

# Informative Missingness to Generate Irregular Clinical Time Series

Hadi Mehdizavareh<sup>1\*</sup>, Gabriele Santangelo<sup>2</sup>, Giovanna Nicora<sup>2</sup>, Simon Lebech Cichosz<sup>1</sup>, Arianna Dagliati<sup>2</sup>, Arijit Khan<sup>3,1</sup>, and Riccardo Bellazzi<sup>2,1</sup>

<sup>1</sup> Aalborg University, Aalborg, Denmark  
mhme@cs.aau.dk

<sup>2</sup> University of Pavia, Pavia, Italy

<sup>3</sup> Bowling Green State University, Bowling Green, Ohio, USA

**Abstract.** Laboratory tests in electronic health records are collected irregularly, and the absence of a test order can be as informative as the measurement itself. Such missingness reflects clinicians’ decisions and patient physiology, making it important to model it directly rather than treat it as a preprocessing artifact. Here we present a diffusion-based approach for generating clinical time series that jointly models laboratory values and their observation patterns using the public Data Analytics Challenge on Missing data Imputation (DACMI) benchmark derived from MIMIC-III. To preserve realistic sampling, we align chart times into 4-hour intervals and segment admissions into 7-day windows, producing trajectories that pair each lab value with a corresponding observation indicator. Standard transformations and normalization are applied to stabilize training. Our method extends the TimeDiff framework to learn continuous lab values and discrete missingness patterns through complementary diffusion objectives. Experiments show that the generated data closely match real patient trajectories across individual lab distributions and joint value-missingness embeddings, demonstrating that diffusion models can capture clinically meaningful dependencies between patient physiology and clinicians’ testing behavior under MNAR-like (missing-not-at-random) missingness. These preliminary results indicate that our model can serve as an initial component toward developing clinical foundation models. By producing synthetic priors that preserve key physiology–missingness relationships, this work motivates the subsequent training of Prior-Data Fitted Networks capable of leveraging informative missingness, which we will investigate in the extended work.

**Keywords:** Clinical time series · Diffusion models · Informative missingness · Synthetic EHR generation

## 1 Background

Electronic health records (EHRs) record irregular sequences of measurements shaped by physiology, patient behavior, and clinical decisions [8]. As an example, labs are rarely collected on a fixed schedule, and the absence of testing can

---

\* Corresponding author.

itself reflect severity or clinical intent. Prior work shows that lab missingness is often informative rather than random and can differ across sites and care settings [6,10]. For generative modeling, this means that producing realistic synthetic lab trajectories requires matching not only lab values, but also when tests occur, how they appear in bursts, and how sampling relates to the underlying values. Many clinical time-series pipelines handle missingness mainly through imputation, aiming to densify sparse records before modeling [3,4]. The DACMI challenge [5], built from MIMIC-III [2], offers a public benchmark for imputation methods with ground truth for 13 common blood labs. However, success on imputation benchmarks does not necessarily show that a generator captures the joint process of values and observations under real EHR sampling. This is most critical in the Missing-Not-At-Random (MNAR) setting, where whether a lab is measured depends on unobserved or partially observed clinical state, making missingness tightly coupled to the data-generating mechanism. Diffusion models are state-of-the-art option for time-series generation [1,9]. TimeDiff [7] is among the first diffusion approaches tailored to EHR sequences and supports mixed continuous and binary variables. Still, much prior evaluation focuses on near-regular signals (e.g., vitals) and reports missingness mainly through marginal rates, leaving open whether models reproduce irregular lab sampling bursts and their coupling with values—an issue that can persist even when univariate distributions look correct.

**Contributions.** We present a generative framework designed as the preliminary step for developing clinical foundation models, providing high-fidelity synthetic priors required to train Prior-Data Fitted Networks [12]. Our contributions are: (i) A diffusion-based synthetic lab generation pipeline on a public benchmark that targets irregular labs and informative missingness, generating both lab values and their observation indicators jointly to support downstream inference. (ii) We modify the numerical (Gaussian) loss in TimeDiff [7] to exclude truly missing lab values, treating missingness as absence rather than noisy values. (iii) We present a practical preprocessing and windowing protocol designed to preserve MNAR-like structure while remaining compatible with diffusion training. (iv) We evaluate distributional similarity using univariate lab distributions, joint embeddings over values and masks, and additional similarity metrics for model selection to assess realism in both physiology and sampling behavior.

## 2 Materials and Methods

**Dataset.** We used the publicly available DACMI imputation challenge dataset [5], derived from MIMIC-III, which provides a standardized train/test split designed for fair comparison of methods on sparse clinical time series. From each admission, we extracted 13 laboratory variables and constructed a corresponding missingness mask directly from the raw tables, where each lab at each time step is marked as observed (0) or missing (1). This yields a paired representation of physiology (continuous lab values) and sampling behavior (binary indicators) that is suitable for studying informative missingness, especially in the MNAR-

like regime where ordering and frequency of tests reflect clinical decisions. Models development and selection are based on the DACMI training split.

**Setup and preprocessing.** We discretized `charttime` (minutes since admission) into 4-hour intervals to minimize artificial missingness, creating 42-step windows (7 days). Admissions exceeding this horizon were segmented into consecutive non-overlapping windows, yielding 16,580 training samples from 8,267 admissions. To approximate Gaussianity, features underwent log or Box-Cox transformations followed by z-score standardization. The model input consists of 26 channels (13 standardized values and 13 binary masks). In addition to the values and binary mask trajectories, we used *delta conditioning* to encode observation recency. The delta conditioning provides an auxiliary conditioning tensor to the denoising network at each diffusion step. Therefore, the model still generates only the lab values and mask variables, but is guided by this extra context during denoising. For each lab and time index, the delta value represents the number of steps since the most recent observation. Specifically, at each time index, we computed the number of time steps that must be traversed backward from the current position to reach the latest time point at which the lab is observed (mask = 0). This yields a (42, 13) matrix per window, normalized to  $[0, 1]$  by division by 42.

**Experimental setup and sampling.** We trained the model for 200,000 steps using the Adam optimizer (learning rate  $8 \times 10^{-5}$ , batch size 64). To balance our two objectives, we applied a weight of  $\lambda = 10^{-5}$  to the discrete missingness loss relative to the continuous value loss, adopting the notation from TimeDiff [7]. For evaluation, we generated 16,000 synthetic windows (42 steps  $\times$  26 channels) by iteratively denoising random noise. This sample size matches the real dataset, allowing us to compare the synthetic distributions and observation patterns against the ground truth.

### 3 Results and Discussions

Figure 1, panels A–B compare real and synthetic univariate distributions for two representative labs with different shapes. Hemoglobin (A) is fairly concentrated with a mild right tail, and the synthetic density closely matches the real histogram around the main mode and across the upper tail. White Blood Cells count (B) is markedly more right-skewed with a long tail driven by occasional high values; despite this more challenging setting, the synthetic samples follow the real distribution well in the high-density region and preserve the overall tail profile. Figure 1 panel C compares the overall distribution of real data with the synthetic ones. The embedding was computed using UMAP with a DTW-based distance metric, for all the 26 channels (13 values and 13 masks). The substantial overlap between the two distributions indicates that the generative approach produces high-quality synthetic samples, without generating instances outside the support of the real data.

Table 1 summarizes per-feature realism across the 13 lab variables. It reports (i) missingness prevalence (fraction of time points with mask= 1) in real windows

**Table 1.** Per-feature comparison across 13 labs values: missingness prevalence (mask=1), univariate EMD on lab-value marginals, and DTW trajectory distance.

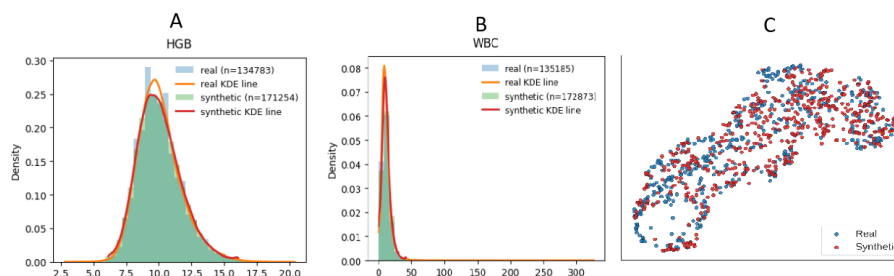
Lab	Missingness prevalence			EMD (univariate)			DTW (trajectory)		
	Real	Synth		R-R	R-S		R-R	R-S	
		w/ $\Delta$	w/o $\Delta$		w/ $\Delta$	w/o $\Delta$		w/ $\Delta$	w/o $\Delta$
Chloride	0.78	0.72	0.74	0.06	0.09	0.11	37.71	37.63	38.09
Potassium	0.78	0.72	0.74	0.03	0.06	0.09	38.44	38.27	38.48
Bicarb	0.78	0.72	0.74	0.07	0.14	0.14	37.90	37.80	38.13
Sodium	0.78	0.72	0.74	0.05	0.12	0.09	37.91	37.77	38.20
Hematocrit	0.80	0.75	0.77	0.06	0.06	0.03	38.34	38.83	38.82
Hemoglobin	0.81	0.76	0.78	0.02	0.11	0.07	38.42	38.97	39.00
MCV	0.81	0.76	0.78	0.03	0.16	0.20	37.73	38.00	38.12
Platelets	0.81	0.76	0.77	0.08	0.20	0.24	37.72	37.62	37.83
WBC count	0.81	0.76	0.77	0.08	0.07	0.08	37.88	38.01	38.15
RDW	0.81	0.76	0.78	0.04	0.09	0.13	37.53	37.53	37.88
BUN	0.78	0.72	0.74	0.08	0.10	0.15	36.41	36.70	37.10
Creatinine	0.78	0.72	0.74	0.02	0.06	0.08	35.99	35.87	36.40
Glucose	0.78	0.72	0.74	0.03	0.07	0.04	37.51	37.43	37.72
Mean	0.79	0.74	0.76	0.05	0.10	0.11	37.65	37.72	37.99

and in synthetic windows generated with and without delta conditioning, (ii) univariate Earth Mover’s Distance (EMD) [11] between lab-value marginals using Real–Real as a same-domain reference and Real–Synthetic for each generator, and (iii) Dynamic Time Warping (DTW) trajectory distances computed per feature on a random subsample of  $N = 100$  windows. Both EMD and DTW are computed in a missingness-aware manner by restricting comparisons to time points marked as observed according to the real or generated masks.

Overall, delta conditioning yields slightly better agreement with the real data: the mean EMD and DTW are lower than the corresponding no-conditioning setting and closer to the Real–Real reference. In contrast, the synthetic missingness prevalence is consistently below the observed prevalence, suggesting that the model still underestimates the frequency of unobserved measurements and that the missingness generation process could be improved. Finally, we note that these results should be interpreted as preliminary evidence that the model captures important aspects of the coupled physiology–sampling process, rather than as a complete validation of downstream clinical utility. In the extended work, we evaluate how the resulting synthetic data can be used to train Prior-Data Fitted Networks for imputation and forecasting under distribution shifts, while explicitly leveraging informative missingness.

## 4 Conclusions

This preliminary work shows how TimeDiff can be adapted to generate laboratory trajectories while jointly modeling informative missingness on a public MIMIC-III-derived benchmark. With practical preprocessing, the model produces synthetic windows that closely match real data in marginal lab distributions and in the joint structure of values and missingness masks. Our results suggest that diffusion models are a promising foundation for realistic synthetic lab generation under MNAR-like missingness, and they motivate future work that leverages the learned generative prior for downstream inference and forecasting in foundation models.



**Fig. 1. Real vs. synthetic comparisons:** (A) Hemoglobin and (B) White Blood Cells count univariate densities; (C) UMAP embedding of windows using all 26 channels (labs + masks), computed on a random subsample of 500 real and 500 synthetic windows.

## 5 Acknowledgment

Hadi Mehdizavareh and Arijit Khan acknowledge support from the Novo Nordisk Foundation grant NNF 22OC0072415.

## References

1. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS), pp. 6840–6851. Curran Associates, Inc., Red Hook, NY (2020)
2. Johnson, A.E.W., et al.: MIMIC-III, a freely accessible critical care database. *Scientific Data* **3**, 160035 (2016)
3. Kazijevs, M., Samad, M.D.: Deep imputation of missing values in time series health data: A review with benchmarking. *Journal of Biomedical Informatics* **144**, 104440 (2023)
4. Li, J., Yan, et al.: Imputation of missing values for electronic health record laboratory data. *npj Digital Medicine* **4**(1), 14 (2021)
5. Luo, Y.: Evaluating the state of the art in missing data imputation for clinical data. *Briefings in Bioinformatics* **23**(1), bbab489 (2022)
6. Tan, A.L.M., et al.: Informative missingness: What can we learn from patterns in missing laboratory data in the electronic health record? *Journal of Biomedical Informatics* **139**, 104306 (2023)
7. Tian, M., et al.: Reliable generation of privacy-preserving synthetic electronic health record time series via diffusion models. *JAMIA Open* **5**(4), ooac229 (2022)
8. Xiao, C., Choi, E., Sun, J.: Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association* **25**(10), 1419–1428 (2018)
9. Yang, Y., et al.: A survey on diffusion models for time series and spatio-temporal data. *ACM Computing Surveys* **58**(8), 196:1–196:39 (2026)
10. Cichosz, S.L., et al.: Biases in glucose metrics are directly related to low coverage of continuous glucose monitoring: insights from diverse populations. *Diabetes Technology & Therapeutics* **28**(2), 180–184 (2026).
11. Farahani, M.A., et al.: Time-series forecasting in smart manufacturing systems: An experimental evaluation of the state-of-the-art algorithms. *Robotics and Computer-Integrated Manufacturing* **95**, 103010 (2025).
12. Müller, S., et al.: Transformers can do Bayesian inference. In: International Conference on Learning Representations (ICLR) (2022).