GAPTRACE: CONTROLLED ATTRIBUTION OF PREDICTION DISPARITIES IN CLINICAL TIME SERIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine learning is increasingly used in clinical decision-making, raising concerns about both fairness and interpretability. These are usually studied separately: attribution methods explain which features influence predictions, while fairness analyses measure disparities without identifying the model behavior associated with them. We introduce GapTrace, a framework for auditing prediction disparities in time series models. Given a gap in predicted risk between patient groups, GapTrace measures how much of the gap remains after adjustment for correlated attributes, identifies features or feature clusters used differently across groups with false-discovery-rate control, and tests whether intervening on the selected features reduces the gap more than matched control interventions. GapTrace is agnostic to the attribution method and can be applied on top of existing tools. On synthetic and semi-synthetic datasets with known subgroup-specific mechanisms, GapTrace identifies the planted features (F1 up to 0.92, compared with 0.35 for random selection). On 30-day mortality prediction in MIMIC-IV, we find three different patterns. The largest prediction gap is across age groups (29.7 percentage points), with 92% remaining after adjustment for other demographics. GapTrace identifies 35 of 525 feature clusters with significant differences in attribution across age groups. Intervention on the selected features reduces the gap more than matched controls. In contrast, most of the marital-status gap disappears after adjustment, while the remaining insurance gap cannot be consistently attributed to a specific set of features. GapTrace turns a measured disparity into a testable explanation: which inputs the model relies on differently across groups, and whether they account for the gap, delivering fairness analysis and explainability together.

1 INTRODUCTION

Machine learning is increasingly used in healthcare to support clinicians in prediction, data interpretation, and decision-making. However, concerns about fairness remain a major barrier to the responsible use of these systems (Ueda et al., 2024). Clinicians need to understand potential biases in model behavior and assess whether predictions are trustworthy across the patient populations in which these systems are used (Matos et al., 2024).

Fairness in healthcare ML is a complex problem that extends beyond reporting differences in predictive performance. Bias can arise at multiple stages of the AI development pipeline (Nakayama et al., 2024), and conclusions about fairness depend strongly on how sensitive groups and fairness criteria are defined (Passi & Barocas, 2019). Moreover, purely technical definitions of fairness may fail to capture important aspects of the broader clinical and social context (Selbst et al., 2019). Consequently, metrics such as differences in true positive rates (TPR), false positive rates (FPR), or prediction distributions can reveal that a disparity exists but provide limited information about the model behavior associated with that disparity.

Interpretability provides a complementary perspective by explaining how model inputs influence predictions. Interpretable clinical models have revealed unexpected relationships between patient characteristics, clinical practice, and predicted outcomes (Caruana et al., 2015). Post-hoc attribution methods (Turbé et al., 2023; Bento et al., 2021) extend this capability to complex models by estimating the contribution of individual features to predictions. For time-series models, interpretability methods can additionally characterize when particular features influence a prediction (Ismail et al., 2020;

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.

Tonekaboni et al., 2020). However, these methods are primarily designed to explain predictive importance rather than systematic differences in model reliance across patient subgroups.

Recent work has recognized the connection between fairness and interpretability (Ferry et al., 2025; Zhang et al., 2024). Nevertheless, the two are commonly evaluated as separate objectives. Existing approaches may compare feature importance across demographic groups or evaluate performance with and without sensitive demographic variables (Meng et al., 2022). A feature that is highly predictive need not distinguish model behavior across groups, whereas a moderately predictive feature may do so if the model relies on it differently across subgroups.

This motivates our central question: given an observed difference in model predictions across groups defined by a sensitive attribute, which feature and temporal patterns are associated with that difference, and are those contributions supported by statistical and intervention-based evidence? We propose **GapTrace**, a subgroup-aware post-hoc framework that quantifies subgroup prediction differences, evaluates their attenuation after adjustment for correlated attributes, identifies features or temporal clusters with differential attribution, and tests whether perturbing selected features changes the prediction gap more than matched control perturbations.

Evaluating such explanations on real-world data is challenging because the true mechanisms underlying subgroup-specific behavior are unknown. We therefore first evaluate GapTrace on controlled synthetic and semi-synthetic datasets with known subgroup-specific dependencies, then apply it to 30-day mortality prediction using MIMIC-IV. Our contributions are:

- A formulation of disparity-aware interpretability that treats differential model behavior across subgroups, rather than individual predictive importance alone.
- GapTrace, a post-hoc framework that explains subgroup prediction disparities through feature- and time-level attribution, false-discovery-rate-controlled selection, and matched feature-repair validation.
- Controlled synthetic benchmarks with known subgroup-specific dependencies, enabling quantitative evaluation of the recovery of planted disparity-generating mechanisms.
- Real-world demonstration of GapTrace on high-dimensional longitudinal clinical data, showing its ability to distinguish disparity sources from correlates, localize features contributing to the disparity at scale, and validate them through controlled interventions.

2 Related work

Fairness and interpretability. Fairness and interpretability address complementary aspects of trustworthy machine learning, but their interaction remains challenging (Ferry et al., 2025). In healthcare, fairness is particularly important because biases in data, model development, and deployment can produce systematic differences in performance across patient populations (Rajkomar et al., 2018; Chen et al., 2023). At the same time, fairness criteria can conflict with one another or with predictive utility (Zhao & Gordon, 2022), making it important not only to measure disparities but also to understand the model behavior associated with them.

Several studies have examined fairness and interpretability jointly. Meng et al. (2022) evaluate both properties on MIMIC-IV across demographic groups such as age and race. However, their analysis treats these largely as parallel model properties rather than localizing the input patterns associated with an observed subgroup prediction difference. Ghosh (2023) combine interpretable rule-based classifiers with formal verification of fairness constraints. Such approaches provide explicit guarantees for constrained model classes, but are less directly applicable to post-hoc analysis of complex temporal neural networks. FairShap (Arnaiz-Rodriguez et al., 2023) uses Shapley-based data valuation to reweight training samples for group-fairness objectives. GapTrace instead operates post hoc on a fixed trained model and analyzes differential attribution across its clinical inputs.

Post-hoc attribution for temporal models. Feature-attribution methods estimate how individual inputs contribute to a model prediction. Integrated Gradients (IG) (Sundararajan et al., 2017) integrates gradients between a reference input and the observed input, while DeepLIFT (Shrikumar et al., 2017) propagates activation differences relative to a reference through the network. DeepLIFTSHAP combines ideas from DeepLIFT and SHAP to approximate Shapley-style feature contributions (Lundberg & Lee, 2017). These methods provide the attribution mechanisms underlying our analysis, but

are conventionally designed to explain predictions rather than differences in prediction distributions across sensitive groups. AICO (Giesecke et al., 2025) provides statistical tests for feature importance through masking; in contrast, we test for differential attribution across sensitive subgroups.

Applying attribution methods to temporal data introduces additional challenges because explanations must capture both feature and temporal importance. Turbé et al. (2023) show that post-hoc explanations for time-series classifiers can vary substantially across attribution methods, motivating quantitative evaluation rather than relying only on qualitative inspection. Similarly, Ismail et al. (2020) benchmark interpretability methods for time-series models and emphasize the need to evaluate whether explanations correctly recover relevant features. Motivated by these concerns, we evaluate disparity-aware explanations on controlled datasets in which the subgroup-specific features and temporal dependencies responsible for prediction differences are known.

Group-aware and contrastive explanations. Several approaches move beyond standard feature attribution by explaining groups of variables or differences between predictions. Grouped feature importance methods quantify the joint importance of sets of related variables (Au et al., 2022; Bosschieter et al., 2026), while class-contrastive explanations identify features that distinguish one predicted class from an alternative (Wang & Wang, 2022). G-DeepSHAP (Chen et al., 2022) is particularly relevant because it considers grouped attribution of sensitive variables and proxies. By grouping correlated features, it can characterize direct and indirect contributions of sensitive information to model outputs. Our setting differs in the quantity being explained. Rather than measuring how sensitive information contributes to individual predictions, we seek to identify clinical features associated with an observed prediction disparity across sensitive groups and test whether perturbing those features changes the gap.

3 METHODOLOGY

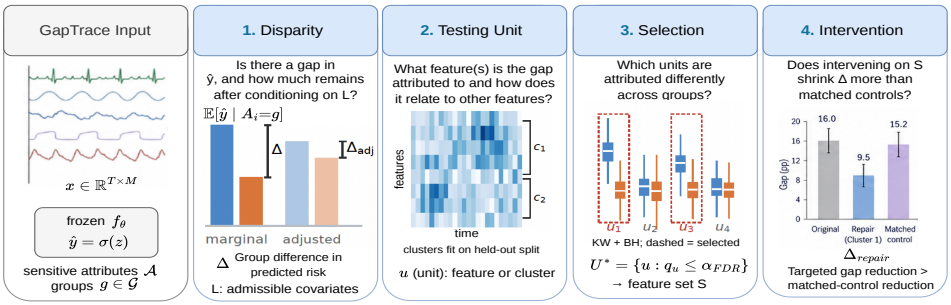


Figure 1: Overview of the GapTrace framework for identifying and experimentally probing the feature contributions driving prediction disparities across sensitive subgroups.

GapTrace starts from an observed difference in model predictions between groups defined by a sensitive attribute A and asks which input signals explain the prediction gap. This requires addressing several sources of ambiguity. An observed gap may reflect correlated sensitive attributes; attribution to predictive signal may be distributed across redundant features; differences may arise by chance when many pathways are tested; perturbing any sufficiently model-relevant feature may alter the gap.

To address the potential sources of ambiguity, GapTrace proceeds in four stages, Fig. 1. We first quantify the prediction disparity and characterize how it changes after conditioning on correlated attributes (§ 3.1). We then compute model attributions and define the units for disparity testing: individual features or, for high-dimensional temporal data, clusters of features with similar attribution patterns (§ 3.2 and 3.3). We intervene on the resulting feature set to determine whether reducing group differences in these features reduces the prediction gap more than matched control (other model-relevant features) (§ 3.4). Finally, features that statistically significantly reduce the gap compared to random control are flagged to the user.

Problem setup: Let f_θ be a trained and frozen classifier that maps a multivariate time series $\boldsymbol{x} \in \mathbb{R}^{T \times M}$ to a predicted probability $\hat{y}_i = \sigma(z_i)$, where z_i is the model logit, M is the number

of input features, and T is the number of time steps. Each sample has an outcome y_i and one or more sensitive attributes $\mathcal{A} = A_1, \ldots, A_K$, where continuous attributes are binned. We measure prediction disparities on the probability scale because they represent differences in predicted risk, but compute attributions on the logit z_i. Probability-space gradients contain the multiplicative factor $\hat{y}(1-\hat{y})$, which depends on predicted risk and can therefore induce systematic differences in attribution magnitude between groups with different score distributions. Proposition A.2.9 establishes the analogous relationship for softmax outputs.

3.1 Disparity and the choice of test statistic

For groups $g \in \mathcal{G}$, we quantify the magnitude of the prediction disparity as:

$$\Delta = \max_g \mathbb{E}[\hat{y} \mid A{=}g] - \min_g \mathbb{E}[\hat{y} \mid A{=}g], \tag{1}$$

reported in percentage points with a bootstrap 95% interval, together with threshold-based statistical parity and disparate impact (Feldman et al., 2015).

The size of a disparity, Δ, does not indicate whether it arose by chance. We therefore also test $H_0 : \hat{Y} \perp A$ by permuting A across samples. Because the appropriate dependence statistic depends on the distribution of $\hat{y}$, we choose it from $\hat{y}$ alone, without reference to A, before any test is run. Let $s = \Pr(\hat{y} < 0.01 \text{ or } \hat{y} > 0.99)$ denote the fraction of saturated scores. We use a χ^2 statistic when $\hat{y}$ takes at most two distinct values; plug-in mutual information on decile-binned scores when $s > 0.2$, since kNN estimators (Kraskov et al., 2004) are unreliable when scores concentrate near 0 and 1; and kNN mutual information otherwise. Permutation p-values are corrected across the K sensitive attributes using Benjamini–Hochberg. Regardless of the statistic used for testing, we report mutual information for every attribute as a descriptive measure, since it registers distributional differences that Δ does not, including differences that leave mean predicted risk unchanged.

Separating inherited from independent disparity A marginal disparity associated with A may reflect its association with other sensitive attributes. We therefore characterize how much of the marginal disparity remains after conditioning on a prespecified set L of other demographic attributes,

$$\hat{y}_i = \beta_0 + \sum_{g \geq 2} \beta_g \mathbb{1}[A_i = g] + \sum_k \gamma_k L_{ik} + \varepsilon_i, \tag{2}$$

fitted with HC3 heteroskedasticity-robust standard errors (MacKinnon & White, 1985) and a joint F-test of $\{\beta_g\}$. Here, β_g represents the adjusted difference in mean predicted risk between group g and the reference group. Taking the reference group's coefficient $\beta_1 = 0$, the adjusted gap is

$$\Delta_{\text{adj}} \;=\; \max_g \beta_g - \min_g \beta_g, \tag{3}$$

in the same percentage-point units as Δ. $1 - \Delta_{\text{adj}}/\Delta$ is the share of the observed gap attenuated after adjustment for L (Kamiran et al., 2013). The adjustment set L contains potential confounders but excludes variables considered potential mediators between A and the model prediction, because conditioning on such variables could remove associations mediated through proxy features that the subsequent attribution analysis is intended to identify. Details of covariate specification and handling of collinearity are given in § A.2.8.

3.2 Attribution and construction of testing units

For each sample i, an attribution method produces $\phi_i \in \mathbb{R}^{T \times M}$, where $\phi_{i,t,j}$ denotes the attribution of feature j at time t to the model logit. The framework is modular with respect to the attribution method; in this study, we evaluate DeepLIFT, Integrated Gradients (IG), DeepLIFTSHAP, and GradientSHAP. Rather than using an all-zero reference, baselines are sampled from the training distribution, either from the full training population or from a designated reference group. For temporal inputs, we aggregate each feature's attribution over time, $\Phi_j = \sum_t \phi_{t,j}$.

The subsequent statistical analysis operates on a generic testing unit u. A testing unit may be an individual feature or a cluster of features representing a shared attribution pattern. For an individual feature j, $\Phi_{i,u} = \Phi_{i,j}$; for a cluster c, $\Phi_{i,u} = \Phi_{i,c} = \sum_{j \in c} \Phi_{i,j}$. This additive aggregation allows signals distributed across redundant channels to be tested jointly.

Constructing Temporal Attribution Clusters Feature-wise testing can lose power when features are correlated. For high-dimensional temporal data, we therefore group features into clusters. Let $\Phi_{\cdot,j} \in \mathbb{R}^N$ denote the sample-level attribution vector for feature j, and let $h_j \in \mathbb{R}^D$ denote a standardized descriptor of the temporal attribution trajectory of feature j. We define the temporal attribution distance

$$D_{jk} = w\,(1 - |\mathrm{corr}(\Phi_{\cdot,j}, \Phi_{\cdot,k})|) + (1-w)\frac{\|h_j - h_k\|_2}{d_{0.95}}, \tag{4}$$

clipped to $[0,1]$, where $d_{0.95}$ is the 95th percentile of pairwise descriptor distances. The first term captures whether two features receive attribution on the same samples; the second captures similarity in when attribution occurs within the sequence. The descriptor h_j is 7 dimensional including three normalized Legendre coefficients capturing the shape of the mean signed attribution profile, temporal centroid, normalized entropy of absolute attribution mass, sign-consistency ratio, and lag-one autocorrelation. We use $w = 0.7$ and average-linkage hierarchical clustering. The construction of clustering cutoffs, and safeguards against inappropriate merges are given in §A.2.4. The choice of testing unit is prespecified for each experimental setting.

3.3 Statistical selection of disparity-associated units

For each testing unit u, we compute its mean attribution within group g, $\bar{\Phi}_u^{(g)} = \frac{1}{n_g}\sum_{i:A_i=g}\Phi_{i,u}$, and quantify the magnitude of differential attribution by the spread across group means,

$$s_u = \max_g \bar{\Phi}_u^{(g)} - \min_g \bar{\Phi}_u^{(g)}. \tag{5}$$

Because s_u captures magnitude but not direction, we also retain the signed mean attribution of each testing unit within each sensitive group. These indicate whether a unit's differential attribution aligns with or opposes the overall prediction gap. To distinguish systematic group differences from variation expected by chance, we compare the attribution distributions of each testing unit across groups using a Kruskal–Wallis test. False discoveries across all testing units are controlled using the Benjamini–Hochberg procedure at level α_{FDR}. Effect magnitudes are reported separately using standardized attribution differences (§3.4). The selected set is $\mathcal{U}^* = \{u : q_u \le \alpha_{\mathrm{FDR}}\}$, where q_u denotes the FDR-adjusted p-value for unit u. For selected units, we additionally report which groups have mean attribution elevated or reduced relative to the pooled mean. As a robustness analysis, we perform stability selection over n_s half-samples and report units selected in at least a fraction π of resamples; stability selection does not alter the definition of $\mathcal{U}^*$. An empty $\mathcal{U}^*$ is a valid outcome and is interpreted relative to the minimum detectable attribution shift described in § A.2.6. This stage identifies testing units with statistically supported differences in model attribution across groups.

From selected testing units to intervention features. Interventions operate on input features rather than attribution clusters. An individually selected feature enters the intervention set directly. For each selected cluster, we retain its m member features with the largest mean absolute attribution, ranked by $\mathbb{E}_i[|\Phi_{i,j}|]$. The union of these features defines the intervention set $\mathcal{S}$. Limiting each cluster to m members keeps the intervention interpretable at the feature level while preserving a pool of unselected features from which matched controls can be constructed (§3.4).

3.4 Counterfactual gap reduction

Differential attribution establishes that the selected testing units are attributed differently across groups, but not that they account for the observed prediction disparity. We therefore intervene on S with the trained model fixed. The intervention is a probe of the model: we ask how the prediction gap would change if the model received inputs in which group differences in $\mathcal{S}$ were removed, with all other features held fixed.

For feature j and group g, let $F_{g,j}$ and $F_{g,j}^{-1}$ denote the empirical distribution and quantile functions, and let $F_{\star,j}^{-1}(\tau) = \mathrm{median}_g\, F_{g,j}^{-1}(\tau)$ be a common reference. Following the rank-preserving repair of Feldman et al. (2015), for intervention strength $\lambda \in [0,1]$ we set

$$\widetilde{F}_{g,j}^{-1}(\tau) = (1-\lambda)\,F_{g,j}^{-1}(\tau) + \lambda\,F_{\star,j}^{-1}(\tau), \tag{6}$$

and replace each value of a selected feature by the value at the same within-group quantile,

$$\tilde{x}_{i,j} = \widetilde{F}_{A_i,j}^{-1}\big(F_{A_i,j}(x_{i,j})\big), \qquad j \in \mathcal{S}. \tag{7}$$

Thus $\lambda = 0$ leaves the inputs unchanged, and $\lambda = 1$ gives every group the same marginal distribution on S while preserving within-group ranks. Evaluating the frozen model on $\tilde{x}$ gives the counterfactual gap $\Delta_\lambda(S)$ and the gap reduction $R(\mathcal{S}) = \Delta - \Delta_\lambda(\mathcal{S})$, in percentage points.

A small reduction under marginal alignment does not imply that the selected features are irrelevant, since the disparity may depend on higher moments, joint dependencies, or more complex use of $\mathcal{S}$. When the primary intervention reduces the gap by less than 5% of Δ, we evaluate stronger interventions: (I1) moment-aware alignment using feature-specific mean, scale, or quantile transformations; (I2) joint transplantation, replacing the target group's values on $\mathcal{S}$ with jointly sampled values from the pooled non-target groups; and (I3) ablation, replacing $\mathcal{S}$ with the pooled mean trajectory for all samples. We report the first intervention in this sequence that produces a detectable reduction; failure even under I3 indicates that the selected set is insufficient to account for the observed gap. Details are given in § A.2.7.

Because any sufficiently strong perturbation can change model predictions, each targeted intervention is compared with the same intervention applied to n_c control sets $B_1, \ldots, B_{n_c}$: equally sized sets of unselected features with similar attribution magnitude. We report $R(\mathcal{S})$ together with the mean and standard deviation of each of the control sets $R(B_b)$, and the empirical significance level $p_{\text{emp}} = \frac{1+\sum_{b=1}^{n_c} \mathbb{1}\left[R(B_b) \geq R(\mathcal{S})\right]}{1+n_c}$. This tests whether perturbing the attribution-selected features reduces the gap more than a comparable perturbation of other model-relevant inputs. Control construction, finite-resolution considerations, and degenerate cases are detailed in § A.2.7.

4 DATASETS

Simulation We first test GapTrace on synthetic time series where the disparity mechanism is known by construction. Each dataset contains ($M = 6$) signals of length ($T = 500$). Four signals are informative: they are sine waves with a high-frequency burst beginning at a randomly selected time. The remaining two signals contain either a square or a flat segment. The binary label is determined by the combined burst frequencies of the informative signals, using the median as the classification threshold. We then create two groups that differ only in which signals determine the label. In the reference group, all four informative signals contribute to the label. In the target group, only $(4-d)$ of these signals contribute. Thus, the (d) excluded signals carry label information in the reference group but not in the target group, providing a known ground truth for group-specific label dependence.

We recalibrate the classification threshold separately within each group so that the two groups have similar class prevalence. This avoids deliberately introducing an outcome-rate difference and allows us to focus on differences in feature reliance. The resulting mean predicted-rate gap is only 4.0 percentage points. We consider ($d \in 1, 2$), enumerate all possible subsets of excluded signals, and repeat each setting across five random seeds. This produces 50 runs, each with ($N = 10{,}000$) samples. Across these experiments, the Attentive BiLSTM achieves 0.85 accuracy and 0.94 AUC, with an accuracy gap of 0.03 between groups.

Semi-synthetic Benchmark Dataset To bridge the gap between fully simulated and real-world settings, we construct a semi-synthetic Human Activity Recognition (HAR) benchmark based on real-world wearable data from the Parkinson's Disease Smartwatch (PADS) dataset (Varghese et al., 2024); see § A.7 for details. We introduce controlled subgroup-specific perturbations to existing sensor channels, providing known ground truth while preserving the structure of real-world signals. The perturbed channel and injection magnitude are selected to ensure that the planted disparity propagates to model predictions. In this scenario, since the ground truth mechanism is defined at the feature level, the testing units are individual features.

Validation when the disparity mechanism is known We consider group-conditional mean shifts on informative channels, additive noise, temporal warping and dropout, and train-only label noise, with each injection specification recorded as machine-readable ground truth (App. A.8). Because detection can fail either because the planted disparity does not propagate to model predictions or because GapTrace fails to recover it, we evaluate these cases separately. We report the propagation rate, the fraction of runs in which the planted disparity is detectable in $\hat{y}$, and recall conditional on propagation, the precision and recall of the selected set against the planted channels among runs in which the disparity propagates. For each trained model, we also compute an occlusion reference by permuting each channel in turn and recording the resulting change in the group gap. Because

this requires only forward passes, it provides an architecture-independent reference for whether the trained model uses the planted channel, helping distinguish attribution failure from failure of the model to learn the injected disparity.

MIMIC-IV The primary real-world dataset used in this study is the Medical Information Mart for Intensive Care (MIMIC-IV) dataset (Johnson et al., 2023; Johnson, 2020), which provides comprehensive electronic health records (EHRs) for patients admitted to intensive care units. We extend the existing data-processing pipeline (Gupta et al., 2022) to train the BiLSTM model. We audit 3,822 ICU stays from adult patients aged 18 years or older, using the first 24 hours of each admission to predict 30-day mortality. Each admission is represented as an hourly multivariate time series with 856 channels capturing clinical information from multiple sources. These include vital signs and laboratory measurements, administered medications such as antibiotics and vasopressors, procedures such as mechanical ventilation and imaging tests, charted events, and output measurements such as fluid and urine output. We use this setting to evaluate GapTrace on disparities arising in a real-world clinical prediction task. In this scenario, inputs contains many potentially redundant temporal signals, features are clustered to be used as testing units.

5 EXPERIMENTS

GapTrace can be applied with different attribution methods; here, we evaluate DeepLIFT, Integrated-Gradients (IG), DeepLIFTSHAP, and GradientSHAP, §A.2.1. We selected these widely used methods to span complementary reference-based attribution approaches. To interpret feature-level disparity attributions, we distinguish between positive and negative contributions. Positive contributions act in the direction of the observed disparity, whereas negative contributions offset it. We then test these associations by intervening on the identified features and measuring the resulting gap reduction $R(S)$ in percentage points (pp). Positive values indicate a smaller prediction gap after intervention. To assess whether this reduction is specific to the identified features, we compare each targeted intervention with the same intervention on matched sets of other model-relevant features. We perform this analysis for each attribution method and each attribute whose marginal disparity passes FDR correction. Thus, attribution identifies where the disparity is represented in the model, while intervention tests whether perturbing those (sets of) features reduces it.

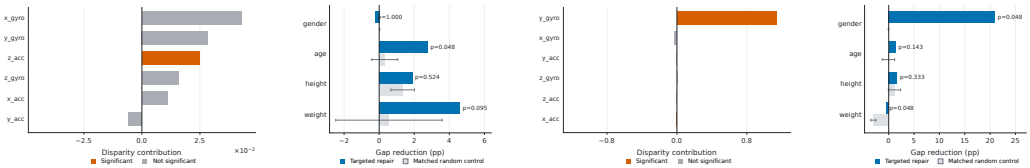
5.1 RESULTS

Simulated Data Analysis Table 1 evaluates whether GapTrace recovers the known disparity-driving features. We compare four GapTrace variants, differing only in their attribution method, against the planted feature-level ground truth and a random baseline that independently selects each feature with probability 0.5. The evaluation measures both recovery of disparity-driving features and rejection of non-driving features. Importantly, the negative set includes both uninformative features and features predictive for both groups, testing whether GapTrace identifies differential attribution rather than generic feature importance. Across attribution methods, GapTrace achieves balanced accuracy of $0.80 - 0.94$, compared with 0.50 for random selection, and F1 scores of $0.75 - 0.92$, compared with 0.35 for random selection. GapTrace-IntegratedGradients shows the strongest overall recovery in this benchmark, 0.92 F1 score.

Table 1: Feature-recovery performance on the controlled synthetic benchmark. Results are reported as mean $\pm$ standard deviation across runs.

Method	TPR ↑	TNR ↑	Precision ↑	F1 ↑	Bal. Acc. ↑
Random	0.50	0.50	0.27	0.35	0.50
GapTrace-DeepLIFT	0.80 ± 0.05	0.96 ± 0.05	0.89 ± 0.11	0.84 ± 0.05	0.88 ± 0.03
GapTrace-IntegratedGradients	$\mathbf{0.90 \pm 0.07}$	$\mathbf{0.98 \pm 0.02}$	$\mathbf{0.95 \pm 0.06}$	$\mathbf{0.92 \pm 0.06}$	$\mathbf{0.94 \pm 0.04}$
GapTrace-GradientSHAP	0.63 ± 0.10	0.98 ± 0.02	0.93 ± 0.07	0.75 ± 0.07	0.80 ± 0.05
GapTrace-DeepLIFTSHAP	0.80 ± 0.05	0.96 ± 0.05	0.89 ± 0.11	0.84 ± 0.05	0.88 ± 0.03

Semi-Synthetic Benchmark Results Considering age (Fig. 2a), the injected z-acc channel is identified as the only significant contributor, with a positive contribution to the disparity, while the



(a) Age: Disparity contribution

(b) Age: gap reduction after intervention

(c) Gender: Disparity contribution

(d) Gender: gap reduction after intervention

Figure 2: Recovery of planted disparities on semi-synthetic data (TCN, GapTrace-DeepLIFT, seed 0). Left: age, injected into z_acc (amount 0.1); right: gender, injected into y_gyro (amount 1.0). Feature-analysis panels show feature-level contributions and significant selections; intervention panels compare targeted repair with n_c matched random-feature controls.

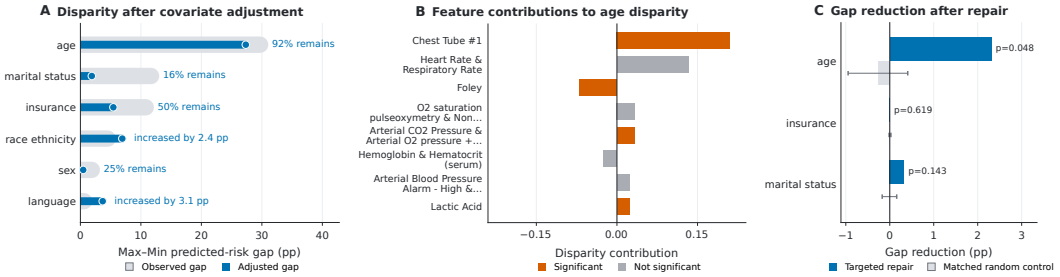


Figure 3: MIMIC-IV analysis of demographic disparities and targeted feature repair. (A) Max–Min differences in mean predicted risk before and after covariate adjustment; annotations show the percentage of the observed gap remaining after adjustment, or the absolute increase in pp when the gap increases. (B) Top feature-group contributions to the age-associated disparity, ranked by absolute magnitude; red denotes statistically significant groups. (C) Gap reduction following targeted repair compared with matched random-feature controls; error bars show control variability and p-values compare targeted and control interventions.

remaining channels are non-significant. In the corresponding intervention analysis, targeting z-acc reduces the age-associated gap by approximately 2.3 pp and is the only intervention that exceeds matched controls ($p = 0.048$); the interventions for gender, height, and weight do not ($p = 1.000$, $p = 0.524$, and $p = 0.095$, respectively). Considering gender (Fig. 2c), the disparity-attribution analysis identifies y-gyro as the only significant contributor and the only channel with a positive contribution, while the remaining five channels are non-significant and predominantly negative. Consistently, targeting y-gyro reduces the gender-associated gap by approximately 21 pp, exceeding the matched controls ($p = 0.048$); interventions for age ($p = 0.143$) and height ($p = 0.333$) show substantially smaller effects. Together, these results show that the disparity-attribution analysis identifies the planted disparity features and that interventions targeting these features produce gap reductions exceeding those of matched controls.

MIMIC-IV Results Model-predicted probability differed most strongly across age groups (Fig. 3A). Mean predicted risk differed by 29.7 pp between patients aged 18–44 and those aged ≥ 85 years (95% CI 24.4–34.9), and 92% of this difference remained after adjustment for the other demographic attributes. In contrast, 84% of the 11.7-point marital-status gap attenuated after adjustment and the residual was not significant ($q = 0.97$). For insurance, 50% of the 10.8-point unadjusted gap remained after adjustment, leaving a significant 5.4-point residual gap ($q = 0.011$). No significant marginal or within-age disparity was detected for race, sex, or language. The adjusted point estimates for race and language nevertheless increased from 4.5 to 6.9 points (+2.4 pp) and from 0.6 to 3.7 points (+3.1 pp), respectively, reflecting the suppression effect discussed in §A.2.8.

The choice of disparity statistic substantially affected the detection. With 53% of predicted risks at the boundaries of $[0, 1]$, the kNN mutual-information estimator yielded $q = 0.395$ for age, whereas the binned estimator selected by the rule in § 3.1 yielded $q = 0.004$ on the same data. At a prevalence-

matched decision threshold, selection rates increased monotonically from 5.2% in the 18–44 group to 13.1% in the ≥ 85 group. The TPR was also 6pp lower in the 18–44 group than in the ≥ 85 group.

At the feature level, 35 of 525 attribution clusters were selected for age at a 5% false discovery rate (Fig. 3B). Chest Tube had the largest positive signed differential attribution (0.21 logit units). A cluster containing heart rate and respiratory rate had a positive differential attribution of 0.13 logit units but was not selected, whereas Foley had a significant negative differential attribution (-0.06 logit units). Intervention on the age-associated feature set reduced the age disparity by approximately 2.3 percentage points, significantly more than matched controls ($p = 0.048$). This reduction corresponds to approximately 8% of the observed 29.7-point age gap, indicating that the selected features account for a measurable but limited portion of the disparity under the intervention, while most of the gap remains.

For insurance, only one attribution cluster was selected ($q = 0.04$), and its identity was not stable across the two attribution methods, GapTrace-IntegratedGradients and GapTrace-DeepLIFT, (Esmolol versus Troponin-T). Overall, targeted repair produced evidence of a gap reduction beyond matched controls for age, but not for marital status ($p = 0.143$) or insurance ($p = 0.619$), indicating that only the age-associated feature set showed clear intervention-specific effects.

6 Discussion and Conclusion

GapTrace treats an observed prediction disparity as a starting point for a sequence of testable questions rather than as an endpoint of the analysis. A disparity may reflect associations with correlated attributes, be associated with specific differences in model behavior, or persist without a stable feature-level explanation. The controlled experiments validate the method while also clarifying these distinctions: differential attribution is not equivalent to predictive importance, and matched interventions provide a functional test of whether perturbing selected features changes the subgroup gap beyond comparable perturbations of other model-relevant inputs.

On MIMIC-IV, these analyses reveal three qualitatively different patterns: a disparity that persists after conditioning, localizes to stable attribution differences, and responds to targeted intervention; a disparity that largely attenuates after conditioning; and a disparity that persists after conditioning but lacks a stable attribution pattern or intervention response. The age disparity is large, persists after conditioning on the other demographic attributes, localizes to statistically supported differences in model attribution, and is reduced beyond matched controls by intervention on the selected features. Chest Tube shows the strongest positive differential attribution, whereas Foley shows a significant negative differential attribution, illustrating that identified features can align with or oppose the overall prediction gap. In contrast, most of the marital-status disparity attenuates after conditioning, and intervention on its selected features does not reduce the gap beyond matched controls. The insurance disparity retains a significant residual after conditioning, but its selected features are not stable across attribution methods and intervention does not outperform matched controls.

Overall, GapTrace provides a framework for moving from detecting a prediction disparity to asking what lies behind it: whether it persists after conditioning, which inputs receive different attribution across groups, and whether perturbing those inputs changes the gap beyond matched controls. Although we focus on demographic disparities, the same framework can in principle be applied to predefined populations such as institutions, acquisition settings, or disease phenotypes, and with different attribution methods.

Limitations and Future Work. GapTrace characterizes associations in model behavior rather than causal effects. Conditioning on measured covariates can characterize how much of a subgroup prediction difference remains, but residual differences may still reflect unmeasured clinical, socioeconomic, or data-collection factors. Attribution-based localization also depends on the attribution method and reference distribution; disagreement across methods should therefore be interpreted as evidence against a stable feature-level explanation. Moreover, interventions may disrupt cross-feature dependencies and produce inputs outside the observed data manifold, so gap reduction measures model reliance rather than the effect of a realizable clinical intervention. Future work will develop interventions that better preserve multivariate temporal structure, extend the conditioning analysis to a broader set of clinical covariates, and evaluate GapTrace across additional architectures, institutions, and clinical tasks.

AI USE STATEMENT

Generative AI tools were used to assist with editing and improving the clarity of the manuscript, refining the presentation of experimental results, and providing feedback during the development and documentation of the experimental methodology. AI tools were also used to assist with code editing and debugging. All AI-assisted content, analyses, and code were reviewed and verified by the authors. The authors take full responsibility for the methodology, experiments, results, claims, and final content of the paper.

ETHICS STATEMENT

This work studies subgroup differences in machine-learning predictions using demographic and clinical data. A measured subgroup score gap is not interpreted as evidence of discrimination or unfairness by itself, and demographic categories are used as audit variables rather than biological explanations. The MIMIC-IV analysis uses de-identified data available to credentialed PhysioNet users under the required training and Data Use Agreement.

REPRODUCIBILITY STATEMENT

To support reproducibility, the source code required to reproduce the experiments will be provided as supplementary material with the submission and will be made publicly available on GitHub following the review process. The real-world datasets used in this work are available from their original repositories subject to their respective access requirements; MIMIC-IV requires credentialed PhysioNet access and completion of the applicable Data Use Agreement. The data preprocessing and experimental procedures are described in the paper and supplementary material. The code used to construct the controlled datasets and inject synthetic disparities is also included in the supplementary material. The generated synthetic data will be made publicly available alongside the code. Additional implementation details, experimental settings, and hyperparameters required to reproduce the reported results are provided in the appendix.

REFERENCES

Adrian Arnaiz-Rodriguez, Francisco Escolano, and Nuria Oliver. Fairshap: a data re-weighting approach for algorithmic fairness based on shapley values. *arXiv preprint arXiv:2303.01928*, 1(4), 2023.

Quay Au, Julia Herbinger, Clemens Stachl, Bernd Bischl, and Giuseppe Casalicchio. Grouped feature importance and combined features effect plot. *Data Mining and Knowledge Discovery*, 36(4): 1401–1450, 2022.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

João Bento, Pedro Saleiro, André F Cruz, Mário AT Figueiredo, and Pedro Bizarro. Timeshap: Explaining recurrent models through sequence perturbations. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 2565–2573, 2021.

Tomas Bosschieter, Luis França, Jessica Wolk, Yiyuan Wu, Bella Mehta, Joseph Dehoney, Orsolya Kiss, Fiona C Baker, Qingyu Zhao, Rich Caruana, et al. The most important features in generalized additive models might be groups of features. *Scientific Reports*, 16(1):14371, 2026.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21st ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1721–1730, 2015.

Hugh Chen, Scott M Lundberg, and Su-In Lee. Explaining a series of models by propagating shapley values. *Nature communications*, 13(1):4512, 2022.

Richard J Chen, Judy J Wang, Drew FK Williamson, Tiffany Y Chen, Jana Lipkova, Ming Y Lu, Sharifa Sahai, and Faisal Mahmood. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature biomedical engineering*, 7(6):719–742, 2023.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 797–806, 2017.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21st ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268, 2015.

Julien Ferry, Ulrich Aïvodji, Sébastien Gambs, Marie-José Huguet, and Mohamed Siala. Taming the triangle: On the interplays between fairness, interpretability, and privacy in machine learning. *Computational Intelligence*, 41(4):e70113, 2025.

Bishwamittra Ghosh. Interpretability and fairness in machine learning: A formal methods approach. In *IJCAI*, pp. 7083–7084, 2023.

Kay Giesecke, Enguerrand Horel, and Chartsiri Jirachotkulthorn. Aico: Feature significance tests for supervised learning. *arXiv preprint arXiv:2506.23396*, 2025.

Mehak Gupta, Brennan Gallamoza, Nicolas Cutrona, Pranjal Dhakal, Raphael Poulain, and Rahmatollah Beheshti. An extensive data processing pipeline for mimic-iv. In *Machine Learning for Health*, pp. 311–325. PMLR, 2022.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96, 2019.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Aya Abdelsalam Ismail, Mohamed Gunady, Hector Corrada Bravo, and Soheil Feizi. Benchmarking deep learning interpretability in time series predictions. *Advances in neural information processing systems*, 33:6441–6452, 2020.

Alistair Johnson. Mimic-iv v1.0, 2020. URL `https://physionet.org/content/mimiciv/1.0/`.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Faisal Kamiran, Indrė Žliobaitė, and Toon Calders. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and information systems*, 35(3):613–644, 2013.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 69(6):066138, 2004.

William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

James G MacKinnon and Halbert White. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of econometrics*, 29(3):305–325, 1985.

João Matos, Jack Gallifant, Anand Chowdhury, Nicoleta Economou-Zavlanos, Marie-Laure Charpignon, Judy Gichoya, Leo Anthony Celi, Lama Nazer, Heather King, and An-Kwok Ian Wong. A clinician's guide to understanding bias in critical clinical prediction models. *Critical Care Clinics*, 40(4):827–857, 2024.

Chuizheng Meng, Loc Trinh, Nan Xu, James Enouen, and Yan Liu. Interpretability and fairness evaluation of deep learning models on mimic-iv dataset. *Scientific Reports*, 12(1):7166, 2022.

Luis Filipe Nakayama, João Matos, Justin Quion, Frederico Novaes, William Greig Mitchell, Rogers Mwavu, Claudia Ju-Yi Ji Hung, Alvina Pauline Dy Santiago, Warachaya Phanphruk, Jaime S Cardoso, et al. Unmasking biases and navigating pitfalls in the ophthalmic artificial intelligence lifecycle: A narrative review. *PLOS Digital Health*, 3(10):e0000618, 2024.

Robert M O'brien. A caution regarding rules of thumb for variance inflation factors. *Quality & quantity*, 41(5):673–690, 2007.

Samir Passi and Solon Barocas. Problem formulation and fairness. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 39–48, 2019.

Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12): 866–872, 2018.

Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 59–68, 2019.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pp. 3145–3153. PMLR, 2017.

Michael W Sjoding, Robert P Dickson, Theodore J Iwashyna, Steven E Gay, and Thomas S Valley. Racial bias in pulse oximetry measurement. *New England Journal of Medicine*, 383(25):2477–2478, 2020.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Robert J Tibshirani and Bradley Efron. An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57(1):1–436, 1993.

Sana Tonekaboni, Shalmali Joshi, Kieran Campbell, David K Duvenaud, and Anna Goldenberg. What went wrong and when? instance-wise feature importance for time-series black-box models. *Advances in Neural Information Processing Systems*, 33:799–809, 2020.

Hugues Turbé, Mina Bjelogrlic, Christian Lovis, and Gianmarco Mengaldo. Evaluation of post-hoc interpretability methods in time-series classification. *Nature Machine Intelligence*, 5(3):250–260, 2023.

Daiju Ueda, Taichi Kakinuma, Shohei Fujita, Koji Kamagata, Yasutaka Fushimi, Rintaro Ito, Yusuke Matsui, Taiki Nozaki, Takeshi Nakaura, Noriyuki Fujima, et al. Fairness of artificial intelligence in healthcare: review and recommendations. *Japanese Journal of Radiology*, 42(1):3–15, 2024.

Julian Varghese, Alexander Brenner, Michael Fujarski, Catharina Marie van Alen, Lucas Plagwitz, and Tobias Warnecke. Machine learning in the parkinson's disease smartwatch (PADS) dataset. *npj Parkinson's Disease*, 10(1):9, 2024.

Bo Wang, Lin Huang, Yuke Zhu, Anshul Kundaje, Serafim Batzoglou, and Anna Goldenberg. Vicus: Exploiting local structures to improve network-based analysis of biological data. *PLoS computational biology*, 13(10):e1005621, 2017.

Yifei Wang, Liqin Wang, Zhengyang Zhou, John Laurentiev, Joshua R Lakin, Li Zhou, and Pengyu Hong. Assessing fairness in machine learning models: A study of racial bias using matched counterparts in mortality prediction for patients with chronic diseases. *Journal of biomedical informatics*, 156:104677, 2024.

Yipei Wang and Xiaoqian Wang. "why not other classes?": Towards class-contrastive back-propagation explanations. *Advances in Neural Information Processing Systems*, 35:9085–9097, 2022.

Haoran Zhang, Walter Gerych, and Marzyeh Ghassemi. A data-centric perspective to fair machine learning for healthcare. *Nature Reviews Methods Primers*, 4(1):86, 2024.

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

Table 2: Notation used throughout. Percentage points (pp) are absolute differences in predicted risk, so a gap of 0.297 in probability is reported as 29.7 pp.

Category	Symbol	Definition
Data	$\boldsymbol{x} \in \mathbb{R}^{T \times M}$	Input sequence; T time steps, M channels (T=1 for tabular data)
	N	Number of samples in the evaluation split
	$y \in \{0, 1\}$	Observed outcome
	A_k, $\mathcal{A}$	The k-th sensitive attribute; the set of K attributes
	$g \in \mathcal{G}_A$	A group of attribute A; $G = \lvert\mathcal{G}_A\rvert$
	n_g	Number of samples in group g
	L	Admissible (legitimate) covariates used for adjustment
Model	f_θ	Trained classifier under audit
	$z, \hat{y}$	Logit and corresponding predicted probability
	p_k, m_k	Softmax probability and one-vs-rest logit margin for class k
	C	Number of output classes
	$\hat{y} = \sigma(z)$	Predicted risk; all reported disparities use this scale
	C_f	Cost of one forward/backward pass
Disparity	$m_g = \mathbb{E}[\hat{y} \mid A{=}g]$	Mean predicted risk in group g
	$\Delta = \max_g m_g - \min_g m_g$	Statistical parity difference in expectation (pp)
	Δ_{adj}	The same gap from Eq. 2, adjusting for L
	$1 - \Delta_{\text{adj}}/\Delta$	Explained share of the gap
	r_g, $\min_g r_g / \max_g r_g$	Selection rate in group g; disparate-impact ratio
	TPR_g, FPR_g	Group-conditional error rates at threshold t
Inference	$\text{I}(A; \hat{y})$	Mutual information between attribute and prediction
	$\mathcal{S}$, $\mathcal{S}_\pi$	Test statistic; its value under a permuted labelling
	n_π, n_b, n_c	Permutations; bootstrap resamples; matched controls
	p, q	Permutation p-value $\frac{1+\#\{\mathcal{S}_\pi \geq \mathcal{S}\}}{1+n_\pi}$; BH-adjusted
	α, $\delta_{\min}$	FDR level; minimum standardised effect for selection
	SMD	Standardised mean difference between groups
Attribution	$\phi \in \mathbb{R}^{T \times M}$	Per-element attribution of the logit
	$\Phi_j = \sum_t \phi_{t,j}$	Time-aggregated attribution of channel j
	$\bar{\Phi}_j^{(g)}$	Mean of Φ_j within group g
	$s_j = \max_g \bar{\Phi}_j^{(g)} - \min_g \bar{\Phi}_j^{(g)}$	Disparity contribution of channel j
	$\boldsymbol{x}'$, $\mathcal{D}$	Baseline input; baseline distribution
	$\mathcal{M}$, c_m	Set of attribution methods; model passes per sample per method
	$\mathcal{S}$	Selected feature (or cluster) set
Cluster	c, K_c	An attribution cluster; the number of clusters, $C \ll M$
	$\Phi_{i,c} = \sum_{j \in c} \Phi_{i,j}$	Merged cluster attribution (preserves completeness)
	τ	Sign-cancellation threshold for splitting a cluster
Repair	$\lambda \in [0, 1]$	Repair level; 0 = original data, 1 = full repair
	$F_{g,j}^{-1}$, $F_{A_i,j}^{-1}$	Group and median-across-groups quantile functions
	$\tilde{F}_{g,j}^{-1} = (1{-}\lambda)F_{g,j}^{-1} + \lambda F_{A_i,j}^{-1}$	λ-partial geometric repair
	$\Delta_{\text{before}} - \Delta_{\text{after}}$	Gap reduction (pp), reported against n_c matched controls

A Appendix

A.1 Notation Table

A.2 Methodology Additional Explanation

A.2.1 Preliminaries

Let $F : \mathbb{R}^{T\times M} \to \mathbb{R}^C$ denote a neural network mapping a multivariate time-series input to a C-class output. Given $x \in \mathbb{R}^{T\times M}$ with T time steps and M features, $F(x) = [F_1(x), \ldots, F_C(x)]$, where $F_c(x)$ is the logit for class c. Attribution methods quantify the contribution of each input element $x_{t,m}$ to a target class c relative to a reference (baseline) input $x' \in \mathbb{R}^{T\times M}$.

Gradient-based saliency measures the sensitivity of the target output with respect to the input,

$$\text{Saliency}_c(x) = \frac{\partial F_c(x)}{\partial x}, \tag{8}$$

assigning an importance score to each time-feature pair (t, m). Saliency is computationally efficient but sensitive to noise, lacks a reference point, and may suffer from gradient saturation.

Integrated Gradients (IG) integrates gradients along a straight-line path from x' to x:

$$\text{IG}_c(x) = (x - x') \odot \int_0^1 \nabla_x F_c(x' + \alpha_{IG}(x - x'))\, d\alpha. \tag{9}$$

DeepLIFT (Shrikumar et al., 2017) compares neuron activations with those induced by a reference input x' and propagates contribution scores through the network using layer-wise rules. For input element (t, j), the resulting attribution to target logit F_c is

$$\phi^{\text{DL}}_{t,j} = (x_{t,j} - x'_{t,j})\, m^{F_c}_{x_{t,j}}, \tag{10}$$

where $m^{F_c}_{x_{t,j}}$ denotes the DeepLIFT multiplier obtained by propagating multipliers through the network according to the rescale rule. DeepLIFT-SHAP averages DeepLIFT attributions over a background distribution $\mathcal{D}$ of reference inputs:

$$\phi^{\text{DLSHAP}}_{t,j}(x) = \mathbb{E}_{x'\sim\mathcal{D}}\left[(x_{t,j} - x'_{t,j})\, m^{F_c}_{x_{t,j}}(x, x')\right]. \tag{11}$$

GradientSHAP samples references from $\mathcal{D}$, interpolates, and adds Gaussian noise before differentiating:

$$\text{GradSHAP}_c(x) = \mathbb{E}_{x'\sim\mathcal{D},\, \alpha_G,\, \epsilon}\left[(x - x') \odot \nabla_x F_c(x' + \alpha_G(x - x') + \epsilon)\right], \tag{12}$$

with $\alpha_G \sim \mathcal{U}[0, 1]$ and ϵ Gaussian. The number of interpolation samples materially affects the variance of the estimator; we use $n_{sg} = 50$ rather than the library default of 5, since the downstream per-feature test is variance-sensitive.

A.2.2 GapTrace pipeline

Algorithm 1 summarizes the GapTrace pipeline, from disparity detection and attribution-based feature selection to targeted intervention and comparison against matched controls.

A.2.3 Model-fit assessment

A disparity audit is interpretable only on a model that has learned the task. We therefore report, before any fairness quantity, the accuracy at a prevalence-matched threshold, its excess over the majority-class baseline, the AUC and average precision, the fraction of predictions outside $[0.01, 0.99]$ (saturation), and the per-group accuracy and AUC gaps. A model at or below the majority baseline can still exhibit a large $\text{I}(A; \hat{y})$: it has learned group membership rather than the outcome. Such runs are reported with their fit statistics and excluded from aggregate claims. Per-group AUC gaps are reported alongside the disparity, since unequal discrimination across groups is a distinct failure from unequal risk.

Algorithm 1 GapTrace

Require: frozen f_θ; data $\{(\boldsymbol{x}_i, y_i)\}$ in clustering and test splits; attributes $\mathcal{A} = \{A_1, \ldots, A_K\}$; admissible covariates L; attribution method; α_{FDR}, w, m, λ, n_c
Ensure: for each attribute: Δ, Δ_{adj}, $\mathcal{U}^*$, $\mathcal{S}$, $R(\mathcal{S})$, p_{emp}

$\hat{y}_i \leftarrow \sigma(z_i)$ with z_i the logit of $f_\theta(\boldsymbol{x}_i)$ ▷ test split
choose the dependence statistic from $\hat{y}$ alone (Sec. 3.1)
for $A \in \mathcal{A}$ **do**
    compute Δ (Eq. 1) and permutation p-value p_A for $H_0 \colon \hat{Y} \perp A$
end for
BH-correct $\{p_A\}$; keep attributes with $q_A \le \alpha_{\mathrm{FDR}}$
$\boldsymbol{\phi}_i \leftarrow$ logit attributions; $\Phi_{i,j} \leftarrow \sum_t \phi_{i,t,j}$
if clustering is used **then**
    build clusters from D_{jk} (Eq. 4) on the clustering split
    units $u \leftarrow$ clusters; $\Phi_{i,u} \leftarrow \sum_{j \in u} \Phi_{i,j}$
else
    units $u \leftarrow$ individual features j; $\Phi_{i,u} \leftarrow \Phi_{i,j}$
end if
for each kept attribute A **do**
    $\Delta_{\mathrm{adj}} \leftarrow$ adjusted gap (Eqs. 2–3) ▷ characterizes; does not gate
    for each u: s_u (Eq. 5) and Kruskal–Wallis p-value across groups
    $\mathcal{U}^* \leftarrow \{u : q_u \le \alpha_{\mathrm{FDR}}\}$ after BH correction
    if $\mathcal{U}^* = \emptyset$: report minimum detectable shift and go to next A
    $\mathcal{S} \leftarrow$ selected features $\cup$ top-m members (by $\mathbb{E}_i|\Phi_{i,j}|$) of each selected cluster
    for intervention in (primary, I1, I2, I3) **do**
        $R(\mathcal{S}) \leftarrow \Delta - \Delta_\lambda(\mathcal{S})$ after applying the intervention to $\mathcal{S}$
        draw new $B_1, \ldots, B_{n_c}$ matched on attribution magnitude; compute $R(B_b)$ and p_{emp}
        stop if primary and $R(\mathcal{S}) \ge 0.05\,\Delta$
    end for
    report $\Delta, \Delta_{\mathrm{adj}}, \mathcal{U}^*, \mathcal{S}, R(\mathcal{S}), p_{\mathrm{emp}}$ and the intervention level used
end for

Proxy detection Disparities can arise only if group membership is recoverable from the model's inputs. Per-channel marginal dependence $\mathrm{I}(X_j; A)$ identifies single features that leak group membership; a random-forest probe predicting A from the time-averaged input, evaluated by cross-validated balanced accuracy, flags proxy sets that encode A only jointly. We report the excess over chance rather than raw accuracy, since chance level $1/|\mathcal{G}_A|$ differs across attributes, and fold counts are capped by the smallest group. Where the attribute's own column is a model input it is excluded from the probe. For temporal inputs the probe operates on time-averaged channels and its score is therefore a lower bound: proxies encoded purely in dynamics may evade it while surfacing in the disparity-attribution below.

Temporal localization Because attributions are computed per time step before aggregation, the decomposition extends to when the disparity arises. For group g against reference g_{ref} we form $\Delta(t, j) = \bar{\phi}_{t,j}^{(g)} - \bar{\phi}_{t,j}^{(g_{\mathrm{ref}})}$, whose double sum recovers the aggregate gap for reference-based methods. Point-wise inference over the $T \times M$ map uses a max-statistic permutation scheme, controlling the family-wise error rate across the map while respecting temporal autocorrelation; for multi-group attributes, per-pair maps are corrected across pairs. The analogous map on the inputs localizes when the groups' data differ; comparing the two distinguishes where the disparity enters the data from where the model responds to it, noting that recurrent architectures may lag or smear attribution relative to the input difference.

A.2.4 Attribution Clustering: Implementation and Alternative Backends

Clustering is defined on the attributions rather than on the inputs, because dilution is a property of what the model credits together. Any grouping that yields a partition of the channels can be substituted; the machinery below is shared across backends, and the backends differ in the distance they define.

Preprocessing (all backends). Two steps are applied before any distance is computed. First, each row of $\mathbf{\Phi}$ is centred, removing a per-sample component common to all of a sample's attributions; such an offset correlates every pair of channels and collapses the clustering into a single component. In our clinical cohort, omitting this step produced a largest cluster spanning half the feature space at the same cutoff. Second, constant and numerically silent channels are excluded: their correlations are undefined, and their trajectory descriptors are identically zero, so they occupy the same point in descriptor space and merge into one artefactual cluster. Excluded channels re-enter the partition as singletons, so the cluster set still covers every feature.

Size control and diagnostics (all backends). A distance cutoff alone does not bound cluster size. Any cluster exceeding a cap is re-cut at successively tighter cutoffs until it splits or a floor is reached; clusters that will not split at the floor are reported as unresolved rather than silently retained, since a cluster spanning a large fraction of the feature space cannot support an interpretable claim. We report the size distribution, the branch composition where the input has multiple modalities, and a coherence diagnostic: within-cluster mean $|\mathrm{corr}|$ against a random-pair baseline. Clusters no more internally correlated than random pairs are merge artefacts and are flagged as such. Because summation cancels opposing group effects, any multi-member cluster whose merged spread falls below $\tau_{split} \max_{j \in c} |s_j|$ ($\tau_{split} = 0.5$) is split; a flagged cluster is itself informative, indicating channels the model credits together but in opposite directions across groups.

Validity (all backends). Clusters are estimated on attributions from a split disjoint from the one on which disparities are tested, so the permutation inference on the test split remains exact. Cluster granularity is treated as a hyperparameter and results are reported across a grid of cutoffs.

- Backend 1: temporal (default). The distance of Eq. 4 combines co-credit across patients with agreement in within-stay timing. Implementation settings: $n_{\mathrm{basis}} = 3$ Legendre coefficients, yielding a seven-dimensional descriptor, $w = 0.7$, average linkage, cutoff 0.25, maximum cluster size 45. Average linkage is used because the distance is precomputed and not Euclidean; Ward and centroid linkage assume Euclidean coordinates and, although SciPy will run them on a condensed distance matrix, the result has no interpretation.
- Backend 2: co-credit only ($w = 1$). Setting $w = 1$ in Eq. 4 reduces the distance to $1 - |\mathrm{corr}(\mathbf{\Phi}_{\cdot,j}, \mathbf{\Phi}_{\cdot,k})|$, grouping channels the model credits on the same patients regardless of when within the stay. This is the natural ablation for the temporal term and is the only backend applicable when $T = 1$.
- Backend 3: spectral embedding for large M. For feature spaces where the pairwise distance matrix is unwieldy, the similarity $S_{jk} = |\mathrm{corr}(\mathbf{\Phi}_{\cdot,j}, \mathbf{\Phi}_{\cdot,k})|$ can be embedded with a context-aware spectral method (Vicus) (Wang et al., 2017) and clustered in the embedded space with Ward linkage, which is valid there because the embedding provides Euclidean coordinates. The cluster count is set rather than a cutoff, which bounds the number of clusters but not their size, so the size cap above still applies.
- Backend 4: supplied grouping. Where a physical grouping is available, sensor $\times$ axis for wearable data, or an ontology for clinical measurements, it may be supplied directly in place of any computed distance. This is the most transparent option when the redundancy structure is known in advance, and it serves as a reference point for whether the attribution-derived groupings recover it.

A.2.5 The mutual-information null floor

This appendix substantiates two claims from §3.1: that mutual information can be bounded below the noise floor on a saturated score, and that a permutation p-value cannot be corrected across many hypotheses at a feasible permutation budget.

Why the estimate is positive under independence. Both the k-nearest-neighbour and plug-in estimators of $\mathrm{I}(A; \hat{y})$ are biased upward in finite samples: with G groups and B score bins the plug-in estimator has bias approximately $(G-1)(B-1)/(2N)$ nats, and the kNN estimator has an analogous positive bias that grows with G. Consequently $\mathbb{E}_\pi[\hat{\mathrm{I}}_\pi] > 0$ under permuted labels, and the relevant comparison is against the permutation null rather than against zero. We therefore report the null mean, its 95th percentile, and the bias-corrected excess $\hat{\mathrm{I}} - \mathbb{E}_\pi[\hat{\mathrm{I}}_\pi]$.

Thresholding provides a lower bound on score information. Let $B = \mathbf{1}[\hat{y} > t]$ denote the thresholded prediction, with group-wise selection rates r_g and overall rate $\bar{r} = \sum_g w_g r_g$. Then

$$I(A; B) = \sum_g w_g D_{\mathrm{KL}}(\mathrm{Bern}(r_g) \,\|\, \mathrm{Bern}(\bar{r})) \leq I(A; \hat{y}), \tag{13}$$

where the inequality follows from the data-processing inequality because B is a deterministic function of $\hat{y}$. Thus, thresholded predictions can retain only part of the dependence between group membership and the continuous prediction score.

Why permutation inference cannot be corrected at scale. A permutation p-value is bounded below by $p_{\min} = 1/(n_\pi + 1)$. Benjamini–Hochberg at level α_b over m hypotheses rejects the smallest p only if $p \leq \alpha_b/m$, so rejection is possible at all only when

$$n_\pi \;\geq\; \frac{m}{\alpha_b} - 1. \tag{14}$$

For the $K = 6$ attribute-level screens this requires $n_\pi \geq 119$, which is comfortably met. For per-unit selection over $m \approx 525$ clusters it requires $n_\pi \geq 10{,}499$ per attribute, and over $m = 856$ raw channels $n_\pi \geq 17{,}119$.

Calibrated normal tail as an alternative. Where a permutation null is required at scale, the floor can be avoided by using the permutation distribution only to calibrate location and scale and taking a normal tail probability,

$$z = \frac{S - \mathbb{E}_\pi[S_\pi]}{\mathrm{sd}_\pi[S_\pi]}, \qquad p = 2\,\Phi(-|z|), \tag{15}$$

which has no lower bound. This trades the exactness of the permutation test for an assumption of approximate normality in the tail of the null, which we verify by comparing the empirical and normal quantiles of S_π.

A.2.6 Sensitivity of the feature selection

Standardized attribution differences. Statistical significance and effect magnitude are reported separately. For each testing unit u, we define the standardized attribution difference as

$$\delta_u = \frac{s_u}{\mathrm{SD}(\Phi_{\cdot,u})} = \frac{\max_g \bar{\Phi}_u^{(g)} - \min_g \bar{\Phi}_u^{(g)}}{\mathrm{SD}(\Phi_{\cdot,u})}, \tag{16}$$

where the denominator is the standard deviation of the unit's attribution across all samples in the evaluation set. Thus, δ_u expresses the max–min difference in group mean attribution in units of the overall attribution standard deviation.

In the reported experiments, we set $\delta_{\min} = 0$, so effect magnitude does not impose an additional selection criterion beyond the Kruskal–Wallis test with Benjamini–Hochberg correction. We therefore report δ_u descriptively for every selected testing unit, alongside s_u and the BH-adjusted q-value. This separates statistical evidence for a between-group attribution difference from the magnitude of that difference, which is particularly important in large samples where small effects may be statistically detectable.

Minimum detectable attribution shift. An empty selected set is difficult to interpret without knowing what magnitude of attribution difference the selection procedure could reliably detect. We therefore complement the observed effect sizes above with an empirical sensitivity analysis.

Group labels are first permuted so that the attribution matrix retains its empirical covariance structure but contains no systematic association with group membership. We then select n_{inj} testing units at random and add a shift of δ attribution standard deviations to one group for those units. The complete selection procedure is then reapplied, including the Kruskal–Wallis tests and Benjamini–Hochberg correction. We sweep δ over a prespecified grid and repeat the experiment across random injections.

For each value of δ, we report the proportion of injected units recovered and the number of non-injected units selected. We define the minimum detectable attribution shift as the smallest injected

standardized shift for which at least 80% of injected units are recovered. This quantity characterizes the empirical sensitivity of the selection procedure and is distinct from the observed effect size δ_u: δ_u describes the magnitude of an observed between-group attribution difference, whereas the minimum detectable shift describes the magnitude that the procedure can reliably recover under the injection experiment.

The sensitivity bound applies to individual testing units. It does not exclude a disparity that is distributed across many units, each carrying an attribution difference below the detection threshold. Accordingly, an empty selected set should be interpreted as an absence of individually detectable units at the sensitivity of the procedure, rather than as evidence that no group-dependent attribution structure is present.

A.2.7 Intervention: controls, degenerate cases, and diagnostics

This appendix collects the operational detail of the intervention of §3.4: how controls are drawn, how the procedure behaves when the control pool is inadequate, and which reductions are treated as uninterpretable.

Intervention levels. The primary intervention is the rank-preserving quantile alignment defined in Eq. 6, applied with $\lambda = 1$. This gives every group the same marginal distribution on the selected feature set $\mathcal{S}$ while preserving within-group ranks. A small reduction under marginal alignment does not imply that the selected features are irrelevant, since the disparity may depend on higher moments, joint dependencies, or more complex use of $\mathcal{S}$. We therefore consider three stronger interventions of increasing strength, applied to the fixed model on a copy of the evaluation data.

- **I1 Moment-aware alignment**: feature-specific mean, scale, or quantile transformations are applied to remove additional group differences not captured by the primary marginal alignment.
- **I2 Joint transplantation**: samples in the target group $g^\star$ receive the values of all features in $\mathcal{S}$ from a single randomly drawn sample of the pooled non-target groups, preserving the joint structure of the transplanted features while removing their original group-specific values.
- **I3 Ablation**: every feature in $\mathcal{S}$ is replaced by its pooled mean trajectory for all samples, removing the information carried by $\mathcal{S}$. Failure to obtain a material gap reduction even under I3 indicates that the selected set is insufficient to account for the observed gap under these interventions.

The target group is the group whose mean prediction lies furthest from the cohort mean, $g^\star = \arg\max_g \left|\mathbb{E}[\hat{y} \mid A = g] - \mathbb{E}[\hat{y}]\right|$.

Escalation rule. We call a reduction *material* when $\Delta - \Delta_{\text{int}} \geq \kappa\Delta$, with $\kappa = 0.05$ fixed in advance. The primary quantile intervention is always applied first. If it does not produce a material reduction, I1, I2, and I3 are evaluated in turn, and we report the first intervention that does. If none produces a material reduction, we report the primary intervention and conclude that $\mathcal{S}$ is insufficient to account for the observed gap under the interventions considered. Materiality concerns only the magnitude of the reduction; whether the reduction exceeds what a comparable perturbation achieves is assessed separately against matched controls. When k intervention levels are tested, we report the empirical p-value of the reported intervention together with its Bonferroni-adjusted value $\min(1,\, k\, p_{\text{emp}})$.

Matched controls. Any perturbation of sufficient magnitude can alter predictions, so a gap reduction is meaningful only relative to that produced by a comparable perturbation of features not selected by GapTrace. For each intervention, we draw n_c control sets of $|\mathcal{S}|$ features from the unselected pool and apply the same transformation with the same parameters. Controls are matched to the selected features on both attribution magnitude and between-group input divergence. The unselected pool is stratified according to these quantities, and each control set is sampled without replacement, with fallback to the nearest non-empty stratum when necessary.

Uniform sampling would preferentially draw from the inert channels that dominate a large feature space; in our clinical cohort, roughly a fifth of channels are constant. Such controls would therefore

perturb features the model does not use and could make the targeted intervention appear favourable regardless of its specificity. Controls are constructed separately for each intervention bound and undergo the same transformation as the selected features, so the targeted gap reduction is always evaluated against a comparably matched control intervention.

Resolution of the empirical p-value. The reported $p = (1 + \#\{\text{control} \geq \text{targeted}\})/(1 + n_c)$ is bounded below by $1/(n_c + 1)$, so a targeted repair that exceeds every control attains $p = 1/(n_c + 1)$ and no smaller value. We therefore set n_c from the resolution required rather than leaving it at a default, and report it alongside every p-value. Each additional control costs one repair pass and one forward pass over the evaluation set, so the cost is linear in n_c (Table 12).

Degenerate control pools. Two cases arise on small feature spaces. When the pool admits fewer than n_c distinct draws of size $|\mathcal{S}|$, we enumerate the available control sets exhaustively and report the pool size with the result: resampling a small pool with replacement produces near-identical control sets and understates the control spread, which inflates the apparent excess over random. When the pool is smaller than $|\mathcal{S}|$ no control set exists at all, and the comparison is reported as unavailable rather than as a null result, an important distinction, since an absent control and a control the targeted repair failed to beat render identically in a figure. This case arises whenever selection retains most of a small feature space, and we cap the per-cluster expansion of $\mathcal{S}$ partly to avoid it.

Reductions that are not interpretable. Three outcomes are reported without an effect size. A reduction measured against a pre-intervention gap that does not itself exceed its permutation null is reported as null, since the quantity being reduced is indistinguishable from noise. A reduction on an attribute for which $\mathcal{S}$ is empty is not computed, as the fallback to top-k features would describe a set the selection did not endorse. And a repair that increases the gap is diagnosed as off-manifold extrapolation rather than as evidence about $\mathcal{S}$: Eq. 6 equalises a feature's marginal within each group while ignoring its dependence on the remaining features, so a repaired sample can fall outside the region the model was fitted on, and the model's response there carries no information about the feature's role.

Expected behaviour under the validation protocol. A correct audit yields no intervention effect distinguishable from matched controls on uncorrupted data and, after a controlled injection, an effect concentrated on the injected attribute. The injection designs of §A.8 probe the intervention bounds differently: a label-independent shift on an informative channel alters the group-wise marginal distribution and should therefore produce a material gap reduction under I1, whereas a label-dependent shift induces a group-specific predictive relationship that may not be captured by marginal alignment and should appear primarily in the I2 bound. These controlled injections therefore validate whether the intervention bounds respond as expected to known disparity mechanisms, rather than inferring the mechanism from the observed intervention response.

A.2.8 Conditional disparity across correlated attributes

Why a marginal disparity is not attributable. Sensitive attributes in clinical cohorts are correlated by construction: Medicare eligibility begins at 65, widowhood tracks age, and minority patients in our cohort are younger than White patients (Fig. 5). A marginal gap for attribute A therefore mixes a contribution specific to A with one inherited from a correlated attribute Z, and the two are indistinguishable without conditioning. The chain rule makes this disparity-attribution explicit:

$$\mathrm{I}(A, Z; \hat{y}) = \underbrace{\mathrm{I}(Z; \hat{y})}_{\text{carried by } Z} + \underbrace{\mathrm{I}(A; \hat{y} \mid Z)}_{\text{specific to } A}, \tag{17}$$

so the quantity of interest is the conditional term: what A tells us about the prediction beyond what is already captured by Z. A marginal disparity that vanishes under conditioning is attributable to the association between A and Z; one that persists reflects information associated with A beyond Z.

The estimator we report. Eq. 17 motivates the conditioning operation, but direct estimation presents practical limitations. A stratified estimator requires cells large enough to support a within-stratum estimate, and with K attributes of G groups the number of cells grows as G^K. Coverage therefore falls quickly as the conditioning set expands, and the resulting estimate may describe only a

shrinking subset of the cohort. We therefore report the regression form of Eq. 2, which conditions on admissible covariates without discretising them, preserves the units of Eq. 1, and yields a confidence interval directly. The stratified conditional-MI estimator is retained in the implementation as a robustness check and agrees with the regression on attributes for which cell coverage exceeds 60%.

Choice of conditioning set. For the MIMIC-IV analysis, the conditioning set L for each audited attribute consists of the remaining five demographic attributes. Thus, when auditing age, we condition on race/ethnicity, sex, language, insurance, and marital status; for each of the other audited attributes, L analogously consists of the remaining demographic attributes. These conditioning analyses are descriptive rather than causal: variables in L may be statistically or structurally related to the audited attribute, so attenuation or amplification after conditioning should be interpreted as shared predictive association rather than as a causal decomposition of the disparity. Explicit specification of L is important because conclusions about conditional disparities depend on which variables are conditioned on; as Corbett-Davies et al. (2017) note, conditional fairness criteria can become uninformative if every observed variable is treated as legitimate for conditioning. We therefore report the conditioning set explicitly and interpret the resulting conditional disparities relative to that set.

Assumptions and limitations. Conditioning is an associational operation. Eq. 17 identifies what A adds beyond Z in the observed data; it does not identify a causal effect of A, and no adjustment recovers one without additional assumptions that we do not make. Unmeasured common causes of A and $\hat{y}$ may remain in the residual association, so a surviving disparity should not be interpreted as a causal effect specific to A. Eq. 2 is fitted jointly only when the maximum variance inflation (O'brien, 2007) factor across $A \cup L$ is below 10; otherwise, each attribute is adjusted for one covariate at a time and the most attenuating adjustment is reported. This pairwise procedure is immune to multivariable collinearity and supports the conservative claim by reporting the adjustment that explains the largest share of the marginal disparity. An adjusted gap exceeding the unadjusted one is reported as inflated rather than as negative explained variance. In our cohort, this suppression effect appears for race and language, whose younger age composition masks part of their marginal gap until age is held fixed.

A.2.9 PROPOSITION 1

For a softmax classifier, probability-space gradients are related to logit-margin gradients by a prediction-dependent scaling factor. This relationship also extends to the loss gradients used in the optional boundary-regression formulation.

Proof. Let $z \in \mathbb{R}^C$ denote the vector of logits and let z_k be the logit corresponding to class k. Define the logit margin

$$m_k = z_k - \log \sum_{j \neq k} e^{z_j}.$$

Note that this can be rewritten as

$$m_k = \log \frac{e^{z_k}}{\sum_{j \neq k} e^{z_j}}.$$

Recall that the softmax probability for class k is

$$p_k = \frac{e^{z_k}}{\sum_{j=1}^{C} e^{z_j}} = \frac{1}{1 + \sum_{j \neq k} e^{z_j - z_k}}.$$

Taking the logit of p_k, we obtain

$$\log \frac{p_k}{1 - p_k} = z_k - \log \sum_{j \neq k} e^{z_j} = m_k.$$

Thus,

$$p_k = \sigma(m_k).$$

By the chain rule, the gradient with respect to the input x satisfies

$$\nabla_x p_k = p_k(1 - p_k) \nabla_x m_k.$$

Table 3: Comparison of GapTrace with related approaches. A checkmark indicates that the corresponding capability is explicitly supported by the method or evaluation framework. GapTrace treats subgroup prediction disparities as the object of explanation, localizing them across features and time and evaluating the identified pathways through statistical testing and controlled interventions.

Method	Subgroup gap analysis	Feature-level differential attribution	Time-level differential attribution	Statistical testing	Intervention validation	Ground-truth validation
DeepSHAP (Lundberg & Lee, 2017)	–	–	✓	–	–	–
Time-series XAI benchmarks (Ismail et al., 2020; Turbé et al., 2023)	–	–	✓	✓	–	✓
Grouped feature importance (Au et al., 2022)	–	–	–	–	–	–
Class-contrastive explanations (Wang & Wang, 2022)	–	–	✓	–	–	–
G-DeepSHAP (Chen et al., 2022)	✓	–	–	–	–	–
Fairness and interpretability analysis (Meng et al., 2022)	✓	–	–	–	–	–
FairShap (Arnaiz-Rodriguez et al., 2023)	✓	–	–	–	–	–
GapTrace (ours)	✓	✓	✓	✓	✓	✓

Hence, as in the binary case, probability-space gradients are rescaled relative to logit-margin gradients by the prediction-dependent factor $p_k(1 - p_k)$.

For the optional boundary-regression formulation, consider the squared margin loss

$$\mathcal{L}_m = (m_k - \tau)^2.$$

Its gradient with respect to the logits z is

$$\nabla_z \mathcal{L}_m = 2(m_k - \tau)\nabla_z m_k.$$

Now consider the squared probability loss

$$\mathcal{L}_p = (p_k - \alpha)^2,$$

whose gradient is

$$\nabla_z \mathcal{L}_p = 2(p_k - \alpha)\nabla_z p_k.$$

Since

$$\nabla_z m_k = \frac{1}{p_k(1 - p_k)} \nabla_z p_k,$$

we obtain

$$\nabla_z \mathcal{L}_m = \frac{2(m_k - \tau)}{p_k(1 - p_k)} \nabla_z p_k.$$

Therefore, $\nabla_z \mathcal{L}_m$ and $\nabla_z \mathcal{L}_p$ are scalar multiples of the same probability gradient. With $\alpha = \sigma(\tau)$, the two objectives have corresponding decision boundaries and induce collinear gradients away from their common boundary. □

A.3 Further Comparison with Related Work

Table 3 summarizes the distinction between GapTrace and closely related interpretability and fairness approaches. Standard attribution methods explain individual model predictions, while existing fairness analyses primarily quantify disparities at the group level. GapTrace instead treats the subgroup prediction gap itself as the quantity to be explained, decomposing it across features and time steps and subsequently evaluating the identified contributions through statistical testing and targeted interventions.

A.4 Case study: measurement bias in a single input channel

Peripheral oxygen saturation (SpO_2) measured by pulse oximetry has been reported to overestimate arterial oxygen saturation more often in patients with darker skin pigmentation, so that occult hypoxaemia is missed at a higher rate (Sjoding et al., 2020). The channel is not a proxy for race in the usual sense: it is a physiological measurement whose error is group-dependent, and a model trained on it inherits that error without any group label being present in the input.

This is the mechanism our framework is designed to localise, and it is structurally the case our controlled injection reproduces (§A.8): a systematic offset applied to one group on a channel the model already relies on. The audit should (i) register an outcome-level disparity, (ii) attribute it to the affected channel rather than to correlated ones, and (iii) close the gap when that channel's group-wise distribution is repaired, and the level of the repair ladder at which it closes distinguishes a shifted distribution from a group-specific decision rule.

A.5 Implementation Details

All analyses are performed using a fixed, pre-trained model so that the measured disparities reflect the learned decision function rather than changes introduced during optimisation. Model weights are restored from a checkpoint, after which the model is frozen and evaluated in inference mode throughout, with dropout disabled and normalisation statistics fixed. We verify that all parameters are successfully restored and abort the analysis otherwise. This check is important because a partially loaded checkpoint may execute without error while retaining parameters at their initial values, potentially producing a complete but invalid audit. As an additional sanity check, we require the restored model to outperform the majority-class baseline on the evaluation split before reporting any disparity measures.

Reported disparities are computed on the probability scale, $\hat{y} = \sigma(z)$, so that the resulting effects can be interpreted directly as differences in predicted risk. Attributions, however, are computed with respect to the corresponding logit z (using a one-vs-rest formulation for multi-class outputs). Computing attributions directly on the probability scale would introduce an additional multiplicative factor of $\hat{y}(1-\hat{y})$. Because this factor depends strongly on the predicted probability, it can systematically rescale attribution magnitudes across groups whenever the groups differ in predicted risk—the very quantity being investigated. Computing attributions on the logit scale avoids this additional source of group-dependent scaling.

Architectures. We evaluate the proposed audit framework across three neural architectures selected to match the characteristics of the corresponding datasets. For the semi-synthetic dataset, we use a Temporal Convolutional Network (TCN) (Bai et al., 2018). For the simulated dataset, we use an attention-based bidirectional LSTM (Attentive BiLSTM) (Hochreiter & Schmidhuber, 1997; Bahdanau et al., 2014). For the MIMIC dataset, we use an LSTM-based architecture (Gupta et al., 2022; Harutyunyan et al., 2019).

To ensure that comparisons primarily reflect differences in attribution behaviour rather than unrelated architectural choices, the TCN and Attentive BiLSTM models use a consistent attention-pooling mechanism, per-sample normalisation, and dropout configuration where applicable.

Attribution. Baselines are drawn from the training distribution rather than an all-zero reference: $n_{\text{base}} = 30$ samples in expected mode, or samples from a designated reference group in group-reference mode. Integrated Gradients uses 32 steps averaged over 4 baselines with an internal batch size of 16; without an internal batch, Captum expands the batch to (batch $\times$ steps), which on the transformer raised peak activation memory by more than an order of magnitude. GradientSHAP uses $n_s = 50$ interpolation samples rather than the library default of 5, which is too noisy for a variance-sensitive per-feature test. The cuDNN fused RNN kernel is disabled only for recurrent backbones, where it does not support the second derivative Captum requires; leaving it disabled for convolutional backbones is a substantial and unnecessary slowdown.

Statistics. The dependence screen uses $n_\pi = 500$ label permutations with the test statistic selected from the shape of the score before any p-value is computed (Sec. 3.1). Effect sizes and their confidence intervals use $n_{\text{boot}} = 1000$ resamples. Feature and cluster selection uses a Kruskal–Wallis test (Kruskal & Wallis, 1952) with Benjamini–Hochberg correction(Benjamini & Hochberg, 1995) at $\alpha = 0.05$ across all testing units. We use the Kruskal–Wallis test with BH correction as the prespecified per-unit statistic. Temporal maps use $n_\pi = 1000$ permutations of the max statistic over the $T \times M$ map, controlling the family-wise error rate across the map. The proxy probe is a 200-tree random forest with 5-fold cross-validation, scored as excess over the majority-class rate. Groups smaller than $n_{\min} = 50$ are merged into a residual category before testing, and intersectional cells smaller than 20 are excluded and the resulting coverage reported.

Clustering and intervention. For the reported MIMIC-IV experiments, attribution clustering uses the temporal distance of Eq. 4 with $\omega = 0.7$, average-linkage hierarchical clustering, a distance cutoff of 0.25, and a maximum cluster size of 45. Clusters are constructed using the training split, separately from the evaluation split used for subsequent disparity testing.

A.6 Simulated Data: Experimental Setup and Additional Results

A.6.1 Data generation

Each simulated dataset contains $N = 10{,}000$ samples of $M = 6$ channels over $T = 500$ time steps. Every channel is a sine carrier with frequency drawn uniformly from $\{2, \ldots, 5\}$ and amplitude 0.5. A channel is made informative by adding a sine burst of length 100, unit amplitude and frequency drawn from $\{10, \ldots, 50\}$, at a uniformly random onset. The remaining channels instead receive a unit-amplitude square-wave segment, or nothing at all, chosen at random per sample, so that they carry comparable structure while contributing no label information. All channels then receive independent Gaussian noise with $\sigma = 0.15$.

The binary label sums the burst frequencies of the sample's informative channels and thresholds that sum at the median of its reference distribution, estimated by Monte Carlo over 10^4 draws; class 1 is the above-median side and is the class for which attributions are computed. The sensitive attribute is a balanced binary gender variable and is not an input to the model.

The disparity mechanism assigns different informative sets to the two groups. All four candidate channels are informative for the reference group (S_{ref}), while only a nested subset is informative for the target group ($S_{\text{tgt}} \subset S_{\text{ref}}$, with $|S_{\text{ref}} \setminus S_{\text{tgt}}| = d$; the target group is the female group in the generator). The d channels in $S_{\text{ref}} \setminus S_{\text{tgt}}$, which is equivalently the symmetric difference of the two sets since they are nested, are the ground truth: each carries a label-bearing burst for the reference group and is generated by the uninformative rule for the target group. The label threshold is computed separately for each group from that group's number of informative channels, so class prevalence is matched and the mechanism plants no outcome gap by construction.

We enumerate every subset at $d = 1$ (four configurations) and $d = 2$ (six configurations), each over five seeds, giving 50 runs. Each run is split $70/15/15$ into training, validation and test. The Attentive BiLSTM is trained for 100 epochs with inverse-frequency class-weighted cross-entropy and the best-validation checkpoint is retained; across the 50 runs it reaches 0.85 test accuracy and 0.94 AUC for class 1, with a 0.03 accuracy gap between groups, so attributions are computed on a model that has learned both groups' rules. For the simulated arm, attribution baselines use $n_{\text{base}} = 30$ training samples.

A.6.2 Scoring

The unit of evaluation is one channel in one run, so each run contributes $M = 6$ decisions and the 50 runs contribute 300. The d ground-truth channels are the positives (80 in total) and the remaining $6 - d$ are the negatives (220 in total). A channel counts as selected if the selection procedure of Sec. 3.2 marks it significant after Benjamini–Hochberg correction at $\alpha = 0.05$. Counts are accumulated over all runs and micro-averaged, so a $d = 2$ run contributes twice as many positives as a $d = 1$ run. The random baseline selects each channel independently with probability 0.5, giving $\text{TPR} = \text{TNR} = 0.50$ and a precision equal to the ground-truth prevalence, $80/300 = 0.27$.

The negatives are of two kinds, and the distinction is the one the framework rests on. Of the 220 negative decisions, 120 are channels informative for both groups: the model depends on them for prediction, but they carry no group difference. The remaining 100 are the uninformative channels. Selection rejects the two kinds almost equally well. For DeepLIFT 0.96 on the shared informative channels against 0.97 on the uninformative ones, and for Integrated Gradients 0.99 against 0.97. So a channel being important for prediction does not by itself make it a contributor to the disparity.

A.6.3 Recovery across settings

Because class prevalence is matched, the mechanism produces only a small marginal disparity, with a mean difference in predicted positive rate between groups of $2.9\,\text{pp}$ at $d = 1$ and $4.8\,\text{pp}$ at $d = 2$. Nevertheless, the ground-truth group-specific channels are recovered with a TPR of 0.80 for DeepLIFT and 0.90 for Integrated Gradients (Table 1). This shows that feature-level differential attribution can remain detectable even when its marginal expression in the predicted-score distribution is small.

Table 4 breaks this down by d. This confirms that the pooled values in Table 1 are not an artifact of averaging across substantially different settings: each of the four reference-based methods remains

Table 4: Recovery on simulated data by the number of channels d whose label-bearing status differs between groups, micro-averaged over 20 runs at $d = 1$ and 30 at $d = 2$, $\pm$ the standard deviation across the five seeds. Only TPR is shown: TNR and precision vary little between the two settings (0.94–0.99 and 0.87–0.97 across the four reference-based methods), so the pooled values in Table 1 are representative of both.

Method	TPR at $d = 1$	TPR at $d = 2$
GapTrace-DeepLIFT	0.85 ± 0.22	0.78 ± 0.05
GapTrace-Integrated Gradients	0.85 ± 0.22	0.92 ± 0.06
GapTrace-GradientSHAP	0.60 ± 0.22	0.63 ± 0.14
GapTrace-DeepLIFTSHAP	0.85 ± 0.22	0.78 ± 0.05

within 0.05 of its pooled TPR in both settings, and all remain above the random baseline of 0.50. The relative performance of the two strongest methods does differ by setting. DeepLIFT and Integrated Gradients tie at 0.85 when a single channel differs between groups, whereas Integrated Gradients performs better at $d = 2$ (0.92 versus 0.78). Because $d = 2$ contributes 60 of the 80 ground-truth positive decisions, the pooled margin between the two methods is driven primarily by this setting rather than holding uniformly across configurations. GradientSHAP is the least sensitive of the four reference-based methods but the most stable across settings, with TPRs of 0.60 and 0.63. The wider seed-to-seed spread at $d = 1$ follows from there being only four ground-truth decisions per seed in that setting, which restricts per-seed TPR to multiples of 0.25; the pooled value over all 20 runs is the one to compare.

A.7 PADS DATASET

The PADS dataset (Varghese et al., 2024) consists of six accelerometer and gyroscope features and includes multiple neurological conditions with varying degrees of similarity to Parkinson's disease (PD). Each session contains 20.48 seconds of accelerometer and gyroscope measurements sampled at 100 Hz, corresponding to 2,048 time points per channel. The dataset is notably imbalanced, with most samples belonging to PD (276) and healthy controls (79), followed by Other Movement Disorders (OMD; 60), Essential Tremor (ET; 28), Atypical Parkinsonism (AP; 15), and Multiple Sclerosis (MS; 11). We restrict our analysis to the "Relaxed" and "RelaxedTask" sessions and use signals recorded from the left wrist.

From a modeling perspective, these conditions present different levels of classification difficulty. 1. Parkinson's disease (PD): The primary in-distribution class, characterized by resting tremor and motor impairment. 2. Healthy controls: Represent typical movement patterns and provide a baseline for distinguishing pathological cases. 3. Essential Tremor (ET): Characterized primarily by tremor during voluntary movement rather than at rest. Despite this distinction, ET can resemble PD in sensor data, making it difficult to differentiate and a potential source of misclassification. 4. Atypical Parkinsonism (AP): Encompasses disorders with Parkinson-like symptoms but different underlying etiologies and typically faster progression. These conditions often exhibit patterns that overlap with PD, making them particularly challenging to distinguish and relevant for out-of-distribution detection. 5. Multiple Sclerosis (MS): Associated with coordination and balance impairments that may produce movement patterns distinct from those of PD, potentially making MS samples easier to identify as anomalous. 6. Other Movement Disorders (OMD): A heterogeneous category comprising various conditions with diverse movement patterns, adding further variability and complexity to the dataset.

A.8 CONTROLLED DISPARITY INJECTION

To evaluate whether the audit recovers a known disparity, we plant one in data whose ground truth we control. A planted corruption is only a useful positive control if it reaches the model's predictions, and this is not automatic: a shift applied to a channel the model does not use produces no outcome gap at any magnitude, and a shift that carries no label information gives a well-fit model no reason to attend to it. We therefore select the injection target and its magnitude from a model trained on uncorrupted data, and verify propagation before running the audit.

Table 5: Injection designs. g denotes protected-group membership, y the label, σ_j the median within-window standard deviation of channel j.

Design	Rule (for $i \in g$)	Channel predictive?	Group mean moves?
(i) shift	$x_{ij} \mathrel{+}= \delta\sigma_j$	no	yes
(ii) shortcut	$x_{ij} \mathrel{+}= \delta(2y_i - 1)\sigma_j$	yes	no
(iii) asym	$x_{ij} \mathrel{+}= \delta\sigma_j$ if $y_i = c$	yes	yes

Target selection. We first train a clean model and rank input channels by permutation importance, measuring the mean absolute change in predicted probability when each channel is shuffled across samples. The injection targets a highly ranked channel, since a shift on an ignored channel cannot move the outcome. We scale the shift by the median within-window standard deviation σ_j of the target channel rather than its pooled standard deviation; in PADS the two differ by a factor of sixteen because a small number of high-amplitude windows dominate the pooled estimate, and using the pooled value yields an injection several times larger than a typical window's entire signal.

Injection designs. Let g denote membership in the protected group and $y \in \{0, 1\}$ the label. For a window i and channel j we consider three mechanisms, which differ in whether the corrupted channel carries label information and whether it moves the group mean:

$$\text{(i)} \quad x_{ij} \leftarrow x_{ij} + \delta\,\sigma_j, \qquad i \in g, \tag{18}$$

$$\text{(ii)} \quad x_{ij} \leftarrow x_{ij} + \delta\,(2y_i - 1)\,\sigma_j, \qquad i \in g, \tag{19}$$

$$\text{(iii)} \quad x_{ij} \leftarrow x_{ij} + \delta\,\sigma_j, \qquad i \in g,\ y_i = c, \tag{20}$$

where $c \in \{0, 1\}$ selects which label side is shifted. Design 18 is measurement bias: the channel's values are systematically offset for one group, and because the model already uses the channel no new learning is required. Design 19 makes the channel predictive within the group, so reliance grows with training, but the shift is symmetric in y and the group mean is unchanged by construction. Design 20 keeps the channel predictive within the group while moving the group mean, so the disparity is visible in the outcome; it is our primary positive control. Designs 18 and 19 serve as mechanism contrasts: the first tests repairability, the second produces a disparity that is invisible to any outcome-gap metric.

Confound control. Protected attributes are often associated with the label independently of any injection; in PADS the positive rate differs by 23 percentage points between gender groups. Where such an association exists, we restrict design 18 to a subset $S \subseteq g$ chosen so that $P(y{=}1 \mid i \in S)$ matches the split's global positive rate. The shift then carries no label information by construction, and the corruption is not confounded with the pre-existing association. Group coverage $|S|/|g|$ is reported per split.

Consistency and verification. The same corruption, identical channel, magnitude, and scaling constant σ_j, is applied to the training, validation, and test splits. All splits are standardised with a scaler fitted once on clean training data, so the injection is identical in scaled units across arms and the loss is computed with class weights derived from the clean label distribution. Before any audit we confirm that the corruption reached the predictions, measuring the change in the group mean prediction, the per-group AUC, and the per-group selection rate against the clean arm. For designs 19 and 20, which alter per-group discrimination without necessarily altering group means, the mean gap alone is not a sufficient check. Corruption specifications are serialised and reloaded by the auditing stage, which reconstructs the identical dataset the model was trained on.

A.8.1 Semi-synthetic Data Creation

The dose–response experiments were used to determine injection magnitudes for the subsequent controlled audits (Tables 8 and 9). The objective was not to use the same numerical magnitude for each protected attribute, but rather to select an amount that produces a controlled and detectable disparity for that attribute while preserving meaningful model behaviour. This is important because the same perturbation magnitude need not have the same effect across subgroup definitions: the groups differ in their baseline composition, label association (Table 6), and model response to the injected feature.

Table 6: Group composition and label association.

Dataset	Attribute	Split	$\lvert g\rvert$	N	$P(y{=}1 \mid g)$	$P(y{=}1 \mid \neg g)$
PADS	gender	train	1204	2982	0.640	0.874
PADS	gender	test	322	994	0.739	0.875
PADS	age	train	952	2982	0.765	0.786
PADS	age	test	280	994	0.900	0.804

Table 7: Seed variation with the injected channel pinned. All runs use channel 2 ($\sigma_2 = 0.00222$); Δ is measured against the clean arm of the same seed. Gender's clean-arm gap varies by 5.3 pp across seeds while the injected gap varies by 1.1 pp, so effects are reported as differences rather than raw gaps. The per-group AUC difference changes sign between amounts 0.10 and 0.15 for age in every seed.

		clean	amount 0.15			amount 0.25		
Attribute	Seed	Gap (pp)	Gap (pp)	Δ gap	ΔAUC	Gap (pp)	Δ gap	ΔAUC
gender	0	-15.13	$+10.24$	$+25.4$	$+0.277$	$+14.62$	$+29.8$	$+0.346$
gender	1	-10.70	$+10.37$	$+21.1$	$+0.290$	$+12.39$	$+23.1$	$+0.289$
gender	2	-16.02	$+9.24$	$+25.3$	$+0.206$	$+12.93$	$+28.9$	$+0.269$
gender	mean	-13.95	$+9.95$	$+23.9$	$+0.258$	$+13.31$	$+27.3$	$+0.301$
gender	sd	2.85	0.62	2.5	0.045	1.17	3.6	0.040

		clean	amount 0.10			amount 0.15		
Attribute	Seed	Gap (pp)	Gap (pp)	Δ gap	ΔAUC	Gap (pp)	Δ gap	ΔAUC
age	0	$+9.13$	$+12.78$	$+3.7$	-0.015	$+26.48$	$+17.4$	$+0.113$
age	1	$+11.07$	$+21.57$	$+10.5$	-0.038	$+31.55$	$+20.5$	$+0.089$
age	2	$+9.23$	$+23.32$	$+14.1$	-0.105	$+37.40$	$+28.2$	$+0.113$
age	mean	$+9.81$	$+19.22$	$+9.4$	-0.053	$+31.81$	$+22.0$	$+0.105$
age	sd	1.09	5.61	5.3	0.047	5.46	5.5	0.014

Table 8: Design (iii) dose–response, PADS gender, TCN seed 0, channel 2. The per-group AUC difference and val AUC both separate from the clean arm between amounts 0.10 and 0.15.

Amount	Gap (pp)	Δ gap	AUC$_g$	AUC$_{\neg g}$	ΔAUC	sel$_g$	sel$_{\neg g}$	val AUC
clean	-9.30	—	—	—	—	—	—	0.672
0.05	-6.79	$+2.5$	0.708	0.647	$+0.060$	0.276	0.374	0.673
0.10	-4.39	$+4.9$	0.739	0.647	$+0.092$	0.323	0.385	0.686
0.15	$+11.38$	$+20.7$	0.936	0.670	$+0.267$	0.761	0.649	0.740
0.20	$+10.64$	$+19.9$	0.952	0.618	$+0.334$	0.798	0.744	0.755
0.25	$+14.57$	$+23.9$	0.952	0.657	$+0.295$	0.767	0.621	0.771
0.50	$+22.36$	$+31.7$	0.978	0.604	$+0.374$	0.720	0.496	0.752
1.00	$+26.49$	$+35.8$	0.991	0.565	$+0.425$	0.742	0.484	0.755
2.00	$+20.09$	$+29.4$	0.994	0.636	$+0.359$	0.789	0.592	0.755
4.00	$+26.14$	$+35.4$	1.000	0.589	$+0.411$	0.795	0.613	0.736

For age, we use an injection amount of 0.10. As shown in Table 9, at this magnitude the perturbation is deliberately subtle: the prediction gap increases only modestly relative to the clean arm, from 9.13 pp to 12.78 pp for seed 0, while the per-group AUC difference remains negative ($\Delta\mathrm{AUC} = -0.015$). Increasing the amount to 0.15 produces a much sharper transition, with the gap increasing to 26.48 pp and $\Delta\mathrm{AUC}$ changing sign to $+0.113$. The same transition in $\Delta\mathrm{AUC}$ between 0.10 and 0.15 is observed across seeds (Table 7). We therefore select 0.10 as a more challenging positive control, allowing us to test whether the proposed audit can recover the planted disparity-driving channel before the injected effect becomes dominant in conventional group-level performance measures.

Table 9: Design (iii) dose–response, PADS age, TCN seed 0, channel 2. ΔAUC changes sign between 0.10 and 0.15.

Amount	Gap (pp)	Δ gap	AUC_g	$\text{AUC}_{\neg g}$	ΔAUC	sel_g	$\text{sel}_{\neg g}$	val AUC
clean	+9.13	—	—	—	—	—	—	0.674
0.05	+7.65	−1.5	0.620	0.698	−0.078	0.404	0.310	0.669
0.10	+12.78	+3.7	0.672	0.687	−0.015	0.464	0.294	0.702
0.15	+26.48	+17.4	0.839	0.726	+0.113	0.904	0.580	0.774
0.20	+25.26	+16.1	0.884	0.688	+0.196	0.943	0.682	0.788
0.25	+44.47	+35.3	0.886	0.709	+0.177	0.936	0.391	0.807
0.50	+51.33	+42.2	0.936	0.713	+0.223	0.939	0.352	0.811
1.00	+51.11	+42.0	0.986	0.730	+0.256	0.939	0.368	0.814
2.00	+43.01	+33.9	0.993	0.747	+0.246	0.961	0.499	0.821
4.00	+46.03	+36.9	0.998	0.768	+0.230	0.961	0.444	0.821

Table 10: Direction control for design (iii). Shifting the positive label side of the protected group increases the prediction gap, whereas shifting the negative label side reverses its direction and inverts the selection-rate ordering. PADS gender, amount 1.0, seed 0.

Arm	Gap (pp)	Δ gap	AUC_g	$\text{AUC}_{\neg g}$	sel_g	$\text{sel}_{\neg g}$
clean	−15.13	—	—	—	—	—
asym, c=1 (positives)	+23.96	+39.1	0.988	0.780	0.727	0.455
asym, c=0 (negatives)	−17.35	−2.2	0.992	0.738	0.425	0.622

Table 11: Seed variation for design (iii) on PADS gender, comparing a moderate injection, amount 0.25, with the stronger amount 1.0 used in the controlled audit. The injected channel is fixed to channel 4 across seeds.

Seed	Amount	Clean gap	Injected gap	Δ gap	AUC_g	$\text{AUC}_{\neg g}$	ΔAUC
0	0.25	−15.13	+10.23	+25.4	0.903	0.748	+0.156
1	0.25	−10.70	+4.25	+15.0	0.899	0.766	+0.133
2	0.25	−16.02	+5.42	+21.4	0.842	0.693	+0.148
0	1.00	−15.13	+23.96	+39.1	0.988	0.780	+0.208
1	1.00	−10.70	+23.61	+34.3	0.985	0.776	+0.210
2	1.00	−16.02	+18.13	+34.2	0.969	0.706	+0.262

For gender, we use a stronger injection amount of 1.0. As shown in Table 6, gender has a substantial pre-existing association with the outcome: in the PADS training split, $P(y = 1 \mid g) = 0.640$ compared with $P(y = 1 \mid \neg g) = 0.874$. Consequently, the clean model already exhibits a sizeable gender prediction gap, making a stronger injection useful for constructing an unambiguous positive control on top of this baseline disparity. At amount 1.0, Table 8 shows that the injected arm produces a 26.49 pp gap, corresponding to a 35.8 pp change relative to the clean arm, together with a large per-group AUC difference of 0.425. The validation AUC remains 0.755, indicating that the planted subgroup-specific signal does not simply arise from a collapse in overall predictive performance. Thus, the age experiment tests recovery under a relatively subtle planted effect, whereas the gender experiment provides a stronger positive control in the presence of a substantial pre-existing group–label association.

Table 10 provides a direction control for the asymmetric injection at the gender amount used in the controlled audit. When the positive-label side of the protected group is shifted ($c = 1$), the clean gap of −15.13 pp changes to +23.96 pp, corresponding to a +39.1 pp change, and the protected group has the higher selection rate (0.727 versus 0.455). Shifting the negative-label side instead ($c = 0$) reverses this behavior: the gap becomes −17.35 pp and the selection-rate ordering is inverted (0.425 versus 0.622). This directional reversal provides an additional control that the observed disparity responds to the intended label-dependent injection rather than merely to the presence of a perturbation in the selected channel. Table 11 further examines the stability of the positive-side injection across seeds at amounts 0.25 and 1.0. A disparity is already induced at the moderate amount of 0.25, but its

magnitude varies across seeds. At the selected amount of 1.0, the injected gap is consistently positive across all three seeds despite substantial variation in the clean gender gap, with Δgap ranging from $+34.2$ to $+39.1$ pp and ΔAUC remaining positive across all runs. We therefore retain amount 1.0 for the gender controlled audit as a strong and reproducible positive control, while the results at 0.25 demonstrate that the identified effect is not restricted to the stronger injection magnitude.

A.9 MIMIC IV IDENTIFIED GROUPS

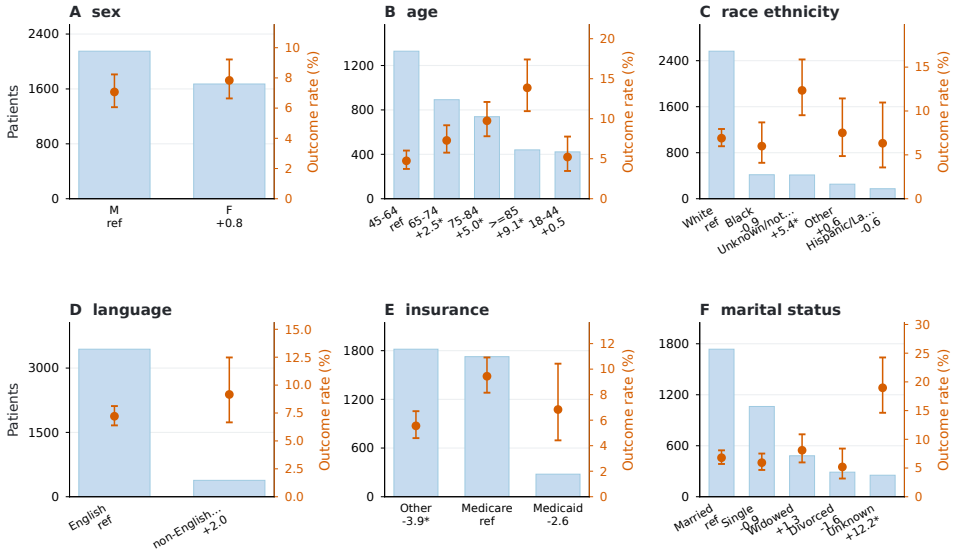


Figure 4: **Cohort composition and observed mortality by demographic group.** Each panel shows one attribute. Blue bars give the number of patients in each group (left axis); orange points give that group's observed 30-day mortality with a 95% Wilson interval (right axis). The two axes are on different scales and are not comparable to one another: bar height reflects only how common a group is. Groups are ordered by size, and the value printed beneath each group is its mortality difference from the reference group in percentage points, with `ref` marking the reference (male; age 45-64; White; English; Medicare; married). An asterisk indicates that the 95% Newcombe interval on that difference excludes zero; unmarked differences are within sampling variation at this sample size.

Fig. 4 summarises the cohort and its observed mortality. Age shows the clearest and only monotone gradient: relative to the 45-64 reference, mortality is 2.5 points higher at 65-74, 5.0 points higher at 75-84 and 9.1 points higher at 85 and above, all with intervals excluding zero. This is consistent with age as the dominant risk factor in an intensive care population. This is consistent with age as the dominant risk factor in an intensive care population and, as § 5.1 shows, conditioning on age materially changes several of the disparities associated with other demographic attributes.

Two attributes show a difference that survives the interval check, and in both cases the group concerned is a missing-data category rather than a demographic one. Patients whose race or ethnicity is unknown or not reported have 5.4 points higher mortality than White patients, and patients whose marital status is unknown have 12.2 points higher mortality than married patients. Undocumented demographics are plausibly associated with the circumstances of admission, emergency or transfer arrivals, and patients who die before intake is completed, so we read these as evidence about data provenance rather than as disparities affecting an identifiable population. They nevertheless matter for a fairness audit, because a model trained on these records may learn to treat missing documentation as a risk signal.

The remaining comparisons are within sampling variation. Female patients are fewer than male patients and their observed mortality is 0.8 points higher, but the interval on that difference covers zero, so the data do not establish a sex difference in mortality in this cohort; we therefore do not attempt to attribute one to biological or access-related mechanisms. Mortality is highest among Medicare patients, with patients covered by other insurance 3.9 points lower; this is expected, since Medicare eligibility begins at age 65 and the insurance categories are largely an age stratification

(Fig. 6). The Medicaid group sits 2.6 points below Medicare, an interval that includes zero, and the roughly 1.3-point gap between Medicaid and other insurance is a contrast between two non-reference groups that this figure does not test. Non-English-speaking patients show 2.0 points higher mortality, again with an interval covering zero and on a group of only a few hundred patients.

For the disparity analysis that follows we retain all six attributes rather than restricting to the largest groups, since an attribute with no marginal difference may still exhibit one within a stratum, and report group sizes throughout so that the precision of each comparison is visible. Where a group falls below 50 patients its estimates are reported but not interpreted.

Confounding factors, such as the overlap between sensitive attributes (e.g., race, gender) and non-sensitive factors like socioeconomic status or comorbidities, often complicate fairness evaluations (Wang et al., 2024). Traditional group-based comparisons may fail to account for these confounders, potentially leading to misleading conclusions about fairness or unfairness. To address this, a matched counterpart approach can be employed, where patients from different groups (e.g., Black and White) are matched on key characteristics such as age, comorbidities, and socioeconomic status. This method isolates the effects of sensitive attributes, offering actionable insights for clinicians rather than presenting fairness as an open-ended problem. Therefore, in Fig. 5 and 6, we look into the confounders overlap to narrow down the fairness analysis to more insightful groups.

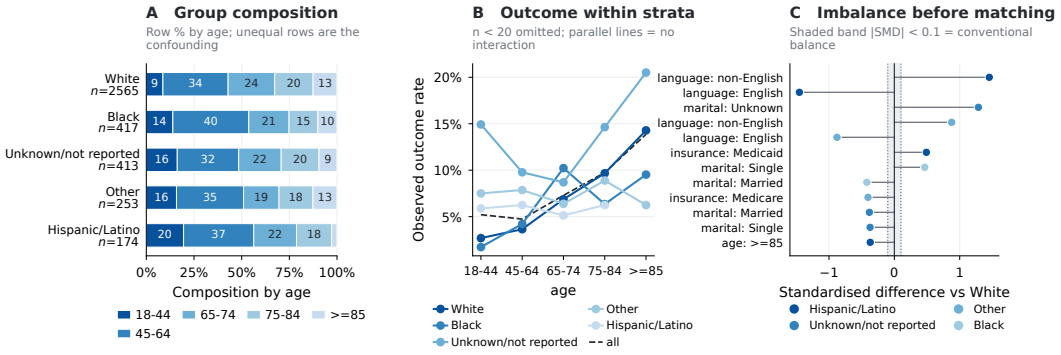


Figure 5: Race/ethnicity groups differ in the covariates that drive mortality. A, Age composition of each race/ethnicity group (row percentages, group sizes inset). B, Observed 30-day mortality within each age band; the dashed line is the pooled rate and cells with fewer than 20 patients are omitted. C, Standardised mean differences against the White reference group for age, insurance, marital status and primary language; the shaded band marks the conventional balance threshold $|\mathrm{SMD}| < 0.1$. All quantities are computed from the observed data and the recorded outcome, before any model is involved, so the imbalance in C is a property of the cohort rather than of the classifier.

Fig. 5 and Fig. 6 show that the sensitive attributes overlap substantially with the covariates that carry the outcome disparity, so a raw between-group comparison is not interpretable on its own. Race/ethnicity groups differ markedly in age (Fig. 5A): 20% of Hispanic/Latino and 14% of Black patients are under 45, against 9% of White patients, while the White group carries the largest share aged 85 or older. Because 30-day mortality rises steeply with age in this cohort, from 3.8% in the 18–44 band to 12.9% at 85 and above, any unadjusted racial comparison largely reflects this composition rather than differential outcomes.

Stratifying by age makes the point directly (Fig. 5B). Within most age bands the group-specific mortality curves track the pooled rate closely; the visible separations occur in the oldest band, where the Black group reaches 20.5% against 14.3% for the White group. These cells contain fewer than 50 patients, however, and a formal test of interaction between race/ethnicity and age was not significant ($p = 0.44$, Benjamini–Hochberg corrected across attribute pairs), so we treat the crossing curves as suggestive rather than established. The imbalance is not confined to age (Fig. 5C): language, marital status and insurance all exceed the conventional $|\mathrm{SMD}| < 0.1$ balance threshold against the White reference.

Insurance category illustrates the same confounding in a form that is structural rather than incidental (Fig. 6). Eligibility rules place 88% of Medicaid patients below age 65 while Medicare enrolment begins at 65, so the two groups have almost no age overlap (Fig. 6A) and age dominates the covariate

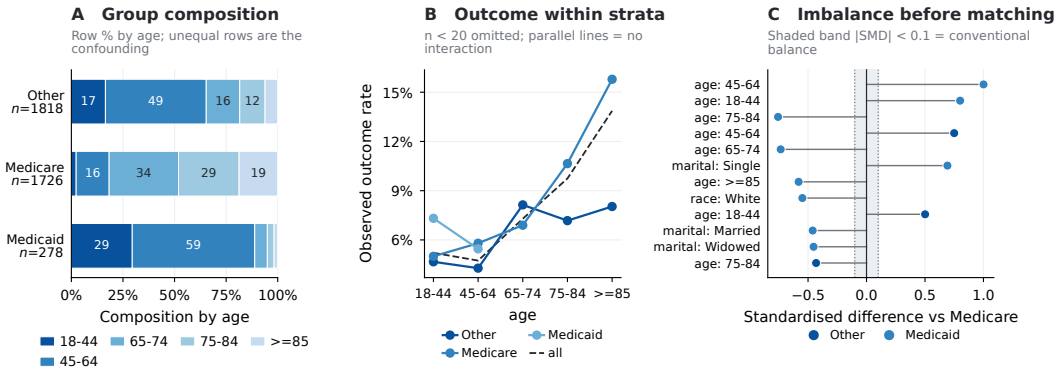


Figure 6: Insurance category is largely a proxy for age. Panels as in Figure 5, with insurance as the grouping variable and Medicare as the reference. A, Medicaid patients are concentrated in the younger bands while Medicare eligibility begins at age 65, so the two groups barely overlap in age. B, Mortality within age band. C, Age dominates the imbalance, with standardised differences approaching 1.

imbalance (Fig. 6C). Consistent with this, half of the observed insurance disparity in predicted risk is explained once age is held fixed, although a residual gap of 5.4 percentage points remains and is significant after correction, indicating that insurance is not only an age proxy.

These observations motivate the covariate-adjusted analysis of § A.2.8 and, where cell sizes permit, a matched comparison: they identify age as the covariate that any race, insurance or marital-status comparison must control for, and they show which strata contain enough patients to support such a comparison.

A.9.1 Per-timestep attribution differences across age groups

We next examined whether age-related differences in GapTrace-DeepLIFT attributions were localized to particular clinical features and periods of the admission window (Fig. 7). Overall, the magnitude of the age effect was modest. The proportion of attribution variance explained by age group was small across individual feature-timestep combinations ($\eta^2 \lesssim 0.01$).

Despite these small effect sizes, the temporal maps exhibited structured rather than uniformly null patterns. Several physiological and intervention-related variables showed localized age dependence, including chest-tube output, arterial blood-gas measurements, respiratory rate, Foley output, and blood-pressure measurements. These effects were often restricted to particular portions of the admission window rather than persisting uniformly across all 24 timesteps.

Comparison with the Spearman rank correlations further indicated that the observed between-group differences were not uniformly monotonic with age. Some features exhibited positive or negative age gradients over restricted time intervals, whereas others showed weak or sign-changing rank correlations despite detectable between-group variance. Thus, an appreciable η^2 was not necessarily accompanied by a strong rank correlation, consistent with non-monotonic differences among the five age groups.

Taken together, these results suggest that age is associated with small but structured differences in the temporal attribution patterns learned by the LSTM. These differences are feature- and time-specific rather than reflecting a single global shift in attribution with increasing age. Importantly, the observed patterns characterize the behavior of the trained LSTM and its GapTrace-DeepLIFT explanations, and therefore should not be interpreted as model-independent clinical effects of age.

A.10 Computational Analysis

The table 12 treats C_f as a constant, but it varies by orders of magnitude across the architectures we audit. Recurrent and convolutional models are linear in the sequence length, whereas self-attention is quadratic: with T steps, H heads and L layers, one forward pass materialises $O(T^2HL)$ attention

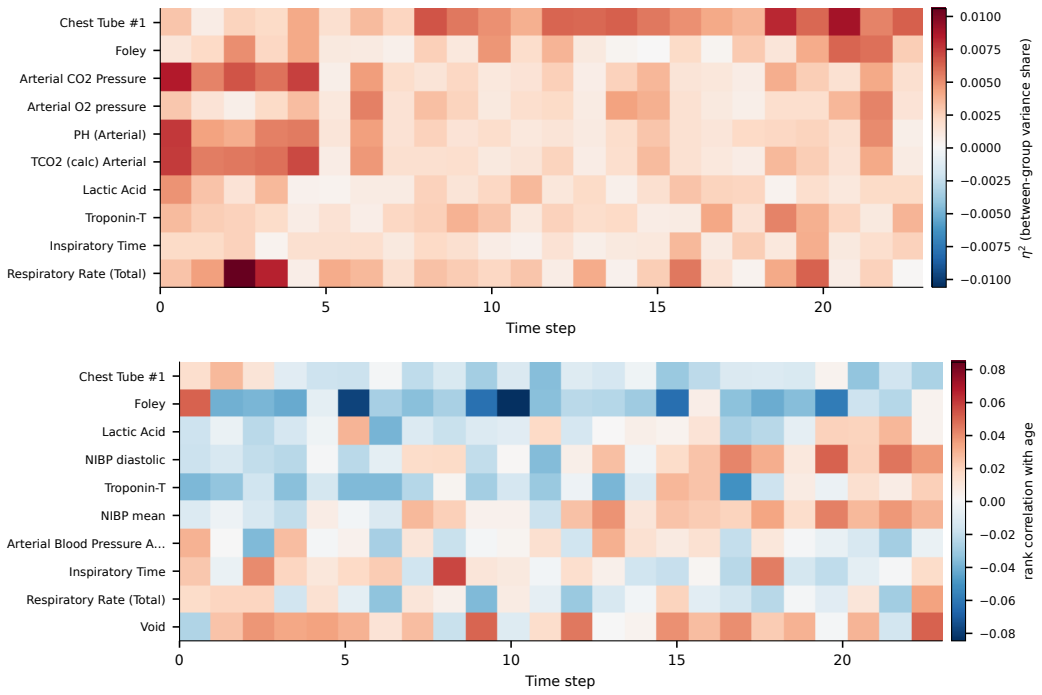


Figure 7: Temporal age dependence of GapTrace-DeepLIFT attributions over the admission window. Results are shown for MIMIC-IV 30-day mortality ($n = 3{,}822$ stays stratified into five age groups), with each cell representing one feature at one timestep (hour). Top Panel: η^2, the proportion of attribution variance associated with between-age-group differences, irrespective of whether those differences follow a monotonic age trend. Age-related effects were small overall ($\eta^2 \lesssim 0.01$), but exhibited structured feature- and time-specific patterns. Several physiological and intervention-related variables showed localized between-group differences, while their corresponding rank correlations were often weak or varied in direction across the admission window. Bottom Panel: Spearman rank correlation between age group and attribution, quantifying the direction and strength of monotonic age-related changes. Red indicates attributions that increase with age, whereas blue indicates attributions that decrease with age.

Table 12: Computational cost of the audit. N samples, T time steps, M channels, K attributes with $\leq G$ groups, $|\mathcal{M}|$ attribution methods, c_m model passes per sample per method, n_π permutations, n_b bootstrap resamples (Tibshirani & Efron, 1993), n_c matched controls per intervention, C_f the cost of one forward/backward pass, $P = \binom{K}{2}$ candidate pairs and $P' \leq P$ pairs surviving the screen. $K_{\text{sig}} \leq K$ is the number of attributes surviving FDR and $|\mathcal{M}_{\text{iv}}| \leq |\mathcal{M}|$ the methods for which the intervention is run. R is the cost of one repair pass over the selected features, i.e. computing and applying the group-wise quantile map to $|\mathcal{S}|$ channels across N samples. Attributions, cluster structure and baselines are computed once and reused.

Stage	Cost per unit	Marginal units	+ Intersectional
Attributions (shared)	$\lvert\mathcal{M}\rvert\, c_m\, N\, C_f$	1	0
Attribution clustering (shared)	$NM^2 + M^2 \log M$	1	0
Dependence screen + null	$n_\pi\, N \log N$	K	P
Effect size (bootstrap)	$n_b\, N$	K	P
Attribution + selection	$\lvert\mathcal{M}\rvert\, NM \log N$	K	P'
Covariate adjustment	$n_b\,(Nd^2 + d^3)$	K	—
Intervention + control	$\boldsymbol{(1{+}n_c)(NC_f + R)}$	$K_{\text{sig}}\, \lvert\mathcal{M}_{\text{iv}}\rvert$	P'
Temporal maps (optional)	$\lvert\mathcal{M}\rvert(G_\cap{-}1)\, n_\pi\, NTM$	K	P'

Table 13: Wall-clock time by stage for GapTrace-DeepLIFT on MIMIC-IV with a batch size of 128. The run used a bidirectional LSTM for mortality prediction and required 3GB of GPU memory.

	Attributions	Clustering	Screen + effect size	Selection	Adjustment	Intervention + control
MIMIC-IV	26.1s	1.8m	56.7s	3.1s	57.1s	7.3m

Table 14: Parameters controlling cost, in the order we recommend reducing them, with the statistical quantity each one trades against.

Parameter	Effect on cost	Trades against
Intervention scope	linear in $K_{\mathrm{sig}}\lvert\mathcal{M}_{\mathrm{iv}}\rvert$	nothing, if restricted to attributes that survive FDR: repairing features selected for a non-significant attribute answers no question
n_c (controls)	linear	resolution of $p_{\mathrm{vs\,random}}$, bounded below by $1/(n_c{+}1)$
IG baselines $\times$ steps	linear	attribution variance
n_b (bootstrap)	linear	confidence-interval precision only
n_π	linear	the achievable p-value floor
Attribution batch size	memory, not time	nothing

scores per sample. Attribution multiplies this further, because Integrated Gradients evaluates n_s interpolation points per baseline and expands the effective batch to $B\,n_s$ unless an internal batch size is set. Note that a cost model expressed only in N, T and M does not transfer across architectures.

A permutation p-value is bounded below by $1/(n_\pi + 1)$. When testing m units with multiplicity correction at level α, resolving p-values on the order of α/m therefore requires n_π on the order of m/α. Importantly, these permutations operate on the fixed attribution matrix and do not require additional model or attribution passes. Our primary per-unit inference instead uses the prespecified Kruskal–Wallis test with Benjamini–Hochberg correction.

Table 13 shows that GapTrace has a modest computational cost relative to the additional analysis it provides. The most computationally intensive stage is the intervention and control analysis, while the remaining stages require substantially less computation time.