Multimodal Alignment for Synthetic Clinical Time Series

Arinbjörn Kolbeinsson
K01
Reykjavík, Iceland
arinbjorn@K01.is

Benedikt Kolbeinsson
K01
Reykjavík, Iceland
benedikt@K01.is

Abstract

Privacy concerns limit sharing of clinical data, motivating synthetic data generation. However, multimodal clinical time series are challenging due to cross-modal physiological dependencies. We investigate what mechanisms enable generative models to preserve these relationships by comparing three autoregressive conditional mean models using metrics for statistical similarity, clinical rule violations and trend consistency. Analysis on ICU sepsis data suggests that better feature ordering may improve feature-level metrics, yet does not improve cross-feature performance despite strong signals in the training data. This suggests architectural choices limit clinical plausibility, highlighting the need for joint generative architectures or explicit constraint generation for synthetic healthcare time series.

1 Introduction

Privacy concerns often limit the sharing of real clinical data, slowing research and hindering the method development [Beaulieu-Jones et al., 2019, Price and Cohen, 2019]. Synthetic data might be able to alleviate some of these limitations, particularly in exploratory research and validation, as it allows a wide range of scenarios to be tested without compromising privacy [Chen et al., 2021, Yoon et al., 2020]. However, multimodal clinical time series are challenging to synthesise as they consist of heterogeneous measurement modalities, asynchronous and different temporal resolutions and clinical assessments where each feature can have distinct dynamics and cross-modal physiological dependencies [Ghassemi et al., 2020].

The focus of many existing synthetic data methods has been on similarity with single source modalities [Choi et al., 2017, Yoon et al., 2019, Naseer et al., 2023]. However, clinical data requires not only statistical realism but also multimodal alignment to preserve the relationships between measurement types (e.g., fever in vital signs typically accompanies inflammatory markers in lab values). This gap is partly due to a lack of evaluation frameworks that assess cross-modal clinical plausibility alongside statistical similarity [Alaa et al., 2022].

In this paper, we investigate the challenge of *multimodal alignment* in synthetic clinical time series generation. We focus on understanding which model properties are required to preserve cross-modal relationships. Using an autoregressive generation framework that combines conditional mean prediction with residual modelling, we systematically compare how different conditional mean models, from a simple baseline to Ridge regression and Gradient Boosting capture cross-modal dependencies. We introduce an evaluation framework that combines statistical (conditional MMD [Gretton et al., 2012]), clinical plausibility (physiological rule violations across modalities) and trend consistency metrics. The empirical analysis on ICU sepsis data strengthens the belief that architectural limitations, and not just modal ordering, is key for cross-modal alignment. This finding raises important questions about optimal conditioning structures for multimodal clinical data and provides insights for future generative modelling research in healthcare.

EurIPS 2025 workshop on Multimodal Representation Learning for Healthcare.

2 Methods

To investigate the preservation of cross-modal alignment with different model features and ordering, we use an autoregressive generation framework that systematically varies the complexity of cross-modal dependency modelling. Here we describe our problem setup, generation approach and evaluation metrics.

2.1 Problem Setup

We consider multimodal clinical time series $X_t = [V_t, L_t]$ comprising vital signs V_t (e.g., HR, BP, temperature) and lab values L_t (e.g., lactate, WBC, pH), along with static covariates Z (e.g., gender, age). Our goal is to generate synthetic data that preserve: (1) marginal distributions, (2) temporal dynamics and (3) cross-modal physiological relationships, which is the main focus of our investigation. We ask: *What model mechanisms are necessary to preserve cross-modal alignment?*

2.2 Autoregressive Generation Framework

We generate features sequentially: $x_{j,t} = \mu_j(x_{<j,t}, Z, t) + \epsilon_{j,t}$, where μ_j is a conditional mean model and $\epsilon_{j,t}$ is an AR(1) residual. To isolate the effect of cross-modal dependency modelling, we compare three conditional mean models: (1) **AR-only**: constant mean $\bar{\mu}_j$ (tests if multivariate residual covariance alone suffices), (2) **Ridge**: linear model with L2 regularization, and (3) **GBM**: non-linear ensemble (100 trees, depth 3).

Residuals follow $\epsilon_{j,t} = \phi_j \epsilon_{j,t-1} + \eta_{j,t}$ with multivariate innovations $\boldsymbol{\eta}_t \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ (Ledoit-Wolf covariance [Ledoit and Wolf, 2004]). Features are ordered by measurement type (vitals, then labs) in the main experiments. We then perform order permutations which are presented in the Appendix.

2.3 Evaluation Framework

We assess multimodal alignment with three metrics:

Conditional MMD: Statistical similarity within patient subgroups (by gender) using RBF kernel two-sample test. Lower is better. **Clinical rule violations:** We define 10 expert-derived physiological rules (e.g., "hypotension ⇒ tachycardia", see Table 1) and measure violation rates. Real data exhibits 30-80% violations due to medications and individual variation. We aim to *match* real rates, not achieve 0%. **Trend consistency:** We compare regression slopes β_{real} vs β_{syn} for covariate-feature relationships: score $= \text{sign}(\beta_{\text{real}}\beta_{\text{syn}}) \times \min(|\beta|)/\max(|\beta|) \in [-1, 1]$. Higher indicates better direction and magnitude agreement.

Table 1: Clinical Rule Definitions

Rule Name	Implication	Clinical Basis
Hypotension→Tachycardia	MAP $< 65 \Rightarrow$ HR > 80	Compensatory response to low blood pressure Vincent et al. [2018]
Severe Hypotension→Tachycardia	MAP $< 55 \Rightarrow$ HR > 100	Severe shock compensation Cecconi et al. [2014]
Fever→Tachycardia	Temp > 38.3°C $\Rightarrow$ HR > 90	Metabolic demand increase O'Grady et al. [2008]
High Fever→Tachycardia	Temp > 39.5°C $\Rightarrow$ HR > 100	Severe febrile response O'Grady et al. [2008]
Tachypnea→Leukocytosis	Resp $> 22 \Rightarrow$ WBC > 11	Inflammatory response Singer et al. [2016]
Hypoxia→Tachycardia	$SpO_2 < 90\% \Rightarrow$ HR > 90	Hypoxic stress response Calverley [2003]
Severe Hypoxia→Compensation	$SpO_2 < 85\% \Rightarrow$ HR $> 100 \vee$ Resp > 24	Critical hypoxia response Calverley [2003]
Acidosis→Hyperventilation	pH $< 7.30 \Rightarrow$ Resp > 20	Respiratory compensation Berend et al. [2014]
Hyperlactatemia→Acidosis	Lactate > 4 mmol/L $\Rightarrow$ pH < 7.35	Lactic acidosis Kraut and Madias [2014]
Shock→Hyperlactatemia	MAP $< 60 \wedge$ HR $> 110 \Rightarrow$ Lactate > 2	Circulatory shock Hernández et al. [2019]

2.4 Experiments

We use the PhysioNet Computing in Cardiology Challenge 2019 dataset [Reyna et al., 2020] for sepsis prediction. It contains approximately 40,000 ICU patients with hourly recordings of 40 clinical features including vital signs (heart rate, blood pressure, temperature, respiratory rate, oxygen saturation) and laboratory values (lactate, white blood cell count, pH, etc).

We preprocess the data by aligning all measurements to hourly intervals, applying imputation (median) for missing values and z-score normalising each feature using training set statistics. We split the data into 70% training, 15% validation, and 15% test sets. The training set is used to fit all model

parameters, the validation set for hyperparameter selection, and the test set to sample conditioning variables for synthetic generation.

For the conditional mean models, we implement Ridge regression with regularization parameter $\alpha = 1.0$ and Gradient Boosting with 100 estimators, maximum depth of 3, learning rate of 0.1, and subsample ratio of 0.5. Each model is trained separately for each feature using all previously ordered features, static covariates, and time as inputs. The AR(1) coefficients are estimated per-feature from training residuals, and the innovation covariance matrix is estimated using Ledoit-Wolf regularisation to ensure positive definiteness.

We generate 500 synthetic patient trajectories for each method, with static covariates sampled from the test set distribution to match the real data composition. Each experiment is repeated with 3 different random seeds. For evaluation, conditional MMD uses an RBF kernel with median heuristic bandwidth selection. Clinical rules are evaluated on all timesteps where the antecedent condition holds and trend consistency scores are computed using linear regression on gender as the primary covariate.

3 Results

Table 2 shows that conditional MMD improves with model complexity, with GBM achieving the lowest average distance (0.304 ± 0.003) compared to Ridge (0.328 ± 0.007) and AR-only (0.351 ± 0.003). This suggests that modeling cross-modal dependencies through the conditional mean improves similarity within subgroups.

Table 2: Conditional MMD Comparison (mean $\pm$ std over 3 seeds, lower is better)

Method	Gender=0	Gender=1	Average
AR(1)-only	0.342±0.006	0.358±0.009	0.351±0.003
Ridge+AR(1)	0.321±0.016	0.333±0.001	0.328±0.007
GBM+AR(1)	0.293±0.003	0.313±0.007	0.304±0.003

However, clinical rule violations (Table 3) reveal a more complex picture. While some rules show expected patterns, the Fever→Tachycardia relationship remained mostly unlearned across models. Further results are shown in the Appendix, where we experiment with different feature orders, which in theory should allow the model to generate more realistic Fever→Tachycardia relationships. Even with this, the models were unable to improve their clinical realism. This was despite there being a strong signal in the training set (mean HR with fever: 99.1 bpm vs without: 84.3 bpm, Cohen's d = 0.84, $p < 0.001$).

This suggests that our chosen models face inherent difficulty in learning these relationships and generating realistic trajectories. The fact that the AR model performs marginally better, despite not having explicit conditionals suggesting that some parts of the signal are captured in the residuals.

Table 3: Clinical Rule Violations (closer to Real is better)

Rule	Real	AR-only	Ridge	GBM
Hypotension→Tachy	0.51	0.464±0.013	0.403±0.010	0.410±0.017
Severe Hypotension→Tachy	0.86	0.892±0.006	0.855±0.016	0.831±0.038
Fever→Tachy	0.33	0.570±0.029	0.689±0.011	0.689±0.019
High Fever→Tachy	0.37	0.601±0.114	0.889±0.157	0.958±0.059
Tachypnea→Leukocytosis	0.53	0.437±0.012	0.452±0.018	0.467±0.014
Hypoxia→Tachy	0.48	0.581±0.020	0.595±0.043	0.612±0.037
Severe Hypoxia→Comp	0.54	0.167±0.236	0.000	0.333±0.471
Acidosis→Hypervent	0.64	0.607±0.044	0.518±0.038	0.618±0.037
Hyperlactatemia→Acidosis	0.54	0.637±0.035	0.830±0.032	0.814±0.059
Shock→Hyperlactatemia	0.78	0.524±0.093	0.612±0.155	0.671±0.087

Trend consistency scores (Table 4) further illustrate the complexity. Ridge achieves strong positive consistency for blood pressure features (MAP: +0.79±0.09, DBP: +0.76±0.17) and pH (+0.80±0.09), while GBM shows comparable performance on some features but worse on others (Temp: +0.22±0.30 vs Ridge +0.45±0.48). Notably, the AR-only baseline shows near-zero or negative consistency for most features, confirming that explicit cross-modal modeling is necessary.

These results suggest that different aspects of multimodal alignment are not uniformly improved by model complexity. These patterns are further complicated by the uneven performance gains of optimising feature orders, highlighting the need for improved joint modality modelling for generative realism.

Table 4: Trend Consistency (higher is better)

Feature	AR-only	Ridge	GBM
HR	+0.08±0.04	+0.45±0.15	+0.28±0.22
O2Sat	-0.22±0.31	+0.41±0.11	+0.52±0.15
Temp	-0.61±0.40	+0.45±0.48	+0.22±0.30
SBP	+0.01±0.25	+0.07±0.28	-0.12±0.60
MAP	+0.10±0.64	+0.79±0.09	+0.68±0.18
DBP	-0.40±0.21	+0.76±0.17	+0.79±0.12
Resp	-0.54±0.48	-0.06±0.64	-0.13±0.67
Lactate	+0.06±0.36	-0.18±0.05	-0.29±0.17
WBC	+0.27±0.38	+0.46±0.09	+0.53±0.27
pH	+0.18±0.42	+0.80±0.09	+0.21±0.29

4 Discussion

Our investigation uncovers patterns suggesting an underexplored factor in multimodal clinical time series generation: feature ordering in autoregressive models may act as an inductive bias affecting which cross-modal relationships can be learned. Consider the Fever→Tachycardia rule. In our original experiments, heart rate appears before temperature, meaning the conditional mean model predicts HR without access to temperature information. This architectural constraint could have explained why models fail to capture fever→tachycardia. However, in our additional systematic experiments (see Appendix) where we vary feature order we did not observe improvement for cross-feature or cross-modal clinical rules, suggesting the root cause is deeper than feature ordering. Improvements were, however, observed in feature-level metrics.

This finding connects to broader work in autoregressive generation and modeling. Autoregressive models in vision and language implicitly encode ordering assumptions through raster scan or left-to-right generation [van den Oord et al., 2016, Vaswani et al., 2017], though some work has explored learned orderings and parallel generation to mitigate this [Uria et al., 2016, Niu et al., 2020]. In causal inference, variable ordering in structural equation models similarly determines which relationships can be represented [Pearl, 2009, Spirtes et al., 2000]. Our contribution is demonstrating that these architectural choices have measurable consequences for domain-specific plausibility in healthcare data, where the "correct" ordering may correspond to physiological causality rather than data structure.

The practical implications are significant. For practitioners generating synthetic clinical data, our results suggest that distributional metrics alone are insufficient and that domain-specific evaluation is necessary to verify clinical plausibility. Moreover, our findings indicate that simpler models may sometimes be preferable if they preserve critical cross-modal relationships better than complex models that overfit spurious patterns in the chosen feature ordering. For researchers, our work highlights underexplored factors that warrant systematic investigation, possibly through causal discovery methods that optimise for domain-specific constraints.

Our study has the following limitations. We evaluate on a single dataset (ICU sepsis), use one conditioning variable for MMD (gender), and manually define 10 clinical rules. The feature ordering is varied in only three different permutations, and we do not compare to deep generative models like GANs or VAEs. Future work should explore joint generation of all features per timestep, and more comprehensive rule sets possibly discovered through automated mining of clinical literature.

5 Conclusion

We investigate what mechanisms enable generative models to preserve cross-modal physiological relationships in synthetic clinical time series. Through systematic comparison of three conditional mean models evaluated with statistical, clinical plausibility and trend consistency metrics, our results suggest that feature ordering may constrain which cross-modal relationships can be learned. When heart rate is predicted before temperature, the model could struggle to learn that fever causes tachycardia. This suggests that architectures implicitly encode causal assumptions with measurable consequences for domain-specific plausibility. Our results show that distributional metrics alone are insufficient and that specific evaluation is essential for synthetic healthcare data. Future work should explore principled feature orderings through causal discovery or joint generation architectures.

References

Ahmed M. Alaa, Boris Van Breugel, Evgeny S. Saveliev, and Mihaela van der Schaar. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 290–306. PMLR, 2022.

Brett K. Beaulieu-Jones, Zhiwei Steven Wu, Chris Williams, Ran Lee, Sanjeev P. Bhavnani, James Brian Byrd, and Casey S. Greene. Privacy-preserving generative deep neural networks support clinical data sharing. *Circulation: Cardiovascular Quality and Outcomes*, 12(7):e005122, 2019. doi: 10.1161/CIRCOUTCOMES.118.005122.

Kenneth Berend, A. Paul Jan de Vries, and Rijk O. B. Gans. Physiological approach to assessment of acid–base disturbances. *New England Journal of Medicine*, 371(15):1434–1445, 2014. doi: 10.1056/NEJMra1003327.

Peter M. A. Calverley. Respiratory failure in chronic obstructive pulmonary disease. *European Respiratory Journal Supplement*, 47:26s–30s, 2003. doi: 10.1183/09031936.03.00030103.

Maurizio Cecconi, Daniel De Backer, Massimo Antonelli, Richard Beale, Jan Bakker, Christoph Hofer, Roman Jaeschke, Alexandre Mebazaa, Michael R. Pinsky, Jean-Louis Teboul, et al. Consensus on circulatory shock and hemodynamic monitoring. task force of the european society of intensive care medicine. *Intensive Care Medicine*, 40(12):1795–1815, 2014. doi: 10.1007/s00134-014-3525-z.

Richard J. Chen, Ming-Yu Lu, Tiffany Y. Chen, Drew F. K. Williamson, and Faisal Mahmood. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6):493–497, 2021. doi: 10.1038/s41551-021-00751-8.

Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. In *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 286–305. PMLR, 2017.

Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L. Beam, Irene Y. Chen, and Rajesh Ranganath. A review of challenges and opportunities in machine learning for health. *AMIA Summits on Translational Science Proceedings*, 2020:191–200, 2020.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773, 2012.

Glenn Hernández, Rinaldo Bellomo, and Jan Bakker. The ten pitfalls of lactate clearance in sepsis. *Intensive Care Medicine*, 45(1):82–85, 2019. doi: 10.1007/s00134-018-5213-x.

Jeffrey A. Kraut and Nicolaos E. Madias. Lactic acidosis. *New England Journal of Medicine*, 371 (24):2309–2319, 2014. doi: 10.1056/NEJMra1309483.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. doi: 10.1016/j.jmva.2003.11.009.

Ahmed Ammar Naseer, Benjamin Walker, Christopher Landon, Andrew Ambrosy, Marat Fudim, Nicholas Wysham, Botros Toro, Sumanth Swaminathan, and Terry Lyons. Scoehr: generating synthetic electronic health records using continuous-time diffusion models. In *Machine Learning for Healthcare Conference*, pages 489–508. PMLR, 2023.

Chenhao Niu, Yang Song, Jiaming Song, Shengjia Zhao, Aditya Grover, and Stefano Ermon. Permutation invariant graph generation via score-based generative modeling. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 4474–4484. PMLR, 2020.

Naomi P. O'Grady, Philip S. Barie, John G. Bartlett, Thomas Bleck, Karen Carroll, Andre C. Kalil, Peter Linden, Dennis G. Maki, David Nierman, William Pasculle, and Henry Masur. Guidelines for evaluation of new fever in critically ill adult patients: 2008 update from the american college of critical care medicine and the infectious diseases society of america. *Critical Care Medicine*, 36 (4):1330–1349, 2008. doi: 10.1097/CCM.0b013e318169eda9.

Judea Pearl. *Causality*. Cambridge University Press, Cambridge, 2 edition, 2009.

W. Nicholson Price and I. Glenn Cohen. Privacy in the age of medical big data. *Nature Medicine*, 25 (1):37–43, 2019. doi: 10.1038/s41591-018-0272-7.

Matthew A. Reyna, Christopher S. Josef, Russell Jeter, Supreeth P. Shashikumar, M. Brandon Westover, Shamim Nemati, Gari D. Clifford, and Ashish Sharma. Early prediction of sepsis from clinical data: the physionet/computing in cardiology challenge 2019. *Critical Care Medicine*, 48 (2):210–217, 2020. doi: 10.1097/CCM.0000000000004145.

Mervyn Singer, Clifford S. Deutschman, Christopher W. Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R. Bernard, Jean-Daniel Chiche, Craig M. Coopersmith, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA*, 315(8):801–810, 2016. doi: 10.1001/jama.2016.0287.

Peter Spirtes, Clark N. Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2 edition, 2000.

Benigno Uria, Marc-Alexandre Côté, Karol Gregor, Iain Murray, and Hugo Larochelle. Neural autoregressive distribution estimation. *Journal of Machine Learning Research*, 17(205):1–37, 2016.

Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1747–1756. PMLR, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jean-Louis Vincent, Nathan D. Nielsen, Nathan I. Shapiro, Margaret E. Gerbasi, Aaron Grossman, Robin Doroff, Feng Zeng, Paul J. Young, and James A. Russell. Mean arterial pressure and mortality in patients with distributive shock: a retrospective analysis of the mimic-iii database. *Annals of Intensive Care*, 8(1):107, 2018. doi: 10.1186/s13613-018-0448-9.

Jinsung Yoon, Daniel Jarrett, and Mihaela van der Schaar. Time-series generative adversarial networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jinsung Yoon, Lydia N. Drumright, and Mihaela van der Schaar. Anonymization through data synthesis using generative adversarial networks (ads-gan). *IEEE Journal of Biomedical and Health Informatics*, 24(8):2378–2388, 2020. doi: 10.1109/JBHI.2020.2980262.

A Appendix

We perform further experiments to investigate the impact of ordering on our metrics. In addition to the default order (vitals→labs), we try reverse order (labs→vitals) and temperature first (temperature and HR before other features) to target the fever+tachycardia relationship directly. We test this across all models with three random seeds, taking the mean and standard deviation between runs.

The real data signal is strong between fever→tachycardia, mean HR with fever: 99.1 bpm vs without: 84.3 bpm, Cohen's d = 0.84, $p < 0.001$.

Temperature first improves the learned temperature trend across models (Table 7), indicating that improved ordering can improve feature-level metrics in some cases. This is possibly due to relaxed constraints or inductive bias.

However, placing the temperature generation step before HR did not lead to meaningful and consistent improvement on the clinical rule violation metric. This suggests that simple reordering that theoretically allows more realistic generation is insufficient for improving cross-feature relationships. Other rules show minimal sensitivity to changed ordering (Table 6. This pattern is consistent across all models tested (Table 8), further supporting the hypothesis that simple reordering is insufficient.

Distributional similarity showed minimal variance across the orders we tested (Table 5).

Table 5: Distributional Similarity (Conditional MMD) by feature ordering for the Ridge Model (lower is better). We see that Conditional MMD with mean±std across three random seeds. Ordering has minimal impact on distribution properties.

Metric	**Default**	**Temp first**	**Reversed**	**Δ (Def→Temp)**
Average MMD	0.336±0.003	0.325±0.004	0.329±0.002	-0.011

Table 6: Clinical Rule Violations by feature ordering for the Ridge model. Results show violation rates with mean±std across three random seeds. The Fever→Tachycardia rule shows minimal sensitivity to ordering despite temp first placing Temperature before HR during learning and generation steps.

Rule	**Real**	**Default**	**Temp first**	**Reversed**	**Δ**
Hypotension→Tachy	0.51	0.403±0.010	0.400±0.013	0.401±0.008	**-0.003**
Fever→Tachy	**0.33**	**0.689±0.011**	**0.697±0.042**	0.713±0.029	**+0.008**
Tachypnea→Leukocytosis	0.53	0.452±0.018	0.456±0.018	0.446±0.008	**+0.004**
Acidosis→Hypervent	0.64	0.518±0.038	0.531±0.018	0.542±0.020	**+0.013**
Hyperlactatemia→Acidosis	0.54	0.830±0.032	0.853±0.041	0.842±0.015	**+0.023**

Table 7: Trend Consistency by feature ordering for the Ridge model. Results show trend consistency scores with mean±std across three random seeds. Temp first ordering substantially improves Temperature trend consistency, demonstrating ordering affects feature-covariate alignment.

Feature	**Default**	**Temp first**	**Reversed**	**Δ (Def→Temp)**
HR	+0.45±0.15	+0.43±0.23	+0.44±0.14	-0.02
Temp	**+0.45±0.48**	**+0.75±0.16**	-0.56±0.43	**+0.30**
MAP	+0.79±0.09	+0.48±0.30	+0.69±0.14	-0.31
pH	+0.80±0.09	+0.51±0.24	+0.73±0.10	-0.29
Resp	-0.06±0.64	+0.49±0.18	-0.06±0.29	+0.55

Table 8: Ordering Effects on Key Metrics Across All Conditional Mean Models. Summary of Temperature trend consistency improvements and Fever→Tachycardia rule insensitivity across AR-only, Ridge and GBM models. Values show mean±std across three random seeds. All models show consistent patterns: Temp trend improves with temp first ordering, while Fever→Tachy rule violations remain largely unchanged.

Model	Metric	Default	Temp first	Δ
AR-only	Temp Trend	-0.61±0.40	+0.33±0.40	**+0.94**
	Fever→Tachy	0.570±0.029	0.548±0.023	-0.022
Ridge	Temp Trend	+0.45±0.48	+0.75±0.16	**+0.30**
	Fever→Tachy	0.689±0.011	0.697±0.042	**+0.008**
GBM	Temp Trend	+0.22±0.30	+0.64±0.28	**+0.42**
	Fever→Tachy	0.689±0.019	0.693±0.044	**+0.003**