IS INVARIANCE ALL YOU NEED FOR ALGORITHMIC FAIRNESS? REMOVING DEMOGRAPHIC INFORMATION CAN CREATE NEW BIAS

Anonymous authors
Paper under double-blind review

ABSTRACT

Encoded demographic information in internal model representations is a commonly assumed risk factor for algorithmic bias, with demographic representation invariance often being touted as the ideal state. However, while demographic shortcut learning is a genuine threat, some degree of encoding is necessary when demographics correlate with target labels. Here, we show, mathematically and empirically, that enforcing demographic invariance can actually *hamper* bias mitigation and even create new biases. We distinguish marginal from class-conditional representation invariance, and show that they imply the standard group fairness notions of demographic parity and equalized odds, respectively. We evaluate the effects on predictive performance and fairness of enforcing both invariance types, both theoretically and empirically across five tabular and two chest X-ray imaging datasets. Our findings support our mathematical argument that demographic representation invariance is neither *desirable* nor *sufficient* for fairness.

Full code, configs, and random seeds are made available anonymously.[1]

1 INTRODUCTION

Demographic information can be recovered from latent representations produced by medical imaging models, even when these models are neither explicitly trained to predict demographic attributes, nor trained using them (Glocker et al., 2023; Yang et al., 2024; Duffy et al., 2022; Gichoya et al., 2022; Yi et al., 2021; Lotter, 2024; Quarta et al., 2026). This raises the valid concern that a patient's group membership, once encoded in the representation, may be exploited by the model, e.g. as shortcuts used instead of task-relevant features (Stanley et al., 2025b; Lin et al., 2024; Brown et al., 2023; Yang et al., 2024), or to reproduce historical biases encoded in systematically biased labels (Akintande et al., 2025; Parikh et al., 2026).

In this paper, we argue that this intuitive concern is a misunderstanding. We show, mathematically and empirically, that *demographic decodability*: the ability to recover group membership from representation, by itself, does not explain unfairness or performance disparities between groups (Jones et al., 2023; Stanley et al., 2025a; Jones et al., 2025). It is entirely possible for a representation to only encode task-specific features, yet still have high demographic decodability, simply because the task-specific features correlate with group membership, e.g., when disease symptoms or prevalence differ between groups. Removing all demographic information therefore risks removing task-specific features associated with group membership and thus harming performance. In the context of different true prevalences, enforcing demographic invariance results in systematic differences in error rates between the groups, harming equalized odds fairness.

These mathematical arguments stand in contrast to previous studies reporting positive empirical effects of applying alignment methods such as adversarial learning (Edwards & Storkey, 2016; Zhang et al., 2018) or distribution matching (Louizos et al., 2016) to minimize internal encoding of demographic information (Zemel et al., 2013; Creager et al., 2019; Zhao & Gordon, 2022; Wang et al., 2024; Rashid et al., 2026; Sadri et al., 2025; Sarridis et al., 2024). The theoretical guarantees and fundamental limitations of these methods have, however, received little attention (Zhao et al., 2022).

[1] Available at: https://anonymous.4open.science/r/invariance-project-931C/README.md

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

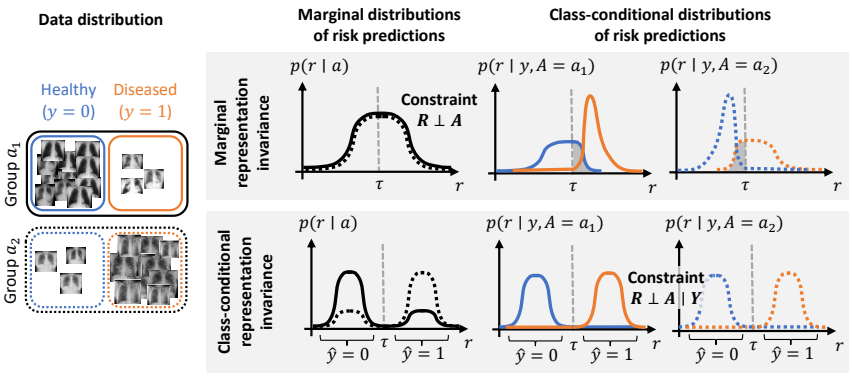


Figure 1: **Two forms of representation invariance and the effects of enforcing those** in a case study predicting disease for two groups (a_1 and a_2) with different prevalence. **Top:** $Z \perp A$ gives identical marginal risk distributions across groups. Under unequal prevalence, the class-conditional risk distributions must differ, giving different error rates across groups. **Bottom:** $Z \perp A \mid Y$ gives equal error rates, but still has drawbacks: it is incompatible with calibration by groups and requires equalizing away potentially important differences between unknown disease subtypes.

Our experiments support Cerrato et al. (2024) showing that removing demographic information is surprisingly hard, and that most well-known methods fail at accurately removing demographic information from internal representations. This suggests that the regularizing effect of demographic decoding found in previous work is likely caused by failure to actually remove demographic information as intended.

We show that marginal representation invariance, which is the intuitive and common target of such studies, implies *demographic parity*, also known as the independence fairness criterion. This is important, as demographic parity is well known to be incompatible with other fairness criteria often deemed more appropriate in real-life applications (Barocas et al., 2023). We distinguish marginal representation invariance, in practice the most commonly considered form of invariance, from *class-conditional* representation invariance, which we show implies the well-known *equalized odds*, or *separation*, fairness criterion (Barocas et al., 2023).

Our experiments compare a range of different methods for enforcing representation invariance, and we use the one that actually *does* work. We clearly demonstrate how marginal invariance leads to failure in satisfying equalized odds fairness. Our experiments *also* show that while class-conditional invariance enforces equalized odds well, it also comes with the known risk of leveling performance down (Mittelstadt et al., 2023) for the best groups in order to obtain fairness.

2 Problem Setting

We denote random variables by uppercase letters (X, Y, Z) and their realizations by lowercase letters (x, y, z). Consider a supervised prediction task with input $X \in \mathcal{X}$, outcome $R \in \mathcal{R}$ sometimes thresholded to a categorical $Y \in \mathcal{Y}$, and sensitive attribute $A \in \mathcal{A} = \{a_1, \ldots, a_m\}$ with $P(A = a) > 0$ for all a. The input X can be tabular features, an image, or any other data type. The outcome can be binary ($Y \in \{0, 1\}$), multiclass ($Y \in \{0, 1 \ldots, K-1\}$), a regressed score $R \in \mathcal{R}$, or even a more complex output such as a segmentation mask. The attributes A identify groups such as sex, ethnicity, or age group, and may be multi-valued or intersectional.

A deterministic feature extractor $g \colon \mathcal{X} \to \mathcal{Z}$ produces a representation $Z = g(X)$, which is mapped to predictions $R \in \mathcal{R}$ by a second deterministic component (e.g., a classifier head) $h \colon \mathcal{Z} \to \mathcal{R}$, producing a risk/score vector $R = h(Z)$. If required, R may be discretized to obtain binary or categorical outcome labels $\hat{Y}$. Neither g, h, nor τ take A as input.

We consider a model to *encode*[2] group membership A in its latent representation Z if group membership is *decodable* from Z above chance (i.e. $Z \not\perp A$). We use *decodability* as a measurement (or a proxy) of invariance throughout. Notice that decodability depends on the evaluation dataset $\mathcal{D}$ and is not a property of the model alone.

3 What does Representation Invariance Guarantee?

While the propositions in this section are not new individually, the way we connect them with two forms of representation invariance (illustrated in Fig. 1) and derive their consequences for fairness and performance is new. This connection is important to avoid common misunderstandings that may increase bias and decrease performance. We cite prior work where proofs are available. Sec. 4 then introduces an experimental setup that allows us to empirically test these propositions in Sec. 5.

3.1 Marginal Representation Invariance

Marginal representation invariance asks that the distribution of Z be identical across groups:

$$p(z \mid a) = p(z) \quad \forall a \qquad \iff \qquad Z \perp A \tag{1}$$

Under this condition, the learned representation Z is statistically independent of A, denoted by $Z \perp A$: the model does not *encode* group membership on this dataset.

Proposition 1. *Marginal representation invariance implies demographic parity/independence.*

Proof. If the risk prediction R and the decision $\hat{Y} = \mathbb{1}[R \geq \tau]$ depend only on Z, then both predictions inherit this independence of A from Z:

$$Z \perp A \iff p(z \mid a) = p(z)\, \forall a \implies p(r \mid a) = p(r), \quad p(\hat{y} \mid a) = p(\hat{y}), \quad \forall a. \tag{2}$$

That is, if $Z \perp A$, then we get $R \perp A$ and $\hat{Y} \perp A$: the distributions of predicted risk scores and predicted decisions are identical across sensitive groups, which implies the fairness criterion known as *demographic parity*, or *independence*. □

For a binary attribute $A \in \{0, 1\}$ this gives $P(\hat{Y} = 1 \mid A = 0) = P(\hat{Y} = 1 \mid A = 1)$, whose violation is measured by the demographic parity gap

$$\Delta_{\mathrm{DP}} = \Big| P(\hat{Y} = 1 \mid A = 0) - P(\hat{Y} = 1 \mid A = 1) \Big|. \tag{3}$$

Therefore, exact marginal representation invariance implies $\Delta_{\mathrm{DP}} = 0$ if the classification head and the decision threshold do not use A (Zemel et al., 2013; Zhao et al., 2022; Madras et al., 2018). Its finite-sample estimate may differ from zero ($\Delta_{\mathrm{DP}} \approx 0$) due to sampling variability.

Unequal prevalence across groups causes problems for marginal invariance

Outcomes Y often vary across demographic groups: Income differs across age groups or sex; crime rates differ across populations due to historical and institutional factors; disease prevalence depends on e.g. sex and age, with breast cancer, which affects some but very few men, as an extreme example.

Proposition 2. *Under unequal group prevalences, marginal invariance creates an upper bound on predictive performance.*

Proof. While marginal invariance requires $p(\hat{y} \mid A = 0) = p(\hat{y} \mid A = 1)$, an accurate classifier should approximately reproduce the group-specific outcome distributions $p(\hat{y} \mid a) \approx p(y \mid a)$. However, when $p(Y \mid A = 0) \neq p(Y \mid A = 1)$, these requirements, i.e., demographic parity and perfect (desired) prediction, cannot both hold. Marginal representation invariance must thus introduce prediction errors in at least one group. □

[2]We use the term *encoding* for consistency with the literature. Above-chance prediction of A from Z does not necessarily mean the model explicitly represents demographic identity or uses it as a shortcut.

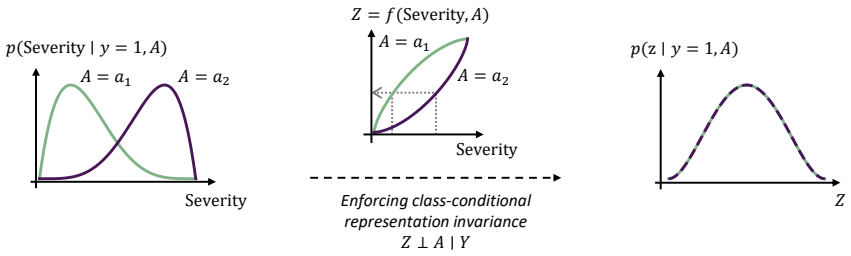


Figure 2: **Effects of enforcing Class-conditional representation invariance in the presence of intra-class variation.** When the within-class distribution of a clinically relevant factor (here disease severity among $Y{=}1$ patients) differs between groups, enforcing $Z \perp A \mid Y$ requires mapping patients with differing disease severities to the same latent representation, depending on which group a patient belongs to (center). The aligned representation then hides a real difference in disease presentation, and the same z no longer means the same severity in both groups (see also Lohaus et al. (2022)).

For a binary outcome, the group-conditional representation distribution is

$$\begin{aligned} p(z \mid a) &= \sum_{y \in \{0,1\}} p(z \mid y, a)\, p(y \mid a) \\ &= \pi_a\, p(z \mid Y = 1, A = a) + (1 - \pi_a)\, p(z \mid Y = 0, A = a), \end{aligned} \tag{4}$$

where $\pi_a = P(Y = 1 \mid A = a)$ denotes the positive-label prevalence in group a. If the within-class distributions are shared across groups, $p(z \mid y, a) = p(z \mid y)$, and $Z \not\perp Y$: the difference in group-wise $p(z \mid a)$ originates in π_a.

For a binary prediction task, let $e_a = P(\hat{Y} \neq Y \mid A = a)$ denote the prediction error for group a. Any predictor satisfying demographic parity obeys $e_0 + e_1 \geq |\pi_0 - \pi_1|$ (McNamara et al., 2019; Zhao et al., 2022). For multi-class tasks, an iterative division into binary one-versus-rest prediction problems can be used to derive a similar bound. The bound constrains the combined error across groups; the entire cost may fall on one group.

3.2 Class-Conditional Representation Invariance

Class-conditional representation invariance instead aligns the representation distributions separately within each outcome class:

$$p(z \mid y, a) = p(z \mid y) \quad \forall a, y \quad \iff \quad Z \perp A \mid Y. \tag{5}$$

This condition allows the marginal representation distributions $p(z \mid a)$ to differ when the groups have different outcome prevalences. Thus, class-conditional representation invariance does not imply marginal representation invariance ($Z \perp A \mid Y \not\Longrightarrow Z \perp A$).

Proposition 3. *Class-conditional representation invariance implies the separation/equalized odds fairness criterion.*

Proof. If the risk prediction R and thresholded prediction $\hat{Y}$ depend only on Z, then they inherit the conditional independence of Z implying $R \perp A \mid Y$ and $\hat{Y} \perp A \mid Y$. This property is known as *separation* (Hardt et al., 2016; Madras et al., 2018); separation on the classification level ($\hat{Y}$) in the case of binary classification is also known as *equalized odds*. □

We note that when $Y = 1$ we get equal TPRs (*equal opportunity* fairness criterion), whereas with the additional $Y = 0$ we also get equal FPRs. Therefore, class-conditional representation invariance implies $\Delta_{\mathrm{EOpp}} = 0$ and $\Delta_{\mathrm{EO}} = 0$.

3.3 Invariance-Fairness Incompatibilities

Combining Props. 1 and 3 with known fairness incompatibility theorems, we obtain the following invariance-fairness incomparability results for classification:

Corollary 1. *Assume that R and Y are not independent of A. In this case: (i) Assume that Y is binary; then marginal representation invariance is incompatible with the separation fairness criterion. If group-wise prevalences differ, representation invariance is incompatible with separation even for multiple classes. (ii) Marginal representation invariance is incompatible with the sufficiency fairness criterion, as well as with calibration by groups. (iii) Class-conditional representation invariance is incompatible with the independence fairness criterion. (iv) Assume that all events in the joint distribution of (A, R, Y) have positive probability. Then, class-conditional representation invariance is incompatible with the sufficiency fairness criterion, as well as with calibration by groups.*

Proof. (i) If marginal invariance holds, the independence criterion holds by Prop. 1. For binary Y, (i) follows from (Barocas et al., 2023, Ch. 3, Prop. 3). If prevalences depend on A and marginal independence holds, then constant predictive rates lead to differences in either TPR, FPR, or both, across those groups where prevalences differ. As a consequence, separation does not hold.

(ii) If marginal invariance holds, the independence criterion holds by Prop. 1, and thus sufficiency cannot hold by (Barocas et al., 2023, Ch. 3, Prop. 2). If calibration by groups were to hold, then sufficiency would hold as well, which it cannot; this proves the claim for calibration by groups.

(iii) If class-conditional invariance holds, then equalized odds holds by Prop. 3. Thus, independence cannot hold by (Barocas et al., 2023, Ch. 3, Prop. 3).

(iv) If class-conditional invariance holds, then equalized odds holds by Prop. 3. Thus, sufficiency cannot hold by (Barocas et al., 2023, Ch. 3, Prop. 4). If calibration by groups were to hold, then sufficiency would hold as well, which it cannot; this proves the claim for calibration by groups. □

These results all follow from known incompatibility results as stated in (Barocas et al., 2023); we note that even stronger or more detailed fairness incompatibility results from the literature will generalize into invariance-fairness incompatibility results following the same line of reasoning.

We also note that groupwise prevalence differences, which are used as a running challenging example in this paper, are a special case of A and Y not being independent, which is an assumption for all these incompatibility results. Fig. 2 shows a further concern about class-conditional invariance conflicting with within-class differences in disease severity across groups.

3.4 What happens under prevalence shifts?

Consider a fixed model evaluated on two populations P and Q, with unchanged within-class distribution (i.e., $P(X \mid Y, A) = Q(X \mid Y, A)$) but a change in group-specific prevalence ($P(Y \mid A) \neq Q(Y \mid A)$). This could, e.g., happen if a disease over time becomes more prevalent in one group due to, e.g., updated diagnostic criteria that fit the group better, or behavioral changes that affect one group but not the other (e.g., a new contraceptive pill entering the market affecting women, but not men). It also commonly occurs when intentionally up- or downsampling certain classes in training and evaluation datasets. In such settings, we show:

Proposition 4. *Class-conditional representation invariance is preserved under group-specific prevalence shifts, while marginal representation invariance is not.*

Proof. As the model is fixed, its conditional representation distribution remains unchanged ($P(Z \mid Y, A) = Q(Z \mid Y, A)$). Therefore, if representations across groups for each outcome Y are aligned for P, they also remain aligned for Q (Asiaee & Aryan, 2026):

$$Z \perp_P A \mid Y \quad \Longrightarrow \quad Z \perp_Q A \mid Y.$$

For binary outcomes, let $\pi_a^Q = Q(Y = 1 \mid A = a)$. The representation distribution for Q becomes

$$Q(Z \mid A = a) = \pi_a^Q P(Z \mid Y = 1, A = a) + (1 - \pi_a^Q) P(Z \mid Y = 0, A = a).$$

Changing π also changes the proportions of positive and negative representations in each group. Their overall distributions can therefore become different, even when they were aligned under P:

$$Z \perp_P A \quad \not\Longrightarrow \quad Z \perp_Q A.$$

This finalizes our proof of both claims. □

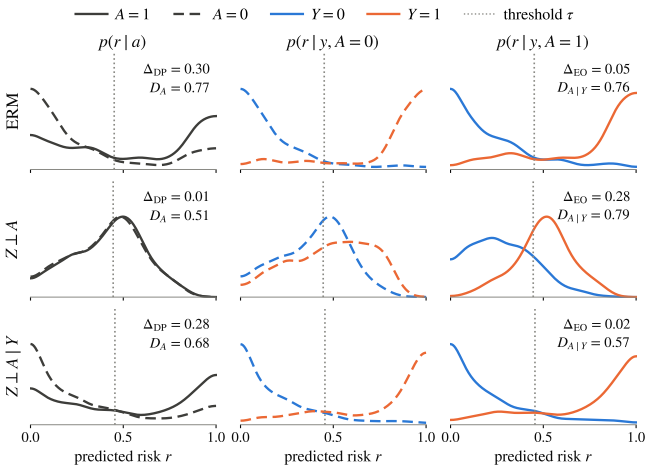


Figure 3: **Fig. 1 estimated from data** (Dutch, shift-50, seed 42). Rows: ERM, $Z \perp A$, $Z \perp A \mid Y$. Left: marginal risk densities $p(r \mid a)$ $A \in \{0, 1\}$, with Δ_{DP} and D_A. Middle and right: class-conditional densities $p(r \mid y, a)$ within each group, with Δ_{EO} and $D_{A|Y}$. Dotted: Youden threshold. Marginal alignment makes the left column coincide by pulling the within-class curves apart across groups (P_1); class-conditional alignment makes the middle and right columns coincide and leaves the left column at the base-rate gap (P_3).

These results show that **marginal representation invariance depends on the population it is evaluated on** (Zhao & Gordon, 2022). Class-conditional representation alignment, on the other hand, is preserved under group-specific prevalence shift.

4 Experimental Design

Datasets and prevalence conditions: We examine five tabular benchmarks (Adult, Dutch, Communities & Crime, Taiwan credit, COMPAS) (Kohavi, 1996; Calders & Verwer, 2010; Redmond & Baveja, 2002; Yeh & Lien, 2009) and two CheXpert tasks (edema, cardiomegaly) (Irvin et al., 2019); depending on the dataset, sensitive attributes are sex, race, or racial composition. Tabular datasets are used to experiment with different group prevalence disparities: natural, equalized, and removal of 50% or 75% of positive examples from the lower-prevalence group (*shift-50, shift-75*), resulting in substantial prevalence disparities. In CheXpert, we simply remove 50% of female-positive cases and compare with natural prevalence. Each condition is a fixed subsample, so every train/test split shares the same prevalence under the given condition (See App. 6).

Methods: All methods are compared against empirical risk minimization (ERM) baselines. On tabular data, we use categorical Fair Normalizing Flows (FNFs) (Balunović et al., 2022) sweeping over $\gamma \in \{0, .02, .1, .2, .5, .9, 1\}$, two MADE hidden widths $\{50, 100\}$, and two feature sets, across four prevalence conditions, and three seeds $\{42, 123, 456\}$ to yield $5 \times 4 \times 3 \times (7 \times 2 \times 2 + 2) = 1{,}800$ runs. For CheXpert, we compare the invariant methods (FNFs, MMD regularizer, adversarial) across two tasks, with the same three seeds, $\gamma \in \{.05, 0.50, 0.95\}$, yielding $2 \times 3 \times (3 \times 3 + 1) = 60$ runs. Data splits are 60/10/10/20% (tabular) and 60/5/5/10/20% (CheXpert) for training, validation, probe-validation, and test partitions; thresholds are selected on the validation set.

Evaluation: *Decodability* D_A is measured with independent probes (logistic regression, gradient-boosted tree, 3 MLPs) trained on held-out probe-validation representations. We evaluate predictive performance using the AUROC score. For fairness evaluation, we report group differences: Δ_{DP}, Δ_{EO} and Δ_{EOpp}, along with ΔPPV and ΔNPV. *Calibration* is measured by calibration bias ($\Delta_{\mathrm{bias}} = |(\bar{y}_1 - \bar{p}_1) - (\bar{y}_0 - \bar{p}_0)|$) and group-wise recalibration intercepts $\Delta_{\mathrm{cal}} = |a_1 - a_0|$ (Cox, 1958) (unlike ECE, both stay unbiased under unequal group sizes (Petersen et al., 2023)). All results are reported as mean $\pm$ standard deviation over training seeds. Paired bootstrapping uses the same samples for each method and is matched to the ERM baseline: 1,000 resamples for tabular, and 2,000 resamples for CheXpert. Additional results for robustness, including 95% percentile intervals, Bonferroni-adjusted confidence intervals for probe-AUROC, sensitivity analysis over γ, and calibration-bin tests, are in Appendixes A and B.

5 Results

Table 1 summarizes the *shift-50* case; the rest, including full-hyperparameter sweeps and per-seed scores, are in the Appendix A.4. We find that D_A drops for every alignment method, relative to

		Utility	Decodability of A		Indep.	Separation	Sufficiency	Group calibration
	Method	AUROC	D_A	$D_{A\|Y}$	Δ_{DP}	Δ_{EO}	Δ_{PPV} / Δ_{NPV}	Δ_{bias} / Δ_{cal}
Adult	ERM	0.89	0.93	0.99	0.40	0.28	0.18 / 0.05	0.02 / 0.32
	$Z \perp A$	0.79	0.58	0.85	**0.04** ✓	0.40	0.49 / 0.04	0.23 / 2.12
	$Z \perp A \mid Y$	0.86	0.62	0.55	0.17	**0.08** ✓	0.40 / 0.07	0.15 / 1.80
Dutch	ERM	0.90	0.77	0.77	0.30	0.05	0.27 / 0.28	0.22 / 2.00
	$Z \perp A$	0.71	0.52	0.79	**0.01** ✓	0.31	0.63 / 0.21	0.44 / 2.32
	$Z \perp A \mid Y$	0.89	0.68	0.56	0.27	**0.03** ✓	0.34 / 0.27	0.23 / 2.05
C & C	ERM	0.85	0.79	0.66	0.43	0.20	0.50 / 0.30	0.23 / 2.10
	$Z \perp A$	0.67	0.55	0.73	**0.05** ✓	0.28	0.71 / 0.31	0.55 / 2.87
	$Z \perp A \mid Y$	0.85	0.74	0.57	0.40	**0.28**	0.53 / 0.37	0.29 / 2.34
Taiwan	ERM	0.75	0.63	0.62	0.07	0.08	0.19 / 0.06	0.08 / 0.68
	$Z \perp A$	0.68	0.52	0.69	**0.01** ✓	0.14	0.28 / 0.04	0.11 / 0.86
	$Z \perp A \mid Y$	0.75	0.54	0.53	0.04	**0.01** ✓	0.18 / 0.07	0.09 / 0.79
COMP.	ERM	0.71	0.68	0.65	0.21	0.15	0.23 / 0.15	0.13 / 0.78
	$Z \perp A$	0.64	0.56	0.63	**0.02** ✓	0.11	0.33 / 0.18	0.24 / 1.19
	$Z \perp A \mid Y$	0.69	0.56	0.54	0.10	**0.04** ✓	0.27 / 0.18	0.20 / 1.09
Edema	ERM	0.84	0.98	0.98	0.15	0.10	0.08 / 0.02	0.01 / 0.09
	$Z \perp A$	0.83	0.62	0.63	**0.01** ✓	0.05	0.16 / 0.06	0.10 / 0.80
	$Z \perp A \mid Y$	0.83	0.60	0.59	0.05	**0.02** ✓	0.13 / 0.05	0.07 / 0.58
Cardio.	ERM	0.85	0.98	0.98	0.13	0.09	0.09 / 0.02	0.01 / 0.19
	$Z \perp A$	0.84	0.64	0.65	**0.01** ✓	0.05	0.17 / 0.03	0.07 / 0.96
	$Z \perp A \mid Y$	0.84	0.62	0.62	0.04	**0.01** ✓	0.15 / 0.03	0.05 / 0.74

Table 1: **Fairness criteria under the *shift-50*** on test split, mean over 3 seeds. $Z \perp A$ is marginal FNF at $\gamma = 1$ (CheXpert: $\gamma = 0.95$; strongest setting), $Z \perp A \mid Y$ class-conditional FNF at the same γ, both with the feature set and width selected by the lowest probe AUROC. D_A and $D_{A|Y}$ are the maximum probe AUROC for recovering A from Z marginally and within class. ✓ denotes the criterion each invariance target guarantees ($\mathrm{P}_{1,3}$); every other Δ is not-stable relative to ERM. $\pm$std are omitted for space and reported, together with the full γ and feature-set sweep and the remaining CheXpert configurations, in Appendix A.4 and A.3.

ERM, but no method achieves exact statistical invariance: the best methods, i.e., with the lowest D_A values, reach AUROC 0.52–0.58, while the within-class probes reach 0.53–0.62. The results reflect approximate invariance, not exact invariance; therefore, we advise reading fairness conclusions with residual dependence in mind. In the following, we refer to propositions from Section 3 using P_n.

Marginal alignment achieves *independence* but not *separation* or *sufficiency* ($\mathrm{P}_{1,2}$). Marginal alignment methods reduce ΔDP to $0.01 - 0.05$ across all datasets (P_1). However, due to inherent incompatibility, it degrades separation (ΔEO). On the Dutch dataset, probe-AUC reduces $0.77 \rightarrow 0.52$ and the ΔDP $0.30 \rightarrow 0.01$, while ΔEO increases $0.05 \rightarrow 0.31$. The predictive performance cost reflects the $e_0 + e_1 \geq |\pi_0 - \pi_1|$ bound, resulting in task-AUROC drop of 0.07–0.19 points across datasets (P_2) (Detailed in Appendix A.1). Finally, forcing identical score distributions across groups breaks *sufficiency* and *calibration*:ΔPPV increases in four of five datasets (Dutch $0.27 \rightarrow 0.63$; Adult $0.18 \rightarrow 0.49$; C&C $0.50 \rightarrow 0.71$; Credit $0.19 \rightarrow 0.28$), and the group calibration-bias gap Δ_{bias} nearly doubles (Dutch $0.22 \rightarrow 0.44$; C&C $0.23 \rightarrow 0.55$).

Class-conditional alignment implies *separation* (P_3). Class-condition representation invariance produces the smallest ΔEO (Dutch $0.05 \rightarrow 0.03$, Credit $0.08 \rightarrow 0.01$, COMPAS $0.15 \rightarrow 0.04$, Adult $0.28 \rightarrow 0.08$; the exception is C&C whose stratum for $A = 0$ has < 80 rows), implying *separation* (P_3). Unlike marginal alignment, task-AUROC is maintained within $0.01 - 0.03$ of ERM. Calibration and *sufficiency* metrics slightly degrade (Dutch ΔPPV: $0.27 \rightarrow 0.34$; Δ_{bias}: $0.22 \rightarrow 0.23$), showing conflict between score-level separation and calibration. Notice this only applies to *risk score separation* ($R \perp A \mid Y$); calibrated scores can still yield equal error rates under thresholded decisions (Lazar Reich & Vijaykumar, 2021).

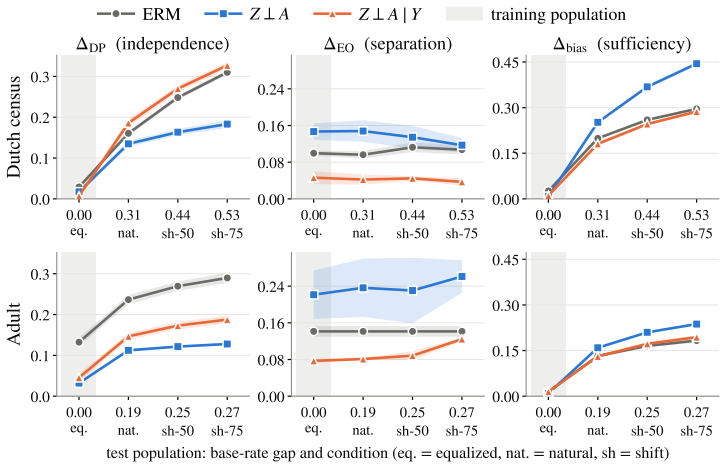


Figure 4: **Guarantees under a change of evaluation population** (P_4). Trained on the equalized condition (shaded) and evaluated on increasing base-rate gap: $P(Y \mid A)$ changes, $P(X \mid Y, A)$ fixed. $Z \perp A$ loses parity as the gap grows; $Z \perp A \mid Y$ keeps Δ_{EO} flat and loses parity by exactly the base-rate gap. Δ_{bias} grows. Mean $\pm$ std over three seeds.

Prevalence shifts break marginal but not conditional alignment (P_4). Fig. 4 demonstrates evaluation stability under change in population shifts. When base-rate $P(Y \mid A)$ changes while $P(X \mid A, Y)$ remains fixed: marginal alignment methods see increase in Δ_{DP} (0.02 (equalized) $\rightarrow$ 0.19 (shift-75)), hence *independence* is lost. Under the same shift, conditional alignment maintains a stable Δ_{EO} ($0.05 \rightarrow 0.03$), but Δ_{DP} increases ($0.01 \rightarrow 0.32$); confirming P_4.

6 Discussion

Removing demographic information as a fairness objective? The motivating question for this work is whether models should be trained to *not encode* group membership in their latent space. Our analysis shows that representation invariance is neither desirable nor sufficient for fairness. Enforcing strict marginal representation invariance while ensuring demographic parity also removes genuinely useful information and hurts model performance when outcome prevalence differs (Barocas et al., 2023; Dwork et al., 2012; Zhao & Gordon, 2022; Zhao et al., 2022). This goes beyond the well-known *leveling down* effect (Mittelstadt et al., 2023; Zietlow et al., 2022), as we do not just harm performance for the advantaged group; marginal invariance is incompatible with commonly desired fairness criteria. In short, removing information via marginal alignment methods is harmful to the fairness objective: enforcing parity at the cost of performance, separation, and sufficiency.

Removing unnecessarily strong demographic encoding is desirable. As illustrated in Table 1, class-conditional representation invariance reduces decodability without costing AUROC and improves EO relative to ERM. This suggests that some (excess) demographic information learned by ERM can be erased without costing fairness and performance. This presents a different trade-off, removing information about intra-class variations (e.g. comorbidities or disease phenotypes, see Castro et al. (2020); Li et al. (2020); Mahajan et al. (2021); Oakden-Rayner et al. (2020)), and compromises calibration by groups (Barocas et al., 2023; Hedden, 2021), although recent work questions whether this is truly problematic (Kina & Petersen, 2026). Every trade-off we discuss stays aligned with the well-established incompatibility between *independence*, *separation* and *sufficiency* (Kleinberg et al., 2017; Barocas et al., 2023). Even with the reported cost of class-conditional invariance, there is no reason to prefer marginal representation invariance over it.

The strength of constraints matters. Representations, scores, and thresholded decisions impose different requirements. Marginal representation invariance makes the score distribution identical across all groups. However, at a chosen threshold, achieving parity only requires the same positive prediction rate. Similarly, aligning Z within each outcome class aligns the class-conditional score distribution, whereas achieving separation at a threshold requires equal TPRs and FPRs. This shows that representation invariance is *sufficient* but not *necessary* for satisfying decision-level criteria (Madras et al., 2018). This difference also matters for calibration: under unequal prevalences, class-conditional alignment of risk scores generally conflicts with calibration by groups, even though equality of error rates at a chosen threshold does not impose the same restrictions (Lazar Reich &

Vijaykumar, 2021; Petersen et al., 2023). Further relaxations are conceivable and potentially desirable, such as distribution *support* matching (Tong et al., 2022), or matching only the first moment of the conditional score distributions (Kina & Petersen (2026); also see Appendix A.6).

When can representation invariance help? Our results do not imply that invariance-based methods are never useful, and the drawbacks discussed depend on what form is applied. Marginal representation invariance should generally be avoided when group prevalences (π_a) are not near equal. Invariance might also be used as a less restrictive mitigation technique (Wu et al., 2022; Chiu et al., 2023) by using weaker regularization strength (in our experiments, $\gamma < 1$ achieves a relatively better performance-parity tradeoff) or in unsupervised domain adaptation where labels y are unavailable for the target domain (Moyer et al., 2020; Li et al., 2018). Here, representation invariance should be applied strategically, acknowledging the associated risks, and favoring conditional approaches for supervised tasks or optimizing post-hoc thresholds if strict latent alignment is unnecessary (Hardt et al., 2016; Lazar Reich & Vijaykumar, 2021).

Statistical invariance vs. model invariance. We discuss a case where a model treats two patients with similar physiology differently due to their group membership. A *model* is invariant to A if any change in A while keeping everything else fixed leaves $\hat{y}$ unchanged. This still allows $\hat{Y}$ and A to be correlated when the labels Y correlate with A. By contrast, a marginally invariant *representation* requires $Z, \hat{Y} \perp A$ even when Y and A are correlated. Enforcing this can require disparate treatment: if disease presentations differ between groups, the model still maps them to identical distributions by treating similar patients from different groups differently (Lohaus et al., 2022).

To address this, an alternative conceptual goal is to learn invariant mechanisms that are stable across all environments (Arjovsky et al., 2019; Makar et al., 2022; Puli et al., 2022; Schölkopf et al., 2021; Subbaswamy & Saria, 2020). One approach is *counterfactual invariance*, where the attribute A or a specific intervention should not change the model's prediction (Veitch et al., 2021). When A is a demographic attribute, this corresponds to *counterfactual fairness* (Kusner et al., 2017). However, defining these counterfactuals is challenging in medical imaging. Changing A may also change anatomy, physiology, overlapping diseases, or even disease presentation. Therefore, specifying which factors change and which remain fixed can affect the predictions. This makes any fairness guarantees dependent on these modeling choices, their (unknown or undiscovered) confounding, and also the validity of the causal model (Chiappa & Gillam, 2018; Kilbertus et al., 2020).

Limitations of our work. We advise reading our findings with caution. Our models and methods achieved approximate rather than exact statistical invariance, leaving demographic traces in representations. However, this does not change our primary conclusion. Our empirical analysis is limited to classification tasks and simulated prevalence differences between groups, which may not reflect real-world extremes. Fairness gaps and calibration costs are sensitive to operating thresholds, compression, or model selection. **As future work**, we plan to study why *decodability* behaves differently across data domains and methods, and expand our analysis to complex tasks like segmentation.

7 Conclusion

We have shown that representation invariance, in spite of its intuitive appeal for fair AI, is incompatible with most popular fairness criteria when A and Y are not independent. We show, moreover, that previous work showing the benefit of representation independence might have been a result of failure to completely remove demographic information, suggesting that while invariance is often not desirable, reducing the encoding of demographic information can have a regularizing effect on both performance and fairness. In particular, we implore the community to avoid the often strongly highlighted concerns regarding encoded demographic information. While encoded demographics may sound deeply alarming to users and even developers not familiar with the invariance-fairness-performance tradeoff, these encoded demographics are, in many practical applications, necessary to obtain the most equitable and high-performing models. There *is* a real risk of demographic shortcut learning in the case where models encode demographic information more strongly than is necessary for accurate prediction, as was the case for all ERM models in our experiments. In such cases, the impetus must not be to aim for *no* demographic encoding, however, but rather to limit encoding to the smallest level that is sufficient for maintaining accurate, robust, and fair prediction.

Reproducibility statement

All code, dataset and, experimental configurations, and random seeds needed to reproduce our results are released in an anonymous repository[3].

Generative AI Disclosure

Following fair and responsible use of AI for research, we list the use of generative AI for the following sub-tasks:

- **Major use:** Writing boilerplate code for the experimental pipeline, automated logging and tracking runs across thousands of experiments, formatting LaTeX tables for the main paper and appendix (entries were verified).
- **Minor use:** Writing scripts to visualize metrics from raw logs, and mapping different notations across the literature (all mathematical proofs and steps are verified), minor checks for grammar, phrasing, flow, and typo fixes.
- **Not used for:** Ideation, core methodology, deriving theoretical proofs, main manuscript, literature referencing and the bibliography, or structuring the paper.

All AI-assisted outputs were manually verified by the authors. The authors take full responsibility for the final output and claims in this work.

References

Olalekan Joseph Akintande, Siavash Arjomand Bigdeli, and Aasa Feragen. Medicine after death: XAI and algorithmic fairness under label bias. In *Fourth European Workshop on Algorithmic Fairness*, pp. 171–186. ML Research Press, 2025.

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

Amir Asiaee and Kaveh Aryan. Fairness under group-conditional prior probability shift: Invariance, drift, and target-aware post-processing, 2026. URL https://arxiv.org/abs/2602.05144.

Mislav Balunović, Anian Ruoss, and Martin Vechev. Fair Normalizing Flows. In *International Conference on Learning Representations*, 2022.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems*, 36:66044–66063, 2023.

Alexander Brown, Nenad Tomasev, Jan Freyberg, Yuan Liu, Alan Karthikesalingam, and Jessica Schrouff. Detecting shortcut learning for fair medical ai using shortcut testing. *Nature communications*, 14(1):4314, 2023.

Toon Calders and Sicco Verwer. Three naive Bayes approaches for discrimination-free classification. *Data mining and knowledge discovery*, 21(2):277–292, 2010.

Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):3673, 2020.

Mattia Cerrato, Marius Köppel, Philipp Wolf, and Stefan Kramer. 10 years of fair representations: Challenges and opportunities, 2024. URL https://arxiv.org/abs/2407.03834.

Silvia Chiappa and Thomas P. S. Gillam. Path-specific counterfactual fairness, 2018. URL https://arxiv.org/abs/1802.08139.

[3] https://anonymous.4open.science/r/invariance-project-931C/README.md

Ching-Hao Chiu, Hao-Wei Chung, Yu-Jen Chen, Yiyu Shi, and Tsung-Yi Ho. Toward fairness through fair multi-exit framework for dermatological disease diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 97–107. Springer, 2023.

David R Cox. Two further applications of a model for binary regression. *Biometrika*, 45(3/4): 562–565, 1958.

Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*, pp. 1436–1445. PMLR, 2019.

Grant Duffy, Shoa L Clarke, Matthew Christensen, Bryan He, Neal Yuan, Susan Cheng, and David Ouyang. Confounders mediate AI prediction of demographics in medical imaging. *NPJ digital medicine*, 5(1):188, 2022.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Harrison Edwards and Amos J. Storkey. Censoring Representations with an Adversary. In *International Conference on Learning Representations*, 2016.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, et al. AI recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 4(6):e406–e414, 2022.

Ben Glocker, Charles Jones, Mélanie Bernhardt, and Stefan Winzeck. Algorithmic encoding of protected characteristics in chest x-ray disease detection models. *EBioMedicine*, 89, 2023.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19, 2006.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Brian Hedden. On statistical criteria of algorithmic fairness. *Philosophy and Public Affairs*, 49:209, 2021.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Christopher Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 590–597, 2019.

Charles Jones, Mélanie Roschewitz, and Ben Glocker. The role of subgroup separability in group-fair medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 179–188. Springer, 2023.

Charles Jones, Fabio de Sousa Ribeiro, Mélanie Roschewitz, Daniel Castro, and Ben Glocker. Rethinking fair representation learning for performance-sensitive tasks. In *International Conference on Learning Representations*, volume 2025, pp. 29269–29288, 2025.

Niki Kilbertus, Philip J Ball, Matt J Kusner, Adrian Weller, and Ricardo Silva. The sensitivity of counterfactual fairness to unmeasured confounding. In *Uncertainty in artificial intelligence*, pp. 616–626. PMLR, 2020.

Mohamed Amine Kina and Eike Petersen. Prevalence calibration as shortcut mitigation. In *17th International Workshop on Machine Learning in Medical Imaging (MLMI 2026) at MICCAI 2026*, 2026.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67, pp. 43:1–43:23, 2017. doi: 10.4230/LIPIcs.ITCS.2017.43.

Ron Kohavi. Scaling up the accuracy of Naive-Bayes classifiers: a decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pp. 202–207. AAAI Press, 1996.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

Claire Lazar Reich and Suhas Vijaykumar. A Possibility in Algorithmic Fairness: Can Calibration and Equal Error Rates Be Reconciled? In *2nd Symposium on Foundations of Responsible Computing (FORC 2021)*, volume 192 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 4:1–4:21, 2021. doi: 10.4230/LIPIcs.FORC.2021.4.

Bo Li, Yezhen Wang, Tong Che, Shanghang Zhang, Sicheng Zhao, Pengfei Xu, Wei Zhou, Yoshua Bengio, and Kurt Keutzer. Rethinking distributional matching based domain adaptation. *arXiv preprint arXiv:2006.13352*, 2020.

Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Manxi Lin, Nina Weng, Kamil Mikolaj, Zahra Bashir, Morten BS Svendsen, Martin G Tolsgaard, Anders N Christensen, and Aasa Feragen. Shortcut learning in medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 623–633. Springer, 2024.

Michael Lohaus, Matthäus Kleindessner, Krishnaram Kenthapadi, Francesco Locatello, and Chris Russell. Are two heads the same as one? identifying disparate treatment in fair neural networks. *Advances in Neural Information Processing Systems*, 35:16548–16562, 2022.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael Jordan. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31, 2018.

William Lotter. Acquisition parameters influence ai recognition of race in chest x-rays and mitigating these factors reduces underdiagnosis bias. *Nature Communications*, 15(1):7465, 2024.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S. Zemel. The variational fair autoencoder. In Yoshua Bengio and Yann LeCun (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1511.00830.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations, 2018. URL https://arxiv.org/abs/1802.06309.

Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *International conference on machine learning*, pp. 7313–7324. PMLR, 2021.

Maggie Makar, Ben Packer, Dan Moldovan, Davis Blalock, Yoni Halpern, and Alexander D'Amour. Causally motivated shortcut removal using auxiliary labels. In *International Conference on Artificial Intelligence and Statistics*, pp. 739–766. PMLR, 2022.

Daniel McNamara, Cheng Soon Ong, and Robert C Williamson. Costs and benefits of fair representation learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 263–270, 2019.

Brent Mittelstadt, Sandra Wachter, and Chris Russell. The unfairness of fair machine learning: Leveling down and strict egalitarianism by default. *Mich. Tech. L. Rev.*, 30:1, 2023.

Daniel Moyer, Greg Ver Steeg, Chantal MW Tax, and Paul M Thompson. Scanner invariant representations for diffusion mri harmonization. *Magnetic resonance in medicine*, 84(4):2174–2189, 2020.

Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pp. 151–159, 2020.

Aditya Parikh, Stella Frank, Sneha Das, and Aasa Feragen. Towards fairness under label bias in image segmentation: Impact, measurement and mitigation. *arXiv preprint arXiv:2605.06891*, 2026.

Adrián Pérez-Suay, Valero Laparra, Gonzalo Mateo-García, Jordi Muñoz-Marí, Luis Gómez-Chova, and Gustau Camps-Valls. Fair kernel learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 339–355. Springer, 2017.

Eike Petersen, Melanie Ganz, Sune Hannibal Holm, and Aasa Feragen. On (assessing) the fairness of risk score models, 2023. URL `https://arxiv.org/abs/2302.08851`.

Aahlad Manas Puli, Lily H Zhang, Eric Karl Oermann, and Rajesh Ranganath. Out-of-distribution generalization in the presence of nuisance-induced spurious correlations. In *International Conference on Learning Representations*, 2022.

Alessandro Quarta, Filipe Santos, Aldo Marzullo, Joao MC Sousa, Susana M Vieira, Leo Anthony Celi, Laleh Seyyed-Kalantari, and Francesco Calimeri. Hidden in plain sight: Vector embeddings give away demographic information. *IEEE Journal of Biomedical and Health Informatics*, 2026.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

Darakshan Rashid, Raza Imam, Dwarikanath Mahapatra, and Brejesh Lall. Stride-net: Fairness-aware disentangled representation learning for chest x-ray diagnosis. *arXiv preprint arXiv:2602.10875*, 2026.

Shauli Ravfogel, Francisco Vargas, Yoav Goldberg, and Ryan Cotterell. Adversarial concept erasure in kernel space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6034–6055, 2022.

Michael Redmond and Alok Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3): 660–678, 2002.

Amir Reza Sadri, Thomas DeSilvio, and Satish E Viswanath. Mutual information regularization for fairness-aware deep imaging representations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 399–409. Springer, 2025.

Ioannis Sarridis, Christos Koutlis, Symeon Papadopoulos, and Christos Diou. Flac: Fairness-aware representation learning by suppressing attribute-class associations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(2):1148–1160, 2024.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Hossein Shakibania, Jonas Henry Grebe, Tobias Braun, Ege Aktemur, Saleh Aslani, Mehmet G Yiğit, and Marcus Rohrbach. Obliviate: Erasing concepts from autoregressive image generation models. In *European Conference on Computer Vision*, pp. 665–683. Springer, 2026.

Emma AM Stanley, Raghav Mehta, and Mélanie Roschewitz. Exploring the interplay of label bias with subgroup size and separability: A case study in mammographic density. In *Fairness of AI in Medical Imaging: Third International Workshop, FAIMI 2025, Held in Conjunction with MICCAI 2025, Daejeon, South Korea, September 23, 2025, Proceedings*, pp. 74. Springer Nature, 2025a.

Emma AM Stanley, Raissa Souza, Matthias Wilms, and Nils D Forkert. Where, why, and how is bias learned in medical image analysis models? a study of bias encoding within convolutional networks using synthetic data. *EBioMedicine*, 111, 2025b.

Adarsh Subbaswamy and Suchi Saria. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics*, 21(2):345–352, 2020.

Shangyuan Tong, Timur Garipov, Yang Zhang, Shiyu Chang, and Tommi S. Jaakkola. Adversarial support alignment. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=26gKg6x-ie.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. Counterfactual invariance to spurious correlations: why and how to pass stress tests. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, 2021.

Zichong Wang, Zhibo Chu, Ronald Blanco, Zhong Chen, Shu-Ching Chen, and Wenbin Zhang. Advancing graph counterfactual fairness through fair representation learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 40–58. Springer, 2024.

Yawen Wu, Dewen Zeng, Xiaowei Xu, Yiyu Shi, and Jingtong Hu. Fairprune: Achieving fairness through pruning for dermatological disease diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 743–753. Springer, 2022.

Yuzhe Yang, Haoran Zhang, Judy W Gichoya, Dina Katabi, and Marzyeh Ghassemi. The limits of fair medical imaging AI in real-world generalization. *Nature medicine*, 30(10):2838–2848, 2024.

I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications*, 36(2): 2473–2480, 2009.

Paul H Yi, Jinchi Wei, Tae Kyung Kim, Jiwon Shin, Haris I Sair, Ferdinand K Hui, Gregory D Hager, and Cheng Ting Lin. Radiology "forensics": determination of age and sex from chest radiographs using deep learning. *Emergency Radiology*, 28(5):949–954, 2021.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pp. 325–333. PMLR, 2013.

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.

Han Zhao and Geoffrey J Gordon. Inherent tradeoffs in learning fair representations. *Journal of Machine Learning Research*, 23(57):1–26, 2022.

Han Zhao, Chen Dan, Bryon Aragam, Tommi S Jaakkola, Geoffrey J Gordon, and Pradeep Ravikumar. Fundamental limits and tradeoffs in invariant representation learning. *Journal of machine learning research*, 23(340):1–49, 2022.

Dominik Zietlow, Michael Lohaus, Guha Balakrishnan, Matthäus Kleindessner, Francesco Locatello, Bernhard Schölkopf, and Chris Russell. Leveling down in computer vision: Pareto inefficiencies in fair deep classifiers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10400–10411. IEEE, 2022.

A Appendix 1: Tabular Datasets

A.1 Additional Details: Fairness under increasing base-rate gap and prevalence shift conditions

Refer to Fig. 6, Fig. 5 and Table 2 for results of all evaluation metrics against the training base-rate gap for the four prevalence conditions. Class-conditional alignment is similar to ERM's AUROC at every gap, while the marginal flow manages to reduce decodability, and its summed error e_0+e_1 rises with the gap. D_A for ERM follows the gap where A becomes decodable from Y. Δ_{DP} for ERM and class-conditional alignment follows the default base-rate; Δ_{EO} for the marginal method increases with the gap; Δ_{PPV} and Δ_{bias} grow for every method, as anticipated from our propositions. Fig. 7 plots the summed error of all 2,040 evaluations against the bound (McNamara et al., 2019); none falls below it.

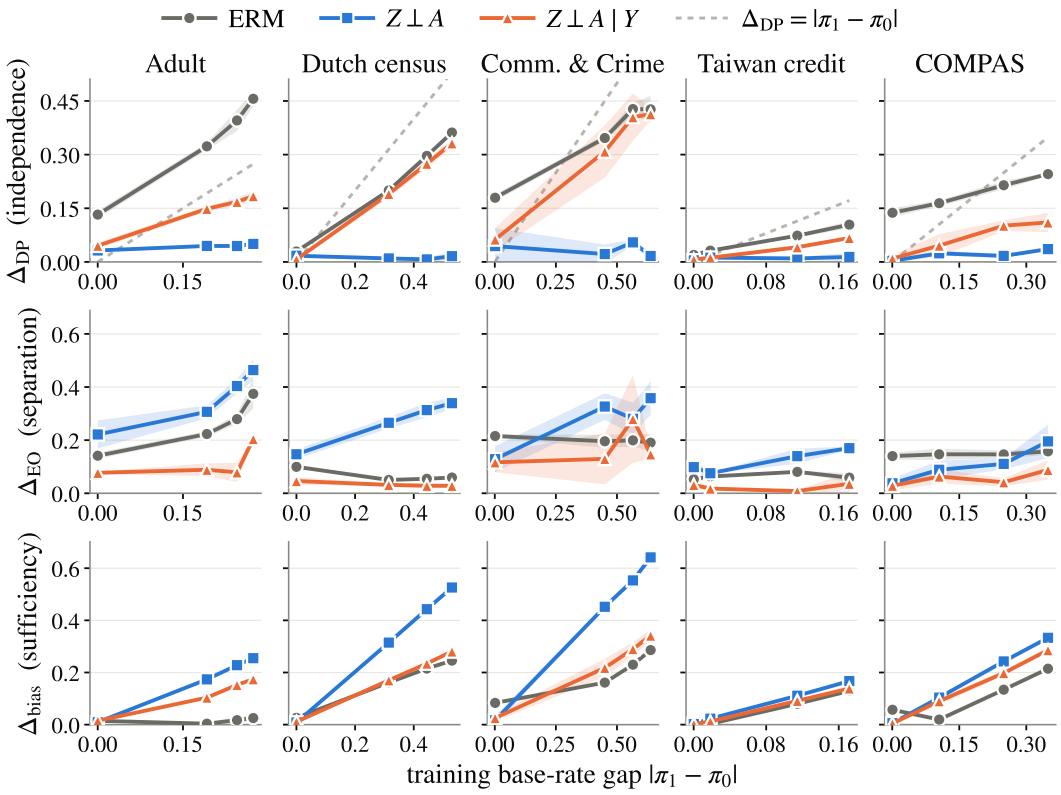


Figure 5: **Fairness gaps increase with the base-rate gap.** Top: Δ_{DP}; Middle: Δ_{EO}; Bottom: Δ_{bias} against the training base-rate gap $|\pi_1 - \pi_0|$ over the four prevalence conditions, per dataset and method. Marginal representation invariance harms *separation* for parity at a rate set by the gap (P_1); class-conditional representation invariance holds Δ_{EO} (P_3) but not Δ_{bias} (P_3). Mean $\pm$ std over three seeds reported.

A.2 Ablations: Probe Ablation

Refer to Fig. 8 and Table 3 for details on probe evaluation. For the marginal alignment methods, the empirical probe AUROC lies below the bound obtained from the exact per-group MADE densities, while the within-class probe stays high or rises above ERM. For the class-conditional alignment methods, the within-class probe stays within the class bound, and the marginal probe rises to the base-rate floor of P_2.

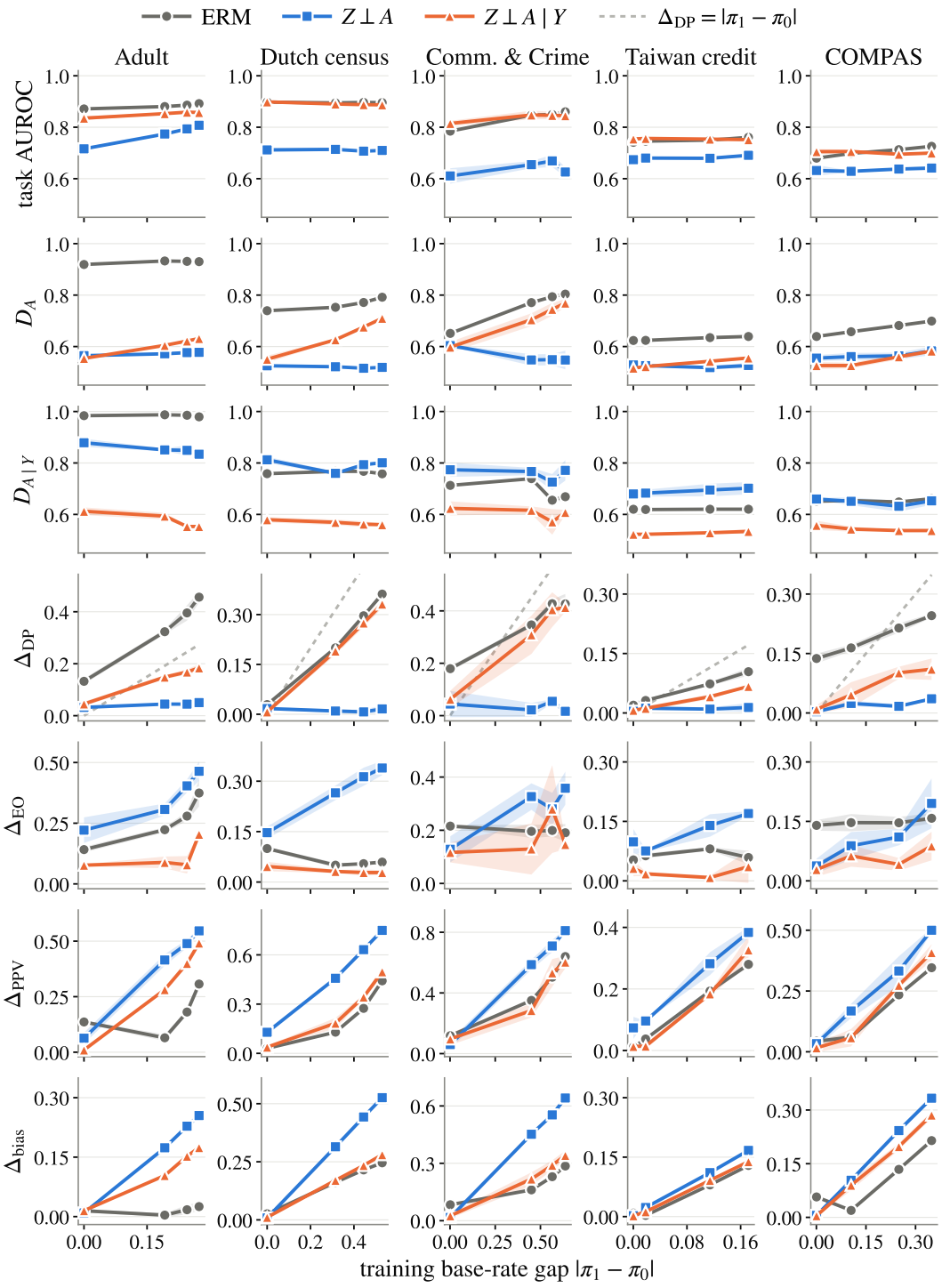


Figure 6: Every metric against the training base-rate gap $|\pi_1 - \pi_0|$ (equalized, natural, shift-50, shift-75; matched test). Dotted line in the Δ_{DP} row: $\Delta_{DP} = |\pi_1 - \pi_0|$. Mean $\pm$ sd over seeds.

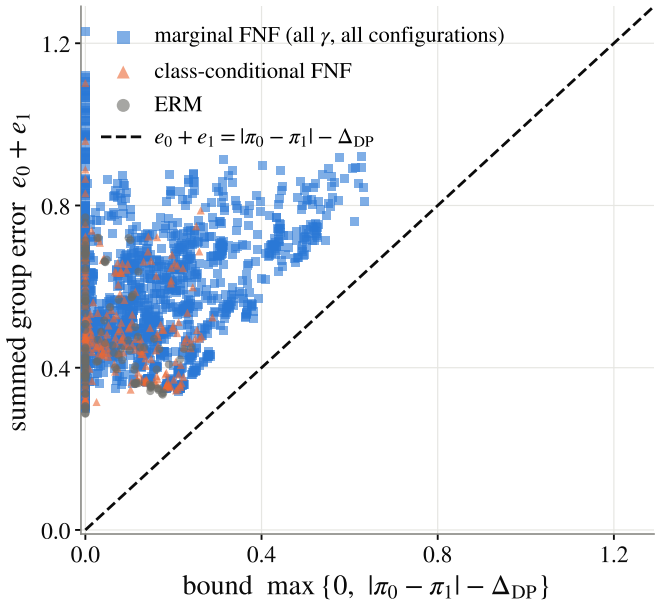


Figure 7: Summed group error $e_0 + e_1$ against the bound $\max\{0, |\pi_0 - \pi_1| - \Delta_{\mathrm{DP}}\}$ (mentioned in P_2) for every model×prevalence condition.

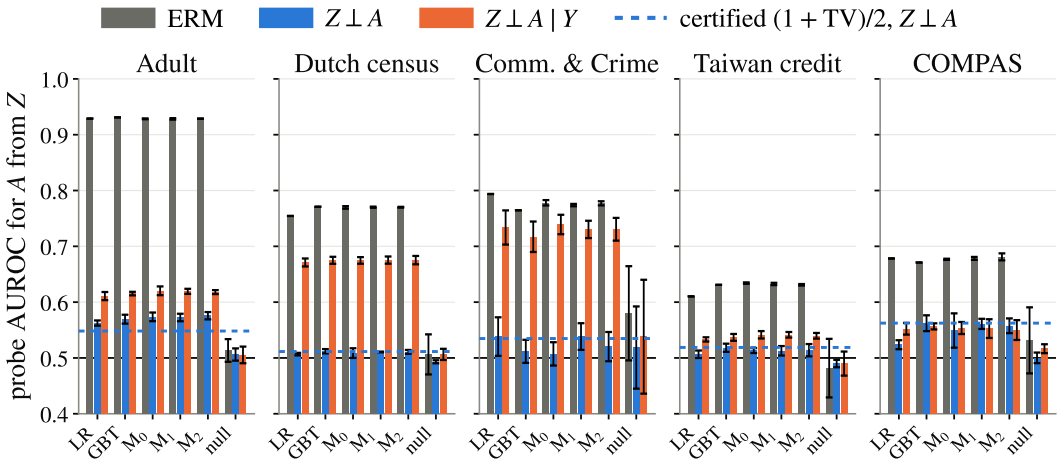


Figure 8: Probe AUROC to *decode* for A from Z under six probe families (*shift-50*, reported mean ± std over all seeds). Dotted: $(1 + \mathrm{TV})/2$ for the marginal flow.

Table 2: **Evaluation metrics by training condition across datasets** (reported as mean ± std over all seeds). $\Delta\pi = |\pi_1 - \pi_0|$ is the base-rate gap. **Conditions**: Eq (equalized), Nat (natural), S50 (shift-50), S75 (shift-75). **Methods**: ERM, **M** (marginal alignment, $Z \perp A$), and **C** (class-conditional alignment, $Z \perp A \mid Y$). Datasets: Adult, Dutch census (Dutch), Communities & Crime (Crime), Taiwan credit (Credit), and COMPAS.

Dataset	Cond.	Method	$\Delta\pi$	AUROC	D_A	$D_{A\|Y}$	Δ_{DP}	Δ_{EO}	Δ_{PPV}	Δ_{bias}	$e_0 + e_1$
Adult	Eq	ERM	0.00	0.87 ± 0.00	0.92 ± 0.00	0.98 ± 0.00	0.13 ± 0.01	0.14 ± 0.01	0.14 ± 0.01	0.01 ± 0.00	0.42 ± 0.02
	Eq	M	0.00	0.72 ± 0.01	0.56 ± 0.00	0.88 ± 0.02	0.03 ± 0.00	0.22 ± 0.05	0.06 ± 0.03	0.01 ± 0.00	0.57 ± 0.06
	Eq	C	0.00	0.84 ± 0.01	0.55 ± 0.00	0.61 ± 0.02	0.05 ± 0.01	0.08 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	0.54 ± 0.05
	Nat	ERM	0.19	0.88 ± 0.00	0.93 ± 0.00	0.99 ± 0.00	0.32 ± 0.01	0.22 ± 0.01	0.06 ± 0.01	0.00 ± 0.00	0.38 ± 0.01
	Nat	M	0.19	0.77 ± 0.00	0.57 ± 0.01	0.85 ± 0.01	0.04 ± 0.00	0.31 ± 0.02	0.42 ± 0.03	0.17 ± 0.00	0.65 ± 0.07
	Nat	C	0.19	0.85 ± 0.00	0.60 ± 0.00	0.59 ± 0.01	0.15 ± 0.01	0.09 ± 0.02	0.28 ± 0.00	0.10 ± 0.00	0.49 ± 0.04
	S50	ERM	0.25	0.89 ± 0.00	0.93 ± 0.00	0.99 ± 0.00	0.40 ± 0.03	0.28 ± 0.02	0.18 ± 0.01	0.02 ± 0.01	0.37 ± 0.03
	S50	M	0.25	0.79 ± 0.00	0.58 ± 0.01	0.85 ± 0.02	0.04 ± 0.00	0.40 ± 0.03	0.49 ± 0.02	0.23 ± 0.00	0.68 ± 0.04
	S50	C	0.25	0.86 ± 0.00	0.62 ± 0.01	0.55 ± 0.00	0.17 ± 0.01	0.08 ± 0.04	0.40 ± 0.01	0.15 ± 0.00	0.51 ± 0.05
	S75	ERM	0.27	0.89 ± 0.00	0.93 ± 0.00	0.98 ± 0.00	0.46 ± 0.02	0.37 ± 0.06	0.31 ± 0.01	0.03 ± 0.01	0.38 ± 0.04
	S75	M	0.27	0.81 ± 0.00	0.58 ± 0.01	0.83 ± 0.01	0.05 ± 0.01	0.46 ± 0.04	0.55 ± 0.01	0.25 ± 0.00	0.63 ± 0.02
	S75	C	0.27	0.86 ± 0.00	0.63 ± 0.01	0.55 ± 0.00	0.18 ± 0.01	0.20 ± 0.00	0.49 ± 0.01	0.17 ± 0.01	0.50 ± 0.02
Dutch	Eq	ERM	0.00	0.90 ± 0.00	0.74 ± 0.00	0.76 ± 0.00	0.03 ± 0.00	0.10 ± 0.01	0.03 ± 0.00	0.03 ± 0.00	0.30 ± 0.00
	Eq	M	0.00	0.71 ± 0.01	0.53 ± 0.00	0.81 ± 0.01	0.02 ± 0.01	0.15 ± 0.02	0.13 ± 0.01	0.01 ± 0.00	0.64 ± 0.03
	Eq	C	0.00	0.90 ± 0.00	0.55 ± 0.01	0.58 ± 0.01	0.01 ± 0.01	0.05 ± 0.01	0.04 ± 0.01	0.01 ± 0.01	0.36 ± 0.01
	Nat	ERM	0.31	0.89 ± 0.00	0.75 ± 0.00	0.77 ± 0.00	0.20 ± 0.01	0.05 ± 0.01	0.13 ± 0.00	0.16 ± 0.00	0.36 ± 0.00
	Nat	M	0.31	0.71 ± 0.01	0.52 ± 0.00	0.76 ± 0.01	0.01 ± 0.00	0.27 ± 0.02	0.46 ± 0.01	0.31 ± 0.00	0.70 ± 0.01
	Nat	C	0.31	0.89 ± 0.00	0.63 ± 0.00	0.57 ± 0.01	0.19 ± 0.01	0.03 ± 0.00	0.18 ± 0.03	0.17 ± 0.01	0.39 ± 0.02
	S50	ERM	0.44	0.90 ± 0.00	0.77 ± 0.00	0.77 ± 0.00	0.30 ± 0.00	0.05 ± 0.00	0.27 ± 0.01	0.22 ± 0.00	0.35 ± 0.01
	S50	M	0.44	0.71 ± 0.00	0.52 ± 0.00	0.79 ± 0.01	0.01 ± 0.01	0.31 ± 0.03	0.63 ± 0.00	0.44 ± 0.00	0.73 ± 0.01
	S50	C	0.44	0.89 ± 0.00	0.68 ± 0.01	0.56 ± 0.01	0.27 ± 0.01	0.03 ± 0.01	0.34 ± 0.01	0.23 ± 0.01	0.38 ± 0.01
	S75	ERM	0.53	0.90 ± 0.00	0.79 ± 0.00	0.76 ± 0.00	0.36 ± 0.01	0.06 ± 0.01	0.44 ± 0.02	0.25 ± 0.00	0.34 ± 0.01
	S75	M	0.53	0.71 ± 0.00	0.52 ± 0.00	0.80 ± 0.01	0.02 ± 0.01	0.34 ± 0.02	0.75 ± 0.01	0.53 ± 0.00	0.75 ± 0.01
	S75	C	0.53	0.89 ± 0.00	0.71 ± 0.00	0.56 ± 0.00	0.33 ± 0.01	0.03 ± 0.01	0.50 ± 0.02	0.28 ± 0.00	0.36 ± 0.02
Crime	Eq	ERM	0.00	0.78 ± 0.00	0.65 ± 0.00	0.71 ± 0.00	0.18 ± 0.01	0.22 ± 0.01	0.12 ± 0.00	0.08 ± 0.00	0.65 ± 0.01
	Eq	M	0.00	0.61 ± 0.03	0.60 ± 0.00	0.77 ± 0.03	0.04 ± 0.05	0.13 ± 0.05	0.06 ± 0.03	0.02 ± 0.02	0.94 ± 0.18
	Eq	C	0.00	0.81 ± 0.01	0.60 ± 0.02	0.62 ± 0.03	0.06 ± 0.03	0.12 ± 0.04	0.10 ± 0.04	0.03 ± 0.01	0.67 ± 0.08
	Nat	ERM	0.45	0.85 ± 0.00	0.77 ± 0.00	0.74 ± 0.00	0.35 ± 0.01	0.20 ± 0.02	0.35 ± 0.01	0.16 ± 0.00	0.47 ± 0.03
	Nat	M	0.45	0.65 ± 0.01	0.55 ± 0.02	0.77 ± 0.01	0.02 ± 0.03	0.33 ± 0.05	0.59 ± 0.03	0.45 ± 0.01	0.82 ± 0.01
	Nat	C	0.45	0.85 ± 0.02	0.70 ± 0.03	0.62 ± 0.01	0.31 ± 0.07	0.13 ± 0.10	0.28 ± 0.05	0.22 ± 0.03	0.47 ± 0.06
	S50	ERM	0.57	0.85 ± 0.00	0.79 ± 0.00	0.66 ± 0.00	0.43 ± 0.02	0.20 ± 0.02	0.50 ± 0.01	0.23 ± 0.00	0.46 ± 0.01
	S50	M	0.57	0.67 ± 0.02	0.55 ± 0.02	0.73 ± 0.03	0.05 ± 0.01	0.28 ± 0.06	0.71 ± 0.04	0.55 ± 0.01	0.80 ± 0.05
	S50	C	0.57	0.85 ± 0.01	0.74 ± 0.02	0.57 ± 0.05	0.40 ± 0.07	0.28 ± 0.16	0.53 ± 0.09	0.29 ± 0.03	0.49 ± 0.06
	S75	ERM	0.64	0.86 ± 0.00	0.80 ± 0.00	0.67 ± 0.02	0.43 ± 0.04	0.19 ± 0.03	0.64 ± 0.03	0.29 ± 0.00	0.45 ± 0.00
	S75	M	0.64	0.63 ± 0.02	0.55 ± 0.04	0.77 ± 0.04	0.02 ± 0.01	0.36 ± 0.06	0.81 ± 0.01	0.64 ± 0.01	0.86 ± 0.03
	S75	C	0.64	0.84 ± 0.02	0.77 ± 0.02	0.61 ± 0.01	0.41 ± 0.02	0.15 ± 0.02	0.60 ± 0.04	0.34 ± 0.02	0.45 ± 0.03
Credit	Eq	ERM	0.00	0.74 ± 0.00	0.62 ± 0.00	0.62 ± 0.00	0.02 ± 0.01	0.05 ± 0.01	0.01 ± 0.00	0.01 ± 0.00	0.57 ± 0.01
	Eq	M	0.00	0.67 ± 0.00	0.53 ± 0.00	0.68 ± 0.01	0.00 ± 0.00	0.10 ± 0.03	0.07 ± 0.04	0.00 ± 0.00	0.62 ± 0.07
	Eq	C	0.00	0.75 ± 0.01	0.52 ± 0.00	0.52 ± 0.01	0.01 ± 0.01	0.03 ± 0.02	0.01 ± 0.01	0.00 ± 0.00	0.52 ± 0.05
	Nat	ERM	0.02	0.75 ± 0.00	0.62 ± 0.00	0.62 ± 0.00	0.03 ± 0.00	0.06 ± 0.01	0.04 ± 0.00	0.00 ± 0.00	0.53 ± 0.02
	Nat	M	0.02	0.68 ± 0.00	0.53 ± 0.00	0.68 ± 0.01	0.01 ± 0.00	0.08 ± 0.00	0.10 ± 0.00	0.02 ± 0.00	0.61 ± 0.03
	Nat	C	0.02	0.76 ± 0.00	0.52 ± 0.00	0.52 ± 0.01	0.01 ± 0.01	0.02 ± 0.01	0.01 ± 0.01	0.01 ± 0.00	0.49 ± 0.03
	S50	ERM	0.11	0.75 ± 0.00	0.63 ± 0.00	0.62 ± 0.00	0.07 ± 0.00	0.08 ± 0.00	0.19 ± 0.01	0.08 ± 0.00	0.49 ± 0.03
	S50	M	0.11	0.68 ± 0.01	0.52 ± 0.01	0.69 ± 0.02	0.01 ± 0.00	0.14 ± 0.03	0.28 ± 0.04	0.11 ± 0.00	0.55 ± 0.06
	S50	C	0.11	0.75 ± 0.00	0.54 ± 0.01	0.53 ± 0.00	0.04 ± 0.01	0.01 ± 0.01	0.18 ± 0.01	0.09 ± 0.00	0.54 ± 0.09
	S75	ERM	0.17	0.76 ± 0.00	0.64 ± 0.00	0.62 ± 0.00	0.10 ± 0.01	0.06 ± 0.02	0.28 ± 0.01	0.13 ± 0.00	0.50 ± 0.09
	S75	M	0.17	0.69 ± 0.00	0.53 ± 0.01	0.70 ± 0.02	0.01 ± 0.01	0.17 ± 0.01	0.38 ± 0.02	0.17 ± 0.00	0.53 ± 0.05
	S75	C	0.17	0.75 ± 0.00	0.56 ± 0.00	0.53 ± 0.01	0.07 ± 0.00	0.04 ± 0.04	0.33 ± 0.05	0.14 ± 0.00	0.41 ± 0.10
COMPAS	Eq	ERM	0.00	0.68 ± 0.00	0.64 ± 0.00	0.65 ± 0.01	0.14 ± 0.01	0.14 ± 0.01	0.04 ± 0.01	0.06 ± 0.00	0.74 ± 0.04
	Eq	M	0.00	0.63 ± 0.02	0.56 ± 0.01	0.66 ± 0.00	0.00 ± 0.00	0.04 ± 0.02	0.03 ± 0.01	0.01 ± 0.01	0.79 ± 0.05
	Eq	C	0.00	0.71 ± 0.01	0.53 ± 0.01	0.56 ± 0.02	0.01 ± 0.01	0.03 ± 0.02	0.02 ± 0.02	0.00 ± 0.00	0.66 ± 0.02
	Nat	ERM	0.11	0.70 ± 0.00	0.66 ± 0.00	0.65 ± 0.00	0.16 ± 0.01	0.15 ± 0.02	0.06 ± 0.01	0.02 ± 0.00	0.68 ± 0.00
	Nat	M	0.11	0.63 ± 0.00	0.56 ± 0.01	0.65 ± 0.02	0.02 ± 0.01	0.09 ± 0.03	0.17 ± 0.03	0.10 ± 0.01	0.84 ± 0.03
	Nat	C	0.11	0.71 ± 0.01	0.53 ± 0.01	0.54 ± 0.01	0.05 ± 0.03	0.06 ± 0.03	0.06 ± 0.04	0.09 ± 0.01	0.66 ± 0.01
	S50	ERM	0.25	0.71 ± 0.00	0.68 ± 0.01	0.65 ± 0.00	0.21 ± 0.01	0.15 ± 0.01	0.23 ± 0.01	0.13 ± 0.00	0.70 ± 0.03
	S50	M	0.25	0.64 ± 0.00	0.56 ± 0.01	0.63 ± 0.02	0.02 ± 0.01	0.11 ± 0.02	0.33 ± 0.06	0.24 ± 0.00	0.78 ± 0.04
	S50	C	0.25	0.69 ± 0.00	0.56 ± 0.01	0.54 ± 0.00	0.10 ± 0.02	0.04 ± 0.02	0.27 ± 0.01	0.20 ± 0.01	0.68 ± 0.03
	S75	ERM	0.35	0.73 ± 0.00	0.70 ± 0.00	0.66 ± 0.00	0.25 ± 0.01	0.16 ± 0.01	0.35 ± 0.01	0.21 ± 0.00	0.67 ± 0.08
	S75	M	0.35	0.64 ± 0.01	0.58 ± 0.02	0.65 ± 0.02	0.04 ± 0.00	0.20 ± 0.06	0.50 ± 0.01	0.33 ± 0.01	0.83 ± 0.03
	S75	C	0.35	0.70 ± 0.01	0.58 ± 0.01	0.54 ± 0.00	0.11 ± 0.03	0.09 ± 0.04	0.41 ± 0.03	0.29 ± 0.01	0.68 ± 0.10

Table 3: **Probe ablation at shift 50.** Methods: ERM, Marginal (**M**: $Z \perp A$), and Conditional (**C**: $Z \perp A \mid Y$). Columns 3–8 report AUC of probe families predicting A from Z: **LR** (logistic regression), **GB** (gradient-boosted trees), **MLP-k** (k hidden layers), and **LR$-A$** (logistic control baseline excluding A). Right columns report the original FNF adversary balanced accuracy (**Adv BA**), the certified balanced accuracy bound $(1 + \mathrm{TV})/2$ (**Bound**), and the maximum within-class probe AUC (**Max** W). Reported as mean $\pm$ sd over three seeds.

		Probe AUC (A from Z)								
Dataset	**Method**	**LR**	**GB**	**MLP-0**	**MLP-1**	**MLP-2**	**LR$-A$**	**Adv BA**	**Bound**	**Max W**
Adult	ERM	0.93 ± 0.00	0.93 ± 0.00	0.93 ± 0.00	0.93 ± 0.00	0.93 ± 0.00	0.51 ± 0.02	0.86 ± 0.00	–	0.99 ± 0.00
	M	0.56 ± 0.00	0.57 ± 0.01	0.57 ± 0.01	0.57 ± 0.01	0.58 ± 0.01	0.51 ± 0.01	0.55 ± 0.01	0.55 ± 0.00	0.85 ± 0.02
	C	0.61 ± 0.01	0.62 ± 0.00	0.62 ± 0.01	0.62 ± 0.00	0.62 ± 0.00	0.51 ± 0.01	0.58 ± 0.01	0.53 ± 0.00	0.56 ± 0.00
Dutch census	ERM	0.75 ± 0.00	0.77 ± 0.00	0.77 ± 0.00	0.77 ± 0.00	0.77 ± 0.00	0.51 ± 0.04	0.68 ± 0.00	–	0.77 ± 0.00
	M	0.51 ± 0.00	0.51 ± 0.00	0.51 ± 0.01	0.51 ± 0.00	0.51 ± 0.00	0.49 ± 0.00	0.50 ± 0.00	0.51 ± 0.00	0.79 ± 0.01
	C	0.67 ± 0.01	0.67 ± 0.01	0.67 ± 0.01	0.68 ± 0.01	0.68 ± 0.01	0.51 ± 0.01	0.60 ± 0.00	0.56 ± 0.00	0.56 ± 0.01
Comm. & Crime	ERM	0.79 ± 0.00	0.76 ± 0.00	0.78 ± 0.00	0.77 ± 0.00	0.78 ± 0.00	0.58 ± 0.08	0.69 ± 0.01	–	0.66 ± 0.00
	M	0.54 ± 0.03	0.51 ± 0.02	0.51 ± 0.02	0.54 ± 0.02	0.52 ± 0.03	0.52 ± 0.07	0.51 ± 0.00	0.54 ± 0.01	0.73 ± 0.03
	C	0.73 ± 0.03	0.72 ± 0.03	0.74 ± 0.02	0.73 ± 0.02	0.73 ± 0.02	0.54 ± 0.10	0.64 ± 0.01	0.55 ± 0.01	0.58 ± 0.04
Taiwan credit	ERM	0.61 ± 0.00	0.63 ± 0.00	0.63 ± 0.00	0.63 ± 0.00	0.63 ± 0.00	0.48 ± 0.05	0.58 ± 0.00	–	0.62 ± 0.00
	M	0.51 ± 0.01	0.52 ± 0.01	0.51 ± 0.00	0.51 ± 0.01	0.51 ± 0.01	0.49 ± 0.01	0.50 ± 0.00	0.52 ± 0.00	0.69 ± 0.02
	C	0.53 ± 0.00	0.54 ± 0.01	0.54 ± 0.01	0.54 ± 0.00	0.54 ± 0.01	0.49 ± 0.02	0.50 ± 0.00	0.52 ± 0.00	0.53 ± 0.00
COMPAS	ERM	0.68 ± 0.00	0.67 ± 0.00	0.68 ± 0.00	0.68 ± 0.00	0.68 ± 0.01	0.53 ± 0.06	0.60 ± 0.00	–	0.65 ± 0.00
	M	0.52 ± 0.01	0.56 ± 0.01	0.55 ± 0.03	0.56 ± 0.01	0.56 ± 0.01	0.50 ± 0.01	0.52 ± 0.01	0.56 ± 0.00	0.63 ± 0.02
	C	0.55 ± 0.01	0.56 ± 0.01	0.55 ± 0.01	0.55 ± 0.02	0.55 ± 0.02	0.52 ± 0.01	0.52 ± 0.01	0.60 ± 0.01	0.54 ± 0.01

A.3 Ablation: Strength of representation invariance, hyperparameters and thresholds

Refer to Fig. 9 for sweeps γ for the marginal alignment method. D_A, Δ_{DP} and AUROC both decrease with γ; Δ_{EO} do not. On Dutch it's par with ERM level for $\gamma \leq 0.1$ and rises to 0.31 at $\gamma = 1$ as $D_{A|Y}$ increases from 0.61 to 0.79; on Adult it is above ERM for all γ; on COMPAS it stays within seed std of ERM. Refer to Table 4, which demonstrates that changing the feature set does not change the conclusion: $\Delta_{\mathrm{DP}} \leq 0.10$ in every config while Δ_{EO} stays at par or above ERM. Table 5 compares the Youden threshold with a default 0.5: at default, the marginal alignment method predicts few positives under low prevalence on Adult, Credit and COMPAS, so both gaps shrink counterintuitively; on Dutch and Crime it holds for both thresholds.

Table 4: Marginal alignment method at $\gamma = 1$ for every feature set and MADE width in the sweep (*shift 50*, mean $\pm$ std over all seeds).

Dataset	Feature set	MADE	n fits	AUROC	D_A	$D_{A\|Y}$	Δ_{DP}	Δ_{EO}	exact TV
Adult	paper6	50	3	0.79 ± 0.00	0.58 ± 0.01	0.89 ± 0.01	0.08 ± 0.00	0.38 ± 0.03	0.103 ± 0.003
Adult	paper6	100	3	0.79 ± 0.00	0.59 ± 0.01	0.87 ± 0.02	0.09 ± 0.01	0.36 ± 0.04	0.102 ± 0.005
Adult	paper6_age (selected)	50	3	0.79 ± 0.00	0.58 ± 0.01	0.85 ± 0.02	0.04 ± 0.00	0.40 ± 0.03	0.096 ± 0.007
Adult	paper6_age	100	3	0.80 ± 0.00	0.58 ± 0.00	0.84 ± 0.02	0.05 ± 0.01	0.38 ± 0.02	0.093 ± 0.024
Dutch census	core6	50	3	0.72 ± 0.00	0.52 ± 0.01	0.81 ± 0.01	0.01 ± 0.01	0.27 ± 0.02	0.022 ± 0.003
Dutch census	core6	100	3	0.72 ± 0.01	0.52 ± 0.00	0.81 ± 0.00	0.01 ± 0.01	0.30 ± 0.01	0.022 ± 0.003
Dutch census	core6_hh (selected)	50	3	0.71 ± 0.00	0.52 ± 0.00	0.79 ± 0.01	0.01 ± 0.01	0.31 ± 0.03	0.023 ± 0.002
Dutch census	core6_hh	100	3	0.71 ± 0.00	0.51 ± 0.01	0.78 ± 0.00	0.01 ± 0.01	0.32 ± 0.02	0.026 ± 0.004
Comm. & Crime	socio4 (selected)	50	3	0.67 ± 0.02	0.55 ± 0.02	0.73 ± 0.03	0.05 ± 0.01	0.28 ± 0.06	0.070 ± 0.020
Comm. & Crime	socio4	100	3	0.64 ± 0.03	0.56 ± 0.03	0.76 ± 0.02	0.04 ± 0.03	0.28 ± 0.05	0.069 ± 0.015
Comm. & Crime	socio6	50	3	0.65 ± 0.02	0.56 ± 0.03	0.74 ± 0.01	0.09 ± 0.03	0.30 ± 0.09	0.085 ± 0.013
Comm. & Crime	socio6	100	3	0.68 ± 0.02	0.58 ± 0.02	0.73 ± 0.04	0.10 ± 0.07	0.22 ± 0.02	0.063 ± 0.011
Taiwan credit	core5 (selected)	50	3	0.68 ± 0.01	0.52 ± 0.01	0.69 ± 0.02	0.01 ± 0.00	0.14 ± 0.03	0.038 ± 0.005
Taiwan credit	core5	100	3	0.69 ± 0.00	0.53 ± 0.01	0.69 ± 0.01	0.01 ± 0.01	0.12 ± 0.03	0.039 ± 0.005
Taiwan credit	extended8	50	3	0.65 ± 0.01	0.53 ± 0.01	0.70 ± 0.01	0.01 ± 0.01	0.22 ± 0.04	0.022 ± 0.002
Taiwan credit	extended8	100	3	0.65 ± 0.02	0.58 ± 0.01	0.71 ± 0.03	0.01 ± 0.01	0.25 ± 0.03	0.025 ± 0.005
COMPAS	paper7 (selected)	50	3	0.64 ± 0.00	0.56 ± 0.01	0.63 ± 0.02	0.02 ± 0.01	0.11 ± 0.02	0.125 ± 0.007
COMPAS	paper7	100	3	0.63 ± 0.02	0.57 ± 0.01	0.64 ± 0.04	0.02 ± 0.03	0.10 ± 0.03	0.116 ± 0.016
COMPAS	paper7_nosex	50	3	0.63 ± 0.00	0.60 ± 0.01	0.65 ± 0.04	0.04 ± 0.02	0.10 ± 0.02	0.143 ± 0.018
COMPAS	paper7_nosex	100	3	0.63 ± 0.02	0.60 ± 0.01	0.65 ± 0.01	0.04 ± 0.03	0.07 ± 0.02	0.131 ± 0.007

A.4 Additional Details: Datasets and per-seed results

Dataset details, group definitions, and training-split prevalences per condition are reported in Table 6 for accountability.

Table 7 reports every seed for the four methods at the *shift-50* prevalence condition in Table 1. The training seed variation is small relative to the between-method differences on all datasets except

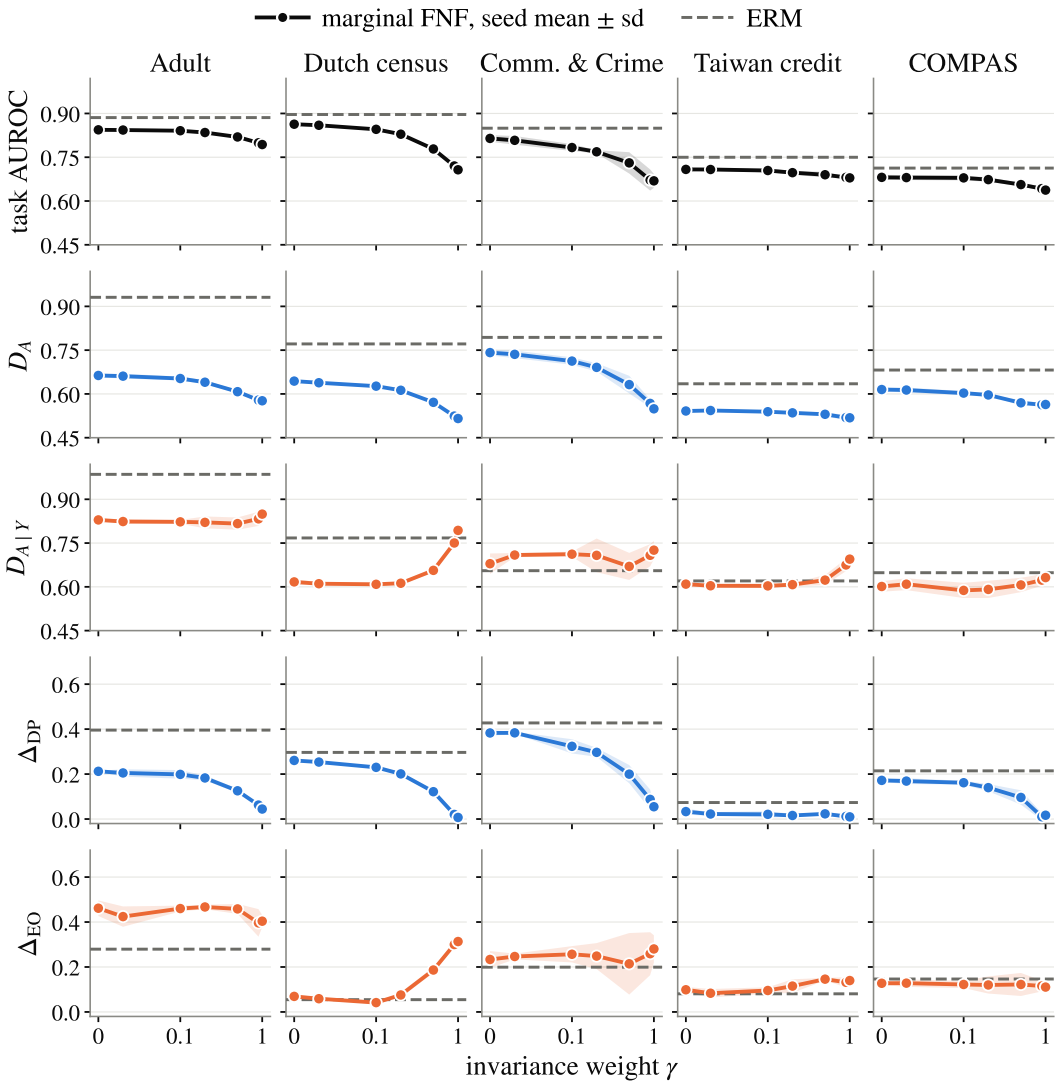


Figure 9: Marginal alignment method as a function of the invariance strength γ (*shift-50*, selected feature set, MADE width 50). Reported mean ± std over seeds.

Table 5: Optimal Youden threshold τ (for validation) vs. a default 0.5 (*shift 50*, reported mean ± std over all seeds).

Dataset	Method	Youden τ	Acc (τ)	Δ_{DP} (τ)	Δ_{EO} (τ)	Acc (0.5)	Δ_{DP} (0.5)	Δ_{EO} (0.5)
Adult	ERM	0.22 ± 0.02	0.78 ± 0.02	0.40 ± 0.03	0.28 ± 0.02	0.84 ± 0.00	0.24 ± 0.01	0.39 ± 0.03
Adult	$Z \perp A$	0.17 ± 0.02	0.69 ± 0.02	0.04 ± 0.00	0.40 ± 0.03	0.78 ± 0.00	0.01 ± 0.01	0.06 ± 0.03
Adult	$Z \perp A \mid Y$	0.17 ± 0.00	0.75 ± 0.02	0.17 ± 0.01	0.08 ± 0.04	0.82 ± 0.00	0.11 ± 0.01	0.06 ± 0.01
Dutch census	ERM	0.43 ± 0.02	0.82 ± 0.00	0.30 ± 0.00	0.05 ± 0.00	0.83 ± 0.00	0.29 ± 0.00	0.06 ± 0.00
Dutch census	$Z \perp A$	0.40 ± 0.04	0.64 ± 0.00	0.01 ± 0.01	0.31 ± 0.03	0.65 ± 0.01	0.01 ± 0.01	0.22 ± 0.03
Dutch census	$Z \perp A \mid Y$	0.41 ± 0.04	0.81 ± 0.00	0.27 ± 0.01	0.03 ± 0.01	0.81 ± 0.00	0.28 ± 0.01	0.02 ± 0.01
Comm. & Crime	ERM	0.45 ± 0.03	0.77 ± 0.01	0.43 ± 0.02	0.20 ± 0.02	0.75 ± 0.01	0.39 ± 0.01	0.27 ± 0.00
Comm. & Crime	$Z \perp A$	0.44 ± 0.03	0.62 ± 0.02	0.05 ± 0.01	0.28 ± 0.06	0.61 ± 0.02	0.04 ± 0.03	0.19 ± 0.04
Comm. & Crime	$Z \perp A \mid Y$	0.50 ± 0.02	0.75 ± 0.03	0.40 ± 0.07	0.28 ± 0.16	0.76 ± 0.03	0.40 ± 0.07	0.25 ± 0.18
Taiwan credit	ERM	0.18 ± 0.02	0.76 ± 0.01	0.07 ± 0.00	0.08 ± 0.00	0.85 ± 0.00	0.05 ± 0.00	0.03 ± 0.00
Taiwan credit	$Z \perp A$	0.20 ± 0.03	0.72 ± 0.03	0.01 ± 0.00	0.14 ± 0.03	0.82 ± 0.00	0.00 ± 0.00	0.02 ± 0.01
Taiwan credit	$Z \perp A \mid Y$	0.17 ± 0.03	0.73 ± 0.05	0.04 ± 0.01	0.01 ± 0.01	0.85 ± 0.00	0.03 ± 0.01	0.02 ± 0.02
COMPAS	ERM	0.39 ± 0.04	0.64 ± 0.01	0.21 ± 0.01	0.15 ± 0.01	0.68 ± 0.00	0.20 ± 0.01	0.14 ± 0.01
COMPAS	$Z \perp A$	0.41 ± 0.03	0.62 ± 0.01	0.02 ± 0.01	0.11 ± 0.02	0.62 ± 0.00	0.02 ± 0.02	0.05 ± 0.03
COMPAS	$Z \perp A \mid Y$	0.39 ± 0.07	0.66 ± 0.01	0.10 ± 0.02	0.04 ± 0.02	0.65 ± 0.00	0.08 ± 0.00	0.05 ± 0.01

Table 6: **Dataset statistics and prevalence shift configurations.** Binary groups ($A{=}0$ vs. $A{=}1$) and prediction targets ($Y{=}1$) are: *Adult*: female vs. male, income >50K; *Dutch*: female vs. male, high-level occupation; *Crime*: Black population share below vs. at or above median, violent crime rate above median; *Taiwan credit*: female vs. male, default next month; *COMPAS*: Caucasian vs. African-American, two-year recidivism. Rightmost columns report the training-split base rates (π_0 / π_1) across synthetic shift regimes. Selected feature sets: Adult `paper6_age` (workclass, education, marital status, occupation, relationship, race, binned age), Dutch `core6_hh` (age, education, economic status, current activity, marital status, household position, household size), Crime `socio4` (median income, poverty, unemployment, two-parent families; quantile-binned), Credit `core5` (education, marriage, binned age, last repayment status, binned limit), COMPAS `paper7` (sex, age category, charge degree, binned priors, three juvenile counts).

			Training Prevalences (π_0 / π_1)			
Dataset	**Samples**	**Natural** π_0/π_1	**Equalized**	**Natural**	**Shift-50**	**Shift-75**
Adult income	48,842 (29,312 / 4,880 / 4,885 / 9,765)	0.11 / 0.30	0.11 / 0.11	0.11 / 0.31	0.06 / 0.31	0.03 / 0.31
Dutch census	60,420 (36,252 / 6,042 / 6,042 / 12,084)	0.33 / 0.63	0.33 / 0.33	0.33 / 0.62	0.20 / 0.62	0.11 / 0.62
Comm. & Crime	1,994 (1,196 / 199 / 199 / 400)	0.27 / 0.71	0.26 / 0.26	0.26 / 0.69	0.15 / 0.69	0.08 / 0.69
Taiwan credit	30,000 (18,000 / 3,000 / 3,000 / 6,000)	0.21 / 0.24	0.21 / 0.21	0.21 / 0.25	0.12 / 0.25	0.06 / 0.25
COMPAS	5,278 (3,166 / 528 / 528 / 1,056)	0.39 / 0.52	0.39 / 0.39	0.39 / 0.54	0.24 / 0.54	0.14 / 0.54

Communities & Crime (400 test rows, 38–151 positives in $A{=}0$)/ It is the only dataset where the training seed variation of Δ_{EO} and Δ_{PPV} is comparable to the method effect.

Table 7: Detailed results for the *shift-50* prevalence condition per selected seeds.

Dataset	Method	Seed	AUROC	D_A	$D_{A\|Y}$	Δ_{DP}	Δ_{EO}	Δ_{PPV}	Δ_{bias}
Adult	ERM	42	0.89	0.93	0.99	0.42	0.30	0.17	0.00
Adult	ERM	123	0.89	0.93	0.99	0.36	0.27	0.19	0.02
Adult	ERM	456	0.89	0.93	0.99	0.40	0.27	0.18	0.02
Adult	M-FNF ($Z \perp A$)	42	0.79	0.58	0.86	0.04	0.37	0.48	0.23
Adult	M-FNF ($Z \perp A$)	123	0.79	0.57	0.86	0.05	0.43	0.47	0.23
Adult	M-FNF ($Z \perp A$)	456	0.79	0.58	0.83	0.04	0.41	0.51	0.23
Adult	C-FNF ($Z \perp A \mid Y$)	42	0.84	0.66	0.83	0.21	0.41	0.47	0.11
Adult	C-FNF ($Z \perp A \mid Y$)	123	0.84	0.65	0.78	0.20	0.41	0.48	0.11
Adult	C-FNF ($Z \perp A \mid Y$)	456	0.84	0.66	0.81	0.21	0.48	0.50	0.11
Dutch census	ERM	42	0.90	0.77	0.76	0.30	0.05	0.26	0.22
Dutch census	ERM	123	0.90	0.77	0.77	0.30	0.05	0.29	0.22
Dutch census	ERM	456	0.90	0.77	0.77	0.30	0.05	0.27	0.21
Dutch census	M-FNF ($Z \perp A$)	42	0.71	0.51	0.79	0.01	0.28	0.63	0.44
Dutch census	M-FNF ($Z \perp A$)	123	0.70	0.51	0.80	0.00	0.32	0.63	0.44
Dutch census	M-FNF ($Z \perp A$)	456	0.71	0.52	0.79	0.01	0.33	0.63	0.45
Dutch census	C-FNF ($Z \perp A \mid Y$)	42	0.87	0.64	0.61	0.26	0.07	0.29	0.27
Dutch census	C-FNF ($Z \perp A \mid Y$)	123	0.86	0.64	0.61	0.26	0.06	0.29	0.27
Dutch census	C-FNF ($Z \perp A \mid Y$)	456	0.86	0.64	0.62	0.26	0.06	0.29	0.27
Comm. & Crime	ERM	42	0.85	0.79	0.66	0.43	0.18	0.50	0.23
Comm. & Crime	ERM	123	0.85	0.79	0.66	0.41	0.23	0.52	0.23
Comm. & Crime	ERM	456	0.85	0.79	0.66	0.44	0.19	0.50	0.23
Comm. & Crime	M-FNF ($Z \perp A$)	42	0.67	0.53	0.76	0.05	0.35	0.74	0.55
Comm. & Crime	M-FNF ($Z \perp A$)	123	0.65	0.57	0.72	0.06	0.27	0.73	0.56
Comm. & Crime	M-FNF ($Z \perp A$)	456	0.69	0.54	0.70	0.05	0.23	0.66	0.55
Comm. & Crime	C-FNF ($Z \perp A \mid Y$)	42	0.81	0.71	0.74	0.36	0.18	0.54	0.32
Comm. & Crime	C-FNF ($Z \perp A \mid Y$)	123	0.82	0.73	0.73	0.40	0.28	0.55	0.31
Comm. & Crime	C-FNF ($Z \perp A \mid Y$)	456	0.82	0.74	0.56	0.39	0.31	0.56	0.30
Taiwan credit	ERM	42	0.75	0.63	0.62	0.07	0.09	0.20	0.08
Taiwan credit	ERM	123	0.75	0.63	0.62	0.08	0.08	0.19	0.08
Taiwan credit	ERM	456	0.75	0.64	0.62	0.07	0.08	0.20	0.08
Taiwan credit	M-FNF ($Z \perp A$)	42	0.67	0.52	0.67	0.01	0.16	0.29	0.11
Taiwan credit	M-FNF ($Z \perp A$)	123	0.68	0.52	0.72	0.01	0.11	0.24	0.11
Taiwan credit	M-FNF ($Z \perp A$)	456	0.69	0.51	0.69	0.01	0.15	0.31	0.11
Taiwan credit	C-FNF ($Z \perp A \mid Y$)	42	0.71	0.54	0.60	0.03	0.07	0.26	0.09
Taiwan credit	C-FNF ($Z \perp A \mid Y$)	123	0.70	0.54	0.62	0.04	0.08	0.25	0.09
Taiwan credit	C-FNF ($Z \perp A \mid Y$)	456	0.72	0.54	0.57	0.03	0.08	0.25	0.10
COMPAS	ERM	42	0.71	0.68	0.65	0.20	0.13	0.23	0.13
COMPAS	ERM	123	0.72	0.69	0.65	0.22	0.16	0.23	0.13
COMPAS	ERM	456	0.71	0.68	0.65	0.22	0.15	0.24	0.14
COMPAS	M-FNF ($Z \perp A$)	42	0.63	0.56	0.65	0.02	0.13	0.38	0.24
COMPAS	M-FNF ($Z \perp A$)	123	0.64	0.58	0.63	0.01	0.10	0.26	0.25
COMPAS	M-FNF ($Z \perp A$)	456	0.64	0.56	0.61	0.02	0.09	0.35	0.24
COMPAS	C-FNF ($Z \perp A \mid Y$)	42	0.66	0.61	0.62	0.17	0.12	0.20	0.18
COMPAS	C-FNF ($Z \perp A \mid Y$)	123	0.69	0.63	0.62	0.18	0.12	0.21	0.17
COMPAS	C-FNF ($Z \perp A \mid Y$)	456	0.68	0.63	0.62	0.18	0.12	0.21	0.17

A.5 Additional Details: Calibration by groups

Table 8 reports the calibration by groups results for the selected models under *shift-50* prevalence condition. Under marginal alignment: the low-prevalence group $A{=}0$ becomes strongly 'over-confident' (Adult b_0: $0.96 \rightarrow 0.06$; Dutch $0.98 \rightarrow 0.51$) and the high-prevalence group 'under-confident' ($b_1 = 1.34$–1.84), the intercepts can be seen in opposite directions ($a_0 < 0 < a_1$), and Δ_{cal} increases for all datasets, as anticipated by P_2. Similar dependency on training prevalence is also discussed by Kina & Petersen (2026) for calibration by groups: a model fitted to a $P(Y \mid A)$ carries that prevalence, and the mismatch may appears as group-specific miscalibration under Q. . Class-conditional alignment brings back both slopes to what they were for ERM but leaves the intercept gap at par with or above ERM (P_4). Under prevalence shift at deployment (Fig. 5), Δ_{bias} increases for every method, including at ERM, because a score calibrated to the training prevalence π_a is off by $\operatorname{logit} \pi_a^Q - \operatorname{logit} \pi_a$ in the large once the test prevalence is π_a^Q; when the shift is group-specific, this gap differs between groups. Per-run values for all $2{,}040$ evaluations are summarized in the table.

Table 8: **Calibration by groups at *shift 50*:** 15-bin ECE, Cox recalibration slope (Cox, 1958) b_g and fixed-slope intercept a_g (logits), signed mean bias $\bar{y}_g - \bar{p}_g$, and the two between-group gaps Δ_{bias} and Δ_{cal}. Methods: ERM, **M** (marginal alignment, $Z \perp A$), and **C** (class-conditional alignment, $Z \perp A \mid Y$). Datasets: Adult, Dutch census (Dutch), Communities & Crime (Crime), Taiwan credit (Credit), and COMPAS. Reported as mean $\pm$ std over all three seeds.

Dataset	Method	$\mathbf{ECE}_0$	$\mathbf{ECE}_1$	b_0	b_1	a_0	a_1	$\bar{y}_0 - \bar{p}_0$	$\bar{y}_1 - \bar{p}_1$	Δ_{bias}	Δ_{cal}
Adult	ERM	0.02 ± 0.00	0.02 ± 0.00	0.96 ± 0.03	0.85 ± 0.01	-0.29 ± 0.08	0.03 ± 0.12	-0.013 ± 0.004	0.004 ± 0.016	0.02 ± 0.01	0.32 ± 0.06
	M	0.13 ± 0.01	0.11 ± 0.01	0.06 ± 0.04	1.34 ± 0.05	-1.36 ± 0.06	0.75 ± 0.06	-0.116 ± 0.007	0.112 ± 0.009	0.23 ± 0.00	2.12 ± 0.01
	C	0.08 ± 0.01	0.08 ± 0.01	0.79 ± 0.03	0.88 ± 0.02	-1.20 ± 0.14	0.60 ± 0.09	-0.077 ± 0.013	0.075 ± 0.010	0.15 ± 0.00	1.80 ± 0.07
Dutch	ERM	0.12 ± 0.00	0.10 ± 0.00	1.01 ± 0.02	0.90 ± 0.01	-1.24 ± 0.05	0.75 ± 0.02	-0.120 ± 0.004	0.095 ± 0.003	0.22 ± 0.00	2.00 ± 0.03
	M	0.23 ± 0.00	0.22 ± 0.01	0.51 ± 0.08	1.65 ± 0.06	-1.25 ± 0.01	1.07 ± 0.04	-0.224 ± 0.003	0.219 ± 0.007	0.44 ± 0.00	2.32 ± 0.03
	C	0.12 ± 0.01	0.11 ± 0.00	0.98 ± 0.03	0.91 ± 0.02	-1.16 ± 0.05	0.89 ± 0.04	-0.121 ± 0.006	0.114 ± 0.005	0.23 ± 0.01	2.05 ± 0.07
Crime	ERM	0.18 ± 0.02	0.12 ± 0.00	0.45 ± 0.01	0.74 ± 0.01	-1.28 ± 0.04	0.82 ± 0.01	-0.126 ± 0.006	0.104 ± 0.002	0.23 ± 0.00	2.10 ± 0.03
	M	0.28 ± 0.02	0.29 ± 0.00	0.43 ± 0.53	1.48 ± 0.18	-1.46 ± 0.01	1.41 ± 0.02	-0.263 ± 0.003	0.290 ± 0.005	0.55 ± 0.01	2.87 ± 0.03
	C	0.17 ± 0.04	0.16 ± 0.01	0.71 ± 0.26	0.74 ± 0.01	-1.21 ± 0.24	1.13 ± 0.01	-0.137 ± 0.026	0.153 ± 0.007	0.29 ± 0.03	2.34 ± 0.24
Credit	ERM	0.03 ± 0.00	0.05 ± 0.00	0.96 ± 0.01	0.95 ± 0.00	-0.34 ± 0.02	0.35 ± 0.02	-0.034 ± 0.003	0.046 ± 0.002	0.08 ± 0.00	0.68 ± 0.01
	M	0.06 ± 0.00	0.05 ± 0.00	0.50 ± 0.05	1.22 ± 0.01	-0.50 ± 0.02	0.36 ± 0.02	-0.058 ± 0.003	0.053 ± 0.003	0.11 ± 0.00	0.86 ± 0.02
	C	0.05 ± 0.00	0.05 ± 0.00	0.94 ± 0.02	0.89 ± 0.01	-0.44 ± 0.02	0.35 ± 0.01	-0.045 ± 0.002	0.046 ± 0.002	0.09 ± 0.00	0.79 ± 0.02
COMPAS	ERM	0.13 ± 0.00	0.06 ± 0.01	0.63 ± 0.01	0.75 ± 0.03	-0.69 ± 0.02	0.09 ± 0.02	-0.116 ± 0.003	0.018 ± 0.004	0.13 ± 0.00	0.78 ± 0.01
	M	0.18 ± 0.01	0.10 ± 0.01	0.52 ± 0.14	0.86 ± 0.07	-0.81 ± 0.02	0.38 ± 0.01	-0.157 ± 0.004	0.086 ± 0.002	0.24 ± 0.00	1.19 ± 0.01
	C	0.13 ± 0.01	0.11 ± 0.01	0.78 ± 0.01	0.62 ± 0.02	-0.72 ± 0.07	0.37 ± 0.04	-0.125 ± 0.013	0.073 ± 0.006	0.20 ± 0.01	1.09 ± 0.05

A.6 Additional Details: Score-level alignment

We constrain the score distribution, whereas alignment methods constrain the representation. We look for (i) whether aligning the score R directly, rather than Z, changes the outcome, and (ii) whether a weaker constraint that matches only the class-conditional *mean* of the score suffices for separation. We evaluate this on Dutch at *shift-50* with the ERM model (Table 9).

Score-level alignment. It reproduce the representation-level results: marginal alignment gives $\Delta_{\mathrm{DP}} = 0.006$ with $\Delta_{\mathrm{EO}} = 0.26$ and $\Delta_{\mathrm{bias}} = 0.44$ (FNF: 0.007, 0.31, 0.44), and within-class alignment gives $\Delta_{\mathrm{EO}} = 0.024$ (C-FNF: 0.028) with Δ_{DP} and Δ_{bias} at the ERM level. Aligning one scalar costs less utility than aligning Z (AUROC 0.81 versus 0.71 for the marginal target), but the fairness conclusions remain unchanged, as expected for statements about $p(r \mid a)$ and $p(r \mid y, a)$.

Conditional mean matching. We retrain the ERM head with the penalty $\lambda \sum_y \big(\mathbb{E}[R \mid Y{=}y, A{=}0] - \mathbb{E}[R \mid Y{=}y, A{=}1]\big)^2$ estimated per mini-batch, $\lambda \in \{1, 10, 100\}$. The class-conditional mean gaps fall from $+0.041$ ($y{=}0$) and -0.028 ($y{=}1$) under ERM to 0.000 on train and test for $\lambda \geq 100$ (also at $\lambda = 10^3, 10^4$). It does not deliver separation: Δ_{EO} stays at 0.06–0.08 (ERM 0.055) and $D_{A|Y}$ does not fall, because the group-specific shape of $p(r \mid y, a)$ beyond its mean is untouched. It does move calibration: Δ_{bias} rises to 0.41.

Table 9: Score-level alignment and class-conditional mean matching on Dutch, *shift-50*, test split, mean over three seeds.

Method	**AUROC**	D_A	$D_{A\|Y}$	Δ_{DP}	Δ_{EO}	Δ_{PPV}	Δ_{NPV}	Δ_{bias}	Δ_{cal}
ERM	0.896	0.70	0.64	0.296	0.055	0.27	0.28	0.22	2.00
Score $R \perp A$	0.813	0.54	0.77	0.006	0.258	0.55	0.35	0.44	3.68
Score $R \perp A \mid Y$ (within-class map)	0.881	0.67	0.59	0.257	0.024	0.36	0.28	0.24	2.11
Mean matching, $\lambda = 1$	0.896	0.70	0.63	0.291	0.051	0.29	0.28	0.22	1.95
Mean matching, $\lambda = 10$	0.889	0.69	0.66	0.270	0.082	0.27	0.30	0.28	1.79
Mean matching, $\lambda = 100$	0.883	0.69	0.66	0.266	0.081	0.30	0.30	0.41	1.87
FNF $Z \perp A$ ($\gamma = 1$), repr. level	0.707	0.52^*	0.79^*	0.007	0.313	0.63	0.21	0.44	2.32
C-FNF $Z \perp A \mid Y$ ($\gamma = 1$), repr. level	0.888	0.68^*	0.56^*	0.275	0.028	0.34	0.27	0.23	2.05

B Appendix 2: CheXpert

B.1 Additional Details: Dataset details and cohorts

We perform a classification task on Edema and Cardiomegaly from frontal CheXpert radiographs using recorded sex ($A = 0$: female; $A = 1$: male). Images with unknown sex are excluded; Uncertain pathology labels are positive, and blank labels are negative. The *shift50* intervention removes half of female-positive images separately within each partition. Both training and evaluation use the shifted distribution; this experiment does not test a new deployment shift. Table 10 reports the retained cohorts.

Table 10: *shift50* cohorts. Patient partitions are fixed across seeds; image-level deletion changes the retained patient counts.

Task	Partition	Images	Patients	π_0	π_1
Edema	Train	107,688	37,580	0.198	0.313
	Checkpoint	9,050	3,134	0.195	0.325
	Development	8,872	3,138	0.202	0.333
	Probe val.	17,186	6,271	0.206	0.310
	Test	35,078	12,551	0.205	0.309
Cardiomegaly	Train	111,980	38,082	0.081	0.159
	Checkpoint	9,363	3,175	0.093	0.161
	Development	9,225	3,168	0.082	0.173
	Probe val.	17,896	6,351	0.085	0.168
	Test	36,525	12,693	0.085	0.160

Similar to the tabular setting, we use three training seeds (42, 123, 456). ImageNet-init DenseNet121 models with 256×256 RGB input and standard normalization. We use BCE, AdamW (lr 10^{-4}, decay of 10^{-5}), a batch size of 32, and at most 30 epochs with early stopping. It uses a patience of 7 epochs after the first 10 epochs and minimizes validation BCE.

We adapt continuous Fair Normalizing Flows (Balunović et al., 2022). Marginal training uses the analogous marginal KL divided by d. We use $\gamma \in \{.05, .50, .95\}$, 60 epochs, batch size 256, AdamW (learning rate .001, weight decay .0001), gradient clipping at five, and a ten-epoch ramp. Each KL direction uses 128 samples from its fitted prior. Checkpoint selection (passing criteria) minimizes validation MMD (worst-label conditional MMD for conditional FNF), subject to retaining at least 90% of ERM validation AUROC. All 36 FNF fits completed without fallback; all fitted GMMs converged. The existing MMD and adversarial-adapter experiments are ancillary comparisons; the marginal-versus-conditional analysis here holds the FNF architecture fixed. Results are detailed in Tables 11 and 12.

Evaluation. The frozen representation Z and scalar score R are audited separately, both marginally and within each outcome class. Probes use up to 40,000 training examples; the RBF-SVM uses at most 2,000 per sex. Hyperparameters and MLP early stopping use only probe-validation balanced BCE. Decision metrics use a common, sex-independent threshold selected by validation Youden's index. Calibration uses 15 equal-width ECE bins and logistic recalibration.

B.2 Additional Details: Probe Results

Strong conditional alignment did reduce conditional representation probe AUROC from $\approx .978$ under ERM to .590 (Edema) and .616 (Cardiomegaly), but none of the 18 conditional representations pass the marginal or conditional near-chance AUROC criterion (which is, in fact is unrealistic for medical images. Figure 11 shows why a single probe family is insufficient. Score-level results are stronger: conditional FNF at $\gamma = .95$ passes all three test criteria for Edema in 3/3 seeds. Neither task passes the conditional AUROC criterion (Table 13), but are sufficient to support our theoretical findings.

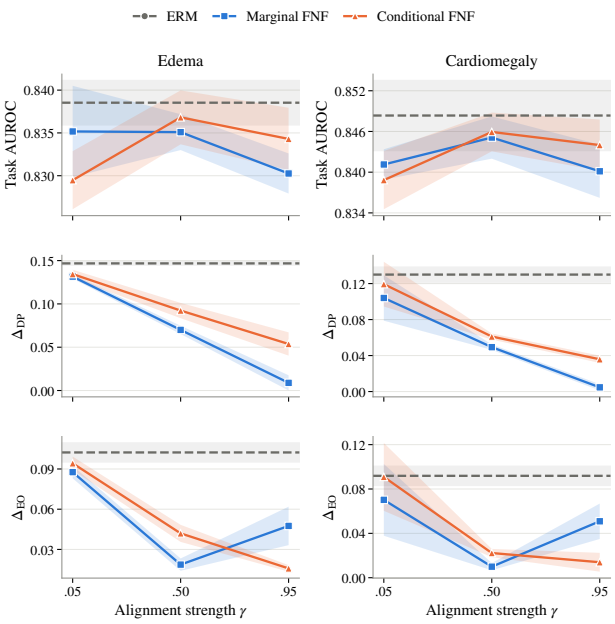


Figure 10: **Invariance strength changes the fairness trade-off.** Lines and bands show means ± std. Conditional alignment lowers EO gaps at strong alignment, while marginal alignment produces smaller DP gaps.

Table 11: **Fairness and calibration for every invariance method strength** (mean ± seed reported). Δ_{bias} is the absolute between-sex difference in mean residual $Y - R$; Δ_{cal} is the absolute difference in fixed-slope calibration intercepts.

Task	Method	Δ_{DP}	Δ_{EO}	Max ECE	Δ_{bias}	Δ_{cal}
Edema	ERM	0.147 ± 0.002	0.102 ± 0.007	0.020 ± 0.009	0.013 ± 0.006	0.092 ± 0.037
	M .05	0.132 ± 0.002	0.088 ± 0.005	0.037 ± 0.003	0.018 ± 0.005	0.160 ± 0.043
	M .50	0.070 ± 0.005	0.019 ± 0.005	0.037 ± 0.002	0.062 ± 0.004	0.504 ± 0.012
	M .95	0.009 ± 0.009	0.048 ± 0.014	0.069 ± 0.006	0.101 ± 0.006	0.797 ± 0.071
	C .05	0.134 ± 0.005	0.094 ± 0.005	0.034 ± 0.009	0.020 ± 0.009	0.165 ± 0.064
	C .50	0.092 ± 0.009	0.042 ± 0.006	0.039 ± 0.004	0.046 ± 0.008	0.387 ± 0.055
	C .95	0.054 ± 0.013	0.016 ± 0.002	0.051 ± 0.006	0.072 ± 0.007	0.582 ± 0.029
Cardiomegaly	ERM	0.130 ± 0.008	0.092 ± 0.009	0.022 ± 0.003	0.013 ± 0.008	0.188 ± 0.102
	M .05	0.104 ± 0.025	0.070 ± 0.032	0.030 ± 0.006	0.019 ± 0.006	0.266 ± 0.087
	M .50	0.049 ± 0.003	0.010 ± 0.003	0.031 ± 0.006	0.044 ± 0.002	0.635 ± 0.032
	M .95	0.005 ± 0.003	0.051 ± 0.016	0.050 ± 0.002	0.070 ± 0.001	0.962 ± 0.018
	C .05	0.119 ± 0.025	0.091 ± 0.031	0.028 ± 0.008	0.016 ± 0.008	0.246 ± 0.110
	C .50	0.061 ± 0.003	0.022 ± 0.004	0.027 ± 0.001	0.041 ± 0.002	0.592 ± 0.018
	C .95	0.036 ± 0.004	0.014 ± 0.008	0.035 ± 0.001	0.052 ± 0.001	0.737 ± 0.008

Table 12: **All settings:** utility and strongest demographic probe AUROC (mean ± sample std over all training seeds). M/C denote marginal/conditional objectives, followed by γ.

Task	Method	AUROC	AUPRC	$D_A(Z)$	$D_{A\|Y}(Z)$	$D_{A\|Y}(R)$
Edema	ERM	0.839 ± 0.003	0.646 ± 0.007	0.978 ± 0.002	0.978 ± 0.002	0.613 ± 0.019
	M .05	0.835 ± 0.005	0.641 ± 0.011	0.693 ± 0.022	0.678 ± 0.022	0.583 ± 0.011
	M .50	0.835 ± 0.002	0.640 ± 0.006	0.631 ± 0.034	0.624 ± 0.029	0.519 ± 0.008
	M .95	0.830 ± 0.002	0.630 ± 0.005	0.623 ± 0.042	0.629 ± 0.043	0.555 ± 0.008
	C .05	0.829 ± 0.003	0.632 ± 0.008	0.675 ± 0.013	0.664 ± 0.024	0.582 ± 0.010
	C .50	0.837 ± 0.003	0.642 ± 0.007	0.613 ± 0.004	0.599 ± 0.013	0.537 ± 0.010
	C .95	0.834 ± 0.004	0.636 ± 0.008	0.596 ± 0.016	0.590 ± 0.023	0.513 ± 0.003
Cardiomegaly	ERM	0.848 ± 0.005	0.531 ± 0.010	0.978 ± 0.003	0.978 ± 0.002	0.602 ± 0.009
	M .05	0.841 ± 0.002	0.514 ± 0.009	0.703 ± 0.019	0.696 ± 0.023	0.607 ± 0.016
	M .50	0.845 ± 0.003	0.524 ± 0.006	0.650 ± 0.021	0.636 ± 0.025	0.551 ± 0.011
	M .95	0.840 ± 0.004	0.507 ± 0.011	0.644 ± 0.028	0.646 ± 0.023	0.533 ± 0.008
	C .05	0.839 ± 0.004	0.507 ± 0.014	0.673 ± 0.012	0.663 ± 0.012	0.592 ± 0.019
	C .50	0.846 ± 0.003	0.526 ± 0.008	0.630 ± 0.005	0.623 ± 0.004	0.537 ± 0.002
	C .95	0.844 ± 0.004	0.521 ± 0.008	0.617 ± 0.002	0.616 ± 0.004	0.515 ± 0.001

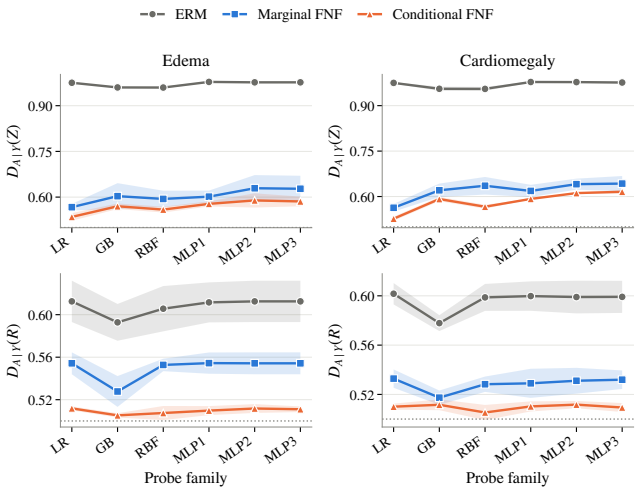


Figure 11: **Conditional probe-family audit at $\gamma = .95$.** Top: representations; bottom: scores.

Table 13: **Conditional score audit.** Upper bounds use the within-scope Bonferroni correction; passing upper bound ≤ 0.55. All three Edema test runs also pass the balanced-accuracy and log-loss interval checks.

Task	Seed	Test $D_{A\|Y}(R)$	Test upper	Dev. upper	Test pass
Edema	42	0.515	0.540	0.591	Yes
	123	0.510	0.541	0.563	Yes
	456	0.514	0.545	0.572	Yes
Cardiomegaly	42	0.514	0.557	0.637	No
	123	0.516	0.559	0.618	No
	456	0.515	0.558	0.606	No

B.3 Additional Details: Calibration and Operating-Point Analysis

At $\gamma = .95$, conditional FNF improves mean AUROC, EO, and worst-group ECE relative to marginal FNF on both tasks, but increases DP gaps. Detailed in Table 14; Fig. 12. The EO comparison depends on the operating point (Fig. 13). All results support our propositions, though extreme thresholds can shrink gaps by predicting the same label for everyone. Stronger alignment is therefore not preferable, and reduced gaps alone do not establish improved predictions.

Table 14: Sex-calibration at $\gamma = .95$ (mean $\pm$ std reported). α_a is fitted separately with slope fixed at one. Negative intercepts and negative mean residuals $\Longrightarrow$ overprediction. Ideal slope is 1; ideal intercept and mean residual are 0.

Task	Method	A	ECE	Slope b_a	Intercept α_a	$\bar{y}_a - \bar{r}_a$
Edema	ERM	0	0.010 ± 0.005	0.978 ± 0.054	0.001 ± 0.034	-0.000 ± 0.004
		1	0.020 ± 0.009	0.926 ± 0.044	0.093 ± 0.071	0.013 ± 0.010
	M .95	0	0.069 ± 0.006	0.875 ± 0.037	-0.561 ± 0.057	-0.069 ± 0.006
		1	0.036 ± 0.004	0.858 ± 0.056	0.236 ± 0.030	0.032 ± 0.003
	C .95	0	0.051 ± 0.006	0.867 ± 0.079	-0.416 ± 0.036	-0.050 ± 0.007
		1	0.029 ± 0.007	0.861 ± 0.076	0.166 ± 0.043	0.023 ± 0.005
Cardiomegaly	ERM	0	0.015 ± 0.004	0.923 ± 0.083	-0.136 ± 0.142	-0.009 ± 0.010
		1	0.018 ± 0.008	0.870 ± 0.102	0.052 ± 0.111	0.005 ± 0.009
	M .95	0	0.050 ± 0.002	0.851 ± 0.030	-0.718 ± 0.031	-0.050 ± 0.002
		1	0.023 ± 0.001	0.864 ± 0.020	0.244 ± 0.025	0.021 ± 0.002
	C .95	0	0.035 ± 0.001	0.857 ± 0.036	-0.534 ± 0.008	-0.035 ± 0.001
		1	0.024 ± 0.003	0.818 ± 0.018	0.204 ± 0.016	0.017 ± 0.001

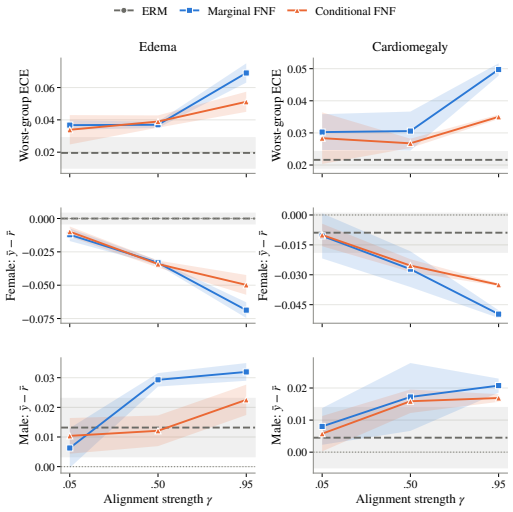


Figure 12: **Calibration costs vs. improved EO.** Means $\pm$ std across strengths.

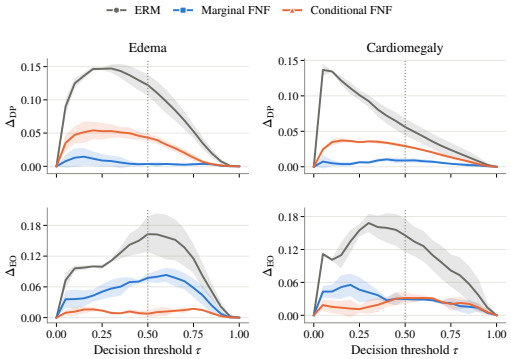


Figure 13: **Fairness gaps dependency on threshold.** ERM and FNF at $\gamma = .95$, evaluated on the same test cohorts. Bands show training seed std; the vertical dotted line marks the default .5 threshold.

Table 15: **Evaluation across pathologies and objectives.** Task AUROC, marginal and class-conditional sex probe AUROCs, demographic parity (DP) gap, and equalized odds (EO) gap. Results are reported as mean ± std.

			Sex Probe AUROC ↓			Fairness Gaps ↓	
Pathology	**Method**	**Task AUROC ↑**	**Marginal**	$Y = 0$	$Y = 1$	**DP Gap**	**EO Gap**
Edema	DANN	0.842 ± 0.003	0.950 ± 0.005	0.950 ± 0.006	0.938 ± 0.008	0.069 ± 0.023	0.033 ± 0.025
	CDANN	0.845 ± 0.003	0.952 ± 0.006	0.955 ± 0.007	0.940 ± 0.012	0.111 ± 0.015	0.074 ± 0.025
Cardiomegaly	DANN	0.842 ± 0.004	0.947 ± 0.006	0.951 ± 0.006	0.928 ± 0.018	0.054 ± 0.031	0.037 ± 0.015
	CDANN	0.843 ± 0.006	0.944 ± 0.030	0.957 ± 0.008	0.924 ± 0.027	0.105 ± 0.028	0.078 ± 0.042

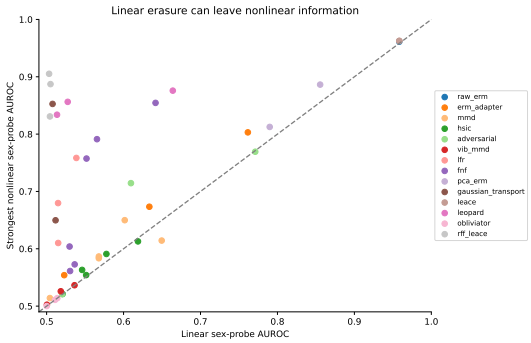
C Appendix 3: Fairness Metrics Do Not Establish Invariance

Failure to achieve demographic erasure using popular methods. We initially used more popular methods, specifically DANN (Ganin et al., 2016) and class-conditional DANN Long et al. (2018). Neither achieved its objective, i.e., to erase demographic information in our CheXpert experiments. Across eight adversary strengths and five seeds, probes were still able to recover demographic information from learned representations. Fairness gaps can sometimes decrease (and even be zero) because diagnostic performance also decreases as a function of demographic erasure. We leave it to future work to investigate why certain methods fail to erase demographic signals even with strong constraints while others succeed, and why it is *easy* on some datasets and not on others. This also shows that improved fairness metrics alone do not establish invariance. Our broader exploratory comparison revealed substantial variation across methods and hyperparameters: some reduced demographic decodability while retaining useful prediction, whereas others retained demographic information or substantially reduced predictive performance. Moreover, nonlinear probes recovered information that linear probes missed, highlighting the importance of evaluating erasure with multiple probe families (also discussed in Appendix 2).

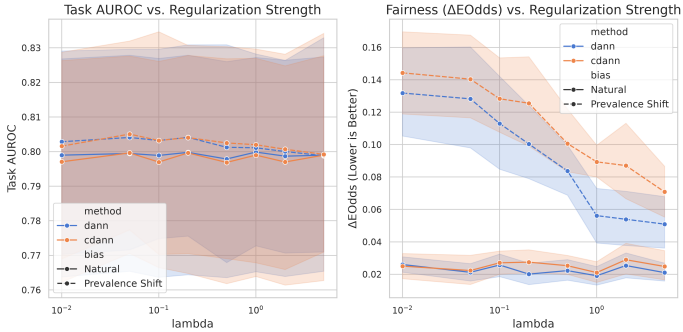
Beyond DANN and CDANN, we also explored MMD (Gretton et al., 2006) and HSIC (Pérez-Suay et al., 2017) penalties during training and their alternative variants, such as adversarial ensemble variational bottlenecks with MMD (Gretton et al., 2006), learning Fair Representations (LFR) (Zemel et al., 2013), Gaussian moment transport, LEACE (Belrose et al., 2023), LEOPARD (Ravfogel et al., 2022), Obliviate (Shakibania et al., 2026), random Fourier features (Rahimi & Recht, 2007) followed by LEACE. Across 51 configurations in the single-seed Edema study, LFR, FNF, Gaussian transport, LEACE, LEOPARD, and RFF–LEACE did not meet our expectation to reduce demographic decodability. As the probe-comparison plot illustrates in Fig C, low linear decodability sometimes hides nonlinear leakage. Some methods like MMD and HSIC did achieve near-chance decodability but with reduced task utility (classification AUROC). Similarly, Obliviator also costs predictive performance; variational bottlenecks and gender-aware channels retained decodable information. An analysis is detailed in Table 16. Note that this analysis does not reflect a limitation or general failure of the methods.

C.1 Additional Details: Clarifications

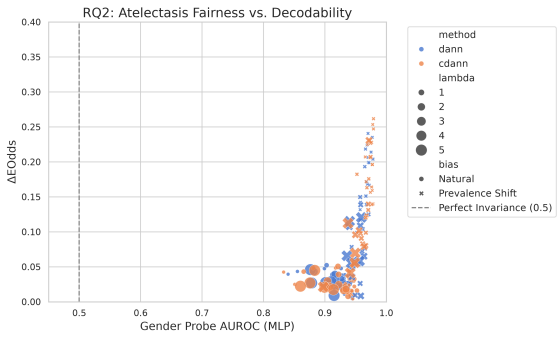
Regarding using sensitive attributes A. In Section. 2. Acquisition variables such as scanner, site, or protocol are nuisance factors that can affect Z, which is unwanted, and are handled by domain-invariant methods. We use A to denote demographic attributes in this work that apply to people; usage and interpretation may differ if we consider nuisance variables.



(a) Nonlinear probes reveal demographic information missed by linear probes.



(b) Diagnostic AUROC and Δ_{EO} across adversarial strengths.



(c) Atelectasis: small fairness gaps coexist with strong demographic decodability.

Figure 14: **Evaluating demographic erasure on CheXpert.** **(a)** Comparison of linear and nonlinear probes in the single-seed exploratory Edema study. **(b–c)** Summary of DANN and CDANN experiments across adversarial strengths and seeds. Chance probe AUROC does not establish independence, and marginal probes alone do not assess conditional invariance.

Method	Configs.	Task AUROC ↑	Probe AUROC ↓	Pass	Main Finding / Limitation
Baselines and Controls					
ERM	1	0.840	0.961	0/1	Strong demographic leakage.
ERM adapters	3	0.833	0.554–0.803	0/3	Compression reduced leakage but did not meet the criterion.
PCA + ERM	2	0.838	0.813–0.886	0/2	Substantial residual leakage.
Constant output	1	0.500	0.502	1/1	Trivial removal through complete task collapse.
Learned Representations					
MMD	5	0.802–0.819	0.514–0.650	1/5	The 2-D setting passed development and test auditing.
HSIC	5	0.774–0.818	0.521–0.619	1/5	The 2-D setting passed, with reduced task utility.
Adversarial ensemble	3	0.812–0.828	0.521–0.771	1/3	Strongest setting passed test but missed development criterion.
Variational bottleneck + MMD	3	0.804–0.806	0.502–0.536	3/3	Single-sample passes; mean and repeated-release audits failed.
LFR	3	0.834–0.836	0.610–0.758	0/3	Retained demographic information.
NF	6	0.771–0.838	0.561–0.854	0/6	Reduced leakage but did not meet strict criterion.
Post-Hoc Representation Transformations					
Gaussian moment transport	2	0.832–0.839	0.650–0.853	0/2	Moment matching left demographic information recoverable.
LEACE	2	0.836–0.838	0.962–0.963	0/2	Strong residual leakage in the tested implementation.
LEOPARD	3	0.829–0.834	0.834–0.876	0/3	Nonlinear probes recovered substantial information.
RFF + LEACE	3	0.818–0.836	0.831–0.905	0/3	Near-chance linear probes concealed nonlinear leakage.
Obliviator	3	0.749–0.759	0.500–0.514	3/3	Only one setting met AUROC-loss budget; AUPRC fell substantially.
Score-Only Mechanisms					
Sex-blind score channel	3	0.825–0.829	0.510–0.512	3/3	Also passed repeated-release auditing; backbone unchanged.
Sex-aware score channel	3	0.836–0.837	0.513–0.517	3/3	Repeated releases exposed demographic information.

Table 16: **Exploratory demographic-erasure experiments on CheXpert Edema** at natural prevalence (one training seed; 51 configurations). Results do not reflect general limitations of the evaluated methods.