REAL-EHR: BENCHMARKING OPEN-LOOP CLINICAL EVENT FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-loop EHR forecasting requires models to predict beyond the observed patient record without receiving future observations, yet most models are evaluated with teacher-forced next-event prediction. Teacher-forced performance may not reflect open-loop forecasting, and open-loop evaluation itself depends on how generated futures are scored. We introduce REAL-EHR, a benchmark that compares matched patients and targets while varying information access, scoring criteria, and rollout budget. We adapt 17 model architectures to a common forecasting interface. At a 32-rollout budget on MIMIC-IV, teacher-forced and estimated open-loop NLL rankings show little agreement ($\rho = 0.13$), with 62 of 136 pairwise model orders reversed. At the same budget, all 17 learned models have higher estimated marginal NLL than a frequency baseline, while 16 improve candidate coverage and window occurrence. These NLL comparisons are budget-sensitive: increasing the rollout budget to 512 moves nine of ten audited models below the frequency baseline by point estimate, while weak agreement with teacher forcing persists within this subset ($\rho = 0.13$). Model implementations and evaluation code are available at `https://anonymous.4open.science/r/REAL-EHR-E237`.

1 INTRODUCTION

Anticipating how a patient's condition will evolve is central to clinical care. EHR models can predict specific outcomes such as acute kidney injury, circulatory failure, and sepsis (Tomašev et al., 2019; Hyland et al., 2020; Adams et al., 2022). More recently, generative models such as Foresight and Delphi have moved beyond individual endpoints toward modeling entire sequences of future clinical events (Kraljevic et al., 2024; Shmatko et al., 2025), an ambition shared by biomedical world models (Noori et al., 2026). This shift raises a basic evaluation question: *if a model is intended to forecast open-loop futures, how should we evaluate those futures themselves?*

Existing evaluations capture important but incomplete aspects of this goal. For example, Delphi measures long-term disease-risk discrimination and tests whether generated trajectories reproduce population-level disease incidence (Shmatko et al., 2025). These evaluations establish useful properties of a forecasting model, but they do not directly assess how well it predicts the intervening patient-level events once future observations are withheld. This matters in clinical settings where intervention depends on when a patient's condition begins to deteriorate and what happens before that point. A forecasting model should therefore be evaluated by rolling it forward over the held-out future and comparing its predicted events with the recorded ones on individual patient level.

Under *teacher forcing*, the model predicts the next event and then receives the true event before making the next prediction. Under *open-loop forecasting*, the model sees the observed history only once and must continue without seeing any of the held-out future, using its own generated events or internal state (Figure 1a). The two settings therefore ask different questions. Teacher forcing measures prediction with repeated updates from the true record, while open-loop forecasting measures how well a model can forecast forward without receiving recorded future updates. A model can perform well in the first setting and poorly in the second. We therefore ask whether model rankings under teacher forcing are preserved under open-loop forecasting.

Open-loop forecasting also complicates probabilistic evaluation. A generative model predicts a distribution over possible futures, but each patient provides only one recorded continuation. There is

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

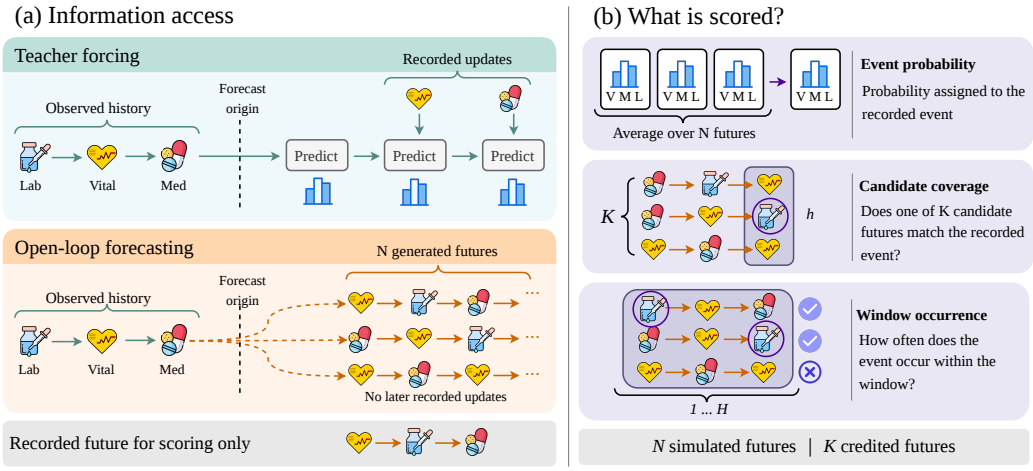

Figure 1: **Evaluating forecasts from the same observed history.** (a) Teacher forcing receives recorded future events between predictions, while open-loop forecasting does not. (b) Open-loop forecasts can be evaluated by event probability, candidate coverage, or window occurrence.

therefore no single obvious way to score the forecast. We can ask how much probability the model assigns to the recorded event, whether that event appears in a small set of sampled futures, or whether it occurs anywhere within a future window (Figure 1b). These measures capture different properties of the forecast and need not rank models in the same way. Even marginal event probability becomes difficult to estimate beyond the first step because it depends on the model-generated histories that lead there. In practice, we estimate it by averaging over sampled rollouts, so the estimate can change with the number of futures generated. We therefore vary both the scoring rule and rollout budget.

We introduce **REAL-EHR** to study these questions under a common experimental setup. All models are evaluated on the same patients, forecast origins, and targets, which allows us to compare teacher-forced and open-loop prediction without changing the prediction task. We adapt 17 model architectures to the same forecasting interface. For open-loop forecasts, we separately evaluate marginal event probability, candidate coverage, and window occurrence, and vary the number of sampled rollouts used for evaluation. We also test how forecasts change when the model receives more history, when observations are removed, and when events are reordered. Our primary study uses MIMIC-IV ICU patients, with additional experiments on eICU and MIMIC-III.

Open-loop forecasting gives a noticeably different view of the same models. On 512 MIMIC-IV patients, teacher-forced and finite-budget open-loop NLL rankings across 17 learned models have a Spearman correlation of only $\rho = 0.13$, with 62 of 136 orders reversed. At 32 rollouts, every learned model has higher estimated marginal NLL than an unconditional frequency baseline, while 16 of 17 improve candidate coverage and window occurrence. Increasing only the rollout budget to 512 changes the NLL comparison for nine of ten audited models.

Our main contributions are:

- A unified benchmark for teacher-forced and open-loop forecasting across 17 model architectures.
- A matched evaluation with rollout-budget audits for open-loop EHR forecasting.
- Analysis for longer, missing, and reordered records, with released model and evaluation code.

2 Related Benchmarks and Forecasting Models

EHR benchmarks and trajectory evaluation. Benchmarks on MIMIC-III, eICU, and HiRID, together with YAIB's multicenter framework, standardize clinical outcome prediction (Harutyunyan et al., 2019; Sheikhalishahi et al., 2020; Yèche et al., 2021; van de Water et al., 2024). EHRSHOT evaluates few-shot transfer, while BEDS-Bench and cross-representation benchmarking examine distribution shift and input representations (Wornow et al., 2023; Avati et al., 2021; Chen et al., 2025). CAREBench assesses the accuracy and temporal stability of adverse-event risk estimates, including

estimates derived from autoregressive generation (Keoliya et al., 2025). Closer to trajectory forecasting, EHRWorld compares next-step prediction with recursive simulation conditioned on supplied clinical inquiries and interventions (Mu et al., 2026), while EHR2Path recursively generates hourly patient states and evaluates events and numerical measurements over future windows (Pellegrini et al., 2026). REAL-EHR asks whether model comparisons remain consistent across observation-access regimes, scoring criteria, and rollout budgets while holding patients, initial histories, and targets fixed.

Architectures for longitudinal EHR modeling. Recurrent models such as Doctor AI and RETAIN summarize histories through recurrent computation (Choi et al., 2016a;b). Attention-based models (Vaswani et al., 2017), including BEHRT, Med-BERT, CEHR-BERT, and Hi-BEHRT, learn contextual representations for clinical prediction (Li et al., 2020; Rasmy et al., 2021; Pang et al., 2021; Li et al., 2023); CEHR-GPT, TransformEHR, ETHOS, Foresight, and Delphi extend generative modeling to future clinical events (Pang et al., 2024; Yang et al., 2023; Renc et al., 2024; Kraljevic et al., 2024; Shmatko et al., 2025). Complementary approaches address irregular dynamics through continuous-time models such as Latent ODE, IVP-VAE, and MIRA (Rubanova et al., 2019; Xiao et al., 2024; Li et al., 2025). Clinical latent-state models and recurrent state-space models represent evolving states (Xue et al., 2019; Hafner et al., 2019), while Mamba and Mamba-2 provide selective state-space sequence architectures (Gu & Dao, 2023; Dao & Gu, 2024). Linear attention, TTT, and LaCT offer recurrent attention computations or test-time memory updates (Katharopoulos et al., 2020; Sun et al., 2025; Zhang et al., 2026). To our knowledge, REAL-EHR is the first benchmark to systematically compare these architecture families under a common open-loop forecasting protocol.

Evaluating open-loop probabilistic forecasts. The distinction between observed and generated histories is familiar from sequence prediction (Bengio et al., 2015); in EHR forecasting, teacher forcing also supplies intervening clinical evidence unavailable at the forecast origin. Proper scoring rules assess predictive distributions (Gneiting & Raftery, 2007), including in probabilistic time-series and weather forecasting (Salinas et al., 2020; Rasul et al., 2021; Price et al., 2025). Best-of-K metrics in motion forecasting and pass@K in program synthesis instead assess whether a candidate set contains a close or successful prediction (Gupta et al., 2018; Salzmann et al., 2020; Kulal et al., 2019; Chen et al., 2021). Such scores can overlook poor candidates (Mohamed et al., 2022); in EHR forecasting, different sampled futures can cover different recorded events without any one future matching the complete trajectory. Distributional comparisons, including KL divergence and MAUVE, assess other aspects of generated outputs (Kullback & Leibler, 1951; Pillutla et al., 2021), but agreement after pooling patients does not establish patient-specific accuracy. These criteria answer different forecasting questions and need not support the same model comparison.

3 REAL-EHR: Forecasting from a Fixed Patient History

REAL-EHR evaluates each forecast from a fixed point in a patient's record. We choose a forecast origin, treat the preceding events as observed history, and reserve the following events for evaluation. Across comparisons, the patient and recorded future stay fixed. We change one aspect at a time: whether the model sees recorded events after forecasting begins, how its generated future is scored, or how much history it receives before the forecast origin.

3.1 Forecasting from a Fixed History

Let $X = (e_1, \ldots, e_c)$ denote the observed history and $y_1, \ldots, y_H$ the next H recorded events. Each event contains an event identity, a timestamp, and, when available, a numerical measurement. The primary scores here evaluate recorded event identity only.

We define forecast horizons by event position rather than elapsed time. EHR observations arrive irregularly. Several measurements may be recorded at the same time, while another part of the record may contain no event for a much longer interval. A fixed clock-time horizon would therefore contain very different numbers of prediction steps across patients. Using event position gives every model the same number of autoregressive steps, while the corresponding elapsed clinical time is reported separately in Appendix B.2.

Teacher forcing and open-loop forecasting differ only after the forecast origin. Under *teacher forcing*, the prediction of y_h conditions on the observed history X and the preceding recorded future

$y_1, \ldots, y_{h-1}$. Under *open-loop forecasting*, none of these future events is revealed. The model reaches position h through its own generated events or latent-state transitions and is still scored against the same recorded event y_h. The gap reflects both error accumulation and the extra clinical information supplied by teacher forcing.

3.2 Scoring Open-Loop Forecasts

A patient provides one realized future, but a generative model represents a distribution over many possible futures. A single sampled trajectory therefore gives an incomplete picture of the forecast. We use three complementary scores to ask how much probability the model gives the recorded future, whether a small set of generated futures covers it, and whether clinically relevant events occur somewhere within the forecast window.

Marginal event probability. At position h, different rollouts may reach the prediction through different generated histories. Let $p_h^{(n)}(v)$ be the next-event distribution after rollout n has reached that position. We first average over these histories and then score the recorded event:

$$\widehat{q}_h(v) = \frac{1}{N}\sum_{n=1}^{N} p_h^{(n)}(v), \qquad \widehat{\ell}_h = -\log \widehat{q}_h(y_h). \tag{1}$$

This asks how much probability the model assigns to y_h after accounting for uncertainty in the path leading to it. The order of these operations matters. Averaging the per-rollout NLLs would score a different quantity.

The marginal probability is not available exactly for the neural models, so we estimate it from N sampled rollouts. Although $\widehat{q}_h$ averages valid conditional probabilities, the logarithm makes finite-N NLL both noisy and upward biased in expectation. Because this bias differs across models, we treat N as part of the evaluation protocol and audit it directly.

Candidate coverage. Candidate coverage instead evaluates a small set of generated futures. A different question arises when a model is used to generate a small set of candidate futures. If m_h of the N sampled continuations contain the recorded token y_h at position h, we estimate the probability that at least one of K candidates matches it as

$$\widehat{C}_h(K) = 1 - \frac{\binom{N-m_h}{K}}{\binom{N}{K}}, \qquad 1 \le K \le N. \tag{2}$$

This is the finite-pool pass@K estimator of Chen et al. (2021). For example, if 8 of 32 sampled futures match the recorded event, then EventPass@5 is approximately 0.79. Here K describes how many candidate futures receive credit, whereas N describes how many rollouts were generated to estimate that coverage. We therefore keep K fixed when studying the effect of increasing N.

Occurrence within a future window. Exact event position can be too strict when the question is whether an event will occur soon rather than at a particular serialized position. For example, predicting a creatinine measurement three events early still captures its presence in the near future even though it is counted as a miss by a position-specific score. For event type u, let $a_u(H)$ indicate whether it appears anywhere in the recorded H-event future and let $\widehat{\pi}_u(H)$ be the fraction of generated futures in which it appears. We measure occurrence with the Brier score

$$\widehat{B}(H) = \frac{1}{|\mathcal{G}_{\text{occ}}|}\sum_{u \in \mathcal{G}_{\text{occ}}} \left(\widehat{\pi}_u(H) - a_u(H)\right)^2. \tag{3}$$

If a measurement occurs in the recorded future and appears in 24 of 32 generated futures, its predicted occurrence probability is 0.75 and its contribution to the score is $(0.75-1)^2$. The primary evaluation uses the 50 most frequent measurement concepts defined from the training data.

These three scores answer different questions about the same forecast. Marginal NLL evaluates probability at an exact position. EventPass@K evaluates coverage by a limited candidate set. Occurrence Brier ignores exact position and asks whether an event appears within the future window.

3.3 Stress-Testing the Observed History

Open-loop forecasts also depend on what the model knows before rollout begins. We therefore modify the observed history while keeping the forecast origin and recorded future fixed. The controls ask whether models can use additional history and whether their predictions depend on information that is missing or arbitrarily ordered.

Longer histories. More history provides additional patient information, but it also gives the model a longer sequence to process. We compare nested prefixes ending at the same forecast origin, so increasing the prefix adds older observations without changing the target. When the prefix exceeds lengths seen during training, the comparison also tests length extrapolation. A separate training control varies the history lengths seen during training so that test-time use of additional context is not confused with lack of training exposure.

Missing observations. Missing information may matter differently depending on where it occurs in the record. We compare two deletions from the same 16-event history while keeping the final observed event and next-event target fixed. The near-end condition removes the three events immediately preceding the final observation. The earlier condition removes three nonconsecutive events from an earlier part of the same history. To make the two deletions comparable, we require the removed events to have the same category composition in eICU, or the same observation source composition in MIMIC-III. In a stricter subset, we also match the clinical-time gap from the forecast origin to the latest retained measurement.

Within-time ordering. EHRs often contain several observations with the same recorded timestamp. Their serialized order can then be determined by preprocessing rather than by a clinically meaningful chronology, although a sequence model must still process one order. We permute complete event tuples within these timestamp ties and keep all other information fixed. We measure both the change in loss and whether the model's highest-probability next event changes, since a small change in average loss can still conceal substantial changes in individual predictions.

4 Datasets and Experimental Design

Datasets. Our primary experiments use adult ICU stays from MIMIC-IV v3.1, focusing on laboratory and vital-sign events (Johnson et al., 2023). Patients are split into disjoint training, validation, and test sets, and all data-dependent preprocessing is fitted on the training set. The primary test cohort contains 512 patients with sufficient history before the forecast origin and at least 64 future events. We use eICU and MIMIC-III as supporting datasets for history and sampling analyses (Pollard et al., 2018; Johnson et al., 2016). Dataset construction and preprocessing are described in Appendix A.2.

Models. We compare three statistical baselines with 17 learned models spanning clinical sequence models, Transformers, state-space models, fast-weight models, and continuous-time latent dynamics. All learned models are adapted to the same forecasting interface, trained from scratch at approximately 10M parameters, and evaluated with three random seeds. This setup controls data, model scale, and training budget rather than reproducing each model's original recipe. Appendix A.3 describes the individual adaptations.

Evaluation setup. The primary evaluation uses 512 MIMIC-IV test patients with at least 128 observed events before the forecast origin and 64 events afterward. All models begin from the same latest 16 observed events and are evaluated on the next 64 events. We fix the initial context at 16 events for the primary comparison and separately vary history length to study how additional context affects forecasting. Teacher-forced and open-loop evaluations use the same patients, forecast origins, and targets. Open-loop metrics are initially estimated from 32 sampled futures per patient. We additionally vary the rollout budget, remove observations, and reorder events in supporting experiments. Further evaluation details are provided in Appendix A.4.

Table 1: **Primary MIMIC-IV results** on 512 patients ($P = 16$, $H = 64$, $N = 32$). Learned models report mean ± SD over three seeds. Pass@5 is in % and Brier is scaled by 100. Bold and underline mark the best and second-best results.

	Teacher-forced	Open-loop		
Model	**NLL ↓**	**NLL ↓**	**Pass@5 ↑**	**Brier ↓**
Statistical references				
Frequency / Recency	3.200	**3.200**	27.7	21.84
Markov (sampled)	0.824	4.216	27.4	20.96
Learned models				
Transformer	**0.381**±0.010	4.252±0.048	34.5±0.4	**13.37**±0.09
CEHR-GPT	0.387±0.010	3.806±0.043	34.0±0.3	15.87±0.37
ETHOS	0.390±0.007	3.827±0.046	34.1±1.2	15.42±0.33
BEHRT	0.392±0.005	4.058±0.088	34.5±0.6	13.74±0.12
TransformEHR	0.394±0.005	3.813±0.063	34.5±0.3	16.63±0.50
LaCT	0.395±0.002	3.940±0.091	30.7±2.7	14.97±0.39
CEHR-BERT	0.397±0.003	3.805±0.056	34.0±1.0	15.99±0.26
Foresight	0.398±0.002	4.193±0.049	34.5±0.5	14.00±0.07
TTT-Linear	0.400±0.013	4.132±0.031	34.2±0.7	13.56±0.11
RSSM	0.401±0.003	3.565±0.016	31.3±0.5	16.04±0.40
Med-BERT	0.403±0.006	4.195±0.096	**34.7**±0.6	13.83±0.11
MIRA	0.404±0.002	4.185±0.049	34.3±0.6	14.13±0.11
Mamba-2	0.408±0.006	4.166±0.010	34.4±0.3	13.61±0.07
Hi-BEHRT	0.420±0.002	3.704±0.093	34.4±0.5	15.80±0.07
Delphi	0.467±0.006	3.904±0.051	34.0±0.8	14.12±0.04
Delphi+Value	0.468±0.012	3.938±0.016	34.0±0.4	14.06±0.18
Latent ODE	0.571±0.003	5.260±0.570	12.4±7.0	27.75±2.30

5 Model Comparisons Change in Open Loop

Table 1 already shows that the 17 learned models do not have a stable ordering across evaluation views. Transformer has the lowest teacher-forced NLL, but ranks sixteenth among the learned models in open loop. RSSM ranks first in open-loop NLL, Med-BERT has the highest EventPass@5, and Transformer has the lowest occurrence Brier. We first examine why teacher-forced rankings change, then ask why the open-loop scores themselves disagree.

5.1 Rankings Differ in Open Loop

Across the 17 learned models, teacher-forced and $N = 32$ open-loop event-64 NLL rankings have Spearman correlation $\rho = 0.13$ and Kendall correlation $\tau = 0.09$. Of the 136 model pairs, 62 reverse order. The pattern is similar across training seeds, with Spearman correlations from 0.09 to 0.15. Transformer moves from first under teacher forcing to sixteenth in open loop, while RSSM moves from tenth to first (Figure 2).

The two regimes differ in more than forecast length. Teacher forcing predicts y_{64} after observing the recorded events $y_{1:63}$. Open-loop forecasting reaches the same target through its own generated history. A model can therefore fit the local next-event conditionals well without maintaining an equally useful predictive distribution once recorded updates stop. Thus, the teacher-forced ordering does not carry over to the $N = 32$ open-loop estimate. Section 5.3 checks how this comparison changes with a larger rollout budget.

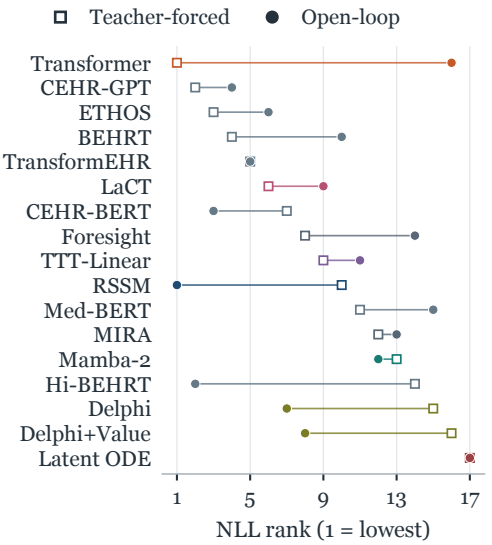

Figure 2: **Teacher-forced and open-loop ranks differ.**

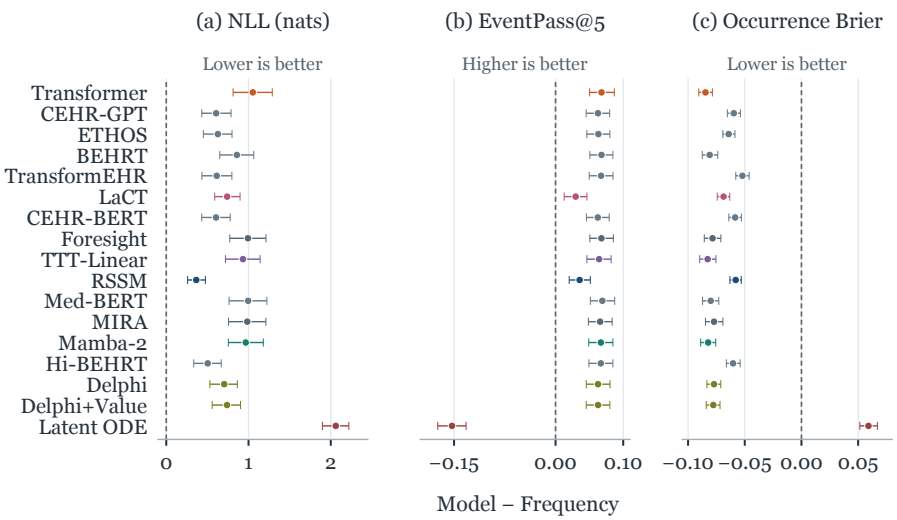

Figure 3: **Comparison with Frequency depends on the score.** Points show mean model–Frequency differences at $N = 32$ on the same 512 patients. Whiskers are 95% paired patient-bootstrap intervals.

5.2 Different Scores Favor Different Models

At $N = 32$, all 17 learned models have higher estimated marginal NLL than Frequency, while 16 of 17 improve both EventPass@5 and occurrence Brier (Figure 3). Latent ODE is the only exception.

Transformer makes the disagreement concrete. Its open-loop NLL is 4.252 nats, compared with 3.200 for Frequency, yet its EventPass@5 is higher (34.5% versus 27.7%) and its occurrence Brier is lower (13.37 versus 21.84 under the $\times 100$ scaling in Table 1). The pattern is not specific to Transformer: RSSM has the lowest open-loop NLL among the learned models (3.565), but its EventPass@5 is only 31.3% and its occurrence Brier is 16.04.

For a single target, EventPass@K is a monotone function of its marginal event probability. The ordering need not survive aggregation, however: EventPass applies a nonlinear transformation before averaging across patients, whereas NLL averages log probabilities (Appendix E). Occurrence Brier departs further because it scores whether a measurement appears anywhere in the forecast window rather than at an exact event position.

5.3 More Rollouts Change the NLL Comparison

Changing only the number of sampled futures is enough to overturn most of the initial model–Frequency comparisons. The patients, targets, and fitted checkpoints remain fixed. Among the ten audited models, nine move from above Frequency at $N = 32$ to below it by point estimate at $N = 512$ (Figure 4). Foresight changes from 4.193 to 3.073 nats, Transformer from 4.252 to 3.125, and RSSM from 3.565 to 3.173. Frequency stays at 3.200 because its marginal probability is available analytically. At $N = 512$, five of the nine negative model–Frequency differences have paired patient intervals excluding zero.

This shift comes from the estimator rather than from retraining. For a neural model, the open-loop marginal probability is approximated by averaging conditional probabilities over sampled histories and then taking a logarithm. Finite Monte Carlo variation therefore produces upward Jensen bias in expected NLL. Its size depends on the model's distribution over generated histories. The same value of N can consequently give very different estimation errors for different models.

The Markov model lets us measure this sampling error directly. Because it is a first-order model, its marginal distribution at any forecast step can be computed analytically by propagating its transition matrix, without sampling trajectories. We can therefore compare this analytic NLL with the NLL obtained by applying the same rollout estimator used for the neural models. Increasing N from 32 to 512 reduces this sampled-minus-analytic difference at event 64 from 0.404 to 0.026 nats on

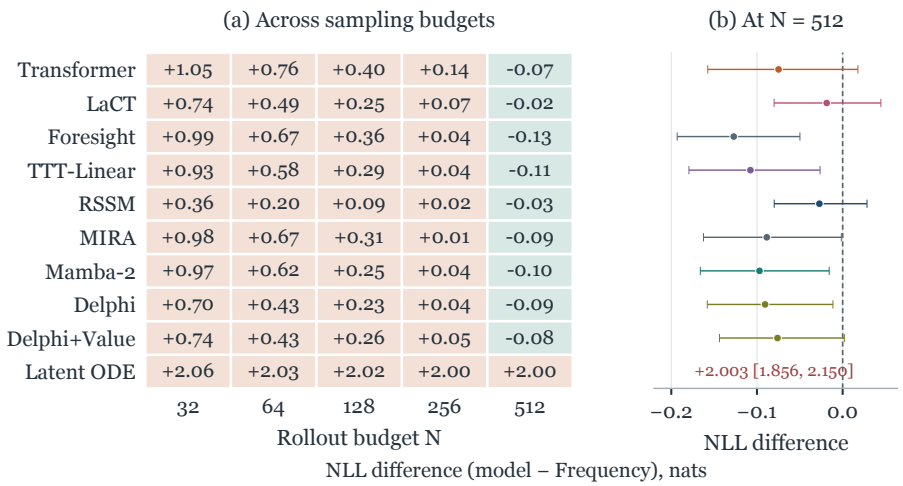

Figure 4: **More rollouts change the estimated NLL comparison.** The same ten models and 512 event-64 targets are compared with analytic Frequency. Panel (b) shows 95% paired patient-bootstrap intervals for the mean model–Frequency difference.

MIMIC-IV, from 0.741 to 0.073 on eICU, and from 1.122 to 0.135 on MIMIC-III (Figure 9). Even at $N = 512$, the sampled estimate remains noticeably above the analytic value on all three datasets.

However, it is worth noting that the weak teacher-forced versus open-loop agreement also persists at larger budgets. Within the ten audited models, Spearman correlation is -0.12 at $N = 32$ and 0.13 at $N = 512$. Estimation error remains substantial at $N = 32$ and smaller, though nonzero, at $N = 512$.

6 How Forecasts Depend on the Observed History

Section 5 kept the observed history fixed. In this section, we conduct analysis on the observed history. More context should provide more information, but it also gives the model a longer sequence to process. Missing or arbitrarily ordered observations pose a different problem because they can change the input without changing the forecast target. We study these effects while keeping the forecast origin and recorded future fixed. The in-range context comparison uses the primary MIMIC-IV cohort. The longer-context and perturbation controls use eICU and MIMIC-III test cohorts.

6.1 More History Does Not Guarantee Better Forecasts

Within the training range, additional history is useful. On primary MIMIC-IV, all 17 learned models improve one-step NLL when the observed history grows from 1 to 64 events, and 14 improve again from 64 to 128. These comparisons use the same target and stay within the 128-event training limit. The models therefore benefit from additional context when the sequence length is similar to training sequence length.

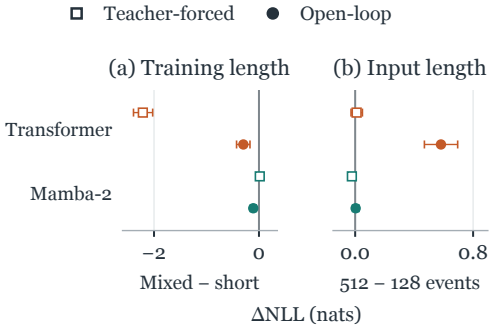

Figure 5: **Training on longer histories and adding more history are different interventions.** Results use 304 eICU event-64 targets with $N = 512$.

A model may fail to benefit from older observations simply because it was never trained to process sequences of that length. To separate this training effect from the value of additional history, we train Transformer and Mamba-2 on eICU using either 128-event histories or a mixture of 128, 256, and 512 events. Both settings use the same targets, three seeds, and 50,000 updates. The mixed-length runs process 2.33× more input events.

Training on longer histories helps when the test input is also long. At $P = 512$, mixed-length training lowers open-loop NLL by 0.298 nats for Transformer and 0.108 for Mamba-2 (Figure 5a). Mixed-length training helps at $P = 512$, but this does not imply that extra test-time history is useful. For the mixed-trained Transformer, increasing the test history from 128 to 512 events barely changes teacher-forced NLL ($+0.010$, 95% patient interval $[-0.028, 0.049]$), but open-loop NLL worsens by 0.582 nats ($[0.470, 0.695]$). Mamba-2 improves slightly under teacher forcing (-0.022), while its open-loop NLL changes little ($+0.003$, $[-0.018, 0.025]$) (Figure 5b).

6.2 Where Information Is Missing Matters

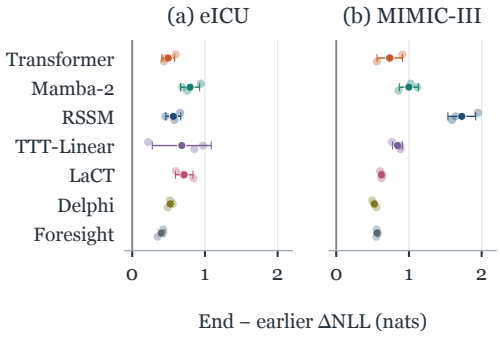

Figure 6: **Removing observations near the input end is more harmful.** Matched one-step controls on eICU and MIMIC-III across three training seeds.

A missing-history experiment can remove the same number of observations while changing very different parts of the record. We start from the same 16-event prefix and keep its final event fixed. The near-end condition deletes events 13–15, whereas the earlier condition deletes three nonconsecutive events from positions 1–12 while preserving the same next-event target. Both retain the final observed event and predict the same next event.

A simple comparison would confound position with the type and timing of the removed information. To avoid comparing different amounts or broad types of missing information, the earlier deletion is chosen to match the category counts of the three near-end events in eICU and their observation-source counts in MIMIC-III. We further restrict the comparison to cases where the two perturbed histories have the same native-time gap to the latest retained eligible measurement. The resulting cohorts contain 89 eICU targets and 345 MIMIC-III targets. In eICU, same-day observations are serialized by token ID, so proximity to the input end should not be interpreted as exact clinical recency. Across all seven evaluated models in both datasets, deleting near the input end causes a larger increase in one-step NLL than deleting earlier (Figure 6). The excess ranges from 0.400 to 0.796 nats on eICU and from 0.523 to 1.727 nats on MIMIC-III.

6.3 Small Loss Changes Can Hide Prediction Changes

The serialized history can also change even when no observation is removed. Several EHR events may share the same recorded timestamp, leaving their within-time order unresolved. Sequence models nevertheless consume one particular ordering. We permute complete observation tuples within these recorded-time ties while leaving the forecast target unchanged. The effect on average NLL can be small even when individual predictions move. For LaCT on eICU, mean one-step NLL changes by only 0.006 nats, yet the highest-probability event changes in 15.7% of cases on average. The full control shows the same distinction across models (Appendix D).

7 Conclusion

REAL-EHR shows that strong next-event prediction does not necessarily translate into strong open-loop forecasting once future observations are withheld. Across 17 models, teacher-forced rankings change substantially in the open-loop setting, while the comparison itself also depends on how forecasts are scored and how many futures are sampled. The history experiments reveal a related problem: forecasts can be sensitive to how much of the record is available and how that record is serialized. Together, these results shift the question from which architecture wins a benchmark to how open-loop clinical forecasts should be evaluated in the first place. REAL-EHR provides a testbed for this next step, including more sample-efficient likelihood estimation, clinically timed and value-aware forecasting, and robustness to incomplete or ambiguously ordered EHR histories.

Ethics Statement

The source datasets contain de-identified clinical records distributed under controlled-access agreements. Access and reuse remain subject to their original requirements. This manuscript reports aggregate results and contains no patient-level examples or identifiers. Forecasts reflect recorded care as well as patient condition and may reproduce documentation practices, treatment disparities, and selection effects. The experiments are retrospective; they do not establish clinical safety, causal effects, or benefit from using generated futures in patient care. Demographic subgroup fairness and prospective clinical utility have not been established.

Reproducibility Statement

Sections 3 and 4 specify the information boundary, sample budget, model panel, and uncertainty estimators. Appendix A provides implementation, preprocessing, model-adaptation, and cohort details. Appendices B–E provide the supporting analyses and metric interpretation. The evaluation artifacts include saved configurations, checkpoint and cohort identities, and scripts that regenerate aggregate results. Clinical records require authorized access and are not part of the manuscript source package.

Statement on the Use of AI

AI assistants, including Codex, were used to help design experimental prompts and analysis workflows, assist with implementation and debugging, summarize literature, inspect and interpret aggregate results, and draft and edit manuscript text, tables, and figures. AI assistance also supported bibliography checks and assembly of the manuscript. The authors remain responsible for the study, the accuracy of its claims and references, and the final submission.

References

Roy Adams, Katharine E. Henry, Anirudh Sridharan, Hossein Soleimani, Andong Zhan, Nishi Rawat, Lauren Johnson, David N. Hager, Sara E. Cosgrove, Andrew Markowski, Eili Y. Klein, Edward S. Chen, Mustapha O. Saheed, Maureen Henley, Sheila Miranda, Katrina Houston, Robert C. Linton, Anushree R. Ahluwalia, Albert W. Wu, and Suchi Saria. Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. *Nature Medicine*, 28(7):1455–1460, 2022. doi: 10.1038/s41591-022-01894-0. URL https://doi.org/10.1038/s41591-022-01894-0.

Anand Avati, Martin Seneviratne, Emily Xue, Zhen Xu, Balaji Lakshminarayanan, and Andrew M. Dai. BEDS-Bench: Behavior of EHR-models under Distributional Shift–A Benchmark, 2021. URL https://arxiv.org/abs/2107.08189.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/e995f98d56967d946471af29d7bf99f1-Abstract.html.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, et al. Evaluating Large Language Models Trained on Code, 2021. URL https://arxiv.org/abs/2107.03374.

Tianyi Chen, Mingcheng Zhu, Zhiyao Luo, and Tingting Zhu. Cross-Representation Benchmarking in Time-Series Electronic Health Records for Clinical Outcome Prediction, 2025. URL https://arxiv.org/abs/2510.09159.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In Finale Doshi-Velez, Jim Fackler, David Kale, Byron Wallace, and Jenna Wiens (eds.), *Proceedings of the 1st Machine*

Learning for Healthcare Conference, volume 56 of *Proceedings of Machine Learning Research*, pp. 301–318, Northeastern University, Boston, MA, USA, 18–19 Aug 2016a. PMLR. URL https://proceedings.mlr.press/v56/Choi16.html.

Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, volume 29, 2016b. URL https://arxiv.org/abs/1608.05745.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 10041–10071. PMLR, 2024. URL https://proceedings.mlr.press/v235/dao24a.html.

Tilmann Gneiting and Adrian E. Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://sites.stat.washington.edu/raftery/Research/PDF/Gneiting2007jasa.pdf.

Albert Gu and Tri Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces, 2023. URL https://arxiv.org/abs/2312.00752.

Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2255–2264, 2018. doi: 10.1109/CVPR.2018.00240. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Gupta_Social_GAN_Socially_CVPR_2018_paper.html.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2555–2565. PMLR, 2019. URL https://proceedings.mlr.press/v97/hafner19a.html.

Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6:96, 2019. doi: 10.1038/s41597-019-0103-9. URL https://www.nature.com/articles/s41597-019-0103-9.

Stephanie L. Hyland, Martin Faltys, Matthias Hüser, Xinrui Lyu, Thomas Gumbsch, Cristóbal Esteban, Christian Bock, Max Horn, Michael Moor, Bastian Rieck, Marc Zimmermann, Dean Bodenham, Karsten Borgwardt, Gunnar Rätsch, and Tobias M. Merz. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nature Medicine*, 26:364–373, 2020. doi: 10.1038/s41591-020-0789-4. URL https://doi.org/10.1038/s41591-020-0789-4.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016. doi: 10.1038/sdata.2016.35.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10:1, 2023. doi: 10.1038/s41597-022-01899-x. URL https://www.nature.com/articles/s41597-022-01899-x.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5156–5165. PMLR, 2020. URL https://proceedings.mlr.press/v119/katharopoulos20a.html.

Mayank Keoliya, Seewon Choi, Rajeev Alur, Mayur Naik, and Eric Wong. Stable Prediction of Adverse Events in Medical Time-Series Data, 2025. URL https://arxiv.org/abs/2510.14286.

Zeljko Kraljevic, Dan Bean, Anthony Shek, Rebecca Bendayan, Harry Hemingway, Joshua Au Yeung, Alexander Deng, Alfred Baston, Jack Ross, Esther Idowu, James T. Teo, and Richard J. B. Dobson. Foresight—a generative pretrained transformer for modelling of patient timelines using electronic health records: a retrospective modelling study. *The Lancet Digital Health*, 6(4):e281–e290, 2024. doi: 10.1016/S2589-7500(24)00025-6. URL https://doi.org/10.1016/S2589-7500(24)00025-6.

Sumith Kulal, Panupong Pasupat, Kartik Chandra, Mina Lee, Oded Padon, Alex Aiken, and Percy Liang. SPoC: Search-based Pseudocode to Code. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.neurips.cc/paper/2019/hash/7298332f04ac004a0ca44cc69ecf6f6b-Abstract.html.

Solomon Kullback and Richard A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. doi: 10.1214/aoms/1177729694. URL https://doi.org/10.1214/aoms/1177729694.

Hao Li, Bowen Deng, Chang Xu, Zhiyuan Feng, Viktor Schlegel, Yu-Hao Huang, Yizheng Sun, Jingyuan Sun, Kailai Yang, Yiyao Yu, and Jiang Bian. MIRA: Medical Time Series Foundation Model for Real-World Health Data. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3296. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/8e12ba543adc673da5b89c9311fcf72c-Abstract-Conference.html.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. BEHRT: Transformer for electronic health records. *Scientific Reports*, 10(1):7155, 2020. doi: 10.1038/s41598-020-62922-y. URL https://doi.org/10.1038/s41598-020-62922-y.

Yikuan Li, Mohammad Mamouei, Gholamreza Salimi-Khorshidi, Shishir Rao, Abdelaali Hassaine, Dexter Canoy, Thomas Lukasiewicz, and Kazem Rahimi. Hi-BEHRT: Hierarchical Transformer-Based Model for Accurate Prediction of Clinical Events Using Multimodal Longitudinal Electronic Health Records. *IEEE Journal of Biomedical and Health Informatics*, 27(2):1106–1117, 2023. doi: 10.1109/JBHI.2022.3224727. URL https://doi.org/10.1109/JBHI.2022.3224727.

Abduallah Mohamed, Deyao Zhu, Warren Vu, Mohamed Elhoseiny, and Christian Claudel. Social-Implicit: Rethinking Trajectory Prediction Evaluation and the Effectiveness of Implicit Maximum Likelihood Estimation. In *Computer Vision – ECCV 2022*, pp. 463–479. Springer, 2022. doi: 10.1007/978-3-031-20047-2_27. URL https://doi.org/10.1007/978-3-031-20047-2_27.

Linjie Mu, Zhongzhen Huang, Yannian Gu, Shengqian Qin, Shaoting Zhang, and Xiaofan Zhang. EHRWorld: A patient-centric medical world model for long-horizon clinical trajectories, 2026. URL https://arxiv.org/abs/2602.03569.

Ayush Noori, Nic Fishman, Ada Fang, Lukas Fesser, and Marinka Zitnik. World models for biomedicine. *Cell*, 189(19):5845–5869, 2026. doi: 10.1016/j.cell.2026.08.032. URL https://doi.org/10.1016/j.cell.2026.08.032.

Chao Pang, Xinzhuo Jiang, Krishna S. Kalluri, Matthew Spotnitz, RuiJun Chen, Adler Perotte, and Karthik Natarajan. CEHR-BERT: Incorporating temporal information from structured EHR data to improve prediction tasks. In *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pp. 239–260. PMLR, 2021. URL https://proceedings.mlr.press/v158/pang21a.html.

Chao Pang, Xinzhuo Jiang, Nishanth Parameshwar Pavinkurve, Krishna S. Kalluri, Elise L. Minto, Jason Patterson, Linying Zhang, George Hripcsak, Gamze Gürsoy, Noémie Elhadad, and Karthik Natarajan. CEHR-GPT: Generating electronic health records with chronological patient timelines, 2024. URL https://arxiv.org/abs/2402.04400.

Chantal Pellegrini, Ege Özsoy, David Bani-Harouni, Nassir Navab, and Matthias Keicher. EHR2Path: Comprehensive Pathway-Level Modeling of Longitudinal Patient Trajectories from Multimodal Electronic Health Records. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=ywa71iOykg.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. MAUVE: Measuring the Gap Between Neural Text and Human Text using Divergence Frontiers. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/260c2432a0eecc28ce03c10dadc078a4-Abstract.html.

Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. The eICU collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5:180178, 2018. doi: 10.1038/sdata.2018.178. URL https://www.nature.com/articles/sdata2018178.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R. Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam, and Matthew Willson. Probabilistic weather forecasting with machine learning. *Nature*, 637:84–90, 2025. doi: 10.1038/s41586-024-08252-9. URL https://doi.org/10.1038/s41586-024-08252-9.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine*, 4(1):86, 2021. doi: 10.1038/s41746-021-00455-y. URL https://doi.org/10.1038/s41746-021-00455-y.

Kashif Rasul, Calvin Seward, Ingmar Schuster, and Roland Vollgraf. Autoregressive Denoising Diffusion Models for Multivariate Probabilistic Time Series Forecasting. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8857–8868. PMLR, 2021. URL https://proceedings.mlr.press/v139/rasul21a.html.

Pawel Renc, Yugang Jia, Anthony E. Samir, Jaroslaw Was, Quanzheng Li, David W. Bates, and Arkadiusz Sitek. Zero shot health trajectory prediction using transformer. *npj Digital Medicine*, 7(1):256, 2024. doi: 10.1038/s41746-024-01235-0. URL https://doi.org/10.1038/s41746-024-01235-0.

Yulia Rubanova, Ricky T. Q. Chen, and David Duvenaud. Latent ODEs for irregularly-sampled time series. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://arxiv.org/abs/1907.03907.

David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. DeepAR: Probabilistic Forecasting with Autoregressive Recurrent Networks. *International Journal of Forecasting*, 36(3):1181–1191, 2020. doi: 10.1016/j.ijforecast.2019.07.001. URL https://doi.org/10.1016/j.ijforecast.2019.07.001.

Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-Feasible Trajectory Forecasting with Heterogeneous Data. In *Computer Vision – ECCV 2020*, pp. 683–700. Springer, 2020. doi: 10.1007/978-3-030-58523-5_40. URL https://doi.org/10.1007/978-3-030-58523-5_40.

Seyedmostafa Sheikhalishahi, Vevake Balaraman, and Venet Osmani. Benchmarking machine learning models on multi-centre eicu critical care dataset. *PLOS ONE*, 15(7):e0235424, 2020. doi: 10.1371/journal.pone.0235424. URL https://doi.org/10.1371/journal.pone.0235424.

Artem Shmatko, Alexander Wolfgang Jung, Kumar Gaurav, Søren Brunak, Laust Hvas Mortensen, Ewan Birney, Tom Fitzgerald, and Moritz Gerstung. Learning the natural history of human disease with generative transformers. *Nature*, 647:248–256, 2025. doi: 10.1038/s41586-025-09529-3. URL https://www.nature.com/articles/s41586-025-09529-3.

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, Tatsunori Hashimoto, and Carlos Guestrin. Learning to (Learn at test time): RNNs with expressive hidden states. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 57503–57522. PMLR, 2025. URL https://proceedings.mlr.press/v267/sun25h.html.

Nenad Tomašev, Xavier Glorot, Jack W. Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, Clemens Meyer, Suman Ravuri, Ivan Protsyuk, Alistair Connell, Cían O. Hughes, Alan Karthikesalingam, Julien Cornebise, Hugh Montgomery, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*, 572:116–119, 2019. doi: 10.1038/s41586-019-1390-1. URL https://doi.org/10.1038/s41586-019-1390-1.

Robin van de Water, Hendrik Schmidt, Paul Elbers, Patrick Thoral, Bert Arnrich, and Patrick Rockenschaub. Yet another ICU benchmark: A flexible multi-center framework for clinical ML. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ox2ATRM90I.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1706.03762.

Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason A. Fries, and Nigam H. Shah. EHRSHOT: An EHR Benchmark for Few-Shot Evaluation of Foundation Models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2933. URL https://papers.nips.cc/paper_files/paper/2023/hash/d42db1f74df54cb992b3956eb7f15a6f-Abstract-Datasets_and_Benchmarks.html.

Jingge Xiao, Leonie Basso, Wolfgang Nejdl, Niloy Ganguly, and Sandipan Sikdar. IVP-VAE: Modeling EHR time series with initial value problem solvers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16023–16031, 2024. doi: 10.1609/aaai.v38i14.29534. URL https://ojs.aaai.org/index.php/AAAI/article/view/29534.

Yuan Xue, Denny Zhou, Nan Du, Andrew Dai, Zhen Xu, Kun Zhang, and Claire Cui. Deep physiological state space model for clinical forecasting, 2019. URL https://arxiv.org/abs/1912.01762.

Zhichao Yang, Avijit Mitra, Weisong Liu, Dan Berlowitz, and Hong Yu. TransformEHR: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nature Communications*, 2023. doi: 10.1038/s41467-023-43715-z. URL https://doi.org/10.1038/s41467-023-43715-z.

Hugo Yèche, Rita Kuznetsova, Marc Zimmermann, Matthias Hüser, Xinrui Lyu, Martin Faltys, and Gunnar Rätsch. HiRID-ICU-Benchmark – A Comprehensive Machine Learning Benchmark on High-resolution ICU Data. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/5878a7ab84fb43402106c575658472fa-Abstract-round1.html.

Tianyuan Zhang, Sai Bi, Yicong Hong, Kai Zhang, Fujun Luan, Songlin Yang, Kalyan Sunkavalli, William T. Freeman, and Hao Tan. Test-time training done right. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/ffbfff81fb78e4bb558273b91e12e318-Abstract-Conference.html.

A Experimental and Evaluation Details

The main text defines the benchmark. Here we give the implementation details needed to reproduce its comparisons and interpret the supporting controls.

A.1 Evaluation implementation

The eligibility mask selects scored reference positions without deleting events from the sequence, shifting horizons, or renormalizing predictive probabilities. Excluded care actions therefore remain in the raw event sequence. Every forecast starts from fresh recurrent state, attention caches, or fast weights. Open-loop continuations are independent draws conditional on the same checkpoint and observed prefix. RSSM additionally samples the posterior state over the prefix and then advances only through prior transitions.

For continuation n, let $p^{(n)}_{i,w,h}(v)$ be the conditional token probability before step h. Marginal NLL uses float64 log-mean-exp,

$$\widehat{q}_{i,w,h}(v) = \frac{1}{N}\sum_{n=1}^{N} p^{(n)}_{i,w,h}(v), \qquad \widehat{\ell}_{i,w,h} = -\log \widehat{q}_{i,w,h}(y_{i,w,h}). \tag{4}$$

This Rao–Blackwellized estimate averages conditional probabilities, not sampled-token match frequencies. Finite N leaves $\widehat{q}$ unbiased for the sampler marginal but introduces upward Jensen bias after the logarithm. Equal rollout budgets can therefore produce different NLL error across models. EventPass@K uses the finite-pool pass@K estimator in the main text and keeps K fixed when N changes.

Window occurrence uses complete raw future windows, including care actions. The primary panel scores 50 training-defined measurement concepts. Cohort agreement maps the primary vocabulary to laboratory, vital-sign, and other categories without smoothing or post-mask renormalization. Patients receive equal weight. Window scores are first averaged within patient and then across patients. Paired 95% intervals use 2,000 patient-bootstrap resamples with all windows from a patient kept together. They condition on fitted checkpoints and cached rollout pools. Seed SDs and independent-pool checks report other sources of variation separately.

A.2 Data preprocessing and partitions

Patient partitioning precedes vocabulary construction and normalization in the three ICU representations below, and all stays from a patient remain in one partition. Forecast cohorts are selected only afterward.

Table 2: Available patient-disjoint ICU partitions after preprocessing. MIMIC-IV denotes the adult-ICU laboratory and vital-sign stream. eICU denotes train-fitted preprocessing. Vocabulary sizes include special tokens. Counts precede forecast-anchor eligibility and cohort sampling.

Dataset	Split	Patients	ICU stays	Recorded events	Vocabulary
MIMIC-IV	Train	43,731	61,486	43,055,225	467
	Validation	5,481	7,863	5,499,094	
	Test	5,484	7,810	5,450,512	
	Total	**54,696**	**77,159**	**54,004,831**	
MIMIC-III	Train	26,152	35,179	40,380,791	2,871
	Validation	3,289	4,474	5,389,340	
	Test	3,245	4,459	4,967,386	
	Total	**32,686**	**44,112**	**50,737,517**	
eICU	Train	111,434	160,731	21,819,356	3,174
	Validation	13,929	19,897	2,697,304	
	Test	13,930	20,134	2,726,780	
	Total	**139,293**	**200,762**	**27,243,440**	

Table 3: Completed learned adaptations in the primary MIMIC-IV adult-ICU panel. Counts include embeddings and prediction heads. All runs finish 50,000 updates. The last column gives the validation-selected update for seeds 41, 42, and 43.

Model	Parameters	Selected updates (41 / 42 / 43)
Delphi	9,973,600	47,750 / 50,000 / 50,000
Delphi+Value	10,094,401	49,250 / 50,000 / 49,000
Foresight	9,880,886	50,000 / 49,000 / 48,750
BEHRT	10,014,302	49,000 / 48,250 / 48,750
Med-BERT	10,013,582	50,000 / 49,500 / 49,750
CEHR-BERT	10,017,902	50,000 / 48,250 / 49,750
CEHR-GPT	10,105,382	50,000 / 49,500 / 48,500
ETHOS	9,884,006	50,000 / 49,000 / 49,750
Hi-BEHRT	10,017,902	50,000 / 49,000 / 49,500
TransformEHR	10,117,558	50,000 / 49,500 / 49,500
Transformer	10,184,706	50,000 / 48,500 / 48,500
TTT-Linear	10,063,962	34,250 / 31,000 / 28,250
LaCT	10,029,490	49,500 / 46,750 / 46,250
Mamba-2	10,006,770	19,500 / 16,250 / 19,000
RSSM	9,634,933	50,000 / 49,000 / 49,250
Latent ODE	10,003,486	46,000 / 48,250 / 49,250
MIRA	10,363,874	49,000 / 48,250 / 48,750

MIMIC-IV adult ICU. The primary stream contains timed laboratory and vital-sign measurements within ICU stays (Johnson et al., 2023). Measurement concepts are defined by source item and normalized unit. Numerical values use the training-set mean and sample SD with an explicit validity mask. Event time is $\max(\texttt{charttime}, \texttt{storetime})$ in fractional hours from ICU admission. Events are ordered by availability time, source rank, native source identifier, and source row. Untimed billing diagnoses are excluded, and training windows never cross ICU stays.

eICU and MIMIC-III. The train-fitted eICU representation (Pollard et al., 2018) retains selected laboratory abnormalities, out-of-range vital signs, and care events. Vocabulary, thresholds, and normalization are fit on training patients. Its model clock is day-rounded and same-day events are ordered by token ID. The MIMIC-III representation (Johnson et al., 2016) retains selected finite measurements and timed care events, uses integer elapsed ICU days for the model, and preserves higher-resolution source times for matching history perturbations.

Historical hospital-wide MIMIC-IV. The cross-dataset sampling audit uses a separate patient-level MIMIC-IV stream with day-rounded timestamps. Its validation patients were previously inspected, and the original extraction revision remains unverified. We therefore use it only to study estimator behavior within that representation. It is not an independent replication of the primary adult-ICU panel.

A.3 Model adaptations and optimization

The 17 learned systems are controlled adaptations trained from scratch on the common primary stream. They do not reproduce every method's original pretraining regime. An IVP-VAE adaptation (Xiao et al., 2024) did not complete the required three-seed primary training and is excluded from the 17-model ranking.

Delphi retains its event-rate objective and waiting-time sampler (Shmatko et al., 2025). Delphi+Value adds value embeddings and a token-conditioned scalar head. Foresight, BEHRT, Med-BERT, CEHR-BERT, CEHR-GPT, ETHOS, Hi-BEHRT, and TransformEHR are local causal forecasting adaptations of their published architectures (Kraljevic et al., 2024; Li et al., 2020; Rasmy et al., 2021; Pang et al., 2021; 2024; Renc et al., 2024; Li et al., 2023; Yang et al., 2023). The general Transformer and Mamba-2 feed generated events back during rollout (Vaswani et al., 2017; Dao & Gu, 2024). RSSM assimilates the factual prefix with a posterior and then rolls forward through its prior (Hafner et al., 2019). TTT-Linear and LaCT retain fast-weight/test-time-update mechanisms (Sun et al., 2025; Zhang et al., 2026). Latent ODE and MIRA preserve continuous-time latent dynamics without access to a future observation grid (Rubanova et al., 2019; Li et al., 2025). Auxiliary time/value objectives

remain model-specific, so the comparison is controlled at the data, scale, update, and evaluation levels rather than isolating architecture alone.

All completed primary runs use seeds 41–43, 50,000 updates, effective batch size 32, histories of at most 128 events, AdamW with peak learning rate 10^{-4}, 2,000 warmup updates, cosine decay to 10^{-5}, and gradient clipping at norm one. Checkpoints minimize observation-token NLL on fixed validation batches evaluated every 250 updates. Runs use approximately 10M parameters and BF16 training with FP32 parameter masters. Forecasts aggregate probabilities in FP64.

A.4 Comparison cohorts and statistical references

The primary cohort contains 512 test patients sampled from 4,747 eligible patients, one fixed anchor per patient, and at least 128 preceding and 64 subsequent events. A $P = 16$ forecast receives only the final 16 preceding events. Open loop uses $N = 32$, temperature one, and the common nonpadding support. Teacher forcing and open loop share patients, anchors, targets, and checkpoints.

Frequency uses training-event frequencies with pseudocount 10^{-3}. Markov counts adjacent within-stay transitions and backs off to Frequency with total mass eight in the primary representation. Recency mixes the latest token with Frequency. Validation selects mixing weight zero on the adult-ICU stream. Exact Frequency and Markov marginals are used only as estimator checks for their fitted statistical models.

Table 4: Map from result figures to cohorts and evaluation budgets. Counts are scored patients unless noted.

Figure	Data and status	Patients	P, H	N	Displayed uncertainty
2	MIMIC-IV adult ICU, test	512	$16, 64$	32	Descriptive ranks
3	Same primary cohort	512	$16, 64$	32	95% patient CI
4	Same cohort, ten models	512	$16, 64$	32–512	95% patient CI at 512
5	eICU, train-fitted, exploratory	304	$128/512, 64$	512	95% patient CI
6	eICU / MIMIC-III, exploratory	89 / 345	$16, 1$	—	Training-seed SD
7	Same primary MIMIC-IV cohort	512	16, 1–64	32	95% patient CI
8a	Primary continuation event counts	512	$16, 64$	—	Descriptive proportions
8b–c	Separate MIMIC-IV duration audit	5,243 anchors	128, 1–128	—	Duration IQR
9	Hospital-wide MIMIC-IV / eICU / MIMIC-III, inspected	1,385 / 1,553 / 1,742	$16, 64$	32–512	Realized error, no interval

The expanded sampling audit uses nested cohorts up to 2,048 patients per representation and $N \in \{32, 64, 128, 256, 512\}$. These supporting cohorts reuse frozen native checkpoints and were previously inspected. They test estimator sensitivity rather than replicate the parameter-matched primary ranking.

B Supporting Evidence for Evaluation Sensitivity

B.1 Forecast horizon profiles

The event-64 comparison is not an isolated endpoint. Figure 7 follows marginal NLL and EventPass@5 across earlier positions using the same $N = 32$ pools. Their trends need not coincide because probability and candidate coverage reward different properties of the forecast.

B.2 Event depth in clinical time

The primary 512-patient continuations contain 32,768 future positions. Their ten most frequent tokens account for 75.2% of events and the top 50 for 97.1%. A separate audit selected 5,243 event-128 anchors without requiring future follow-up. Among them, 4,617 (88.1%) have a 64th subsequent event after a median 6.77 hours with IQR 4.85–8.68. A total of 4,031 (76.9%) have a 128th event

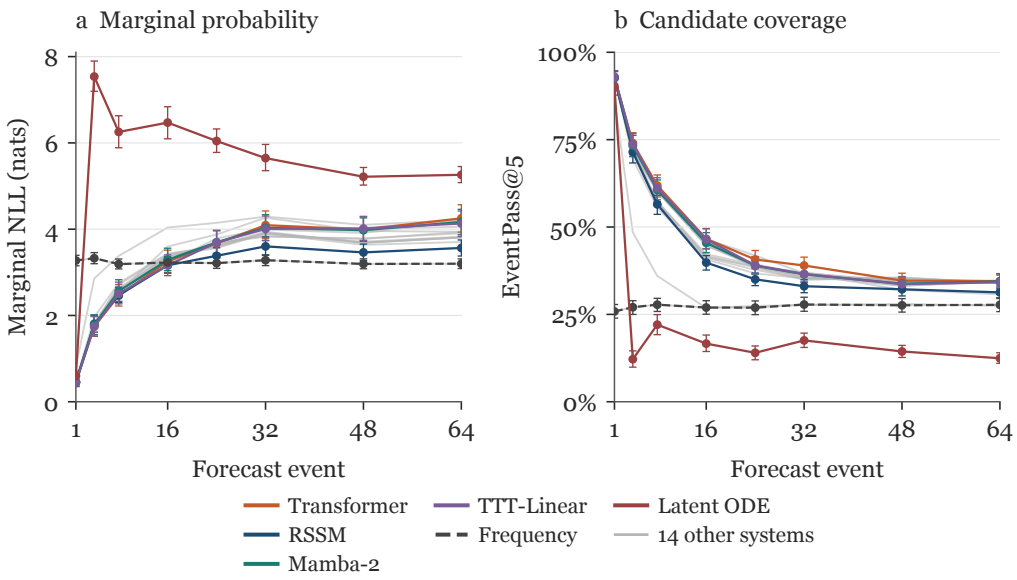

Figure 7: **Probability and candidate coverage across forecast positions.** Primary MIMIC-IV adult-ICU panel, 512 patients, $P = 16$, $H = 64$, and $N = 32$. EventPass@5 evaluates five candidates drawn from each common pool. Colored curves identify six systems, including the Latent ODE outlier. Each gray curve is one of the remaining 14 systems. Neural curves average separately scored seeds 41–43. Error bars give pointwise 95% patient-bootstrap intervals for the highlighted systems, averaging each patient's scores over the fixed seeds before resampling. Frequency uses one fit. Intervals condition on checkpoints and cached pools and omit Monte Carlo variability. All horizons use the same 64-event-eligible patients.

after a median 13.72 hours. Event count therefore corresponds to variable clinical time rather than a fixed-duration endpoint.

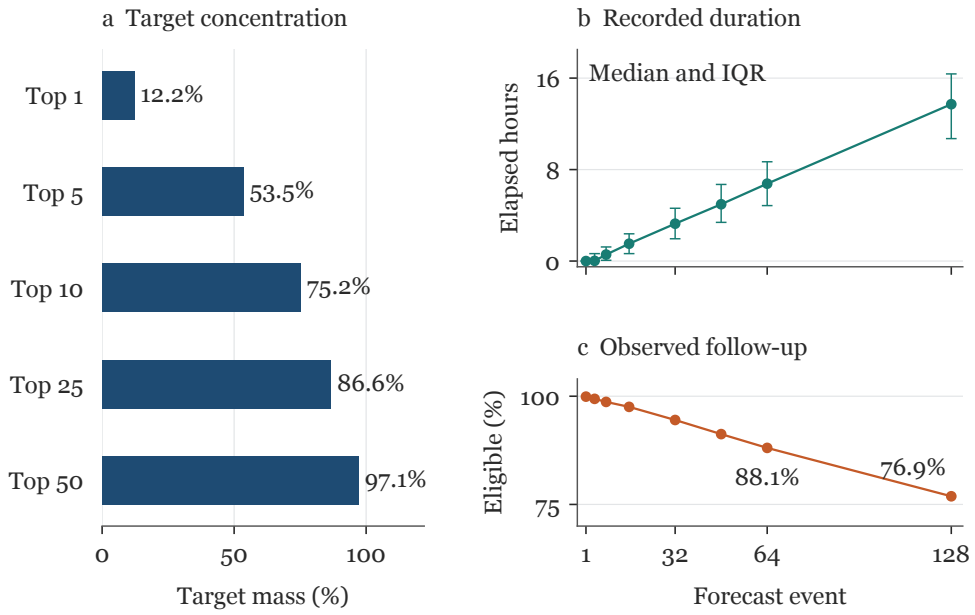

Figure 8: **Event horizons depend on recording concentration and availability.** (a) Cumulative token mass over all 32,768 positions in the primary 512-patient continuations. (b) Median elapsed hours and interquartile range among anchors with the specified future event. (c) The fraction with that event recorded among 5,243 audit anchors selected without requiring future follow-up. Panels (b–c) use a separate event-128 anchor cohort and the native $\max(\texttt{charttime}, \texttt{storetime})$ clock. At event 64, 4,617 anchors remain eligible (88.1%), with median elapsed time 6.77 hours. The IQR describes variation across eligible anchors, not uncertainty in the median.

C Numerical Results for Section 5

C.1 Matched sampling-budget audit

Ten models completed the $N = 32 \rightarrow 512$ audit on the original 512 patients. They are Transformer, RSSM, Mamba-2, Delphi, Delphi+Value, Foresight, TTT-Linear, LaCT, MIRA, and Latent ODE. The two budgets use the same checkpoints, targets, and nested rollout pools. The other seven learned systems are outside this audit.

Within this ten-model subset, teacher-forced and open-loop ranks remain weakly associated at both budgets. Spearman ρ is -0.115 at $N = 32$ and 0.127 at $N = 512$. Kendall τ is -0.111 and 0.156, respectively. Increasing the rollout budget changes likelihood estimates substantially without restoring agreement with teacher forcing.

C.2 Exact Markov estimator check

The cross-dataset audit fixes each fitted Markov model and target set, then compares sampled and analytic event-64 marginals. It scores 1,385, 1,553, and 1,742 eligible targets in the historical hospital-wide MIMIC-IV, train-fitted eICU, and MIMIC-III representations. Sampling error falls with N in all three datasets but remains above the prespecified 0.02-nat reporting tolerance at $N = 512$. This tolerance is not a clinical threshold.

Table 5: **Rollout-budget audit on the same 512 patients and ten fitted models.** Entries are three-seed mean NLL ± SD. The last column gives model minus analytic Frequency at $N = 512$ with its paired 95% patient interval, conditional on fitted models and rollout pools.

Model	$N = 32$	$N = 512$	At $N = 512$ model − Frequency Mean [95% patient interval]
Transformer	4.252 ± 0.048	3.125 ± 0.012	-0.075 [-0.157, $+0.018$]
LaCT	3.940 ± 0.091	3.181 ± 0.017	-0.019 [-0.080, $+0.045$]
Foresight	4.193 ± 0.049	3.073 ± 0.024	-0.127 [-0.193, -0.050]
TTT-Linear	4.132 ± 0.031	3.092 ± 0.005	-0.108 [-0.179, -0.026]
RSSM	3.565 ± 0.016	3.173 ± 0.032	-0.027 [-0.080, $+0.028$]
MIRA	4.185 ± 0.049	3.111 ± 0.009	-0.089 [-0.162, -0.002]
Mamba-2	4.166 ± 0.010	3.103 ± 0.037	-0.097 [-0.166, -0.016]
Delphi	3.904 ± 0.051	3.109 ± 0.038	-0.091 [-0.158, -0.011]
Delphi+Value	3.938 ± 0.016	3.124 ± 0.031	-0.076 [-0.144, $+0.002$]
Latent ODE	5.260 ± 0.570	5.203 ± 0.522	$+2.003$ [$+1.856$, $+2.150$]

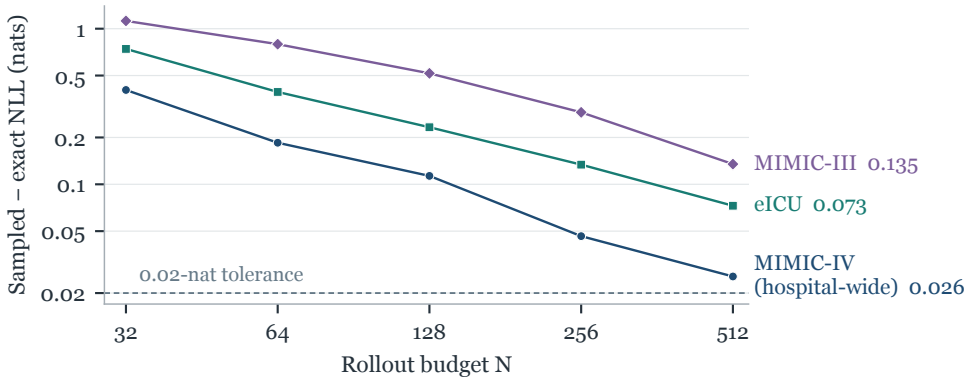

Figure 9: Sampled-minus-exact Markov event-64 NLL as the rollout budget increases. At $N = 512$, all three realized errors remain above the prespecified 0.02-nat tolerance.

C.3 Additional primary-panel metrics

Table 6: **Larger candidate sets and cohort agreement are distinct from likelihood.** Same primary panel as Table 1, with $N = K = 32$. EventPass@32 scores the event-64 token. Cohort KL compares the event-64 laboratory, vital-sign, and OTHER mixture. Learned rows are three-seed means $\pm$ SD. Cohort agreement does not establish patient-specific forecast accuracy. These summaries are descriptive.

Model	EventPass@32 ↑	Cohort KL $\times 10^3$ ↓
Frequency / Recency	0.727	1.713
Markov (sampled)	0.719	1.639
Transformer	0.754 ± 0.011	4.896 ± 1.932
CEHR-GPT	0.734 ± 0.008	14.704 ± 5.029
ETHOS	0.743 ± 0.006	17.502 ± 10.611
BEHRT	0.770 ± 0.011	0.419 ± 0.386
TransformEHR	0.733 ± 0.011	23.368 ± 3.412
LaCT	0.750 ± 0.013	18.989 ± 16.911
CEHR-BERT	0.738 ± 0.013	10.763 ± 2.432
Foresight	0.766 ± 0.005	0.234 ± 0.234
TTT-Linear	0.760 ± 0.009	0.388 ± 0.241
RSSM	0.755 ± 0.015	2.263 ± 1.358
Med-BERT	0.762 ± 0.012	0.583 ± 0.399
MIRA	0.760 ± 0.006	0.402 ± 0.324
Mamba-2	0.759 ± 0.006	1.358 ± 0.596
Hi-BEHRT	0.740 ± 0.003	16.175 ± 8.086
Delphi	0.764 ± 0.005	1.346 ± 1.335
Delphi+Value	0.758 ± 0.009	1.535 ± 0.667
Latent ODE	0.337 ± 0.024	425.794 ± 368.339

Models with similar EventPass@32 can have very different cohort KL. For example, Foresight and Delphi have nearly identical candidate coverage, while their cohort KL differs by more than fivefold. Candidate coverage and aggregate distributional agreement therefore describe different properties of the same forecast pool.

D Additional Evidence for the History Controls

These exploratory controls supplement Section 6. They use previously inspected test cohorts and keep targets fixed within each comparison. Absolute scores are dataset-specific and should not be pooled across cohorts.

D.1 Context length and training exposure

The train-fitted eICU control has 383 fixed anchors, 304 with eligible event-64 targets. Transformer and Mamba-2 are trained either on 128-event inputs or an equal-cycle mixture of 128, 256, and 512 events. Both arms supervise the same final 128 targets from common 512-event windows and complete 50,000 updates, but mixed training processes $2.33\times$ as many input events. Open-loop evaluation uses $N = 512$. Teacher forcing uses the final P factual events before the target. Open loop starts P events before the original forecast origin, so a $P = 512$ rollout has processed 575 events before event 64.

Mixed-length training improves open-loop NLL at a fixed 512-event input for both models, but adding older observations to the mixed-trained Transformer still worsens its open-loop forecast. Training on long inputs can therefore improve operation at that length without making the additional history itself useful.

Table 7: Absolute event-64 NLL on 304 scored train-fitted eICU targets. Entries are means ± sample SD across seeds 41–43. Open-loop estimates use $N = 512$. Factual conditioning windows and open-loop prefix lengths have the different endpoints specified above.

		Teacher forced		Open loop	
Model	Training	$P = 128$	$P = 512$	$P = 128$	$P = 512$
Transformer	Short	1.344 ± 0.055	3.577 ± 0.176	5.104 ± 0.198	5.127 ± 0.122
Transformer	Mixed	1.355 ± 0.037	1.364 ± 0.037	4.247 ± 0.015	4.829 ± 0.012
Mamba-2	Short	1.626 ± 0.026	1.594 ± 0.023	4.371 ± 0.040	4.458 ± 0.129
Mamba-2	Mixed	1.632 ± 0.028	1.609 ± 0.028	4.346 ± 0.014	4.349 ± 0.031

D.2 Matched deletion

The one-step deletion control starts from 16-event prefixes and protects the final event. The near-end arm removes the preceding three events. The earlier arm removes three nonconsecutive positions from the earlier prefix. eICU matches deleted category counts, MIMIC-III matches observation-source counts, and the stricter subset also matches the native-time gap from the forecast origin to the latest retained eligible measurement. This leaves 89 eICU and 345 MIMIC-III scored targets. Exact token identities and numerical values are not matched, so the contrast does not isolate position alone.

Table 8: Absolute next-event NLL on the native-gap-matched subsets. Entries are mean ± sample SD across three training seeds. These SDs are not patient confidence intervals.

Model	Intact	Earlier	Near end
eICU ($n = 89$)			
Transformer	1.814 ± 0.053	2.265 ± 0.087	2.761 ± 0.175
Mamba-2	2.253 ± 0.128	2.417 ± 0.121	3.213 ± 0.027
RSSM	2.608 ± 0.134	2.531 ± 0.138	3.097 ± 0.063
TTT-Linear	1.950 ± 0.242	2.379 ± 0.411	3.061 ± 0.037
LaCT	2.010 ± 0.067	2.179 ± 0.090	2.894 ± 0.103
Delphi	1.841 ± 0.005	2.146 ± 0.026	2.668 ± 0.025
Foresight	2.058 ± 0.061	2.225 ± 0.085	2.625 ± 0.075
MIMIC-III ($n = 345$)			
Transformer	0.778 ± 0.011	0.933 ± 0.041	1.669 ± 0.134
Mamba-2	0.759 ± 0.011	0.924 ± 0.045	1.922 ± 0.093
RSSM	0.791 ± 0.009	1.027 ± 0.023	2.754 ± 0.206
TTT-Linear	0.787 ± 0.021	0.934 ± 0.014	1.776 ± 0.054
LaCT	0.777 ± 0.013	0.926 ± 0.015	1.546 ± 0.008
Delphi	0.781 ± 0.006	0.928 ± 0.015	1.451 ± 0.018
Foresight	0.767 ± 0.008	0.896 ± 0.002	1.457 ± 0.010

Near-end deletion is more harmful than earlier deletion for every evaluated three-seed mean in both datasets. Because exact tokens and values are not matched, the result shows that removal rate alone is insufficient to characterize an incomplete history. It does not identify a causal effect of event position.

D.3 Ordering within exact recorded-time ties

Complete observation tuples are permuted only within adjacent groups sharing the retained recorded-time fields. Actions and the final anchor stay fixed. Four permutations are averaged within each patient. eICU contributes 362 scored patients and MIMIC-III 474. Equality of recorded time identifies ambiguity in serialization, not physiological simultaneity.

Mean loss and prediction identity can tell different stories. LaCT changes mean NLL by only 0.006 nats on eICU, yet its top prediction changes for 15.7% of scored patients on average. Small aggregate loss changes therefore do not imply invariant patient-level predictions.

Table 9: Exact recorded-time permutation, one-step prediction from 16-event prefixes. Entries are three-seed means ± sample SD. Loss changes are permuted minus intact. Flip rates are percentages. These exploratory test cohorts differ from the historical validation panels reported elsewhere.

Dataset	Model	ΔNLL (nats)	Top-1 flip (%)
eICU ($n = 362$)			
	Transformer	$+0.115 \pm 0.035$	12.8 ± 2.3
	Mamba-2	$+0.050 \pm 0.163$	42.3 ± 3.0
	RSSM	$+0.243 \pm 0.066$	25.7 ± 2.3
	TTT-Linear	$+0.111 \pm 0.054$	25.0 ± 1.2
	LaCT	$+0.006 \pm 0.028$	15.7 ± 2.1
MIMIC-III ($n = 474$)			
	Transformer	$+0.147 \pm 0.013$	10.9 ± 0.1
	Mamba-2	$+0.366 \pm 0.025$	21.0 ± 1.5
	RSSM	$+0.776 \pm 0.030$	29.5 ± 0.9
	TTT-Linear	$+0.163 \pm 0.015$	11.8 ± 1.2
	LaCT	$+0.140 \pm 0.010$	10.8 ± 1.1
	Delphi	$+0.140 \pm 0.003$	10.0 ± 1.0
	Foresight	$+0.129 \pm 0.011$	10.5 ± 1.1

D.4 In-range factual context

The primary one-step control predicts a fixed target from nested factual histories. All 17 learned systems improve from $P = 1$ to 64 events, and 14 improve again from 64 to 128. Most of the gain occurs before 64 events. The additional 64 events produce much smaller changes, suggesting diminishing returns within the training range. These are seed-mean scores rather than paired significance tests.

Table 10: Primary MIMIC-IV one-step NLL at a fixed target ($n = 512$). Entries are means across three training seeds. All context lengths are within the 128-event training limit.

Model	$P = 1$	$P = 64$	$P = 128$
Transformer	0.811	0.393	0.383
RSSM	0.829	0.402	0.392
Mamba-2	0.815	0.399	0.387
TTT-Linear	0.809	0.397	0.379
LaCT	0.813	0.388	0.376
Latent ODE	0.882	0.600	0.602
MIRA	0.833	0.408	0.397
Delphi	0.841	0.461	0.462
Delphi+Value	0.842	0.464	0.466
Foresight	0.825	0.399	0.383
BEHRT	0.829	0.394	0.388
Med-BERT	0.827	0.403	0.390
CEHR-BERT	0.826	0.398	0.390
CEHR-GPT	0.814	0.394	0.386
ETHOS	0.825	0.400	0.386
Hi-BEHRT	0.829	0.415	0.408
TransformEHR	0.824	0.406	0.394

E Why the Forecast Criteria Differ

The three forecast criteria answer different questions. The examples below are established score properties, not new scoring rules.

Marginal probability. For true scored-token distribution r_h and exact model probability q_h, expected excess log loss is

$$\mathbb{E}_{Y_h \sim r_h}[-\log q_h(Y_h)] - \mathbb{E}_{Y_h \sim r_h}[-\log r_h(Y_h)] = D_{\mathrm{KL}}(r_h \| q_h) \geq 0. \tag{5}$$

Thus exact NLL is minimized by the correct target probability (Gneiting & Raftery, 2007). This statement applies to the selected target population induced by the eligibility rule. It does not establish calibration of an unfiltered event stream or the full trajectory distribution.

Candidate coverage. For a recorded token with exact marginal probability $q_h(y_h)$, K independent futures give coverage $C_h(K) = 1 - (1 - q_h(y_h))^K$. At one target this is a monotone transform of the same marginal probability, but averaging the nonlinear utility across targets can order models differently from average NLL. For example, under true token probabilities $(0.8, 0.1, 0.1)$, forecasting $(0.6, 0.2, 0.2)$ improves expected EventPass@5 relative to the calibrated forecast $(0.8, 0.1, 0.1)$ while worsening expected NLL. Coverage therefore measures usefulness of a finite candidate set, not probability calibration.

Window occurrence. Position marginals do not determine whether an event occurs anywhere in a window. An equal mixture of (A, A) and (B, B) and an equal mixture of (A, B) and (B, A) both assign probability $1/2$ to each token at each position. The first predicts occurrence probability $1/2$ for each token, whereas the second predicts one. Window Brier can therefore detect dependence differences missed by position-specific NLL and coverage. Even their combination does not certify complete-trajectory accuracy.