A Controlled Study of Fairness Interventions for Temporal Graph Transformers on ICU Mortality Prediction

Anonymous Authors[1]

Abstract

Temporal Graph Transformers (TGTs) have been proposed for prediction on electronic health records (EHRs), but it is unclear whether their graph architecture reduces demographic performance gaps or whether standard fairness mitigation behaves differently on TGTs than on sequential baselines. We present a controlled study on the MIMIC-IV ICU mortality task. We compare three TGT edge configurations against classical, sequential, and transformer baselines, and benchmark two in-training fairness interventions (sample reweighting, variance regularization) and three post-hoc interventions (Platt scaling, isotonic regression, per-subgroup threshold equalization). We find that (i) TGT graph structure alone does not eliminate subgroup AUROC gaps, but the choice of edge type matters: TGTFULL achieves the smallest race AUROC gap of any model under matched training; (ii) single-attribute reweighting reduces the targeted attribute's gap but enlarges the gap on at least one other attribute in every model evaluated; and (iii) per-subgroup threshold equalization on top of Platt-scaled probabilities reduces the TPR gap from 0.20–$0.23 \rightarrow <0.03$ on both LSTM and TGTFULL, while calibration alone leaves AUROC gaps largely unchanged or worse.

1. Introduction

Electronic health records (EHR) provide a longitudinal view of patient states that can support clinical prediction, but EHR time series are difficult to model. Sequence models such as LSTM represent each pair of time steps as adjacent or unrelated. Instead, TGTs represent patient data as graphs whose edges encode temporal adjacency or clinical similarity, applying attention over these edges.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Clinical prediction models exhibit performance disparities across race, age, sex, and socioeconomic strata (Obermeyer et al., 2019; Pfohl et al., 2021; Chen et al., 2018). For sequential models, interventions such as sample reweighting, calibration, and threshold tuning are well studied; but for TGTs, intervention methods are seldom reviewed. Relational inductive biases could either reduce demographic gaps by routing attention through clinical similarity rather than temporal adjacency, or amplify them by converting latent group correlations into weighted edges. Current TGT works report only aggregate metrics, leaving the question: does graph structure reduce, redistribute, or amplify demographic performance gaps?

To evaluate this clinical disparity, we benchmark a classical, a sequential, and three transformer baselines, plus three TGT edge configurations on MIMIC-IV (Johnson et al., 2023) ICU mortality prediction tasks under equal cohort and evaluation protocols across three seeds. We then apply two in-training fairness interventions (inverse-frequency sample reweighting and variance regulation) and three post-hoc fairness interventions (Platt rescaling, isotonic rescaling, and per-subgroup threshold equalization) to our LSTM baseline and TGTFULL models, evaluating fairness metrics on each demographic subgroup (age, race, sex).

Our controlled comparison shows three main insights. First, graph topology affects fairness and accuracy distinctly: TGTFull matches LSTM in overall AUROC and achieves a smaller AUROC gap, with clinical edges reducing the gap and temporal edges widening it. Second, single attribute reweighting reduces the targeted subgroup gap while increasing others in both LSTM and TGTFull, indicating that this tradeoff exits in the intervention, not the model. Third, implementing a per-subgroup threshold with Platt-scaled probability is the best way to mitigate TPR gaps on all models, while global thresholds on isotonic calibration widen them.

2. Related Work

Prior sequence models for EHRs used recurrent networks (Choi et al., 2016); more recent work, such as GT-BEHRT (Poulain & Beheshti, 2024), constructs per-visit graphs and aggregates them with a transformer, but reports only aggregate predictive performance. Fairness in clinical ML has

been studied through subgroup performance investigation (across AUROC, AUPRC, calibration) and group fairness criteria such as equalized odds, with prior work showing that selecting one metric to optimize and applying mitigation techniques to satisfy that metric can degrade clinically relevant performance and is not always compatible with other criteria (Pfohl et al., 2019; Hardt et al., 2016; Pleiss et al., 2017; Pfohl et al., 2021; Chen et al., 2018).No prior benchmark comparews graph and sequencal models on the same cohort, splits, and seeds; Extended discussion of related work appears in Appendix A.1.

3. Methods

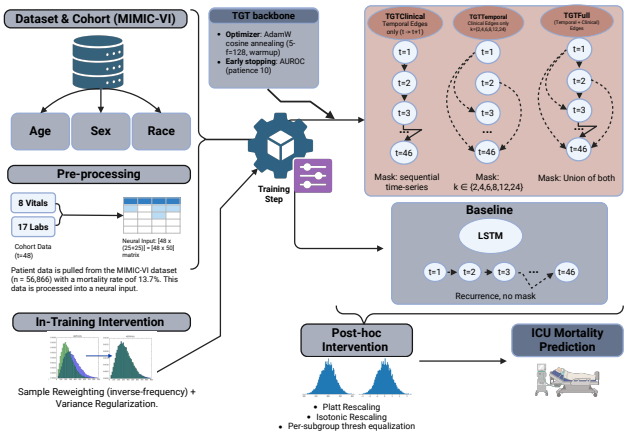

Figure 1. Our pipeline feeds extracted time-series features from MIMIC-IV into an LSTM baseline and three TGT edge configurations under matched training and applies two in-training and three post-hoc fairness interventions.

3.1. Cohort and Data Extraction

From MIMIC-IV (Johnson et al., 2023) we extract 56,866 adult ICU patients with an in-ICU mortality rate of 13.7%, restricted to patients with age ≥ 18 and stays ≥ 48 hours. Subgroups are defined over sex, age, and race; subgroups with <30 patients are excluded.

3.2. Feature Extraction and Temporal Representation

For each patient we extract 25 time-varying features (8 vital signs, 17 labs) over the first 48 ICU hours and concatenate them with a binary missingness mask $\{0,1\}^{48\times 25}$ to form a 48×50 neural input. Patients are split $70\%/15\%/15\%$ into train/validation/test, repeated across three seeds (42, 123, 2024).

3.3. Graph Construction

Each 48-hour history is encoded as a graph in three TGT configurations: TGTTEMPORAL (temporal edges only), TGTCLINICAL (clinical edges only), and TGTFULL (both). Edges are realized as attention masks. The TGT backbone has L=3 layers, H=4 heads, hidden dim d=64 (feed-forward dim 128), $\sim$104K parameters.

3.4. Training Steps

All neural models are trained with **BCEWithLogitsLoss** reweighted by $\min(n_{\text{neg}}/n_{\text{pos}}, 3)$, AdamW (Loshchilov & Hutter, 2019) with gradient clipping at norm 1, 5-epoch linear warmup, cosine annealing, a batch size of 128, and a maximum of 80 epochs. Early stopping uses validation AUROC with patience 10.

3.5. Fairness Evaluation

For each attribute $a \in \{\text{sex}, \text{age}, \text{race}\}$ with subgroups $g \in \mathcal{G}_a$, we report the *max–min gap* $\Delta^{(a)} = \max_g M_g - \min_g M_g$ for $M \in \{\text{AUROC}, \text{AUPRC}\}$ and subgroup standard deviation. For thresholded outputs, we also report the TPR gap $\Delta_a^{\text{TPR}} = \max_g \text{TPR}_g - \min_g \text{TPR}_g$.

3.6. In-Training Interventions

Sample reweighting for an attribute a uses inverse-frequency subgroup weights as a per-sample multiplier on BCE loss, normalized to preserve expected per-batch loss magnitude. Subgroup variance regularization computes per-batch subgroup mean losses $\bar{\ell}_g$ and adds $\lambda \cdot \text{Var}_g[\bar{\ell}_g]$ to the loss, with $\lambda \in \{0.01, 0.05\}$. Both are applied to LSTM and TGT-FULL.

3.7. Post-Hoc Fairness Interventions

Platt scaling (Platt, 1999) fits per-subgroup logistic regression on logits; isotonic regression (Zadrozny & Elkan, 2002) fits a per-subgroup monotone calibrator. Both are fit on validation and applied only to subgroups with $n_{\text{val}} \geq 30$. *Per-subgroup threshold equalization*: given a global threshold $\tau^\star$ tuned on the pooled validation set, we choose $\tau_g = \arg\min_\tau |\text{TPR}_g(\tau) - \text{TPR}^\star|$ for each subgroup so that subgroup TPRs match the median subgroup TPR at $\tau^\star$. We evaluate all combinations of {original, Platt, isotonic} $\times$ {global, per-subgroup} thresholds.

3.8. Graph Structure Ablation

To isolate the effect of graph topology, we train a sequential model, LSTM, and the three TGT variants under identical data, splits, seeds, and training settings; the only difference between models is the attention mask.

4. Results

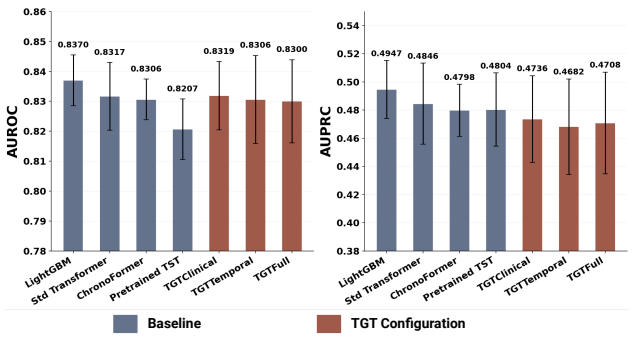

Figure 2. ICU mortality prediction performance across baseline and graph models (mean ± standard deviation over three seeds: 42, 123, and 2024).

LightGBM achieves the best overall performance (Figure 2). Among neural models, the Standard Transformer and the three TGT variants are within 0.005 AUROC of each other; LSTM is comparable. The smallest paired difference between TGTCLINICAL and a transformer baseline is against the Standard Transformer ($\Delta\text{AUROC} = +0.0003 \pm 0.0024$, Appendix Table 1); since the s.d. exceeds the mean by an order of magnitude, this difference is not distinguishable from zero.

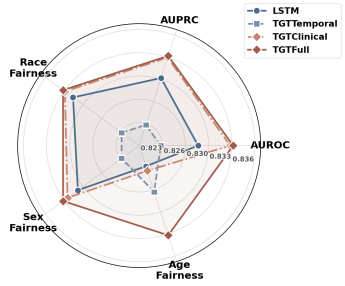

Figure 3. AUROC, AUPRC and attribute fairness across the four comparable models. Each fairness axis is reported so that larger radius indicates a smaller subgroup gap (i.e., more fair). Note: AUROC and AUPRC axes are zoomed to the $[0.82, 0.84]$ and $[0.45, 0.49]$ ranges respectively to make small differences visible. Numerical values for both performance and per-attribute AUROC/AUPRC gaps are provided in Appendix Table 2.

Figure 3 summarizes overall AUROC, overall AUPRC, and fairness across all models. TGTFULL sits on the outermost contour for all three fairness axes and ties TGTCLINICAL for the highest AUROC, while TGTTEMPORAL is the weakest variant on every axis except race fairness. The race gap is the largest disparity in every model (0.13–0.16 on AUROC; see Appendix Table 2), so even TGTFULL does not eliminate the gap, consistent with the central claim that graph structure alone is not sufficient. Per-metric, LSTM narrowly beats TGTFULL on five of the six split AUROC/AUPRC subgroup gaps (TGTFULL wins only on sex AUPRC), but the radar makes clear that this advantage is concentrated in AUPRC; In terms of AUROC fairness, TGTFULL dominates on all attributes.

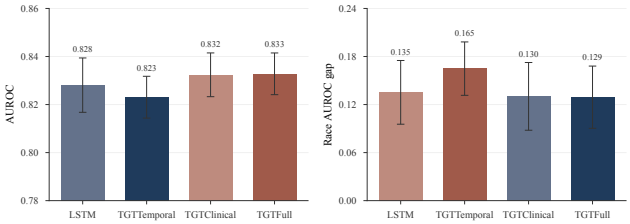

Figure 4. Effect of graph topology on overall AUROC (left) and race-subgroup AUROC gap (right). Error bars: one standard deviation across three seeds.

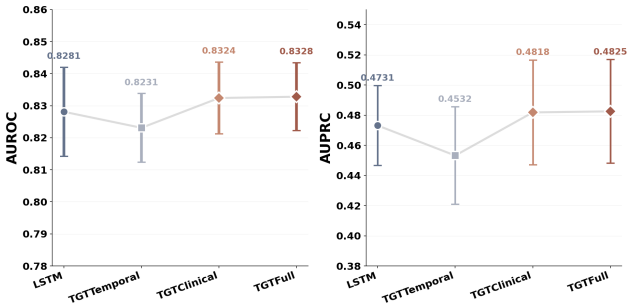

Figure 5. LSTM was compared against TGTTEMPORAL, TGTCLINICAL, and TGTFULL under identical data splits, seeds, and training settings. Overall AUROC and AUPRC are reported from the fairness pipeline and may therefore differ slightly from the benchmark values in Figure 2.

Reweighting for a single attribute reduces the race AUROC gap on TGTFULL ($0.2259 \rightarrow 0.1351$) but the sex and age AUROC gaps grow ($0.0098 \rightarrow 0.0131$ & $0.0709 \rightarrow 0.0880$); LSTM shows the same trade-offs. Isotonic regression raises AUROC ($0.8232 \rightarrow 0.8375$ on LSTM) and reduces the race AUROC gap ($0.1825 \rightarrow 0.1685$); Platt scaling has a much smaller effect on either. Per-subgroup threshold equalization with Platt rescaling yields the lowest TPR gaps (Figure 6): $\Delta^{\text{TPR}} = 0.0273$ for LSTM & 0.0232 for TGTFULL, vs. 0.3003 & 0.3454 under the corresponding global thresholds. Applying a global threshold with isotonic-calibrated probabilities widens TPR gaps relative to the original logits (LSTM: $0.1962 \rightarrow 0.3245$, TGTFULL $0.2254 \rightarrow 0.4536$).

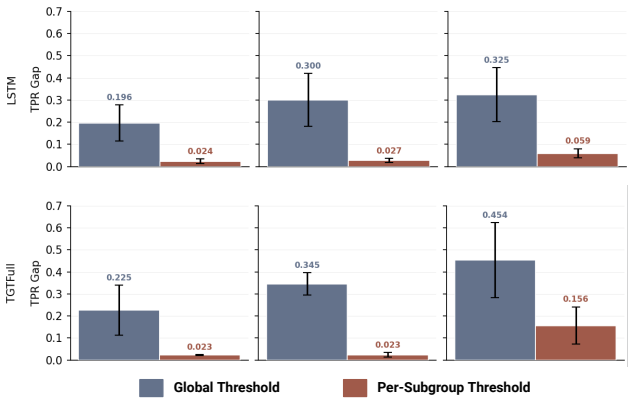

Figure 6. Fairness interventions on overall performance and subgroup gaps. Reweighted values are mean ± standard deviation over three training seeds (42, 123, 2024). Platt and Isotonic scaling are evaluated on pooled test predictions across seeds ($n = 6{,}182$)(Original; left column, Platt; middle column, Isotonic; right column)

5. Discussion

Under matched training (Figure 5), TGTFULL is comparable to LSTM in overall AUROC (0.8328 vs. 0.8281) but has a smaller race AUROC gap (0.1291 vs. 0.1352). Decomposing the change in race gap relative to LSTM, TGTTEMPORAL increases the race gap by $+0.0297$ while TGTCLINICAL decreases it by -0.005; we present this only as a difference relative to LSTM, not as a formal causal attribution. The pattern is consistent with the hypothesis that clinical-similarity edges encode lag relationships that are less correlated with demographic features than raw temporal adjacency, but our experiments do not establish a causal mechanism.

The clinical-edge lag set $k \in \{2, 4, 6, 8, 12, 24\}$ corresponds to the natural hourly to bi-hourly checks for vital signs. Attending across these specific lags emphasizes physiologically coherent temporal patterns (e.g., a 12-hour glucose trend) that are clinically significant regardless of demographics. LSTM and TGTTEMPORAL propagate information through immediate $t \to t{+}1$ neighbors, where recording timing is more sensitive to bedside practice and correlates with demographic attribute. This interpretation aligns with the notion that TGTTEMPORAL widens the race AUROC gap relative to LSTM, TGTCLINICAL narrows it, and TGTFULL inherits the clinical component while retaining the performance benefit of TGTTEMPORAL, achieving the lowest race gap (0.129) and the highest matched-training AUROC (0.833) in Figure 4.

On LSTM, race-weighted loss yielded the largest reduction in race AUROC gap among the in-training interventions but slightly reduced overall AUROC ($0.8281 \to 0.8259$) and AUPRC ($0.4731 \to 0.4716$). The race and sex AUROC gaps shrank ($0.1820 \to 0.1404$ and $0.0026 \to 0.0000$) but the age AUROC gap grew ($0.0640 \to 0.0734$); the same shift appeared for AUPRC. On TGTFULL the race AUROC gap was essentially unchanged ($0.1292 \to 0.1291$), and sex and age gaps both grew. Reweighting for a single attribute should therefore be paired with multi-attribute monitoring, since gap reduction on the targeted attribute consistently came at the cost of larger gaps elsewhere; TGTs were more sensitive to this redistribution than LSTM.

Adding subgroup-loss variance regularization on top of race-weighted training did not change LSTM AUROC meaningfully ($0.8253 \to 0.8251$) and *enlarged* its race gap ($0.1365 \to 0.1409$). On TGTFULL, both overall AUROC ($0.8256 \to 0.8238$) and the race AUROC gap ($0.1259 \to 0.1351$) moved in the wrong direction relative to weighted training alone. Combining reweighting with a variance penalty therefore did not consistently improve either fairness or accuracy in our setting.

Applying a global threshold on top of isotonic-calibrated probabilities widens TPR gaps relative to the original logits (Figure 6). Per-subgroup threshold equalization on top of Platt-scaled probabilities, by contrast, drives TPR gaps below 0.03 on both LSTM and TGTFULL. Platt's smoother calibration preserves rank order for stable per-subgroup threshold selection, making it the most effective intervention tested for TPR equalization.

Limitations. We use a single dataset (MIMIC-IV) drawn from one U.S. academic medical center, only structured EHR features, and three random seeds. TGT memory grows quadratically in window length, limiting scaling to longer ICU stays. Per-subgroup thresholds reduce TPR gaps on our test set but raise legal and ethical questions when implemented clinically (See Impact Statement).

6. Conclusion

We presented a controlled empirical study of standard fairness interventions on Temporal Graph Transformers and a sequential baseline for ICU mortality prediction. Three findings emerge: (i) graph structure does not, by itself, eliminate subgroup AUROC gaps, although the full-graph TGT yields a smaller race AUROC gap than LSTM under matched training; (ii) single-attribute reweighting reduces the targeted attribute's gap but enlarges gaps on at least one other attribute in every model evaluated; and (iii) per-subgroup threshold equalization on top of Platt-scaled probabilities is the most effective intervention in our setting, reducing TPR gaps below 0.03. Future work should extend the analysis to additional ICU datasets such as eICU and HiRID, evaluate rigid fairness criteria, and study fairness of non-neural baselines such as LightGBM in the same controlled setting.

Impact Statement

This paper studies algorithmic bias in clinical prediction. Our main practical recommendation, per-subgroup threshold equalization, raises legal and ethical questions when deployed clinically, including potential conflicts with non-discrimination law and concerns about disparate treatment across protected groups. MIMIC-IV is drawn from a single U.S. academic medical center and our findings should not be assumed to generalize. We do not advocate using race, sex, or age as direct inputs to deployed clinical decision systems; subgroup labels in this paper are used only for evaluation and post-hoc alterations.

References

Chen, I. Y., Johansson, F. D., and Sontag, D. Why is my classifier discriminatory? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Choi, E., Schuetz, A., Stewart, W. F., and Sun, J. Using recurrent neural networks for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, 24(2):361–370, 2016.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016.

Johnson, A. E. W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., Lehman, L.-w. H., Celi, L. A., and Mark, R. G. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Pfohl, S. R., Marafino, B., Coulet, A., Rodriguez, F., Palaniappan, L., and Shah, N. H. Creating fair models of atherosclerotic cardiovascular disease risk. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 271–278, 2019.

Pfohl, S. R., Foryciarz, A., and Shah, N. H. An empirical characterization of fair machine learning for clinical risk prediction. *Journal of Biomedical Informatics*, 113: 103621, 2021.

Platt, J. C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Smola, A. J., Bartlett, P., Schölkopf, B., and Schuurmans, D. (eds.), *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press, 1999.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. On fairness and calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Poulain, R. and Beheshti, R. Graph transformers on EHRs: Better representation improves downstream performance. *arXiv preprint arXiv:2402.16432*, 2024.

Zadrozny, B. and Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 694–699, 2002.

A. Appendix

A.1. Extended Related Work

Temporal Graph Representations Prior clinical sequence models used recurrent networks to capture temporal relationships in EHR data (Choi et al., 2016). More recent approaches such as GT-BHERT (Poulain & Beheshti, 2024) construct graphs per-visit and aggregate across visits with a transformer. These methods report aggregate performance but do not disaggregate across demographic subgroups, leaving open whether graph structure interacts with bias in EHR cohorts.

Fairness in Clinical Machine Learning Subgroup performance (comparing AUROC, AUPRC, or calibration across groups) and group-fairness criteria such as equalized odds (comparing positive prediction across groups) are two avenues of fairness investigation that are common within ML literature. Pfohl et al. (2019) and others have shown that single-criterion mitigation can degrade clinically relevant performance and is not always compatible with other criteria (Hardt et al., 2016; Pleiss et al., 2017; Pfohl et al., 2021; Chen et al., 2018). We focus on subgroup AUROC, AUPRC, and TPR gaps, the metrics most commonly reported for ICU mortality.

Evaluation in Clinical Machine Learning Most clinical ML benchmarks report aggregate AUROC and AUPRC and do not provide a controlled setting for comparing graph-based and sequential models. We address this gap by evaluating baseline and TGT variants on a shared MIMIC-IV ICU mortality benchmark with three random seeds.

A.2. Paired Model Differences

Table 1. Paired differences between TGTCLINICAL and deep learning baselines across three seeds (42, 123, 2024). Values are mean ± standard deviation of per-seed paired differences.

Comparison	ΔAUROC	ΔAUPRC
TGTCLINICAL vs. Standard Transformer	$+0.0003 \pm 0.0024$	-0.0110 ± 0.0043
TGTCLINICAL vs. ChronoFormer	$+0.0013 \pm 0.0030$	-0.0063 ± 0.0082
TGTCLINICAL vs. Pretrained TST	$+0.0012 \pm 0.0011$	-0.0068 ± 0.0017

A.3. Values for Figure 3

Table 2. AUROC subgroup gaps (max–min) underlying the fairness axes of Figure 3. **Bold** marks the best value per column. Values are mean ± standard deviation over three seeds (42, 123, 2024).

Model	Sex	Age	Race
LSTM	0.0053 ± 0.0063	0.0791 ± 0.0041	0.1352 ± 0.0487
TGTTEMPORAL	0.0112 ± 0.0056	0.0768 ± 0.0222	0.1649 ± 0.0408
TGTCLINICAL	0.0039 ± 0.0042	0.0787 ± 0.0092	0.1302 ± 0.0518
TGTFULL	$\mathbf{0.0033 \pm 0.0035}$	$\mathbf{0.0729 \pm 0.0057}$	$\mathbf{0.1291 \pm 0.0476}$

Table 3. AUPRC subgroup gaps (max–min) underlying the fairness axes of Figure 3. **Bold** marks the best value per column. Values are mean ± standard deviation over three seeds (42, 123, 2024).

Model	Sex	Age	Race
LSTM	$\mathbf{0.0291 \pm 0.0293}$	$\mathbf{0.0760 \pm 0.0439}$	$\mathbf{0.2283 \pm 0.0687}$
TGTTEMPORAL	0.0310 ± 0.0441	0.0777 ± 0.0409	0.2664 ± 0.0444
TGTCLINICAL	0.0275 ± 0.0365	0.0933 ± 0.0274	0.2422 ± 0.0602
TGTFULL	0.0287 ± 0.0352	0.0886 ± 0.0306	0.2365 ± 0.0475

A.4. Values for Figure 2

Table 4. ICU mortality prediction performance across baseline and graph models. Values are mean ± standard deviation over three seeds (42, 123, 2024).

Model	AUROC	AUPRC
LightGBM	0.8370 ± 0.0085	0.4947 ± 0.0205
Standard Transformer	0.8317 ± 0.0113	0.4846 ± 0.0287
ChronoFormer	0.8306 ± 0.0068	0.4798 ± 0.0185
Pretrained TST	0.8207 ± 0.0101	0.4804 ± 0.0260
TGTCLINICAL	0.8319 ± 0.0114	0.4736 ± 0.0308
TGTTEMPORAL	0.8306 ± 0.0147	0.4682 ± 0.0339
TGTFULL	0.8300 ± 0.0139	0.4708 ± 0.0361

A.5. Values for Figure 5

Table 5. Graph structure ablation under identical data, splits, seeds, and training settings. Values are mean ± standard deviation over three seeds (42, 123, 2024). **Bold** entries indicate the best-performing model on AUROC and AUPRC. Reported from the fairness pipeline and may differ slightly from Table 4.

Model	AUROC	AUPRC
LSTM	0.8281 ± 0.0139	0.4731 ± 0.0264
TGTTEMPORAL	0.8231 ± 0.0107	0.4532 ± 0.0323
TGTCLINICAL	0.8324 ± 0.0112	0.4818 ± 0.0347
TGTFULL	$\mathbf{0.8328 \pm 0.0106}$	$\mathbf{0.4825 \pm 0.0344}$

A.6. Values for Figure 6

Table 6. TPR gaps under post-hoc calibration and thresholding. Values are mean ± standard deviation over three seeds (42, 123, 2024). **Bold** entries indicate the lowest TPR gap per model.

Model	Method	Global Threshold	Per-Subgroup Threshold
LSTM	Original	0.1962 ± 0.0820	$\mathbf{0.0235 \pm 0.0102}$
LSTM	Platt	0.3003 ± 0.1197	0.0273 ± 0.0098
LSTM	Isotonic	0.3245 ± 0.1211	0.0593 ± 0.0209
TGTFULL	Original	0.2254 ± 0.1134	$\mathbf{0.0226 \pm 0.0017}$
TGTFULL	Platt	0.3454 ± 0.0510	0.0232 ± 0.0105
TGTFULL	Isotonic	0.4536 ± 0.1701	0.1563 ± 0.0848