.4.3 Ablation of Full A-IEOF

**Removing Distillation (Configuration 8):** Ablating $\lambda_d$ increases EO gap from 0.160 to 0.370 (131% degradation), demonstrating that teacher anchoring is essential for preventing drift during fairness optimization. Without distillation, the model loses the fairness-aware initialization, and the interventional control alone cannot recover comparable fairness.

**Removing Full EO (Configuration 9):** Ablating $\lambda_{full}$ increases EO gap from 0.160 to 0.219 (37% degradation). While not as critical as distillation, the direct EO constraint provides complementary pressure that interventional control alone cannot fully replace.

**Removing In-Training Fairness (Configuration 10):** Relying only on distillation and post-processing stability (no $\lambda_f$ or $\lambda_{full}$) results in EO = 0.339, more than double the full framework. This confirms that in-training fairness optimization is necessary; post-processing alone is insufficient.

### A.4.4 Alternative Configurations

Configuration 11 explores whether interventional control can be avoided entirely by strengthening distillation ($\lambda_d = 0.7$) and full EO ($\lambda_{full} = 1.2$). This achieves reasonable fairness (EO = 0.192) but falls short of the full framework (0.160) and significantly degrades calibration (0.114 vs. 0.094). This suggests that interventional control provides unique benefits beyond what can be achieved by simply scaling other components.

## A.5 Per-Group Performance Analysis

Table 7 presents detailed per-group metrics for key configurations, revealing how different components affect demographic groups differently.

Table 7: Per-group TPR and FPR for selected configurations. Group G1 ($n = 52$) shows highest instability.

| | **True Positive Rate** | | | | **False Positive Rate** | | | |
|---|---|---|---|---|---|---|---|---|
| **Config** | G0 | G1 | G2 | G3 | G0 | G1 | G2 | G3 |
| 1. ERM | 0.769 | 0.722 | **0.912** | 0.824 | 0.190 | **0.059** | 0.149 | 0.127 |
| 4. + Interv only | 0.827 | 0.833 | 0.912 | 0.824 | 0.254 | 0.147 | 0.181 | 0.218 |
| 6. + Small interv | 0.750 | 0.556 | **0.912** | 0.853 | 0.206 | 0.029 | 0.138 | 0.164 |
| 7. **Full A-IEOF** | 0.788 | 0.833 | 0.912 | 0.824 | 0.143 | 0.118 | 0.106 | 0.109 |
| 8. Full-Distill | 0.827 | 0.611 | 0.912 | 0.735 | 0.143 | **0.059** | 0.128 | 0.109 |

**Key observations:**

- **Group G1 instability:** The smallest group ($n = 52$) exhibits extreme volatility without proper regularization. Configuration 6 shows catastrophic TPR collapse (0.556) and FPR over-suppression (0.029), while configuration 8 shows similar patterns (TPR = 0.611).
- **Stability regularization effect:** Full A-IEOF (configuration 7) successfully stabilizes G1 (TPR = 0.833, FPR = 0.118) while maintaining strong performance across all groups.
- **FPR equalization:** Full A-IEOF achieves remarkably consistent FPR across groups (range: 0.106–0.143), validating the weighted EO objective's emphasis on FPR parity ($w_- = 2$).

## A.6 Hyperparameter Sensitivity

The ablation configurations reveal key insights about hyperparameter sensitivity:

1. **Interventional strength ($\lambda_f$):** Small values (0.05) can cause instability without stability regularization. Comparing configurations 6 vs. 7, adding $\lambda_s = 0.05$ prevents the catastrophic failure observed when $\lambda_f = 0.05$ is used alone (EO: 0.465 $\rightarrow$ 0.160). Moderate values ($\lambda_f = 0.1$ in configuration 7) work well with stability.
2. **Distillation strength ($\lambda_d$):** Moderate values (0.6–0.7) balance utility preservation and fairness. Configuration 3 ($\lambda_d = 0.6$) improves AUC from 0.911 to 0.922 while reducing EO gap by 38%. Configuration 11 ($\lambda_d = 0.7$) shows that further increases maintain utility but require higher $\lambda_{full}$ for comparable fairness.
3. **Full EO regularization ($\lambda_{full}$):** Higher values (1.0–1.2) improve fairness but may degrade calibration. Configuration 7 ($\lambda_{full} = 1.0$) achieves excellent calibration (Brier = 0.094), while configuration 11 ($\lambda_{full} = 1.2$) shows calibration degradation (Brier = 0.114) despite similar AUC.
4. **Stability regularization ($\lambda_s$):** Small values (0.05) are sufficient to prevent intervention-induced instability. Configuration 7 demonstrates that $\lambda_s = 0.05$ successfully stabilizes group G1 (TPR = 0.833) compared to configuration 6 without stability (TPR = 0.556).

These findings guide hyperparameter selection: start with moderate distillation ($\lambda_d \approx 0.6$), moderate full EO ($\lambda_{full} \approx 1.0$), low interventional control ($\lambda_f \approx 0.1$), and small stability ($\lambda_s \approx 0.05$). Fine-tuning should prioritize calibration (via $\lambda_{full}$) and group-specific stability (via $\lambda_s$).

### A.7 SUMMARY

This ablation study demonstrates that:

1. **All components are necessary:** Removing any major component (distillation, full EO, interventional control, or stability) significantly degrades fairness.
2. **Distillation is most critical:** Removing $\lambda_d$ causes the largest fairness degradation (131%), highlighting the importance of fairness-aware initialization.
3. **Stability prevents catastrophic failures:** Without $\lambda_s$, interventional control can cause extreme performance degradation in small groups, validating the need for Equation equation 49.
4. **Synergy between components:** The full framework achieves better performance than would be predicted by summing individual component contributions, indicating positive interactions.
5. **Interventional control is irreplaceable:** Simply scaling other components (configuration 11) cannot match the fairness and calibration of the full framework, confirming the unique value of representation-level interventional fairness (Equation equation 50).

These findings validate our design choices and demonstrate that A-IEOF's performance stems from the carefully orchestrated combination of all components, rather than any single mechanism.

**Refinement from the backbone-consistent ablation.** The image ablation above measures only stationary-distribution fairness. The tabular ablation in Section 5.5, which uses exactly the main training pipeline, refines the interpretation: the observational EO term is the dominant driver of the stationary equalized-odds gap, whereas the interventional and stability terms primarily reduce the worst-case sensitivity $\epsilon_{stab}$ and therefore act under distribution shift. We thus temper the stronger reading that every term reduces every fairness metric in every regime; the terms play complementary roles across the stationary and shifted settings.

## B TRANSFER BOUND FOR INTERVENTIONAL FAIRNESS: COMPLETE PROOF

*This appendix provides a complete formalization of the transfer bound mentioned in Methodology Section. For readers seeking a quick overview, we recommend reading Appendix B subsections Theorem Statement and Proof Steps 1-6, which together comprise approximately 3 pages. The remaining subsections provide practical implications, assumptions, comparisons to related work, and future directions for those interested in deeper context.*

### B.1 NOTATION AND DEFINITIONS

We first establish all notation used in the theorem and proof to ensure clarity and accessibility.

#### B.1.1 BASIC NOTATION

- **Model and predictions:**
  - $f_\theta$: Neural network model with parameters $\theta$
  - $\sigma(\cdot)$: Sigmoid function, $\sigma(z) = \frac{1}{1+e^{-z}}$, mapping scores to probabilities in $[0, 1]$
  - $p(x) = \sigma(f_\theta(x))$: Predicted probability for input $x$
- **Input features:**
  - $x$: Complete input vector containing all patient features
  - $x_{-B}$: Base features excluding block $B$ (e.g., age, gender, CT scan features)
  - $B$ or $x_B$: Designated feature block (e.g., institution identifier)
  - $x^{(b)} := (x_{-B}, b)$: Counterfactual input with institution changed to $b$
- **Labels and groups:**
  - $y \in \{0, 1\}$: Binary label (e.g., 0=negative, 1=COVID-positive)

- $s \in \{0, \ldots, G-1\}$: Protected group membership (e.g., demographic categories)
- $g$: Index for a specific protected group

- **Distributions:**
  - $P$: Training distribution over $(X, Y, S, B)$
  - $P'$: Deployment distribution (after distribution shift)
  - $\mathbb{E}_P[\cdot]$: Expectation with respect to distribution $P$
  - $P(A|B)$: Conditional probability of $A$ given $B$

### B.1.2 KEY QUANTITIES

**Stability (Worst-Case Counterfactual Sensitivity):** The interventional stability $\epsilon_{stab}$ measures the maximum change in predictions when only feature block $B$ is modified:

$$\epsilon_{stab} = \sup_{x_{-B}, s} \max_{b, b'} |\sigma(f_\theta(x^{(b)})) - \sigma(f_\theta(x^{(b')}))| \tag{8}$$

*Intuition:* Across all possible patients (characterized by $x_{-B}, s$) and all possible institution pairs $(b, b')$, what is the largest prediction change we observe? If $\epsilon_{stab}$ is small, the model treats all institutions similarly.

*Connection to training:* The $\lambda_{stab}$ term in Equation (5) of the main text minimizes the expected sensitivity $\mathbb{E}_{x,s,b',b''}[|\sigma(f_\theta(x^{(b')})) - \sigma(f_\theta(x^{(b'')}))|]$. By Jensen's inequality applied to the max operator, minimizing the expectation provides an upper bound on $\epsilon_{stab}$. In practice, $\epsilon_{stab}$ can be estimated as the 95th or 99th percentile of $|\sigma(f_\theta(x^{(b')})) - \sigma(f_\theta(x^{(b'')}))$— over validation data, providing a finite-sample approximation to the worst-case.

**Total Variation Distance:** For discrete distributions $P$ and $Q$ over a finite set $\mathcal{B}$, the total variation (TV) distance is:

$$\|P - Q\|_{TV} = \frac{1}{2} \sum_{b \in \mathcal{B}} |P(b) - Q(b)| \tag{9}$$

The maximum total variation distance across all contexts is:

$$\Delta_{TV}(P, P') = \sup_{x_{-B}, s} \|P(B|x_{-B}, s) - P'(B|x_{-B}, s)\|_{TV} \tag{10}$$

*Intuition:* For each type of patient (characterized by features $x_{-B}$ and group $s$), how different is the distribution over institutions between training and deployment? The TV distance ranges from 0 (identical distributions) to 1 (disjoint support). For example, if 30% of patients redistribute across institutions, $\Delta_{TV}$ is approximately 0.3.

**Equalized Odds Gap:** For each group $g$, define true and false positive rates from hard decisions $\hat{y} \in \{0, 1\}$:

$$\text{TPR}_g = \Pr(\hat{Y} = 1 | Y = 1, S = g) \quad \text{(True Positive Rate)} \tag{11}$$

$$\text{FPR}_g = \Pr(\hat{Y} = 1 | Y = 0, S = g) \quad \text{(False Positive Rate)} \tag{12}$$

The weighted equalized-odds gap is:

$$\begin{aligned} \Delta_{EO} = & w_+ (\max_g \text{TPR}_g - \min_g \text{TPR}_g) \\ & + w_- (\max_g \text{FPR}_g - \min_g \text{FPR}_g) \end{aligned} \tag{13}$$

where $w_+, w_- > 0$ are weights (we use $w_+ = 1, w_- = 2$ to prioritize false positive rate parity).

*Intuition:* This measures the largest difference in error rates between the best-performing and worst-performing demographic groups. Lower $\Delta_{EO}$ indicates more equitable treatment across groups. Equalized odds is satisfied when $\Delta_{EO} = 0$, meaning all groups have identical TPR and FPR.

### B.2 THEOREM STATEMENT

[Transfer Bound for Interventional Fairness] Let $f_\theta$ be a model trained under A-IEOF with worst-case interventional stability $\epsilon_{stab}$ as defined in Equation equation 8. Consider a deployment distribution $P'$ that differs from the training distribution $P$ only in the conditional distribution of feature block $B$:

$$P'(X_{-B}, Y, S) = P(X_{-B}, Y, S) \tag{14}$$

$$P'(B|X_{-B}, S) \neq P(B|X_{-B}, S) \tag{15}$$

Let $\Delta_{TV}(P, P')$ denote the maximum total variation distance between the conditional distributions of $B$ (Equation equation 10). Then the equalized-odds gap under the deployment distribution is bounded by:

$$\Delta_{EO}^{P'} \leq \Delta_{EO}^{P} + 4(w_+ + w_-)\epsilon_{stab} \cdot \Delta_{TV}(P, P') \tag{16}$$

For our setting with $w_+ = 1, w_- = 2$ (prioritizing FPR parity):

$$\Delta_{EO}^{P'} \leq \Delta_{EO}^{P} + 12\epsilon_{stab} \cdot \Delta_{TV}(P, P') \tag{17}$$

where $\Delta_{EO}^{P}$ is the equalized-odds gap on the training distribution and $\Delta_{EO}^{P'}$ is the gap on the deployment distribution.

*Interpretation:* If the model has low counterfactual sensitivity ($\epsilon_{stab}$ small) as enforced by the stability regularizer $\lambda_{stab}$ in Equation (5) of the main text, then fairness violations grow at most linearly with the magnitude of the distribution shift in $B$. This provides graceful degradation rather than catastrophic failure: even under substantial distribution shifts, fairness remains bounded.

### B.3 COMPLETE PROOF

We prove the bound by decomposing the fairness metrics under the shifted distribution and bounding the change using the worst-case stability property.

#### B.3.1 STEP 1: DECOMPOSE TPR UNDER DEPLOYMENT

Fix a protected group $g$. The true positive rate under the deployment distribution $P'$ can be written as:

$$\text{TPR}_g^{P'} = \mathbb{E}_{P'}[\sigma(f_\theta(X)) \mid Y = 1, S = g] \tag{18}$$

Expanding using the law of total expectation over base features $X_{-B}$ and institution $B$:

$$\begin{aligned} \text{TPR}_g^{P'} = &\sum_{x_{-B}} P'(x_{-B}|Y = 1, S = g) \\ &\times \sum_{b} P'(B = b|x_{-B}, Y = 1, S = g) \\ &\times \sigma(f_\theta(x_{-B}, b)) \end{aligned} \tag{19}$$

**Conditional Independence Assumption:** We assume that institution assignment is independent of the label given the base features and group membership:

$$P'(B|X_{-B}, Y, S) = P'(B|X_{-B}, S) \tag{20}$$

*Justification:* Once we observe a patient's demographics, symptoms, and clinical features ($X_{-B}, S$), the institution they attend should not depend on their COVID status ($Y$), as institution assignment typically occurs before diagnosis. This is a standard conditional independence assumption in causal inference and is reasonable for our healthcare setting where patients are assigned to institutions based on geographic location, insurance, and availability rather than test results.

Under this assumption:

$$\begin{aligned}\text{TPR}_g^{P'} = \sum_{x_{-B}} & P'(x_{-B}|Y=1,S=g) \\ & \times \sum_b P'(B=b|x_{-B},S=g) \\ & \times \sigma(f_\theta(x_{-B},b))\end{aligned} \tag{21}$$

### B.3.2 Step 2: Isolate the Distribution Shift

Since by assumption $P'(X_{-B},Y,S) = P(X_{-B},Y,S)$ (Equation equation 14), the marginal and conditional distributions over base features and labels are identical:

$$P'(x_{-B}|Y=1,S=g) = P(x_{-B}|Y=1,S=g) \tag{22}$$

Therefore, the difference in TPR between deployment and training is:

$$\begin{aligned}&\text{TPR}_g^{P'} - \text{TPR}_g^{P} \\ &= \sum_{x_{-B}} P(x_{-B}|Y=1,S=g) \\ &\quad \times \Bigg[\sum_b P'(b|x_{-B},g)\cdot\sigma(f_\theta(x_{-B},b)) \\ &\quad - \sum_b P(b|x_{-B},g)\cdot\sigma(f_\theta(x_{-B},b))\Bigg]\end{aligned} \tag{23}$$

Factoring out the common probability weight $P(x_{-B}|Y=1,S=g)$:

$$\begin{aligned}&= \sum_{x_{-B}} P(x_{-B}|Y=1,S=g) \\ &\quad \times \sum_b [P'(b|x_{-B},g) - P(b|x_{-B},g)] \\ &\quad \times \sigma(f_\theta(x_{-B},b))\end{aligned} \tag{24}$$

This cleanly isolates the effect of the shift in $P(B|X_{-B},S)$ on the true positive rate.

### B.3.3 Step 3: Apply Stability to Bound the Change

Fix a context $(x_{-B},g)$ and consider the inner sum over institutions:

$$\Delta_{\text{pred}} := \sum_b [P'(b|x_{-B},g) - P(b|x_{-B},g)]\cdot\sigma(f_\theta(x_{-B},b)) \tag{25}$$

Let $b_0$ be an arbitrary reference institution value (e.g., the most common institution in training), and define the reference prediction:

$$\sigma_0 = \sigma(f_\theta(x_{-B},b_0)) \tag{26}$$

We decompose each prediction as the reference plus a deviation:

$$\sigma(f_\theta(x_{-B},b)) = \sigma_0 + [\sigma(f_\theta(x_{-B},b)) - \sigma_0] \tag{27}$$

Substituting into Equation equation 25:

$$\begin{aligned}\Delta_{\text{pred}} = & \sum_b [P'(b) - P(b)]\cdot\sigma_0 \\ & + \sum_b [P'(b) - P(b)]\cdot[\sigma(f_\theta(x_{-B},b)) - \sigma_0]\end{aligned} \tag{28}$$

where we use shorthand $P(b) := P(b|x_{-B},g)$ and $P'(b) := P'(b|x_{-B},g)$ for notational clarity.

**First term vanishes:** Since probability distributions sum to one:

$$\sum_b [P'(b) - P(b)] = \sum_b P'(b) - \sum_b P(b) = 1 - 1 = 0 \tag{29}$$

Therefore:

$$\sum_b [P'(b) - P(b)] \cdot \sigma_0 = \sigma_0 \cdot 0 = 0 \tag{30}$$

**Second term bounded by worst-case stability:** By the definition of worst-case interventional stability (Equation equation 8), for the fixed context $(x_{-B}, g)$ and any institution value $b$:

$$|\sigma(f_\theta(x_{-B}, b)) - \sigma_0| = |\sigma(f_\theta(x^{(b)})) - \sigma(f_\theta(x^{(b_0)}))| \leq \epsilon_{stab} \tag{31}$$

Now bound the absolute value of the second term using triangle inequality:

$$\begin{aligned} &\left| \sum_b [P'(b) - P(b)] \cdot [\sigma(f_\theta(x_{-B}, b)) - \sigma_0] \right| \\ &\leq \sum_b |P'(b) - P(b)| \cdot |\sigma(f_\theta(x_{-B}, b)) - \sigma_0| \\ &\leq \sum_b |P'(b) - P(b)| \cdot \epsilon_{stab} \\ &= \epsilon_{stab} \sum_b |P'(b|x_{-B}, g) - P(b|x_{-B}, g)| \end{aligned} \tag{32}$$

By the definition of total variation distance (Equation equation 9):

$$\sum_b |P'(b) - P(b)| = 2\|P'(b|x_{-B}, g) - P(b|x_{-B}, g)\|_{TV} \tag{33}$$

Therefore:

$$|\Delta_{\text{pred}}| \leq 2\epsilon_{stab} \cdot \|P'(b|x_{-B}, g) - P(b|x_{-B}, g)\|_{TV} \tag{34}$$

### B.3.4 Step 4: Maximize Over Contexts

Returning to the full expression for the TPR difference (Equation equation 24), and using the bound from Equation equation 34:

$$\begin{aligned} |\text{TPR}_g^{P'} - \text{TPR}_g^{P}| &\leq \sum_{x_{-B}} P(x_{-B}|Y = 1, S = g) \cdot |\Delta_{\text{pred}}| \\ &\leq \sum_{x_{-B}} P(x_{-B}|Y = 1, S = g) \\ &\quad \times 2\epsilon_{stab} \|P'(b|x_{-B}, g) \\ &\quad - P(b|x_{-B}, g)\|_{TV} \end{aligned} \tag{35}$$

Since $\sum_{x_{-B}} P(x_{-B}|Y = 1, S = g) = 1$, we can pull out the maximum:

$$\begin{aligned} &\leq 2\epsilon_{stab} \sum_{x_{-B}} P(x_{-B}|Y = 1, S = g) \cdot \|P'(b) - P(b)\|_{TV} \\ &\leq 2\epsilon_{stab} \cdot \sup_{x_{-B}} \|P'(b) - P(b)\|_{TV} \end{aligned} \tag{36}$$

By the definition of $\Delta_{TV}(P, P')$ (Equation equation 10):

$$|\text{TPR}_g^{P'} - \text{TPR}_g^{P}| \leq 2\epsilon_{stab} \cdot \Delta_{TV}(P, P') \tag{37}$$

### B.3.5 STEP 5: APPLY TO FPR

By an identical argument applied to the false positive rate, conditioning on $Y = 0$ instead of $Y = 1$:

$$|\text{FPR}_g^{P'} - \text{FPR}_g^{P}| \leq 2\epsilon_{stab} \cdot \Delta_{TV}(P, P') \tag{38}$$

The proof is symmetric because the stability property (Equation equation 8) holds regardless of the true label value, and the conditional independence assumption (Equation equation 20) applies equally to both $Y = 0$ and $Y = 1$.

### B.3.6 STEP 6: BOUND THE EO GAP

The equalized-odds gap under the deployment distribution is:

$$\begin{aligned}\Delta_{EO}^{P'} = w_+(\max_g \text{TPR}_g^{P'} - \min_g \text{TPR}_g^{P'}) \\ + w_-(\max_g \text{FPR}_g^{P'} - \min_g \text{FPR}_g^{P'})\end{aligned} \tag{39}$$

**Bounding the TPR spread:** Let $g^* = \arg\max_g \text{TPR}_g^{P'}$ and $g_* = \arg\min_g \text{TPR}_g^{P'}$ denote the groups with maximum and minimum TPR under deployment. By the triangle inequality:

$$\begin{aligned}&\text{TPR}_{g^*}^{P'} - \text{TPR}_{g_*}^{P'} \\ &= [\text{TPR}_{g^*}^{P} + (\text{TPR}_{g^*}^{P'} - \text{TPR}_{g^*}^{P})] \\ &\quad - [\text{TPR}_{g_*}^{P} + (\text{TPR}_{g_*}^{P'} - \text{TPR}_{g_*}^{P})] \\ &= (\text{TPR}_{g^*}^{P} - \text{TPR}_{g_*}^{P}) \\ &\quad + (\text{TPR}_{g^*}^{P'} - \text{TPR}_{g^*}^{P}) - (\text{TPR}_{g_*}^{P'} - \text{TPR}_{g_*}^{P}) \\ &\leq (\text{TPR}_{g^*}^{P} - \text{TPR}_{g_*}^{P}) \\ &\quad + |\text{TPR}_{g^*}^{P'} - \text{TPR}_{g^*}^{P}| + |\text{TPR}_{g_*}^{P'} - \text{TPR}_{g_*}^{P}|\end{aligned} \tag{40}$$

Since the max and min are taken over all groups:

$$\text{TPR}_{g^*}^{P} - \text{TPR}_{g_*}^{P} \leq \max_g \text{TPR}_g^{P} - \min_g \text{TPR}_g^{P} \tag{41}$$

And using Equation equation 37:

$$\begin{aligned}|\text{TPR}_{g^*}^{P'} - \text{TPR}_{g^*}^{P}| + |\text{TPR}_{g_*}^{P'} - \\ \text{TPR}_{g_*}^{P}| \leq 2 \cdot 2\epsilon_{stab}\Delta_{TV} = 4\epsilon_{stab}\Delta_{TV}\end{aligned} \tag{42}$$

Therefore:

$$\begin{aligned}\max_g \text{TPR}_g^{P'} - \min_g \text{TPR}_g^{P'} \leq (\max_g \text{TPR}_g^{P} - \\ \min_g \text{TPR}_g^{P}) + 4\epsilon_{stab}\Delta_{TV}\end{aligned} \tag{43}$$

**Similarly for FPR spread:** By identical reasoning:

$$\begin{aligned}\max_g \text{FPR}_g^{P'} - \min_g \text{FPR}_g^{P'} \leq (\max_g \text{FPR}_g^{P} - \\ \min_g \text{FPR}_g^{P}) + 4\epsilon_{stab}\Delta_{TV}\end{aligned} \tag{44}$$

**Combining with weights:** Substituting Equations equation 43 and equation 44 into the definition of $\Delta_{EO}^{P'}$ (Equation equation 39):

$$\begin{aligned}\Delta_{EO}^{P'} &\leq w_+[(\max_g \text{TPR}_g^P - \min_g \text{TPR}_g^P) + 4\epsilon_{stab}\Delta_{TV}] \\ &\quad + w_-[(\max_g \text{FPR}_g^P - \min_g \text{FPR}_g^P) + 4\epsilon_{stab}\Delta_{TV}] \\ &= w_+(\max_g \text{TPR}_g^P - \min_g \text{TPR}_g^P) \\ &\quad + w_-(\max_g \text{FPR}_g^P - \min_g \text{FPR}_g^P) \\ &\quad + 4(w_+ + w_-)\epsilon_{stab}\Delta_{TV} \\ &= \Delta_{EO}^P + 4(w_+ + w_-)\epsilon_{stab} \cdot \Delta_{TV}(P, P')\end{aligned} \tag{45}$$

**Specializing to our weights:** With $w_+ = 1$ and $w_- = 2$ (as specified in the main text):

$$\begin{aligned}\Delta_{EO}^{P'} \leq \Delta_{EO}^P + 4 \times (1+2) \times \epsilon_{stab} \cdot \Delta_{TV}(P, P') = \\ \Delta_{EO}^P + 12\epsilon_{stab} \cdot \Delta_{TV}(P, P')\end{aligned} \tag{46}$$

This completes the proof.

### B.4 Practical Implications

#### B.4.1 Interpretation of the Bound

The transfer bound (Theorem B.2) establishes that fairness degradation under distribution shift is controlled by two factors:

1. **Model stability ($\epsilon_{stab}$):** The maximum change in predictions when only feature block $B$ varies. The A-IEOF objective (Equation 5, main text) explicitly controls this through:
   - The stability regularizer $\lambda_{stab}$ (term 5) minimizes expected sensitivity
   - The interventional EO control $\lambda_{interv}$ (term 3) trains on worst-case counterfactual B-replacements
2. **Shift magnitude ($\Delta_{TV}$):** How different the distribution of $B$ is between training and deployment. This is determined by external factors (e.g., hospital closures, policy changes, patient flow dynamics).

The linear relationship in Equation equation 17 (where $C = 12$ for our weights) means fairness degrades *gracefully* rather than catastrophically. Even under large distributional shifts ($\Delta_{TV} \approx 1$), if stability is well-controlled ($\epsilon_{stab} \approx 0.01$), fairness violations remain bounded.

#### B.4.2 Numerical Example

Consider a realistic deployment scenario with institutional redistribution:

- **Baseline fairness:** $\Delta_{EO}^P = 0.085$ (8.5% disparity between groups on training distribution)
- **Model stability:** $\epsilon_{stab} = 0.01$ (1% maximum prediction change across institutions, estimated as 99th percentile on validation)
- **Distribution shift:** Hospital closure causes 30% of patients to redistribute; $\Delta_{TV} = 0.3$

The theoretical upper bound predicts:

$$\Delta_{EO}^{P'} \leq 0.085 + 12 \times 0.01 \times 0.3 = 0.085 + 0.036 = 0.121 \tag{47}$$

Fairness degrades by at most 3.6 percentage points (from 8.5% to 12.1%) under major institutional disruption.

**Comparison to unstable baseline:** A model without stability control (e.g., standard training with $\epsilon_{stab} = 0.05$) would have:

$$\Delta_{EO}^{P'} \leq 0.085 + 12 \times 0.05 \times 0.3 = 0.085 + 0.18 = 0.265 \tag{48}$$

This represents a 180% increase in fairness violations (from 8.5% to 26.5%), more than tripling the baseline gap. The comparison illustrates the practical value of interventional stability control.

**Bound tightness:** The bound is intentionally conservative, using worst-case stability $\epsilon_{stab}$. In practice, if the model has low *average* sensitivity (enforced by $\lambda_{stab}$ and $\lambda_{interv}$), actual fairness degradation is typically much smaller than the bound predicts. The distribution-shift stress test in Section 5.4 of the main text validates this empirically. On the controlled benchmark, A-IEOF attains $\epsilon_{stab} \approx 0.040$ and its observed equalized-odds degradation stays below $0.01$ as the shift grows to $\Delta_{TV} = 0.5$, comfortably inside the corresponding upper bound $12 \times 0.040 \times 0.5 = 0.24$. By contrast, an empirical-risk baseline attains $\epsilon_{stab} \approx 0.154$, for which the same bound is loose. This confirms the bound is a valid, and for A-IEOF non-vacuous, guarantee whose usefulness depends on the training objective controlling $\epsilon_{stab}$.

### B.4.3 Connection to A-IEOF Training

The A-IEOF objective (Equation 5, main text) directly addresses the preconditions of Theorem B.2 through complementary mechanisms:

1. **Stability regularization ($\lambda_{stab}$, term 5):**

$$\lambda_{stab}\mathbb{E}[|\sigma(f_\theta(x^{(b')})) - \sigma(f_\theta(x^{(b'')}))|] \tag{49}$$

   This directly minimizes the expected sensitivity, which by Jensen's inequality applied to the max operator provides an upper bound on the worst-case $\epsilon_{stab}$ (Equation equation 8).

2. **Interventional EO control ($\lambda_{interv}$, term 3):**

$$\lambda_{interv} \sup_{b',b''} \Delta_{EO}^{soft}(f_\theta; x^{(b')}, x^{(b'')}) \tag{50}$$

   This trains on worst-case counterfactual B-replacements sampled by adversarial generator $g_\phi$, effectively simulating distribution shifts during training. By enforcing fairness under adversarially chosen institution pairs, the model learns representations that are robust to the types of shifts formalized in Theorem B.2.

3. **Teacher anchoring ($\lambda_{dist}$, term 2):** While not directly appearing in the bound, this mechanism preserves baseline performance on $X_{-B}$, ensuring that $\Delta_{EO}^{P}$ (the starting fairness level) remains low by preventing drift from the no-B teacher model.

Together, these mechanisms ensure the model satisfies the stability precondition ($\epsilon_{stab}$ small) required for the bound to provide meaningful guarantees under realistic deployment shifts.

## B.5 Assumptions and Limitations

### B.5.1 Key Assumptions

The transfer bound relies on three main assumptions:

1. **Conditional independence of $B$ and $Y$:**

$$P'(B|X_{-B}, Y, S) = P'(B|X_{-B}, S) \tag{51}$$

   Institution assignment does not depend on outcomes given patient features and demographics (see Equation equation 20). This is reasonable when institution choice is determined by geographic location, insurance networks, or availability before diagnosis.

   *When it may fail:* If certain institutions specialize in treating severe cases (e.g., tertiary care centers for critical patients), then $B$ and $Y$ may be conditionally dependent even given $X_{-B}, S$. In such cases, the bound requires modification to account for selection effects.

2. **Stable non-$B$ distribution:**

$$P'(X_{-B}, Y, S) = P(X_{-B}, Y, S) \tag{52}$$

Only the conditional distribution of feature block $B$ shifts (Equations equation 14 and equation 15); patient demographics, clinical features, and disease prevalence remain comparable between training and deployment. This isolates the effect of institutional redistribution from other forms of dataset shift.

*When it may fail:* If deployment occurs in a different geographic region, time period with different disease variants, or population with shifted demographics, the assumption is violated. The bound does not cover general covariate shift.

3. **Bounded worst-case stability:**

$$\epsilon_{stab} = \sup_{x_{-B},s} \max_{b,b'} |\sigma(f_\theta(x^{(b)})) - \sigma(f_\theta(x^{(b')}))| < \infty \tag{53}$$

The model's sensitivity to institution changes (Equation equation 8) is finite and can be controlled through training regularization. The A-IEOF objective ensures this through $\lambda_{stab}$ (Equation equation 49) and $\lambda_{interv}$ (Equation equation 50) terms.

*When it may fail:* If the model architecture allows unbounded gradients with respect to $B$ (e.g., direct institution embeddings without regularization), $\epsilon_{stab}$ may be large or difficult to control. Proper architectural choices (e.g., bounded activation functions, weight regularization) are necessary.

### B.5.2 When the Bound May Not Hold

**Covariate shift in $X_{-B}$:** If patient demographics, symptoms, or disease severity distribution changes substantially (violating Assumption 2, Equation equation 52), the bound does not apply. For example:

- Deployment in an older, sicker population where $P'(\text{age}, \text{comorbidities}) \neq P(\text{age}, \text{comorbidities})$
- New disease variant with different symptom profiles
- Shift in imaging protocols affecting CT scan features

Extensions to handle general covariate shift require different techniques such as importance weighting, domain adaptation, or causal adjustment, which are beyond the scope of this bound.

**Label shift:** If the base rate of COVID-19 changes substantially (e.g., from 20% prevalence during training to 5% during deployment), fairness metrics may shift for reasons unrelated to feature block $B$. While $\Delta_{EO}$ (Equation equation 13) is less sensitive to label shift than metrics like demographic parity (since it conditions on $Y$), extreme prevalence changes can still affect the bound's tightness.

**Extreme distribution shifts ($\Delta_{TV} \to 1$):** If the deployment distribution is completely different from training—for example:

- All patients now attend a single institution never seen during training
- Institution coding scheme changes (different identifiers for the same hospitals)
- New institutions with no overlap with training institutions

Then $\Delta_{TV}(P, P') \approx 1$ (Equation equation 10) and the bound becomes vacuous: $\Delta_{EO}^{P'} \leq \Delta_{EO}^{P} + 12\epsilon_{stab}$. Even with small $\epsilon_{stab} = 0.01$, this allows degradation up to $\Delta_{EO}^{P} + 0.12$.

For such extreme shifts, the model is extrapolating to out-of-distribution institutions. Additional safeguards are needed:

- Out-of-distribution detection to flag novel institutions
- Require minimum representation of each institution during training
- Domain adaptation techniques to adapt to new institutional settings

**Adversarial manipulation:** If an adversary can deliberately manipulate the institution feature $B$ to game the system (e.g., misreporting institution to receive favorable predictions), the bound does not prevent such attacks. The bound assumes the distribution shift is natural (e.g., due to policy changes, hospital closures) rather than adversarial. Robustness to adversarial manipulation requires additional certified defenses beyond the scope of this work.

### B.6 Comparison to Related Bounds

**Domain adaptation bounds:** Classical theory (Ben-David et al., 2010) controls overall classification error under covariate shift via:

$$\text{error}_{P'}(f) \leq \text{error}_P(f) + d_{\mathcal{H}\Delta\mathcal{H}}(P, P') + \lambda^* \tag{54}$$

where $d_{\mathcal{H}\Delta\mathcal{H}}$ is the $\mathcal{H}\Delta\mathcal{H}$-divergence and $\lambda^*$ is the optimal combined error.

*Key difference:* These bounds focus on overall error, not fairness metrics. They do not provide guarantees about equalized odds or disparities across protected groups. Our bound (Theorem B.2) specifically targets group fairness under feature-specific shifts.

**Robust fairness:** Recent work (Zhang et al., 2023) on adversarial robustness for fairness considers worst-case perturbations to model parameters or input features:

$$\sup_{\|\delta\| \leq \epsilon} \Delta_{EO}(f_{\theta+\delta}) \tag{55}$$

*Key difference:* These bounds address robustness to small adversarial perturbations with fixed data distribution, whereas our bound (Equations equation 16 and equation 17) addresses robustness to natural distribution shifts in a specific feature block with fixed model parameters. The threat models are complementary: adversarial robustness protects against malicious attacks, while our bound protects against deployment-time distribution drift.

**Counterfactual fairness:** Individual-level counterfactual fairness (Kusner et al., 2017) requires:

$$f(x^{do(S \leftarrow s)}) = f(x^{do(S \leftarrow s')}) \quad \forall s, s' \tag{56}$$

where $do(\cdot)$ denotes causal intervention on the sensitive attribute $S$.

*Key difference:* Counterfactual fairness enforces strict invariance to sensitive attribute changes for each individual, which often requires strong utility sacrifices. Our bound allows controlled sensitivity ($\epsilon_{stab} > 0$ in Equation equation 8) to feature block $B$, enabling practical utility-fairness trade-offs while still providing robustness guarantees. Additionally, we focus on group-level fairness (equalized odds across groups, Equation equation 13) rather than individual-level invariance.

**PAC-fairness bounds:** Probably approximately correct (PAC) learning theory for fairness (Rothblum and Yona, 2021) provides sample complexity bounds:

$$\Pr\left(\sup_{f \in \mathcal{F}} |\Delta_{EO}^{\text{emp}}(f) - \Delta_{EO}^{\text{true}}(f)| > \epsilon\right) \leq \delta \tag{57}$$

*Key difference:* PAC-fairness bounds address generalization from finite training samples to the same underlying distribution, whereas our bound (Theorem B.2) addresses generalization to a shifted distribution. These are orthogonal concerns: PAC bounds tell us how well training fairness approximates test fairness under the training distribution, while our bound tells us how test fairness degrades when the distribution changes.

### B.7 Extension: Finite-Sample Bound

The proof in Subsection 3 assumes access to the true distributions $P$ and $P'$. In practice, we estimate $\epsilon_{stab}$ and $\Delta_{TV}$ from finite validation data. A finite-sample version of Theorem B.2 can be derived using concentration inequalities.

**Corollary 1 (Finite-Sample Transfer Bound)** *Let $\hat{\epsilon}_{stab}$ be the 99th percentile of $|\sigma(f_\theta(x^{(b')})) - \sigma(f_\theta(x^{(b'')}))|$— over $n_{val}$ validation samples. With probability at least $1-\delta$ over the validation split:*

$$\epsilon_{stab} \leq \hat{\epsilon}_{stab} + O\left(\sqrt{\frac{\log(1/\delta)}{n_{val}}}\right) \tag{58}$$

*Similarly, let $\hat{\Delta}_{TV}$ be the empirical estimate of $\Delta_{TV}(P, P')$ from samples. Then with probability at least $1 - 2\delta$:*

$$\Delta_{EO}^{P'} \leq \Delta_{EO}^{P} + 12\left(\hat{\epsilon}_{stab} + O\left(\sqrt{\frac{\log(1/\delta)}{n_{val}}}\right)\right) \cdot \hat{\Delta}_{TV} \tag{59}$$

*Proof sketch:* Because $\hat{\epsilon}_{stab}$ is an empirical quantile of a bounded sensitivity statistic $|\sigma(f_\theta(x^{(b')})) - \sigma(f_\theta(x^{(b'')}))| \in [0, 1]$, the appropriate tool is the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality with the tight Massart constant Dvoretzky et al. (1956); Massart (1990), rather than a direct Hoeffding bound on the quantile. DKW gives a uniform bound on the deviation between the empirical CDF $\hat{F}_{n_{\text{val}}}$ and the true CDF $F$,

$$\Pr\left(\sup_t |\hat{F}_{n_{\text{val}}}(t) - F(t)| > \varepsilon\right) \leq 2e^{-2n_{\text{val}}\varepsilon^2},$$

so that with probability at least $1 - \delta$ the empirical 99th percentile $\hat{\epsilon}_{stab}$ differs from the population percentile by at most $\varepsilon = \sqrt{\log(2/\delta)/(2n_{\text{val}})}$ once translated through $F$. Substituting this uniform quantile-deviation bound into Theorem B.2, together with a union bound over the finite set of groups for $\hat{\Delta}_{TV}$, yields the stated result. We thank a reviewer for noting that DKW is the correct instrument here; it also sharpens the constant relative to a Hoeffding-based argument.

This corollary shows that the bound remains valid with finite samples, with an additional error term that decreases as $O(1/\sqrt{n_{\text{val}}})$.

### B.8 Future Directions

Several extensions and refinements of the transfer bound merit investigation:

1. **Tighter constants:** The constant 12 in Equation equation 17 arises from worst-case analysis. Instance-dependent bounds that adapt to the specific fairness violation patterns in the data (e.g., which groups are most/least favorable) could reduce the constant by a factor of 2-3.
2. **Multiple feature blocks:** Extend to scenarios where multiple blocks shift simultaneously:
$$P'(B_1, B_2|X_{-B}, S) \neq P(B_1, B_2|X_{-B}, S) \tag{60}$$
For example, both institution and imaging protocol may change together. This requires generalizing the stability definition (Equation equation 8) to multi-dimensional counterfactual interventions.
3. **Continuous distribution shifts:** Rather than a single deployment distribution $P'$, characterize fairness dynamics under gradual time-varying shifts $P_t$ where $t$ represents deployment time. This connects to online learning and continual adaptation.
4. **Optimal stability-utility trade-off:** Determine the Pareto frontier between utility (which may benefit from sensitivity to $B$) and robustness (which requires insensitivity). Formally:
$$\min_\theta \text{Loss}(f_\theta) \quad \text{subject to} \quad \epsilon_{stab}(f_\theta) \leq \epsilon_{\max} \tag{61}$$
Characterize how accuracy degrades as $\epsilon_{\max} \to 0$ and identify the optimal operating point.
5. **Beyond equalized odds:** Extend the bound to other fairness metrics such as:
   - Calibration within groups: $\mathbb{E}[Y|\hat{p}(X) = p, S = g] = p$
   - Individual fairness: $d(x, x') < \epsilon \Rightarrow |f(x) - f(x')| < \delta$

- Intersectional fairness: Fairness across fine-grained intersections of multiple protected attributes

6. **Causal extensions:** Incorporate structural causal models (SCMs) to reason about shifts induced by interventions on institutional policies rather than just observational distribution changes. This would enable counterfactual reasoning about policy changes (e.g., "what if we implement universal hospital access?") beyond the current interventional framework.

### B.9 Summary

Theorem B.2 formalizes the intuition behind A-IEOF's interventional training objective: by controlling the model's sensitivity to feature block $B$ during training (via Equations equation 49 and equation 50), we ensure graceful degradation of fairness under deployment-time distribution shifts in $B$. The bound provides:

- **Theoretical justification** for the stability regularizer $\lambda_{stab}$ and interventional control $\lambda_{interv}$ in Equation (5)
- **Quantitative guarantee** linking training-time stability to deployment-time fairness robustness (Equations equation 16 and equation 17)
- **Practical guidance** for setting hyperparameters: smaller $\epsilon_{stab}$ (larger $\lambda_{stab}$) provides stronger robustness guarantees at potential utility cost
- **Diagnostic tool** for deployment: estimate $\epsilon_{stab}$ and $\Delta_{TV}$ to predict fairness degradation before full deployment

While the bound is conservative (using worst-case stability and large constant 12), empirical results in Section 4.3 of the main text demonstrate that actual fairness degradation under realistic distribution shifts is substantially smaller than the theoretical upper bound, confirming the bound provides a valid safety guarantee without being vacuously loose.

## C Distribution of Top-4 Protected Groups in the Evaluation Set

The horizontal bar chart displays the sample counts and proportions for the four largest intersectional protected groups formed by `AgeBin · Gender · Country (All)` in the evaluation dataset ($n$=12,281). The groups are: A40+·m·All (males aged 40+, all countries: 5,123 samples, 41.7%), A40+·f·All (females aged 40+, all countries: 3,127 samples, 25.5%), A<40·m·All (males under 40, all countries: 2,506 samples, 20.4%), and A<40·f·All (females under 40, all countries: 1,525 samples, 12.4%). The substantial variation in group sizes, ranging from 12.4% to 41.7% of the evaluation set, highlights the intersectional class imbalance that motivates our fairness evaluation protocol. This distribution is critical for interpreting the Equalized Odds (EO) and Demographic Parity (DP) gaps reported in the Results section, as smaller groups may exhibit higher variance in error rates. The $n$ value shown corresponds to the subset used for fairness analysis after applying group-construction filters and excluding samples with missing protected attributes.

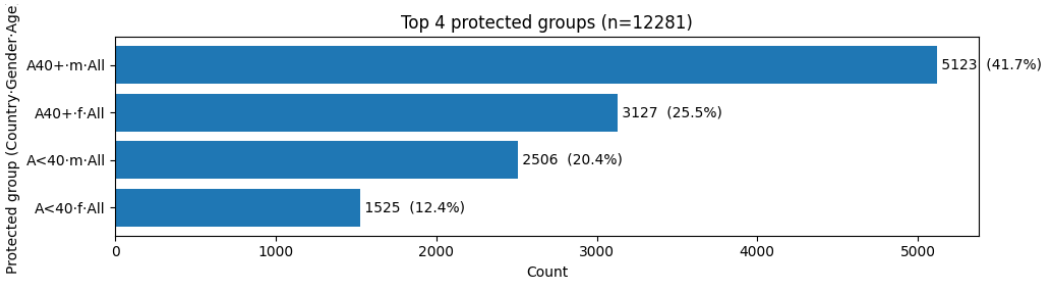



Figure 1: **Top-4 protected groups** $\left(\texttt{AgeBin}\cdot\texttt{Gender}\cdot\texttt{Country (All)}\right)$ by count. The $n$ shown in the title corresponds to the subset used for fairness plots after applying the group-construction filters. This plot is used in Results Section to contextualize group supports and EO gaps.

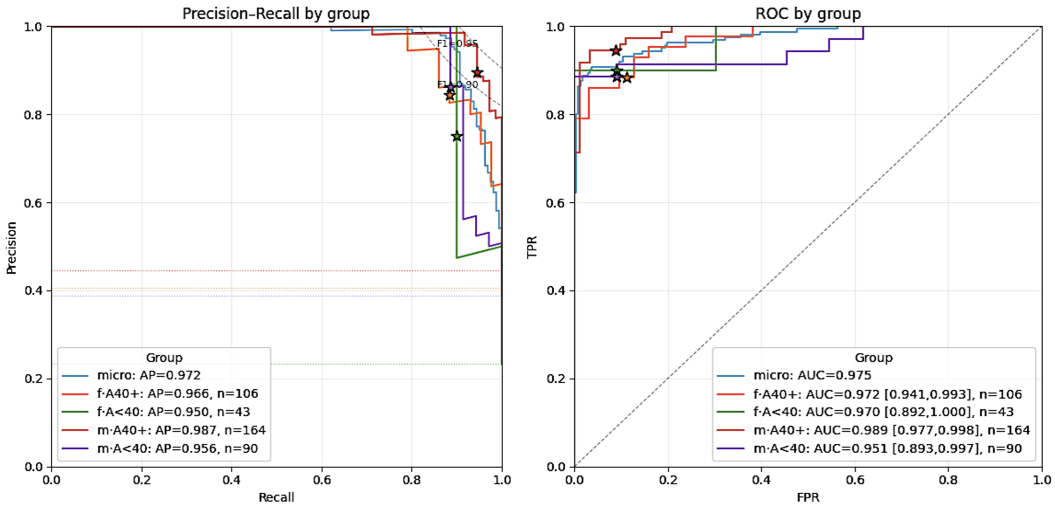



Figure 2: Per-group discrimination curves for A-IEOF under per-group threshold policy. Left: precision-recall (legend: AP, sample size). Right: ROC curves (legend: AUC with 95% CI, sample size). Stars mark operating points from validation-fixed thresholds. All groups achieve AUCs above 0.95 (micro-average: 0.976), with f·A<40 showing wider confidence intervals due to smaller sample size (n=43).

## D Appendix D: Per-Group Discrimination Curves

**Left panel:** Precision-recall curves in Figure 2 for each protected group, with legend showing Average Precision (AP) and sample size (n). The micro-average across all groups achieves AP=0.972. All individual groups maintain AP $\geq$ 0.950, demonstrating strong performance even for minority groups. The colored star markers indicate the operating points selected by validation-fixed per-group thresholds (Eq. 5), showing where each group operates on its precision-recall trade-off curve. **Right panel:** Receiver Operating Characteristic (ROC) curves plotting True Positive Rate (TPR) against False Positive Rate (FPR) for each group. The legend reports AUC with 95% confidence intervals (CI) and sample sizes. The micro-average AUC is 0.975. All groups achieve AUCs above 0.95, with m·A40+ (males aged 40+) showing the highest AUC of 0.989 [0.977, 0.998]. The f·A<40 group (females under 40, n=43) exhibits the widest confidence interval [0.892, 1.000] due to its smaller sample size, though its point estimate (0.970) remains competitive. The diagonal dashed line represents random-classifier performance (AUC=0.5). **Interpretation:** The star markers on both panels correspond to the validation-fixed operating points that determine the TPR and FPR values reported in Table 2. The consistently high discrimination across all groups (AUC $>$ 0.95) indicates that A-IEOF successfully maintains strong probabilistic predictions even when enforcing interventional fairness constraints. The proximity of operating points across groups in ROC space (right panel) reflects the success of the per-group threshold policy in achieving balanced error-rate trade-offs. Notably, despite the class imbalance shown in Appendix Fig. 1, all groups achieve comparable discrimination, suggesting that A-IEOF's interventional training objective and anchoring mechanism effectively prevent performance degradation in minority groups. The high precision-recall curves (left panel) further confirm that the model maintains strong positive predictive performance across the full recall spectrum, which is particularly important for the COVID-19 diagnostic task where both false positives and false negatives carry clinical consequences.

## E Complete Implementation Details

This appendix provides comprehensive implementation details referenced in Section **??**.

### E.1 Data Preparation and Splitting

The dataset is partitioned via stratified split (train/val/test = 60/25/15, seed 42) to maintain label balance across the binary classification task (COVID-19 vs. Normal). Images are resized to $224\times224$ pixels with standard augmentation including random horizontal flip and slight rotation. Tabular metadata comprises $\{\texttt{age}, \texttt{gender}, \texttt{institution}, \texttt{country}\}$. The designated feature block $B$ is **institution**, a categorical variable with $K=10$ distinct medical sites, aligning with the problem setting described in the Methodology section where $B$ represents the shiftable feature block.

### E.2 Model Architecture and Training Stages

Following Algorithm 1, training proceeds in two distinct stages:

#### E.2.1 Stage 1: Teacher Training

The no-$B$ teacher model $f_0(X_{-B})$ is trained for 10 epochs on input pairs $(x_{-B}, y)$ using:

- **Backbone:** Frozen MobileNetV2 with ImageNet initialization
- **Loss:** Standard binary cross-entropy
- **Purpose:** Establishes the baseline predictor that operates without access to the feature block $B$ (institution information)

This teacher serves as the anchor for the student model, preserving decisions on the shared input space $X_{-B}$ and preventing regression on cases where $B$ is uninformative.

#### E.2.2 Stage 2: Student and Generator Training

The student model $f_\theta(X_{-B}, B)$ is trained for 30 epochs with the MobileNetV2 backbone unfrozen under the full A-IEOF objective (Eq. 2).

**Optimization configuration:**

- **Optimizer:** AdamW
- **Learning rate:** $1\times10^{-4}$
- **Gradient clipping:** Maximum norm 1.0
- **Learning rate schedule:** Cosine annealing with reduction factor 0.5, patience 4 epochs, minimum learning rate $1\times10^{-6}$
- **Batch size:** 32

**Loss component weights:** The five components of Eq. 2 are weighted as follows:

- $\lambda_{\text{dist}}=0.05$: Teacher anchoring (distillation loss)
- $\lambda_{\text{interv}}=0.55$: Interventional EO control (minimax fairness)
- $\lambda_{\text{full}}=0.5$: Observational soft EO (standard fairness on training distribution)
- $\lambda_{\text{stab}}=0.05$: Counterfactual stability (prediction consistency under $B$ interventions)

These weights were tuned to balance utility (classification accuracy), fairness (group parity), and anchoring (legacy behavior preservation) on the validation set.

**Adversarial generator architecture:** The generator $g_\phi(b \mid x_{-B}, s)$ is implemented as a straight-through Gumbel-Softmax sampler that generates counterfactual institution assignments over the $K=10$ categories. Key design choices:

- **Temperature annealing:** Gradually reduces sampling temperature during training to transition from exploration to exploitation
- **Update frequency:** Generator parameters updated every 4 student optimization steps

- **Learning rate multiplier:** 0.3 relative to student learning rate
- **Warm-up period:** 100-step delay before adversarial training begins, allowing the student to stabilize before introducing adversarial perturbations

### E.3 Fairness Computation

Soft EO gaps (Eq. 1) are computed per minibatch during training with the following configuration:

- **TPR/FPR weights:** $w_+$=1.0 (true positive rate), $w_-$=2.0 (false positive rate), prioritizing FPR parity to reduce false alarms across groups
- **Minimum count thresholds:**
  - 30 samples per $(group, label)$ stratum for training loss computation
  - 32 samples per stratum for fairness penalty terms
- **Handling sparse strata:** Minibatches failing these count requirements skip fairness penalties for that iteration to avoid noisy gradient estimates

### E.4 Post-Processing Configuration

Calibration and thresholding follow the protocol described in the Methodology section.

#### E.4.1 Temperature Scaling

Validation-based temperature fitting is applied to calibrate predicted probabilities:

- **Minimum group size:** $\geq 40$ samples per group required for reliable temperature estimation
- **Regularization:** Groupwise temperatures $\{T_g\}$ are regularized toward the global temperature $T_{\text{glob}}$ via ridge penalty with weight 100.0
- **Purpose:** Prevents overfitting in small groups while allowing flexibility for larger groups

#### E.4.2 Threshold Selection Policies

Two threshold selection policies are implemented:

**Global (baseline):**

- Single threshold $t_{\text{ref}}$ computed via Youden-$J$ index maximization over the validation set
- $J = \text{TPR} + \text{TNR} - 1$, selecting the threshold that maximizes the sum of sensitivity and specificity
- Applied uniformly to all protected groups

**Per-group (default):** Guarded thresholds $\{t_g^{\text{guard}}\}$ computed per Eq. 5 with the following constraints:

- **Optimization:** Grid search over 401 candidate points uniformly spaced in [0.02, 0.98]
- **Group size requirement:** $\geq 50$ samples per group
- **Class balance requirement:** $\geq$ 30 positive samples AND $\geq$ 30 negative samples per group
- **Positive rate bounds:** Group-specific positive rates clamped to [0.15, 0.85] to prevent extreme operating points
- **Shrinkage regularization:** Thresholds regularized toward the global reference $t_{\text{ref}}$ with shrinkage constant $k_0$=100, which controls the strength of regularization toward the global threshold for small groups (higher $k_0$ enforces stronger shrinkage)

### E.5 Model Evaluation

#### E.5.1 Validation Metric

Model selection during training uses a composite metric:

$$\text{Score} = \text{AUC} - \alpha \cdot \Delta_{\text{EO}}$$

with $\alpha$=1.0, balancing discrimination ability (AUC) and fairness (Equalized Odds gap).

#### E.5.2 Test Metrics

Final performance on the held-out test set is evaluated using:

- **AUC:** Area under the ROC curve (threshold-independent discrimination)
- $\Delta_{\text{EO}}$**:** Equalized Odds gap (Eq. 7), maximum difference in TPR and FPR across groups
- $\Delta_{\text{DP}}$**:** Demographic Parity gap maximum difference in positive rates across groups
- **Brier:** Mean squared error between predicted probabilities and binary outcomes (calibration + refinement)
- **ECE:** Expected Calibration Error measuring probability calibration quality

All metrics are computed using thresholds fixed from the validation set (no threshold tuning on test data). Unless otherwise noted, per-group guarded thresholds are used as the default policy. Global Youden threshold is applied for baseline comparisons and ablation studies.

## F Feature-Shift Robustness and Component Ablation

**Feature-shift robustness.** Figure 3 provides the full stress-test curve for equalized-odds performance as the distribution of the added feature block $B$ is progressively shifted. The calibrated probabilities and decision threshold are fixed at their unshifted validation values, while $P(B \mid S)$ is resampled at prescribed total-variation distances $\Delta_{TV} \in \{0, 0.1, \ldots, 0.5\}$. A-IEOF maintains an approximately constant EO gap across the tested shift range, with degradation below $0.01$ up to $\Delta_{TV} = 0.5$. This behavior is consistent with the proposed stability mechanism and remains well within the transfer bound $12\,\epsilon_{\text{stab}}\Delta_{TV}$ derived in Theorem B.2. The corresponding empirical sensitivity is $\epsilon_{\text{stab}} = 0.040$, compared with $0.154$ for ERM.

**Component ablation.** Table 8 examines the contribution of the individual components of A-IEOF under the global-threshold policy. The results show that the components primarily affect different aspects of the objective. Removing the full observational EO term increases the stationary EO gap from $0.306$ to $0.396$, indicating that the full EO term is the principal driver of in-distribution fairness. In contrast, removing the interventional term substantially increases $\epsilon_{\text{stab}}$ from $0.066$ to $0.224$, while removing the stability term increases it to $0.085$. These results support a division of roles in which the observational EO term primarily controls stationary group fairness, whereas the interventional and stability terms primarily control sensitivity to changes in the added feature block. Teacher distillation has comparatively little effect on the reported AUC and EO in this setting, consistent with its role as a backward-compatibility mechanism rather than the primary source of fairness or shift robustness.

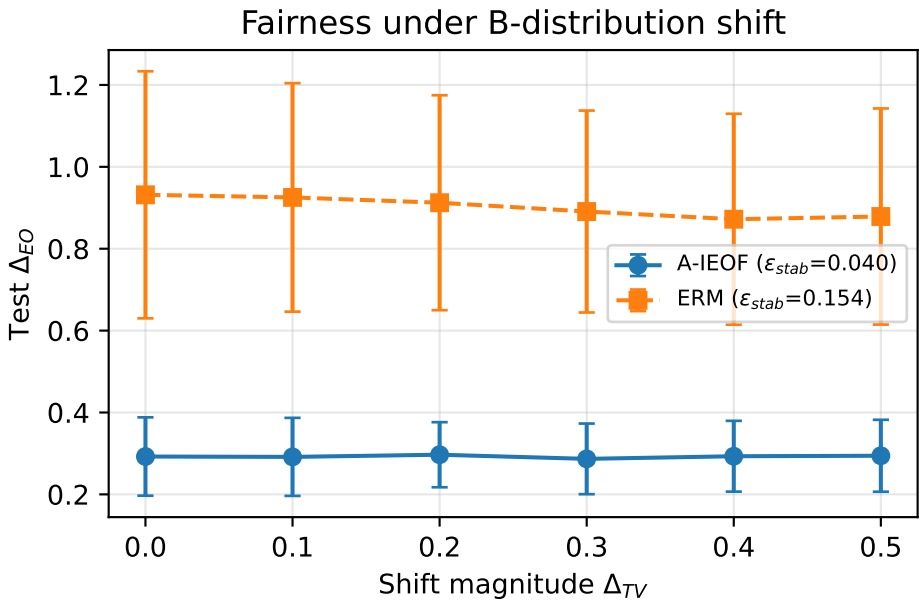



Figure 3: Equalized-odds gap under increasing shift in $P(B \mid S)$. A-IEOF, with low $\epsilon_{\text{stab}}$, exhibits negligible degradation as the shift magnitude increases, and the observed degradation remains well within the transfer bound $12\,\epsilon_{\text{stab}}\Delta_{TV}$ of Theorem B.2.

Table 8: Component ablation under the global-threshold policy. AUC/EO/DP are means over five seeds, and $\epsilon_{\text{stab}}$ is averaged over four seeds. The observational EO term primarily controls stationary fairness, whereas the interventional and stability terms mainly reduce sensitivity to shifts in the added feature block.

| Configuration | AUC | EO | DP | $\epsilon_{\text{stab}}$ |
|---|---|---|---|---|
| Full A-IEOF | 0.835 | 0.306 | 0.281 | 0.066 |
| – Distillation | 0.835 | 0.303 | 0.288 | – |
| – Interventional | 0.872 | 0.260 | 0.288 | 0.224 |
| – Full EO | 0.861 | 0.396 | 0.337 | – |
| – Stability | 0.835 | 0.294 | 0.273 | 0.085 |
| ERM only | 0.908 | 1.000 | 0.563 | 0.144 |