Informative Irregularity as a Diagnostic for Model Robustness

Tamara Krafft [1] **Bernhard Bauer** [1]

Abstract

Electronic health records frequently contain irregularly sampled data. The specific timing of clinical observations itself can contain informative signals regarding patient acuity. To explicitly model this behavior, we extract structural workflow metadata from 3,497 MIMIC-IV-ED stays and introduce a stratification approach to evaluate model robustness across distinct irregularity regimes. Integrating structural metadata yields a $+4.1\%$ AUPRC improvement over the native baseline. Furthermore, feature importance analysis reveals that structural metadata provides a statistically independent diagnostic signal over vital sign data (mean correlation $|\rho| = 0.041$ between top structural and physiological features). Crucially, stratification reveals that the utility of this metadata peaks when sampling does not adhere to local routine rhythms, yielding an $+8.0\%$ AUPRC improvement over the native baseline. These findings demonstrate that quantifying distinct irregularity manifestations provides a diagnostic framework to map how model performance fluctuates across different clinical workflows.

1. Introduction

The early identification of patient deterioration is critical for initiating fast and appropriate interventions and has been associated with improved patient outcomes (Escobar et al., 2020). The availability of structured electronic health record (EHR) data, exemplified by large-scale databases such as MIMIC (Johnson et al., 2024), has enabled the development of advanced early warning systems capable of detecting subtle shifts in a patient's physiological state (Mann et al., 2021). However, translating longitudinal monitoring data into reliable predictive models is complicated by fundamental temporal irregularity, which challenges architectures optimized for evenly spaced sequences (Xie et al., 2022). For patients requiring active surveillance, this sporadic sampling is actively driven by an interplay of the patient's underlying health trajectory, clinical protocols, and staff availability (Sun et al., 2024). Traditionally, irregularity has been treated as a missing data problem, where data is resampled into fixed grids and resulting gaps are addressed using imputation methods (Tipirneni & Reddy, 2022). Naive imputation, such as Last Observation Carried Forward (LOCF), can introduce artificial bias and noise while actively destroying fine-grained temporal details (Tipirneni & Reddy, 2022). Conversely, a growing body of research recognizes that the pattern of missing data is rarely missing completely at random, but exhibits informative missingness, wherein the absence and specific timing of measurements convey meaningful information (Lipton et al., 2016; Lee et al., 2022; Krafft & Bauer, 2026). Beyond that, advanced modeling approaches have been developed that directly leverage structural workflow metadata to combat irregular sampling regimes (Zhang et al., 2023; Karami et al., 2024; Sun et al., 2024; Song et al., 2025). However, these solutions often relegate temporal dynamics to opaque representations, and, although the assessment of data quality is an essential component of the data-centric AI lifecycle (Jarrahi et al., 2023), a systematic, data-centric approach that explicitly quantifies, profiles, and extracts the structure of sampling trajectories remains underdeveloped. We argue that irregularity manifests in distinct regimes, and that a model's vulnerability to non-uniform sampling depends heavily on which regime it is operating within. Rather than treating sampling behavior as a nuisance to be imputed, we propose that it can be explicitly parameterized into a diagnostic instrument. To evaluate this, this paper addresses two research questions:

1. To what extent does explicit structural metadata provide an independent predictive signal that improves performance over naive observation counts and standard observation baselines?
2. How does the utility of this structural signal fluctuate across distinct manifestations of temporal irregularity, and in which specific manifestations are models most vulnerable?

Our contributions include a feature framework to explicitly extract structural workflow metadata and a novel stratifica-

[1]Software Methodologies for Distributed Systems, University of Augsburg, Augsburg, Germany. Correspondence to: Tamara Krafft <tamara.krafft@uni-a.de>.

Proceedings of the Workshop on Structured Data for Health at the 43rd International Conference on Machine Learning, Seoul, South Korea. Copyright 2026 by the author(s).

tion approach to systematically map how distinct irregularity regimes dictate model robustness.

2. Methods

To evaluate the predictive value of structural workflows, we extract temporal sampling behaviors as explicit features, isolate their impact via targeted ablations, and assess model robustness using a leakage-free stratification framework.

2.1. Data and Predictive Task

We utilize the MIMIC-IV and MIMIC-IV-ED (Johnson et al., 2024; 2023) datasets and define the binary classification of clinical deterioration as the composite endpoint of either in-hospital mortality within 48 hours or admission to the intensive care unit (via MIMIC-IV). To ensure a consistent observation window across the cohort, vital sign observations are restricted to the first six hours of each patient's emergency department (ED) stay (via MIMIC-IV-ED). We enforce an inclusion criterion, requiring patients to have ≥ 6 observations within this window, ensuring sufficient temporal density to extract meaningful structural features downstream and avoid confounding effects of near-empty trajectories. Additionally, we exclude any time point where more than 50% of the vital signs are missing simultaneously. The final cohort comprises $N = 3,497$ stays, encompassing a total of $23,979$ timestamps with an outcome prevalence of 37.52%.

2.2. Feature Space

For each stay, vital sign data is represented as a multivariate time series $X = \{(t_i, \mathbf{v}_i)\}_{i=1}^{n}$, where all observations are aligned to a single, shared irregular temporal grid. Each $t_i \in \mathbb{R}^+$ denotes the timestamp of the i-th measurement event, and $\mathbf{v}_i \in \mathbb{R}^6$ represents the corresponding vector of six core vital signs (heart rate, respiratory rate, temperature, oxygen saturation, systolic, and diastolic blood pressure) recorded synchronously. We partition the input vectors into a feature space, explicitly separating the patient's physiological state from the provider's sampling behavior.

Physiological Space ($\mathcal{F}_{phys}$): This subspace models the patient's physiological trajectory directly from the synchronously observed vital signs $\mathbf{v}_i$. To summarize these irregular temporal sequences into tabular inputs, we extract the mean, median, minimum, maximum, and standard deviation, alongside the first and last observed values for each physiological parameter per stay.

Structural Metadata Space ($\mathcal{F}_{meta}$): This subspace captures irregularity by computing descriptors of the measurement timestamps. We define the sequence of inter-observation intervals as $\Delta T = (\Delta t_i)_{i=2}^{n}$, where $\Delta t_i = t_i - t_{i-1}$. From ΔT, we derive (i) interval z-scores $z_{\Delta t}$ to normalize for individual baseline rhythms, (ii) instantaneous acceleration $a_i = \Delta t_i - \Delta t_{i-1}$ for $i \geq 3$, and (iii) the 5-step rolling mean and standard deviation of a_i. Again, these sequences are subsequently aggregated into stay-level summary statistics, including the mean, median, minimum, maximum, and standard deviation.

Finally, patient age and sex are concatenated to the unified feature vector to anchor the physiological and structural signals to the patient's fundamental baseline.

2.3. Ablation Pipelines

Five pipelines are constructed specifically to isolate the predictive value of temporal sparsity and assess the impact of standard imputation techniques: P_{base} models the raw, irregularly sampled vitals using only $\mathcal{F}_{phys}$; P_{count} enhances P_{base} with the total observation count per stay ($\mathcal{F}_{phys} \cup n$); P_{locf} forces measurements into a rigid 30-minute grid utilizing LOCF; P_{meta} preserves native sparsity while integrating the full structural metadata space ($\mathcal{F}_{phys} \cup \mathcal{F}_{meta}$); and $P_{meta+locf}$ computes $\mathcal{F}_{meta}$ using the raw timestamps, but subsequently forces the same 30-minute LOCF grid as P_{locf}.

This ablation design isolates the value of structural metadata beyond naive observation frequency (P_{meta} vs. P_{count}). Furthermore, it explicitly quantifies how standard imputation impacts the innate temporal anchors of clinical workflows, even when structural metadata is preserved (P_{locf} and $P_{meta+locf}$ vs. P_{meta}).

2.4. Two-Phase Stratification

We evaluate model robustness and the impact of crafted features across different irregularity regimes by introducing a two-phase stratification approach.

Phase 1 (Global Axis Selection): To characterize the irregularity of a time series X, we extract ΔT. Each encounter in the full cohort is profiled across three distinct metrics: Quartile Coefficient of Dispersion (QCD), Adherence Rate, and Coefficient of Variation (CV). CV measures the relative dispersion of the temporal gaps, defined as:

$$CV = \frac{\sigma}{\mu}$$

where σ and μ are the standard deviation and mean of the intervals in ΔT, respectively.

QCD measures macro-level dispersion, is robust to outliers, and is defined as:

$$QCD = \frac{q_3 - q_1}{q_3 + q_1}$$

where q_1 and q_3 denote the 25th and 75th percentiles of ΔT. Adherence Rate measures strict adherence to the dominant,

localized rhythm. To identify this underlying cadence, elements in ΔT are first discretized into 60-second bins to extract the modal interval m. Given an allowable tolerance τ, Adherence is defined as the proportion of conforming intervals:

$$Adherence = \frac{1}{|\Delta T|} \sum_{i=2}^{n} \mathbb{I}(|\Delta t_i - m| \leq \tau \cdot m)$$

where $\mathbb{I}$ represents the indicator function. We set $\tau = 0.5$, accepting any interval within $\pm$ 50% of the dominant cadence, ensuring comparability across patients with different monitoring frequencies. To construct an evaluation topography that measures fundamentally distinct aspects of irregularity, we applied a data-driven selection criterion, prioritizing the metric pair with the least absolute Spearman rank correlation ρ. Consequently, Adherence Rate and QCD are selected, exhibiting the lowest collinearity ($\rho = -0.430$) in our cohort, see Table 2 in the Appendix.

Phase 2 (Stratification): To operationalize these axes without introducing target data leakage, the stratification logic is embedded directly within the cross-validation loop. During each fold iteration, the median thresholds for Adherence Rate and QCD are calculated strictly on the training fold. These localized, training-derived thresholds are subsequently applied to the hold-out test fold, projecting the unseen data into four mutually exclusive operational quadrants, which we define as Q_α (Low QCD, High Adherence), Q_β (Low QCD, Low Adherence), Q_γ (High QCD, High Adherence), and Q_δ (High QCD, Low Adherence).

2.5. Evaluation Framework

All pipelines are evaluated using an XGBoost classifier (Chen & Guestrin, 2016), selected for its native support for missing data and high performance on tabular data (Grinsztajn et al., 2022). Baseline hyperparameters (see Appendix Table 3) are fixed and universally applied across all five pipelines without further tuning. Robust performance estimation is achieved via 5x10 repeated stratified K-fold cross-validation, with splits generated strictly at the patient level (subject_id) to prevent longitudinal leakage across recurrent encounters (stay_id).

2.6. Interpretability and Feature Orthogonality

Using the test hold-out from the final fold of P_{meta}, we extract SHapley Additive exPlanations (SHAP) values to identify the most globally impactful features (Lundberg & Lee, 2017). Specifically, we isolate the top-k features of both $\mathcal{F}_{phys}$ and $\mathcal{F}_{meta}$, ranked by their mean absolute SHAP values. Subsequently, we compute pairwise Spearman rank correlation (ρ) across all $k \times k$ feature pairs. Feature independence is evaluated by analyzing the mean absolute correlation ($|\rho|$) and the proportion of statistically significant pairs ($p < 0.05$) to determine whether structural metadata is merely a redundant proxy for physiological extremes.

Table 1. Overall predictive performance across 50 CV runs (Mean $\pm$ 95% CI)

Pipeline	AUPRC	AUPRC Lift	AUROC	Brier Skill Score
P_{base}	0.654 ± 0.006	1.743 ± 0.016	0.717 ± 0.005	0.076 ± 0.012
P_{count}	0.666 ± 0.006	1.774 ± 0.015	0.726 ± 0.005	0.092 ± 0.011
P_{locf}	0.502 ± 0.007	1.338 ± 0.018	0.612 ± 0.006	-0.109 ± 0.013
P_{meta}	$\mathbf{0.681 \pm 0.005}$	$\mathbf{1.815 \pm 0.014}$	$\mathbf{0.745 \pm 0.004}$	$\mathbf{0.117 \pm 0.010}$
$P_{meta+locf}$	0.513 ± 0.007	1.367 ± 0.019	0.616 ± 0.006	-0.102 ± 0.013

3. Results

The evaluation comprises three dimensions: overall ablative performance, robustness across irregularity strata, and feature orthogonality.

Performance metrics across the five ablation pipelines are summarized in Table 1. The native sparsity baseline relying solely on physiology (P_{base}) achieved a mean AUPRC of 0.654. Augmenting this baseline with a temporal observation count (P_{count}) increased the AUPRC to 0.666, representing a +1.8% improvement. The proposed framework (P_{meta}), which integrates the full structural metadata space, achieved a mean AUPRC of 0.681, representing +4.1% over P_{base}. Applying a rigid 30-minute LOCF grid to the physiological observations (P_{locf}) resulted in an AUPRC of 0.502. By augmenting this 30-minute grid with structural metadata ($P_{meta+locf}$), AUPRC increased at 0.513.

Projecting the test-fold instances into the stratification quadrants isolates model performance across distinct observation patterns, see Figure 1 for AUPRC Lift and Appendix Figure 2 for AUROC. In Q_α, characterized by low QCD and high structural adherence, P_{meta} achieved an AUPRC Lift of 1.791 (a +2.1% improvement over P_{base}). However, the impact of the structural metadata emerged in Q_β, which is characterized by low QCD but low adherence, where P_{meta} achieved its highest predictive lift of 1.942, representing an 8.0% improvement over P_{base}. In Q_δ, characterized by high QCD and low adherence, P_{meta} maintained a highly confident AUPRC Lift of 1.828 (+4.8%).

To assess the relationship between structural and physiological signals, we analyzed the test-fold SHAP values and feature correlations. At the aggregate group level, structural metadata commands approximately one-third (30.6% to 33.3%) of the model's total predictive attention across all quadrants (Appendix Figure 3). Based on mean absolute SHAP values, the mean interval length (Δt) and the maximum interval acceleration (a_t) were the first and second most globally impactful $\mathcal{F}_{meta}$ features, scoring 0.3360 and 0.1717, respectively. For comparison, the highest-ranking $\mathcal{F}_{phys}$ features were mean respiratory rate (0.3016) and minimum heart rate (0.2547). The mean absolute correlation

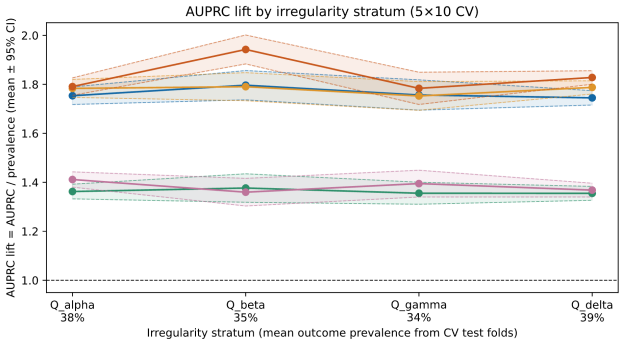

Figure 1. AUPRC Lift by irregularity stratum. Evaluated pipelines are denoted by color: P_{base} (blue), P_{count} (yellow), P_{locf} (green), P_{meta} (red), and $P_{meta+locf}$ (pink).

across the top-10 features of both feature sets (100 feature pairs) was $|\rho| = 0.041$, and 83% (83 pairs) exhibited no statistically significant correlation ($p > 0.05$). A ρ heatmap highlights these results in the Appendix, Figure 4.

4. Discussion

Our findings demonstrate that the temporal irregularity of EHR data can serve as an informative, independent signal. The ablation analysis explicitly quantifies how standard imputation ($P_{locf}, P_{meta+locf}$) can degrade predictive power by erasing the behavioral footprint of clinical staff. While simple observation counts (P_{count}) yield a marginal baseline improvement by capturing the overall volume of surveillance, the proposed pipeline P_{meta} more than doubles this predictive gain by capturing the cadence and acceleration of the sampling trajectory. Evaluation across the stratified irregularity subcohorts reveals that P_{meta} achieves its maximum relative predictive lift not within the most uniform data (Q_α), but in cases of highly volatile irregularity. Specifically, in Q_β, where low adherence and low QCD reflect rapid, clustered triage workflows, P_{meta} outperforms all other pipelines. Noticeably, in Q_γ where irregularity manifests as high QCD but high local adherence, P_{meta} provides only a marginal advantage over the baseline. This contrast implies that structural metadata is uniquely equipped to mitigate the administrative chaos of low-adherence sampling. Volatile time intervals reflect how provider attention shifts dynamically in response to patient acuity. Parametrizing this behavior explicitly enables translating that volatility into a primary signal. Crucially, near-zero mean Spearman correlation and lack of statistical significance across 83% of the top structural and physiological feature pairs demonstrate that sampling behavior operates on a distinct diagnostic axis. Given these findings, we propose our two-phase stratification approach as a diagnostic tool for cohort and model evaluation. Researchers can utilize this approach to systematically profile the innate irregularity of their cohorts prior to deploying black-box models and systematically identify performance degradation, exposing model vulnerabilities linked to specific irregularity regimes.

This work is subject to several limitations. First, continuous vital sign dynamics were flattened into stay-level aggregates. We acknowledge that this discards fine-grained sequential interactions and that deep learning architecture natively handling irregular sequences would likely achieve higher absolute performance. However, applying sequential deep learning would implicitly entangle physiological data with sampling behavior inside an opaque representation. Flattening the sequence into tabular feature spaces was a deliberate methodological choice. It allowed us to utilize XGBoost and SHAP values to strictly isolate the predictive weight of structural metadata, quantify its orthogonality to physiological data, and analyze behaviors across our stratification quadrants. Second, we utilize LOCF to represent imputation. While known to introduce bias and dampen variance (Cook et al., 2004), LOCF serves here as a tool to purge the signal of structural information and highlight the destructive nature of temporal gridding. Finally, our cohort was restricted to patients undergoing active monitoring within a single-center dataset, which may reflect institution-specific protocols. Future research must validate this approach across multi-center datasets and build upon this foundational profiling by integrating structural metadata directly into sequential models to capture full temporal dynamics. Furthermore, the proposed methodology can be extended beyond the irregularity of the underlying time grid to diagnose other EHR meta-signals, including broader informative missingness and informative plausibility violations.

5. Conclusion

We demonstrate that temporal irregularity in EHR data serves as an independent signal when explicitly extracted as structural workflow metadata. Incorporating this metadata outperforms simple observation counts, whereas standard LOCF gridding systematically degrades predictive power. Feature importance analysis further indicates that sampling behavior represents a statistically independent signal. To contextualize these effects, we introduce a two-phase stratification approach that isolates different irregularity manifestations and enables controlled evaluation of model robustness across these regimes. Utilizing this approach, we found that the inclusion of structural metadata achieves its maximum predictive impact during low-adherence workflows. Ultimately, these findings suggest that leveraging irregularity through structural metadata allows for capturing behavioral signals associated with clinical workflow dynamics. Beyond feature engineering, the proposed stratification approach functions as a diagnostic instrument, enabling systematic identification of performance degradation linked to specific

irregularity regimes.

Acknowledgements

Funding. This research was supported by the Bavarian Ministry of Economic Affairs, Regional Development and Energy under grant number DIK0764/02.

Transparency note. The readability of this text has been improved with the support of AI technologies.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning within healthcare settings. By leveraging the temporal irregularity of electronic health records to predict clinical deterioration , this research has the potential to improve early warning systems and patient outcomes. However, because structural workflow metadata inherently reflects clinical staffing and triage behaviors, these patterns may vary significantly across different hospital systems or demographic groups. Future deployments of such models should carefully monitor for potential systemic biases to ensure equitable care.

References

Chen, T. and Guestrin, C. Xgboost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016. doi: 10.1145/2939672.2939785.

Cook, R. J., Zeng, L., and Yi, G. Y. Marginal analysis of incomplete longitudinal binary data: A cautionary note on locf imputation. *Biometrics*, 60(3):820–828, 09 2004. ISSN 0006-341X. doi: 10.1111/j.0006-341X.2004.00234.x.

Escobar, G. J., Liu, V. X., Schuler, A., Lawson, B., Greene, J. D., and Kipnis, P. Automated identification of adults at risk for in-hospital clinical deterioration. *New England Journal of Medicine*, 383(20):1951–1960, 2020. ISSN 0028-4793. doi: 10.1056/nejmsa2001090.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22. Curran Associates Inc., 2022. ISBN 9781713871088.

Jarrahi, M. H., Memariani, A., and Guha, S. The principles of data-centric ai. *Communications of the ACM*, 66(8): 84–92, 2023. ISSN 0001-0782. doi: 10.1145/3571724.

Johnson, A., Bulgarelli, L., Pollard, T., Celi, L. A., Mark, R., and Horng, S. MIMIC-IV-ED. *PhysioNet*, January 2023. doi: 10.13026/5ntk-km72. Version 2.2.

Johnson, A., Bulgarelli, L., Pollard, T., Gow, B., Moody, B., Horng, S., Celi, L. A., and Mark, R. MIMIC-IV. *PhysioNet*, October 2024. doi: 10.13026/kpb9-mt58. Version 3.1.

Karami, H., Atienza, D., and Ionescu, A. TEE4EHR: Transformer event encoder for better representation learning in electronic health records. *Artificial Intelligence in Medicine*, 154:102903, 2024. ISSN 0933-3657. doi: 10.1016/j.artmed.2024.102903.

Krafft, T. and Bauer, B. Toward Data-Centric Experimentation in Prehospital AI: A Customizable Pipeline. *Proceedings of the 19th International Joint Conference on Biomedical Engineering Systems and Technologies*, pp. 351–363, 2026. doi: 10.5220/0014619300004070.

Lee, Y., Jun, E., Choi, J., Suk, H.-I., and Suk, H.-I. Multi-view integrative attention-based deep representation learning for irregular clinical time-series data. *IEEE Journal of Biomedical and Health Informatics*, 26(8):4270–4280, 2022. ISSN 2168-2194. doi: 10.1109/jbhi.2022.3172549.

Lipton, Z. C., Kale, D. C., and Wetzel, R. Modeling missing data in clinical time series with rnns. *arXiv*, 2016. doi: 10.48550/arxiv.1606.04130.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Mann, K. D., Good, N. M., Fatehi, F., Khanna, S., Campbell, V., Conway, R., Sullivan, C., Staib, A., Joyce, C., and Cook, D. Predicting patient deterioration: A review of tools in the digital hospital setting. *Journal of Medical Internet Research*, 23(9):e28209, 2021. ISSN 1439-4456. doi: 10.2196/28209.

Song, Z., Lu, Q., Zhu, H., Buckeridge, D., and Li, Y. Trajgpt: Irregular time-series representation learning of health trajectory. *IEEE Journal of Biomedical and Health Informatics*, PP(99):1–14, 2025. ISSN 2168-2194. doi: 10.1109/jbhi.2025.3620205.

Sun, C., Li, H., Song, M., Cai, D., Zhang, B., and Hong, S. Time pattern reconstruction for classification of irregularly sampled time series. *Pattern Recognition*, 147: 110075, 2024. ISSN 0031-3203. doi: 10.1016/j.patcog.2023.110075.

Tipirneni, S. and Reddy, C. K. Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Transactions on Knowledge Discovery*

from Data (TKDD), 16(6):1–17, 2022. ISSN 1556-4681. doi: 10.1145/3516367.

Xie, F., Yuan, H., Ning, Y., Ong, M. E. H., Feng, M., Hsu, W., Chakraborty, B., and Liu, N. Deep learning for temporal data representation in electronic health records: A systematic review of challenges and methodologies. *Journal of Biomedical Informatics*, 126:103980, 2022. ISSN 1532-0464. doi: 10.1016/j.jbi.2021.103980.

Zhang, J., Zheng, S., Cao, W., Bian, J., and Li, J. Warpformer: A Multi-scale Modeling Approach for Irregular Clinical Time Series. *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3273–3285, 2023. doi: 10.1145/3580305.3599543.

A. Appendix

Table 2. Pairwise Spearman rank correlation (ρ) between stay-level irregularity metrics for Phase 1 Axis Selection. The combination of Adherence Rate and QCD exhibits the lowest collinearity ($|\rho| = 0.430$) and was consequently selected to define the orthogonal stratification topography.

Metric 1	**Metric 2**	**Correlation (ρ)**	$\lvert\rho\rvert$
Adherence Rate	QCD	-0.430	0.430
CV	QCD	0.668	0.668
CV	Adherence Rate	-0.717	0.717

Table 3. XGBoost hyperparameter configuration applied universally across all evaluation pipelines.

Hyperparameter	**Value**
`objective`	`binary:logistic`
`eval_metric`	`aucpr`
`n_estimators`	300
`learning_rate`	0.1
`max_depth`	6
`min_child_weight`	1
`gamma`	0
`subsample`	0.8
`colsample_bytree`	0.8
`tree_method`	`hist`
`scale_pos_weight`	$N_{\text{neg}}/N_{\text{pos}}$ per training fold
`random_state`	$42 +$ fold_index

Table 4. Statistical summary of irregularity metrics across the four operational quadrants (Mean $\pm$ 95% CI) including mean stratum size (N). The metrics characterize the distinct temporal sampling regimes defined by Quartile Coefficient of Dispersion (QCD), Adherence Rate, and the Coefficient of Variation (CV).

Stratum	**N**	**QCD**	**Adherence Rate**	**CV**
Q_α	1152	0.210 ± 0.002	0.733 ± 0.002	0.468 ± 0.003
Q_β	603	0.257 ± 0.002	0.309 ± 0.003	0.729 ± 0.007
Q_γ	495	0.541 ± 0.003	0.608 ± 0.001	0.683 ± 0.004
Q_δ	1251	0.591 ± 0.003	0.357 ± 0.002	0.881 ± 0.004

Table 5. Top 10 physiological and structural feature distributions across patient outcomes, ranked by global mean absolute SHAP value. Values are presented as Median [Q1, Q3]. Outcome 0 represents stable patient encounters, while Outcome 1 indicates clinical deterioration, defined as the composite endpoint of intensive care unit admission or in-hospital mortality within 48 hours.

Feature Group	**Feature**	**Outcome = 0**	**Outcome = 1**
Physiology & Metadata	Resp. Rate (Mean)	18.22 [17.00, 20.20]	19.29 [17.43, 22.17]
	Heart Rate (Min)	78.00 [66.00, 90.00]	82.00 [68.00, 97.00]
	Heart Rate (Last)	84.00 [72.00, 97.00]	90.00 [75.00, 104.00]
	Temperature (Last)	36.78 [36.67, 37.11]	36.72 [36.44, 37.11]
	Heart Rate (First)	91.00 [77.00, 108.00]	95.00 [79.00, 112.00]
	Age at Visit	66.00 [52.00, 78.00]	67.00 [56.00, 79.00]
	Systolic BP (Min)	109.00 [98.00, 124.00]	101.00 [89.00, 114.00]
	Temperature (Median)	36.83 [36.67, 37.17]	36.78 [36.50, 37.22]
	Resp. Rate (Last)	18.00 [16.00, 20.00]	19.00 [17.00, 22.00]
	Systolic BP (First)	129.00 [113.00, 149.00]	120.00 [103.00, 141.00]
Structural	Interval Seconds (Mean)	3470.00 [2948.57, 3936.00]	2921.43 [2328.00, 3552.00]
	Interval Accel. (Max)	3600.00 [2340.00, 5220.00]	2940.00 [1980.00, 4320.00]
	Interval Seconds (Median)	3000.00 [2340.00, 3660.00]	2520.00 [1800.00, 3300.00]
	Interval Accel. (Median)	210.00 [-450.00, 930.00]	135.00 [-300.00, 630.00]
	Interval Accel. (Mean)	216.00 [-336.00, 780.00]	105.00 [-310.00, 525.00]
	Interval Z-Score (Median)	-0.15 [-0.34, 0.07]	-0.19 [-0.35, 0.04]
	Interval Z-Score (Min)	-1.07 [-1.27, -0.89]	-1.09 [-1.28, -0.90]
	Interval Accel. (Min)	-3180.00 [-4980.00, -1860.00]	-2880.00 [-4500.00, -1800.00]
	Interval Seconds (Std)	2189.14 [1621.89, 2895.65]	1773.30 [1273.78, 2454.29]
	Interval Seconds (Min)	900.00 [360.00, 1440.00]	780.00 [360.00, 1260.00]

Table 6. Top 10 physiological and structural feature distributions across irregularity quadrants. Values are presented as Mean $\pm$ 95% CI across 50 cross-validation folds.

Feature Group	**Feature**	Q_α	Q_β	Q_γ	Q_δ
Outcome Prevalence	(in %)	38.39 ± 0.65	35.09 ± 1.07	33.79 ± 1.07	39.28 ± 0.66
Physiology & Metadata	Resp. Rate (Mean)	19.26 ± 0.05	19.36 ± 0.07	19.38 ± 0.09	19.46 ± 0.05
	Heart Rate (Min)	79.99 ± 0.34	80.72 ± 0.38	79.53 ± 0.48	79.07 ± 0.21
	Heart Rate (Last)	87.45 ± 0.30	87.37 ± 0.43	86.18 ± 0.50	86.57 ± 0.28
	Temperature (Last)	35.21 ± 0.12	34.97 ± 0.18	35.63 ± 0.18	34.92 ± 0.15
	Heart Rate (First)	93.46 ± 0.39	94.40 ± 0.50	94.57 ± 0.53	93.93 ± 0.33
	Age at Visit	64.89 ± 0.36	65.11 ± 0.37	64.52 ± 0.56	64.04 ± 0.22
	Systolic BP (Min)	108.58 ± 0.35	107.79 ± 0.57	108.96 ± 0.55	106.78 ± 0.37
	Temperature (Median)	35.32 ± 0.12	35.44 ± 0.17	35.72 ± 0.14	35.06 ± 0.12
	Resp. Rate (Last)	19.09 ± 0.07	18.96 ± 0.08	18.96 ± 0.13	19.00 ± 0.06
	Systolic BP (First)	129.57 ± 0.44	129.06 ± 0.63	129.41 ± 0.78	127.80 ± 0.40
Structural	Interval Seconds (Mean)	3262.79 ± 14.14	3359.95 ± 37.39	3344.87 ± 29.86	3157.22 ± 18.86
	Interval Accel. (Max)	2622.09 ± 23.07	4007.76 ± 59.68	3845.75 ± 39.57	4457.47 ± 34.74
	Interval Seconds (Median)	3274.71 ± 16.75	2588.93 ± 17.10	3562.07 ± 28.39	2273.69 ± 16.55
	Interval Accel. (Median)	160.42 ± 15.46	188.13 ± 23.34	187.45 ± 32.21	164.51 ± 22.29
	Interval Accel. (Mean)	142.37 ± 9.00	-38.21 ± 46.54	182.40 ± 37.14	28.20 ± 23.82
	Interval Z-Score (Median)	0.01 ± 0.00	-0.27 ± 0.00	0.11 ± 0.01	-0.29 ± 0.00
	Interval Z-Score (Min)	-1.29 ± 0.01	-0.97 ± 0.01	-1.15 ± 0.00	-0.95 ± 0.00
	Interval Accel. (Min)	-2411.35 ± 24.54	-4568.87 ± 166.78	-3491.21 ± 114.13	-4741.12 ± 80.31
	Interval Seconds (Std)	1490.36 ± 12.25	2606.71 ± 75.75	2305.33 ± 42.62	2844.19 ± 29.94
	Interval Seconds (Min)	1359.60 ± 12.44	1039.49 ± 18.39	759.71 ± 11.90	603.74 ± 7.61

Table 7. Predictive performance by irregularity stratum across 50 CV runs (Mean ± 95% CI), including Area Under the Precision-Recall Curve, Area Under the Receiver Operating Characteristic Curve, and Brier Skill Score. To provide baseline context for the subgroups, the test-fold outcome prevalence is as follows: Overall (0.375 ± 0.001), Q_α (0.384 ± 0.007), Q_β (0.351 ± 0.011), Q_γ (0.338 ± 0.011), and Q_δ (0.393 ± 0.007).

Stratum	Pipeline	AUPRC	AUPRC Lift	AUROC	Brier Skill Score
Overall Cohort	P_{base}	0.654 ± 0.006	1.743 ± 0.016	0.717 ± 0.005	0.076 ± 0.012
	P_{count}	0.666 ± 0.006	1.774 ± 0.015	0.726 ± 0.005	0.092 ± 0.011
	P_{locf}	0.502 ± 0.007	1.338 ± 0.018	0.612 ± 0.006	-0.109 ± 0.013
	$P_{meta+locf}$	0.513 ± 0.007	1.367 ± 0.019	0.616 ± 0.006	-0.102 ± 0.013
	P_{meta}	$\mathbf{0.681 \pm 0.005}$	$\mathbf{1.815 \pm 0.014}$	$\mathbf{0.745 \pm 0.004}$	$\mathbf{0.117 \pm 0.010}$
Q_α	P_{base}	0.671 ± 0.010	1.753 ± 0.036	0.725 ± 0.009	0.095 ± 0.020
	P_{count}	0.683 ± 0.010	1.783 ± 0.037	0.738 ± 0.010	0.115 ± 0.021
	P_{locf}	0.522 ± 0.011	1.362 ± 0.030	0.619 ± 0.010	-0.098 ± 0.020
	$P_{meta+locf}$	0.540 ± 0.010	1.411 ± 0.031	0.628 ± 0.009	-0.081 ± 0.019
	P_{meta}	$\mathbf{0.685 \pm 0.009}$	$\mathbf{1.791 \pm 0.036}$	$\mathbf{0.744 \pm 0.009}$	$\mathbf{0.105 \pm 0.020}$
Q_β	P_{base}	0.626 ± 0.017	1.797 ± 0.060	0.711 ± 0.012	0.047 ± 0.028
	P_{count}	0.624 ± 0.018	1.790 ± 0.057	0.710 ± 0.013	0.040 ± 0.028
	P_{locf}	0.479 ± 0.018	1.377 ± 0.058	0.609 ± 0.014	-0.124 ± 0.027
	$P_{meta+locf}$	0.474 ± 0.017	1.360 ± 0.056	0.604 ± 0.014	-0.133 ± 0.027
	P_{meta}	$\mathbf{0.676 \pm 0.015}$	$\mathbf{1.942 \pm 0.059}$	$\mathbf{0.749 \pm 0.011}$	$\mathbf{0.131 \pm 0.025}$
Q_γ	P_{base}	0.588 ± 0.016	1.757 ± 0.062	0.687 ± 0.013	0.000 ± 0.025
	P_{count}	0.587 ± 0.016	1.753 ± 0.059	0.688 ± 0.012	0.007 ± 0.024
	P_{locf}	0.456 ± 0.018	1.355 ± 0.045	0.581 ± 0.014	-0.188 ± 0.027
	$P_{meta+locf}$	0.469 ± 0.018	1.395 ± 0.055	0.579 ± 0.014	-0.187 ± 0.028
	P_{meta}	$\mathbf{0.597 \pm 0.015}$	$\mathbf{1.784 \pm 0.066}$	$\mathbf{0.697 \pm 0.013}$	$\mathbf{0.006 \pm 0.025}$
Q_δ	P_{base}	0.684 ± 0.009	1.745 ± 0.029	0.726 ± 0.007	0.088 ± 0.016
	P_{count}	0.701 ± 0.010	1.788 ± 0.027	0.737 ± 0.007	0.113 ± 0.015
	P_{locf}	0.531 ± 0.012	1.355 ± 0.028	0.619 ± 0.011	-0.099 ± 0.021
	$P_{meta+locf}$	0.537 ± 0.013	1.368 ± 0.028	0.623 ± 0.010	-0.091 ± 0.019
	P_{meta}	$\mathbf{0.717 \pm 0.010}$	$\mathbf{1.828 \pm 0.028}$	$\mathbf{0.765 \pm 0.007}$	$\mathbf{0.151 \pm 0.017}$

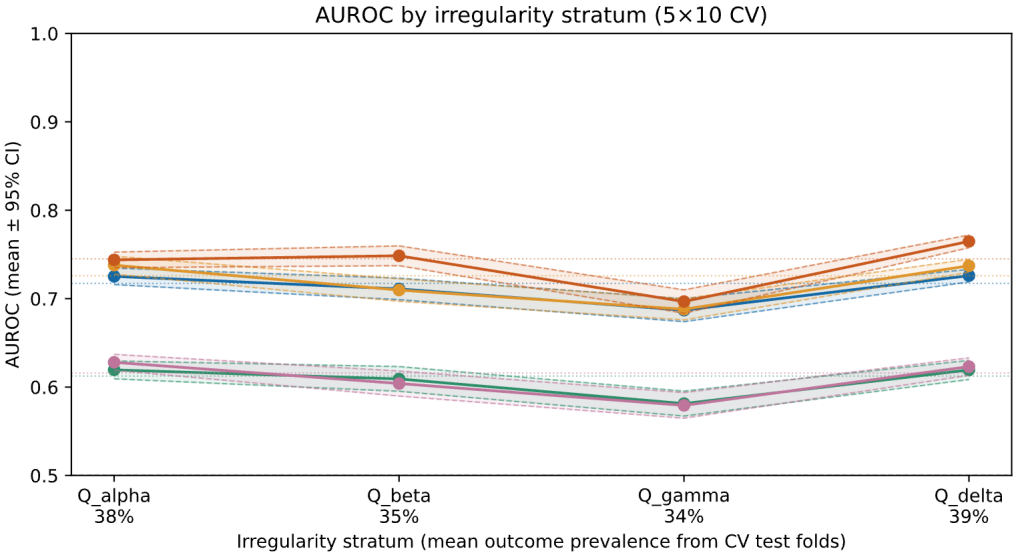

Figure 2. Area Under the Receiver Operating Characteristic Curve by irregularity stratum. Evaluated pipelines are denoted by color: P_{base} (blue), P_{count} (yellow), P_{locf} (green), P_{meta} (red), and $P_{meta+locf}$ (pink). Consistent with the AUPRC Lift trends, the integration of explicit structural metadata (P_{meta}) demonstrates superior performance across all quadrants, while rigid imputation (P_{locf}, $P_{meta+locf}$) systematically suppresses model performance.

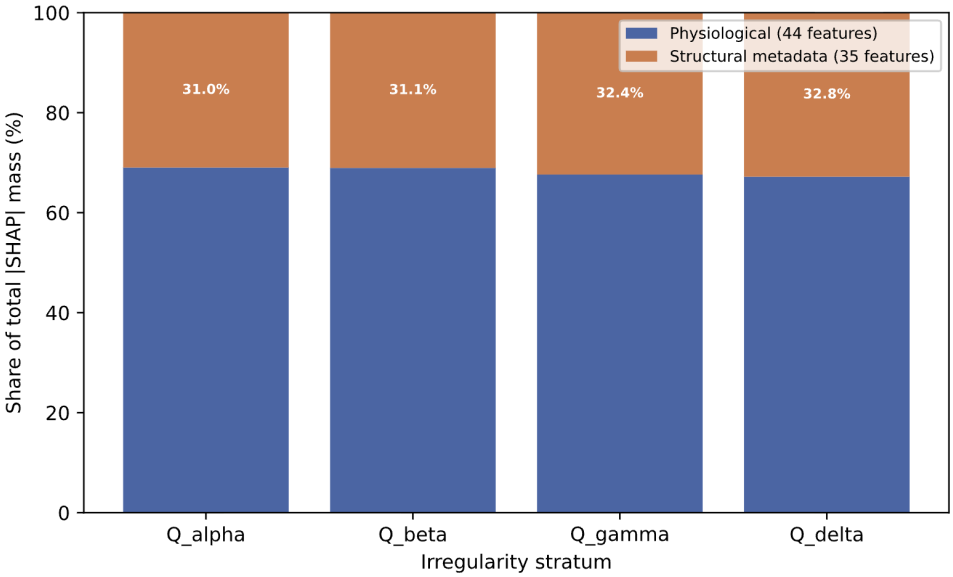

Figure 3. Aggregate predictive influence by feature space and irregularity stratum. The stacked representation illustrates the proportional contribution of the physiological ($\mathcal{F}_{phys}$) and structural metadata ($\mathcal{F}_{meta}$) feature sets, derived from global mean absolute SHAP values. Across all distinct operational regimes (Q_α–Q_δ), the explicit parameterization of sampling behavior consistently commands approximately one-third of the model's total predictive attention.

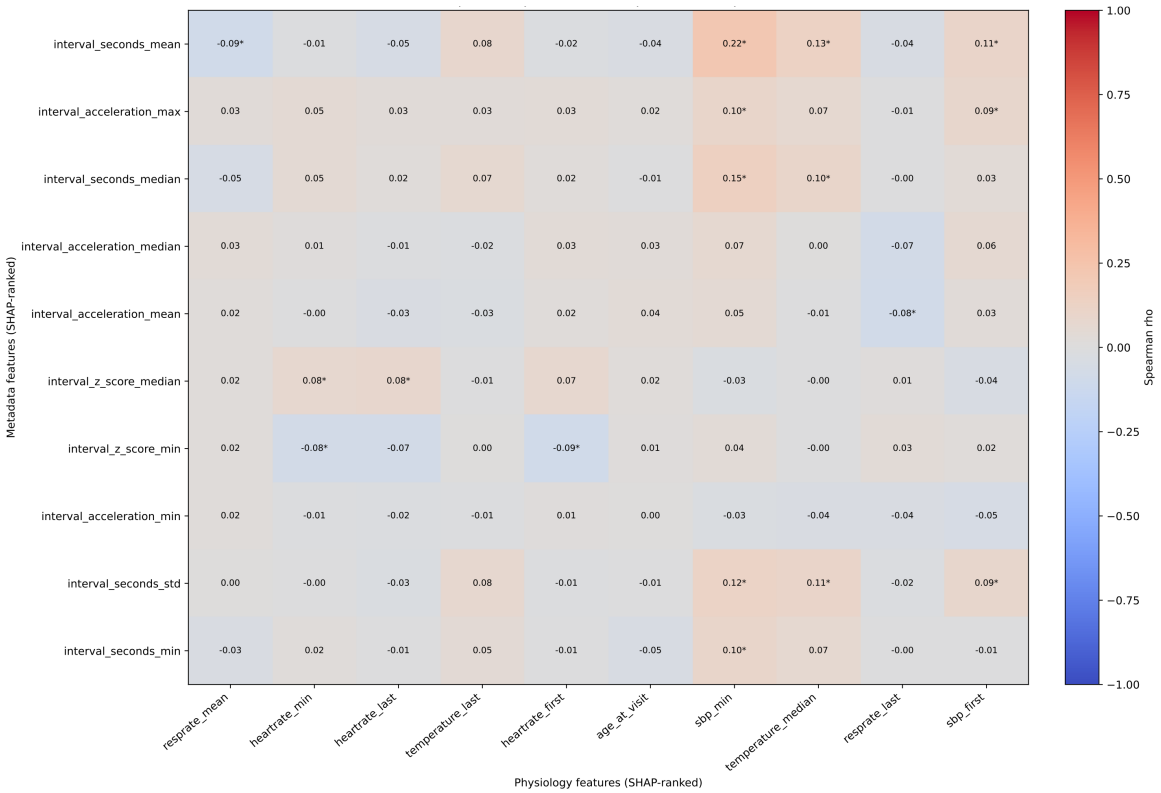

Figure 4. Spearman rank correlation (ρ) heatmap comparing the top-ranked ($k = 10$) physiological and structural metadata features. Statistical significant pairs ($p < 0.05$) are indicated with an asterisk (*). Features are ordered by importance: top to bottom, left to right.