PULSE: Probabilistic Uncertainty-Aware Longitudinal Simulation for EHR Trajectories

Robert Lennart Manschke[1] **Angus Roberts**[1] **Julia Ive**[2]
[1]King's College London, United Kingdom [2]University College London, United Kingdom

Abstract

Realistic intensive-care electronic health record trajectories are difficult to simulate because physiology, treatments, observation times, and patient state co-evolve under sparse clinical measurement. We introduce PULSE, a probabilistic simulator family for intensive care trajectories. PULSE separates treatment from patient-state generation, predicts binary patient state before continuous physiology, and represents uncertainty with patient-dependent output distributions. Covariance groups assign variables to shared decoders, and a monotone B-spline flow models each continuous distribution. On a MIMIC-IV sepsis cohort scored over 37 variables, the B-spline model achieves vector-normalized root mean squared error of 6.556 and 5.805 when forecasting from hours two and six, compared with 6.845 and 6.277 for our G-Transformer reimplementation. Its continuous ranked probability score is 0.4625 and 0.3930, compared with 0.5191 and 0.4725 for that baseline. Paired patient-bootstrap intervals support improvements in both scores at both origins. A continuous normalizing flow achieves similar scores in the same evaluation. Cross-dataset experiments examine transfer to eICU. A controlled benchmark based on CVSim [Heldt et al., 2010] shows that a model can learn treatment responses under known synthetic dynamics.

1 Introduction

Counterfactual prediction from observational electronic health records (EHRs) can be formulated as recursive simulation under a specified treatment policy. In the g-computation framework [Robins, 1986, Hernán and Robins, 2020], a model estimates conditional covariate and outcome distributions from observed histories and then simulates forward under either the observed treatment process or a hypothetical intervention. Intensive-care EHR measurements are, however, collected at variable clinical frequencies, not fixed scientific sampling intervals [Hripcsak et al., 2015, Che et al., 2018]; treatment assignment is confounded by patient state, discharge and death terminate the observed ICU process, and short-horizon physiology contains strong cross-variable dependence. Thus, applying g-computation to ICU trajectory simulation requires more than an expressive sequence encoder: the simulating model must specify how time, observation, treatment, binary state, continuous state, and uncertainty are recursively generated.

G-Transformer [Xiong et al., 2024] established transformer-based g-computation for dynamic ICU treatment regimes, extending earlier recurrent and transformer counterfactual models [Li et al., 2021, Melnychuk et al., 2022]. Its key modeling step is to separate treatment, binary patient-state, and continuous covariate processes during recursive simulation; this separation substantially improves over a plain causal transformer in the reported MIMIC predictive check. The continuous simulator in that model uses predicted conditional means plus empirical residual noise for Monte Carlo rollout. Such residual sampling is simple and empirically useful, but it cannot express patient-state-dependent heteroskedasticity, skew, or local residual shape. It also leaves factual rollout performance difficult to interpret: an all-hours error can arise from continuous dynamics, treatment rollout, binary-state rollout, or a mistimed release/death event.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

We introduce PULSE, a probabilistic event-time simulator for ICU trajectories. The MIMIC feature source used here is hourly binned, but PULSE encodes time in a continuous covariate space: absolute time since ICU admission, elapsed time, log-scaled elapsed time, and sinusoidal clock features are combined with variable-level observation masks and time-since-measurement features. Unobserved variable–hour pairs are represented as absence of a clinical observation in the binned source table, not as newly measured values requiring dense imputation. The model predicts binary/categorical patient state before continuous state at each generated step and separates treatment assignment from patient-state prediction. Some variants use gated anchored residual dynamics around the last observed or simulated value. Probabilistic variants use heteroskedastic Gaussian, monotone B-spline, or continuous-flow residual distributions; covariance-coupled variants further add latent severity pooling and correlated residual sampling within data-driven variable groups.

Contributions First, PULSE is an autoregressive ICU state simulator: treatment, binary patient state, and continuous physiology are generated in rollout order. Unlike previous simulators, event-time covariates represent irregular clinical measurement and gated residual updates are anchored to each patient's carried-forward state. Second, whereas previous work adds static residual noise after prediction, PULSE makes uncertainty conditional on the generated patient state: Gaussian heads learn heteroskedastic scale, B-spline residual flows learn skew and local density shape, and the PULSE-CovSpline variant couples residual samples within data-driven physiologic groups. Third, we evaluate these modeling choices under one 37-variable ICU predictive-check definition, spanning MIMIC-IV factual rollout, simple persistence and population baselines, XGBoost and transformer baselines, absorbing-state error decompositions, calibration and coverage, eICU transfer panels, and CVSim known-effect treatment-response tests. Code will be released at `https://github.com/manscrober/pulse-artifact`.

2 Background and related work

G-computation for longitudinal treatment regimes. Robins' g-formula identifies causal contrasts under time-varying treatment when consistency, exchangeability, positivity, and correct measurement assumptions hold [Robins, 1986, Hernán and Robins, 2020]. Let L_t denote time-varying patient state, A_t treatment, and S static covariates. Following the longitudinal g-formula factorization [Robins, 1986, Hernán and Robins, 2020], a sequential model factorizes the observed process as

$$p_\theta(l_{1:T}, a_{1:T} \mid S) = \prod_{t=0}^{T-1} p_\theta(l_{t+1} \mid \bar{l}_t, \bar{a}_t, S) p_\theta(a_{t+1} \mid \bar{l}_t, \bar{a}_t, l_{t+1}, S), \tag{1}$$

where bars denote history. Under a treatment rule g, the observational treatment term is replaced by $A_{t+1} = g(\bar{L}_t, \bar{A}_t, S)$ and the learned state transition is iterated forward. Causal Transformer introduced a transformer architecture for counterfactual forecasting with balanced representations and cross-attention over time-varying treatments [Melnychuk et al., 2022]. G-Net [Li et al., 2021] and G-Transformer [Xiong et al., 2024] follow a sequential g-computation factorization; G-Transformer extends the transformer setting by explicitly modeling treatment and covariate evolution under the g-formula.

Irregularly observed EHR time series. EHR trajectories are recorded when clinicians measure or intervene, so the absence of a measurement can carry clinical information [Hripcsak et al., 2015, Che et al., 2018]. Fixed temporal resampling can also distort clinical decision processes [Frost et al., 2026]. GRU-D uses masks and time intervals to exploit this structure in clinical time series [Che et al., 2018], and Neural CDEs model irregular observations as continuous-time paths [Kidger et al., 2020]. In PULSE, each variable value is paired with its observation mask and time since measurement; clock time and elapsed time are also encoded continuously, even though the available MIMIC source tables are hourly binned.

Probabilistic residual distributions. Normalizing flows model a complex density by transforming a simple base density through an invertible map [Papamakarios et al., 2021]. Neural spline flows use monotone splines as expressive scalar transforms with tractable sampling and likelihood evaluation [Durkan et al., 2019]; B-splines are a classical smooth spline basis [de Boor, 2001]. In EHR rollout, the role of a scalar flow head is narrower than in image or tabular density estimation: it learns a

patient-state-dependent residual distribution around a recursive mean trajectory. Heteroscedastic Gaussian regression is a calibrated likelihood baseline, but likelihood training can degrade mean estimation or variance calibration without suitable training constraints [Stirn et al., 2023, Immer et al., 2023]. We evaluate Gaussian heads as the baseline conditional residual distribution and B-spline residual flows as the more expressive family.

Data sources. MIMIC-IV is a de-identified critical-care EHR database from Beth Israel Deaconess Medical Center [Johnson et al., 2023b, 2024, Goldberger et al., 2000, Pollard et al., 2026]. We use MIMIC-IV v3.1 and construct a Sepsis-3 ICU cohort following G-Transformer, which itself uses the clinical consensus definition of sepsis as life-threatening organ dysfunction caused by dysregulated host response to infection [Singer et al., 2016, Xiong et al., 2024]. eICU is a multicenter ICU database used for transfer and shared-panel evaluation [Pollard et al., 2018, 2019]. CVSim is a mechanistic cardiovascular simulator that supplies paired known-effect trajectories under treatment perturbations [Heldt et al., 2010]; these paired trajectories let us test whether a simulator can recover treatment-response direction and magnitude when the counterfactual future is known.

3 PULSE

3.1 Autoregressive ICU state factorization

PULSE simulates a patient trajectory by separating patient state from treatment assignment. At hour or event index t, the generated history contains continuous physiology $X_{\leq t}$, binary/categorical patient state $B_{\leq t}$, treatments $A_{\leq t}$, static covariates S, and, in event-time variants, clock features $\tau_{\leq t}$. Consistent with sequential g-computation and the binary-before-continuous G-Transformer rollout convention [Robins, 1986, Hernán and Robins, 2020, Xiong et al., 2024], the next-step transition is factorized as

$$\tilde{B}_{t+1} \sim p_\phi(B_{t+1} \mid X_{\leq t}, B_{\leq t}, A_{\leq t}, S, \tau_{\leq t}), \tag{2}$$

$$\tilde{X}_{t+1} \sim p_\theta(X_{t+1} \mid X_{\leq t}, B_{\leq t}, A_{\leq t}, \tilde{B}_{t+1}, S, \tau_{\leq t}), \tag{3}$$

$$\tilde{A}_{t+1} \sim p_\psi(A_{t+1} \mid X_{\leq t}, B_{\leq t}, A_{\leq t}, S, \tau_{\leq t}). \tag{4}$$

Thus treatments observed up to t inform the binary encoder, the sampled or teacher-forced binary state at $t+1$ informs the continuous encoder for the same generated step, and the treatment-policy model supplies the action used by later steps. During training, recorded binary states condition continuous prediction. Scheduled sampling [Bengio et al., 2015] replaces selected recorded inputs with model outputs; the main B-spline fit uses recorded histories. Death and ICU release are binary patient-state variables with absorbing post-processing; treatment variables are action variables and are not folded into the continuous or binary state heads. Under this indexing, the action recorded at time t conditions the transition to (B_{t+1}, X_{t+1}); a generated or externally imposed $\tilde{A}_{t+1}$ is appended to the state and conditions the following transition. This follows G-Transformer.

Following G-Net by Li et al. [2021] and G-Transformer Xiong et al. [2024], PULSE separates treatment and covariate processes, but changes how the patient state transition is parameterized. Encoders may be shared, assigned per covariate, or grouped according to clinically meaningful variable sets. We refer to the grouped gated-residual mean model as PULSE-Anchor, the B-spline residual-flow model as PULSE-Spline, and the covariance-coupled B-spline model with latent severity and group residual noise as PULSE-CovSpline. The per-covariate G-Transformer configuration is retained as a baseline because it follows the published factorization. The main B-spline model initializes its backbone, location and scale from a pretrained model and trains on recorded histories by negative log-likelihood; validation loss selects the checkpoint. Initialization from a trained two-Gaussian mixture also yields similar performance on a validation subset. The main model uses covariance-grouped decoders and predicts location directly; Figure 1 illustrates its rollout. The G-Transformer baseline uses one encoder per continuous variable. Our grouped model shares an encoder among variables with correlated training-set changes, then predicts separate distribution parameters for each variable. Sampling from these heads supplies a new state for the next step.

3.2 Event-time inputs

We use the MIMIC-IV v3.1 clinical event tables; after cohort construction the available time series are therefore effectively hourly binned. Event-time PULSE variants do not add new measurements;

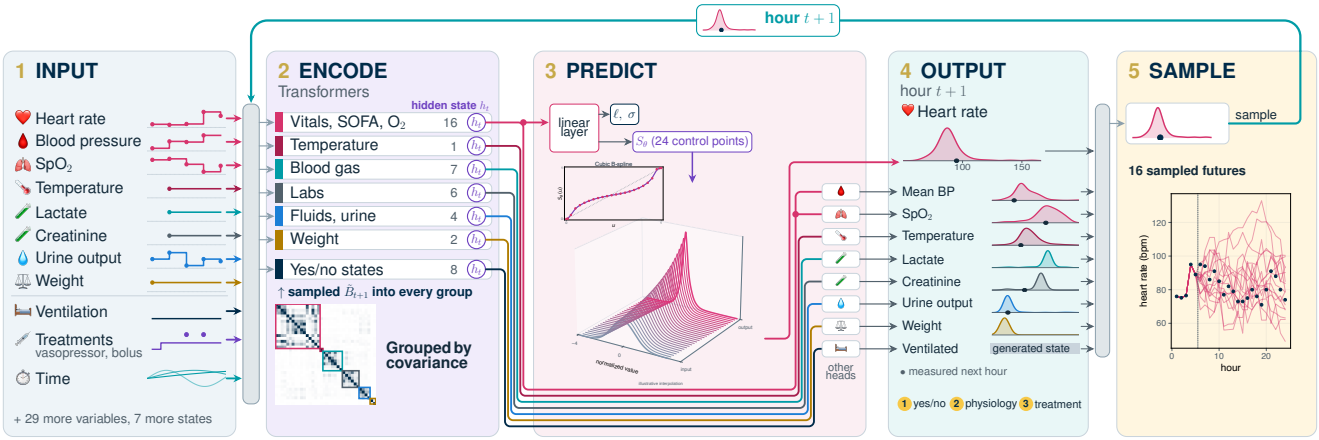


Figure 1: PULSE-Spline rollout. Covariance-grouped encoders predict binary state and continuous distributions, which are sampled recursively. The central spline and its input and output densities use the first heart-rate forecast for a patient in the public MIMIC-IV demo database, conditional on a sampled binary state; intermediate densities illustrate the transition. Other distributions and trajectories show 100 paths generated for the same patient [Johnson et al., 2023a], with 16 paths displayed. Dots mark available measurements; input lines carry recorded values forward for display.

they still operate on this hourly-binned source, but represent the binned sequence in a continuous time coordinate instead of relying only on the integer step index. This removes the need to impute or interpolate dense hourly measurements and leaves waveform-derived vital signs as a natural higher-granularity extension. For each variable v, let $m_{t,v} \in \{0, 1\}$ indicate whether a measurement was made at time t. Following clinical time-series models that expose masks and elapsed time to the network [Che et al., 2018, Kidger et al., 2020] and recent concerns about fixed temporal resampling in clinical decision data [Frost et al., 2026], the value channel is multiplied by $m_{t,v}$ after normalization, and the model also receives $m_{t,v}$ and

$$a_{t,v} = \log(1 + t - \tau_{t,v}), \tag{5}$$

where $\tau_{t,v}$ is the most recent measurement time for variable v. This representation distinguishes a newly observed value from an older carried value and allows the model to learn from clinician measurement cadence without interpreting every hourly slot as a new observation. Time features include hours since ICU admission, elapsed time since the previous event, signed log-scaled time, and sinusoidal positional encodings evaluated on clock time instead of the discrete step index. Auxiliary Δt heads predict the time to the next observation in ablations; event-time results are reported as separate temporal-processing variants.

3.3 Gated anchored residual dynamics

A direct-level decoder predicts $x_{t+1,v}$ from the hidden state without explicitly conditioning the output on the most recent value of variable v. During recursive rollout this can regress toward a population average, especially for slow-moving or cumulative ICU variables. An anchored PULSE variant predicts changes around the last observed or simulated value:

$$\mu_{t+1,v} = x_{t,v}^{\text{anchor}} + g_{\theta,v}(h_t)\Delta_{\theta,v}(h_t), \qquad g_{\theta,v}(h_t) = \sigma(r_{\theta,v}(h_t)). \tag{6}$$

Here $x_{t,v}^{\text{anchor}}$ is the variable-specific carry-forward state, $\Delta_{\theta,v}$ is a learned residual, and $g_{\theta,v} \in [0, 1]$ controls how far the model departs from persistence. Setting $g = 0$ recovers carry-forward; setting $g = 1$ recovers an unconstrained residual model. The learned gate makes the persistence prior explicit and measures whether the model learns nonzero dynamics beyond copying the last value.

3.4 Latent severity and residual grouping

PULSE-CovSpline adds shared latent acuity and group-level residual noise to the continuous residual process; Table 1 defines the added components. All grouping statistics are fit on the training split only.

Table 1: PULSE-CovSpline components.

Component	Definition and role
Latent severity	$z_t = q_\omega(h_t)$ shared across continuous heads; correlations with future release, death, ventilation, and vasopressor events are in Table A.17.
Covariance groups	Train-only clusters $\mathcal{G}_1, \ldots, \mathcal{G}_J$ from correlations of $x_{t+1} - x_t$.
Group residual noise	$\epsilon_{t,\mathcal{G}_j} = L_j u_{t,j} + d_j \odot \eta_{t,j}$, with $u_{t,j} \sim \mathcal{N}(0, I_r)$ and $\eta_{t,j} \sim \mathcal{N}(0, I_{\lvert\mathcal{G}_j\rvert})$, giving shared group noise plus variable-specific noise.
Residual-grouping check	Alternative clusters from $x_{t+1} - \hat{\mu}_t$ after fitting a mean model; reported as a sensitivity analysis.

3.5 Probabilistic continuous heads

For Gaussian heads, PULSE predicts $\mu_{\theta,v}(h_t)$ and $\sigma_{\theta,v}(h_t)$ and minimizes the standard heteroscedastic Gaussian negative log-likelihood used in neural uncertainty regression [Kendall and Gal, 2017, Stirn et al., 2023, Immer et al., 2023],

$$\mathcal{L}_{\mathrm{NLL}} = \frac{1}{2}\left[\frac{(x_{t+1,v} - \mu_{\theta,v})^2}{\sigma^2_{\theta,v}} + 2\log\sigma_{\theta,v} + \log(2\pi)\right]. \tag{7}$$

For spline heads, an invertible monotone transform maps a Gaussian residual to a conditional residual density. The likelihood uses the normalizing-flow change-of-variables rule of Papamakarios et al. [2021]; Durkan et al. [2019] use rational-quadratic spline flows, while our cubic B-spline basis follows de Boor [2001]. With $Z \sim \mathcal{N}(0,1)$ and a monotone B-spline transform $f_{\phi,v}(\cdot; h_t)$,

$$Y_{t+1,v} = \mu_{\theta,v}(h_t) + \sigma_{\theta,v}(h_t) f_{\phi,v}(Z; h_t), \tag{8}$$

and likelihood is computed by the normalizing-flow change-of-variables rule [Papamakarios et al., 2021, Durkan et al., 2019],

$$\log p_\theta(y \mid h_t) = \log p_Z(f^{-1}_{\phi,v}(r)) + \log\left|\frac{\partial f^{-1}_{\phi,v}(r)}{\partial r}\right| - \log\sigma_{\theta,v}, \qquad r = \frac{y - \mu_{\theta,v}}{\sigma_{\theta,v}}. \tag{9}$$

Earlier B-spline experiments use a staged Gaussian-plus-spline head. Mean dynamics are first trained with MSE or Huber loss [Huber, 1964] around the anchored residual mean; Gaussian scale is then fitted with the mean fixed, and the near-identity B-spline adapter is trained on residual likelihood. Stochastic rollout samples Monte Carlo trajectories from this distribution. Conditional-mean rollout gives the most accurate point estimate for several B-spline-family rows because it removes sampling variance before computing RMSE, but it does not represent the simulator's trajectory distribution.

4 Experiments

4.1 MIMIC predictive check

We evaluate on a Sepsis-3 ICU cohort derived from MIMIC-IV [Johnson et al., 2023b, 2024, Singer et al., 2016]. The cohort, first-stay restriction, subject-level split, and 24-hour predictive check are chosen to match the G-Transformer experiments as closely as possible under MIMIC-IV v3.1. For horizon k, the model observes hours before k and simulates hours k through 24; $k = 2$ conditions on hour 1, and $k = 6$ conditions on hours 1–5. The MIMIC-IV v3.1 source contains 94,458 ICU stays and 65,366 unique subjects; the available sepsis cohort contains 12,369 eligible stays from 10,845 subjects. The main split contains 8,934 subjects, with 7,147 training, 893 validation, and 894 test stays. Version and cohort details are reported in Appendix A.1.

The scored panel contains 37 continuous or treated-continuous variables, grouped into vital signs, MAP-derived variables, SOFA and oxygen-requirement changes, laboratory and arterial blood gas values, weight, maintenance fluids, urine output, bolus volume, cumulative fluids, and ICU fluid balance. Binary patient-state variables include edema, pulmonary edema, diuretics, dialysis, mechanical ventilation, death, and ICU release; vasopressor use is treated as a binary action variable. Static variables include age, gender, pre-ICU fluid amount, Elixhauser score, end-stage renal failure, and congestive heart failure. The full variable list and units are given in Table A.2.

4.2 Models and rollout protocols

The main baselines are train population mean, carry-forward, and G-Transformer [Xiong et al., 2024]. Earlier comparisons also include XGBoost autoregression [Chen and Guestrin, 2016] and Causal Transformer [Melnychuk et al., 2022]. XGBoost is a factorized autoregressive tabular baseline: binary next states are predicted from forward-filled current values, observation masks, current binary states, static covariates, and clock features, and continuous next values are predicted after appending the predicted next binary states; the best retained XGBoost row remains far below the sequence baselines. Causal Transformer uses a single causal sequence encoder with MSE continuous heads and empirical validation residual noise for stochastic rollout. The G-Transformer baseline uses the per-covariate factorization from Xiong et al. [2024]. PULSE variants test grouped versus per-covariate encoders, direct versus anchored residual dynamics, event-time encodings with Δt heads, Gaussian NLL, mixture-of-Gaussians (MoG) residual adapters, B-spline residual flows, and a scalar continuous normalizing flow (CNF) residual head.

In the default observational-rollout protocol, future continuous variables, binary states, and death/release are generated by the simulator, with treatments supplied by a shared learned policy. Some error-decomposition protocols replace selected generated variables with factual future values. For example, the true-death/release protocol supplies factual absorbing-state timing to measure how much all-hours RMSE is driven by release or death. These rows use future information and should be interpreted only as error decompositions.

4.3 Metrics

For point evaluation, matching the predictive-check convention used by G-Transformer [Xiong et al., 2024], Monte Carlo trajectories are averaged per patient before RMSE is computed:

$$\bar{x}_{i,t,v} = \frac{1}{M} \sum_{m=1}^{M} \tilde{x}_{i,m,t,v}. \tag{10}$$

The main metric is vector-normalized individual RMSE (VRMSE) over a panel of V scored variables ($V = 37$ for the MIMIC predictive-check panel), using the same aggregate scale as the G-Transformer predictive-check table [Xiong et al., 2024]:

$$e_i^{(k)} = \sqrt{\frac{1}{|\mathcal{T}_k|} \sum_{t \in \mathcal{T}_k} \sum_{v=1}^{V} \left(\frac{\bar{x}_{i,t,v} - x_{i,t,v}}{s_v} \right)^2}, \tag{11}$$

$$\mathrm{VRMSE}_V^{(k)} = \frac{1}{n} \sum_{i=1}^{n} e_i^{(k)}. \tag{12}$$

Here s_v is the train-set standard deviation for variable v. We report VRMSE_{37} for the MIMIC panel to match the numerical scale of the G-Transformer predictive-check table, and scalar normalized RMSE for comparisons across variable panels. Skill is $1 - \mathrm{RMSE}_{model}/\mathrm{RMSE}_{baseline}$.

Probabilistic evaluation uses continuous ranked probability score (CRPS), weighted interval score (WIS), probability integral transform Kolmogorov–Smirnov (PIT KS), and interval coverage. These scores are designed for a predictive distribution evaluated at the realized outcome, which is the observable quantity in factual EHR data. When aggregate Monte Carlo distributions are available, scores are computed from the empirical rollout distribution. We report CRPS, WIS, PIT KS, and coverage for probabilistic rollout rows with saved samples; appendix tables report additional coverage levels. Earlier calibration rows score conditional output heads and are labeled separately from aggregate Monte Carlo mixture results.

4.4 Known-effect synthetic checks

Factual MIMIC trajectories reveal only the observed care path. A synthetic benchmark inspired by CVSim yields paired trajectories under alternative treatment policies, so treatment-response error can be measured directly [Heldt et al., 2010, Li et al., 2021]. We use three CVSim settings. In *clean CVSim*, the model learns treatment response under known physiology. In *corrupted CVSim*, measurement noise, irregular observation, treatment confounding, and censoring are added while

Table 2: MIMIC predictive-check results. Point root mean squared error (RMSE) is computed after averaging Monte Carlo trajectories per patient. Each fitted model uses 100 paths and the shared treatment policy; scoring uses filled and post-event padded targets. Model and rollout settings are defined in Section 4.2; additional variants are in Appendices A.4 and A.8. Model-minus-baseline differences are negative when the model improves on the baseline; intervals are 95% patient-bootstrap intervals, paired for differences. continuous ranked probability score (CRPS) scores distributions; Cov90 targets 0.90. Bold marks the best value and values whose paired 95% confidence interval for the difference from the best includes zero; these share first place in this comparison. Underlining marks the next-best tier using the same rule.

	$k = 2$ ($n = 892$)			$k = 6$ ($n = 891$)		
Model	VRMSE ↓	CRPS ↓	Cov90	VRMSE ↓	CRPS ↓	Cov90
Carry-forward	7.475[6.83,8.22]	0.6813[0.635,0.736]	0.300	6.686[6.01,7.41]	0.5686[0.517,0.626]	0.415
Training population mean	**6.533**[5.88,7.29]	0.6213[0.575,0.674]	0.202	6.522[5.83,7.24]	0.6366[0.585,0.693]	0.179
G-Transformer reimplementation	6.845[6.20,7.56]	0.5191[0.477,0.569]	0.853	6.277[5.59,6.99]	0.4725[0.425,0.524]	**0.866**
PULSE: Gaussian output	6.669[6.03,7.40]	0.4799[0.437,0.529]	0.829	5.841[5.16,6.54]	0.4030[0.356,0.454]	0.829
PULSE: two-Gaussian mixture	**6.592**[5.96,7.33]	**0.4599**[0.418,0.509]	0.827	5.881[5.22,6.57]	**0.3935**[0.348,0.443]	0.833
PULSE-Spline	**6.556**[5.92,7.29]	0.4625[0.420,0.510]	0.829	**5.805**[5.12,6.51]	**0.3930**[0.348,0.442]	0.835
PULSE-CNF	**6.535**[5.90,7.26]	**0.4606**[0.418,0.509]	0.834	**5.767**[5.10,6.46]	**0.3923**[0.346,0.442]	0.836
PULSE: anchored CNF	6.628[5.99,7.35]	0.4732[0.432,0.521]	**0.868**	5.872[5.19,6.57]	0.3985[0.354,0.448]	**0.863**
PULSE-Spline minus G-Transformer reimplementation, with 95% intervals						
VRMSE	-0.289 [-0.366, -0.205]			-0.472 [-0.584, -0.364]		
CRPS	-0.0566 [-0.0611, -0.0522]			-0.0795 [-0.0880, -0.0724]		

simulator-labeled counterfactual outcomes are retained. In *MIMIC-seeded CVSim*, the first observed MIMIC vital signs and treatment context are mapped to a compatible CVSim initial condition before observed, escalated, and withdrawal policies are simulated. These settings separate known-effect response learning from factual MIMIC rollout.

5 Results

5.1 Factual ICU rollout

Table 2 reports the main MIMIC predictive-check results. Our G-Transformer reimplementation is evaluated using the same scoring rules (Appendix A.2); XGBoost remains non-competitive. PULSE-Spline improves over carry-forward and G-Transformer at both horizons in VRMSE_{37} and CRPS. A continuous normalizing flow gives similar scores. We use the B-spline head for its competitive scores and a direct monotone transformation that does not require differential-equation integration. On an Apple M4 Max 40-core GPU, 32-path rollout for eight validation patients takes 0.79 seconds for the flow and 0.92 seconds for the B-spline model on average. The flow integrates a learned differential equation; its computational and distributional trade-offs warrant further study. Model and rollout settings are defined in Section 4.2; broader architecture and probabilistic-head comparisons are reported in Appendices A.4 and A.8.

Table 2 also gives an empirical reason to not use low VRMSE_{37} as the sole metric for simulator validity. At $k = 2$, train population mean has similar VRMSE_{37} to PULSE-Spline and higher CRPS, even though it cannot react to future treatments or generate patient-specific dynamics. Adding held-out residual sampling to G-Transformer makes trajectories stochastic but worsens VRMSE_{37} from 7.124 to 8.525 at $k = 6$ using separate scoring rules (Table A.3), so we treat VRMSE_{37} as a scale-alignment check.

Appendices A.5 and A.6 report external eICU validation and cross-source transfer on shared variable panels. These analyses are interpreted within their panel definitions, not on the 37-variable MIMIC scale.

Ablations in Tables 3, A.3 and A.14 show the trade-off between point prediction and stochastic simulation. Using the scoring rules in Tables A.3 and A.14, grouped anchored residual means give the lowest full-cohort point error. Event-time inputs help B-spline point rollout at $k = 6$, and PULSE-Spline sampling exposes patient-state-dependent trajectory envelopes (Figure A.1). With

those rules, full-cohort sampled VRMSE_{37} is comparable to transformer Monte Carlo baselines but weaker than PULSE-Anchor.

The component ladder in Table 4 shows each model's scores and paired comparisons at $k = 6$.

Table 3: Paired comparisons between fitted models at forecast origins $k = 2$ and $k = 6$. Each difference is the score of the first model minus that of the second; negative values favor the first model. Brackets contain 95% patient-bootstrap confidence intervals. The fixed-noise model predicts a next state and adds Gaussian residual noise estimated from validation data.

Comparison	Δ VRMSE [95% interval]	Δ CRPS [95% interval]
Forecast from hour 2		
Gaussian − fixed-noise	-0.653 [-0.787, -0.531]	-0.0122 [-0.0173, -0.0070]
Two-Gaussian mixture − fixed-noise	-0.730 [-0.860, -0.610]	-0.0323 [-0.0367, -0.0278]
Continuous flow − B-spline	-0.021 [-0.069, +0.023]	-0.0019 [-0.0037, +0.0001]
B-spline − Gaussian	-0.113 [-0.173, -0.053]	-0.0175 [-0.0206, -0.0143]
B-spline − fixed-noise	-0.766 [-0.878, -0.661]	-0.0297 [-0.0340, -0.0252]
Continuous-time model − discrete-time model	+0.226 [+0.189, +0.264]	+0.0140 [+0.0126, +0.0154]
Timing-head model − continuous-time model	+0.550 [+0.427, +0.673]	+0.0040 [-0.0029, +0.0098]
Event-time model − timing-head model	+0.024 [-0.095, +0.144]	+0.0018 [-0.0029, +0.0075]
Anchored flow − flow	+0.094 [+0.051, +0.138]	+0.0127 [+0.0101, +0.0150]
Forecast from hour 6		
Gaussian − fixed-noise	-0.752 [-0.926, -0.585]	-0.0175 [-0.0259, -0.0076]
Two-Gaussian mixture − fixed-noise	-0.712 [-0.908, -0.542]	-0.0270 [-0.0357, -0.0178]
Continuous flow − B-spline	-0.038 [-0.082, +0.001]	-0.0008 [-0.0027, +0.0013]
B-spline − Gaussian	-0.036 [-0.089, +0.019]	-0.0100 [-0.0149, -0.0060]
B-spline − fixed-noise	-0.787 [-0.942, -0.646]	-0.0275 [-0.0343, -0.0207]
Continuous-time model − discrete-time model	+0.160 [+0.113, +0.203]	+0.0085 [+0.0052, +0.0113]
Timing-head model − continuous-time model	+0.321 [+0.187, +0.464]	+0.0032 [-0.0055, +0.0109]
Event-time model − timing-head model	+0.233 [+0.093, +0.387]	-0.0014 [-0.0077, +0.0060]
Anchored flow − flow	+0.104 [+0.033, +0.186]	+0.0063 [+0.0027, +0.0100]

Table 4: Component comparisons for forecasts from hour 6. Each row is a separately fitted model on the same patients and scoring rule. Fixed noise is Gaussian residual noise fitted on validation changes. Gaussian, mixture and continuous-flow rows use the event-time fixed-noise model as reference. For the first three additions, differences compare each row with the preceding row. Other references are named in the table. Intervals are 95% patient-bootstrap intervals. The B-spline model was initialized from a pretrained model. Bold marks the best value and values whose paired 95% confidence interval for the difference from the best includes zero; these share first place in this comparison. Underlining marks the next-best tier using the same rule. These intervals describe patient variation for the fitted models.

Model	VRMSE	Δ VRMSE [95% interval]	CRPS	Δ CRPS [95% interval]
Discrete-time fixed noise	<u>5.878</u>	–	0.4101	–
+ continuous time	6.038	+0.160 [+0.113, +0.203]	0.4186	+0.0085 [+0.0052, +0.0113]
+ measurement-timing head	6.359	+0.321 [+0.187, +0.464]	0.4219	+0.0032 [-0.0055, +0.0109]
+ event-time inputs	6.592	+0.233 [+0.093, +0.387]	0.4205	-0.0014 [-0.0077, +0.0060]
Gaussian vs fixed-noise	<u>5.841</u>	-0.752 [-0.926, -0.585]	<u>0.4030</u>	-0.0175 [-0.0259, -0.0076]
Two-Gaussian mixture vs fixed-noise	<u>5.881</u>	-0.712 [-0.908, -0.542]	**0.3935**	-0.0270 [-0.0357, -0.0178]
B-spline vs continuous flow	**5.805**	+0.038 [-0.001, +0.082]	**0.3930**	+0.0008 [-0.0013, +0.0027]
Continuous flow vs fixed-noise	**5.767**	-0.825 [-1.004, -0.663]	**0.3923**	-0.0282 [-0.0363, -0.0205]
+ anchoring vs continuous flow	<u>5.872</u>	+0.104 [+0.033, +0.186]	<u>0.3985</u>	+0.0063 [+0.0027, +0.0100]

5.2 Probabilistic residual modeling

B-spline heads map a standard normal sample through a learned monotone transform, giving each continuous variable a distribution whose shape depends on patient history. During simulation, the model samples the next state from these distributions and feeds it back to generate later hours. The resulting trajectory distribution includes uncertainty that accumulates across simulated histories. In

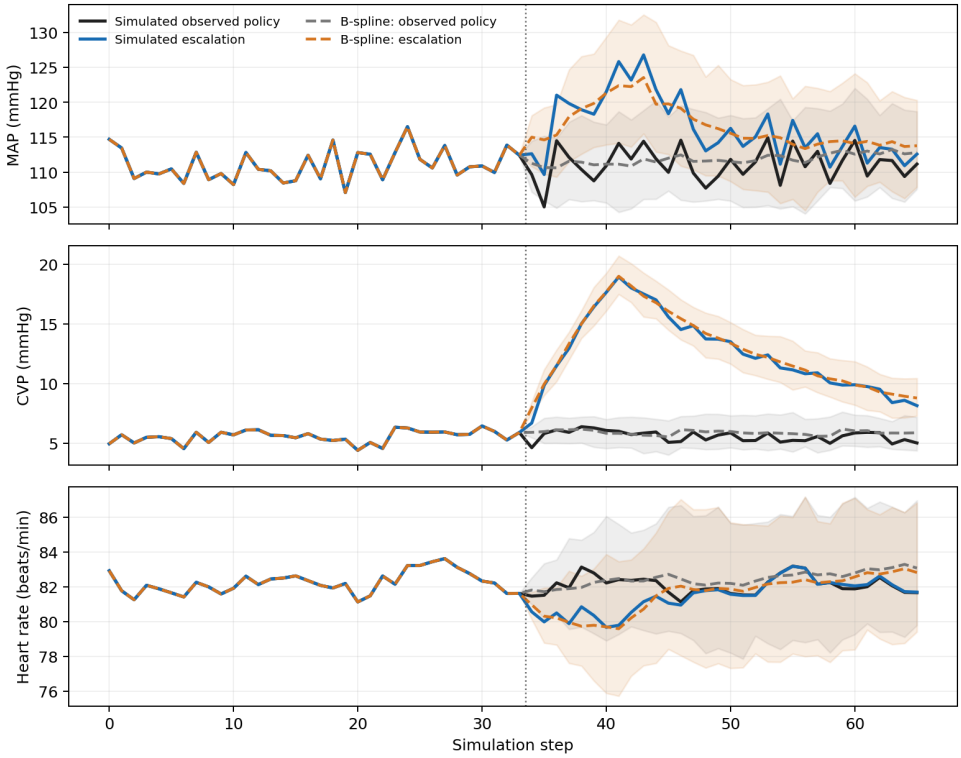


Figure 2: CVSim-inspired known-effect treatment response for a B-spline model under a simulated escalation. The model was fine-tuned on synthetic trajectories, including escalation regimes. Panels show mean arterial pressure (MAP), central venous pressure (CVP) and heart rate. Solid lines show the synthetic simulator; dashed lines and shading show the mean and central 90% range of 64 model paths. Initialization uses a patient from the public MIMIC-IV demo database [Johnson et al., 2023a].

Table 2, the B-spline model has lower VRMSE_{37} and CRPS estimates than the Gaussian model and G-Transformer at both forecast origins; its scores are close to those of the continuous-flow model. The models were fitted separately, and patient-bootstrap intervals describe uncertainty across patients for those fits.

Alternative temporal inputs and covariance-coupled rollouts are evaluated in Tables A.3 and A.15 using the scoring rules defined there. These experiments use different fitted models from the comparison in Table 2. Forecast densities illustrate how uncertainty evolves across simulated patient histories (Figure A.5).

5.3 Known-effect treatment response

A synthetic benchmark inspired by CVSim supplies the known-effect setting for treatment response. The clean-CVSim rows in Table A.20 use the grouped anchored simulator with Gaussian NLL continuous heads and show that continuation on targeted CVSim regimes can improve counterfactual population error for a held-out intervention family. Figure 2 shows the B-spline model under a simulated treatment escalation. Across 16 held-out synthetic patient pairs, its mean arterial-pressure response is 5.34 versus 5.47 mmHg for the simulator, and its heart-rate response is -0.77 versus -0.64 beats/min, using 32 paths per treatment policy. Corrupted CVSim adds EHR-like noise, irregular observation, confounding, and censoring; Table A.21 shows that the advantage becomes less stable. Appendix Table A.22 gives the MIMIC-seeded CVSim transfer experiment, where paired response training improves shared-panel known-effect recovery but individual response overlays remain attenuated.

6 Discussion

PULSE combines anchored residual dynamics, binary-to-continuous autoregression, and treatment-policy rollout in one simulator family. PULSE-Anchor uses a residual gate to make persistence explicit while allowing learned physiologic changes. Patient-specific anchoring and shared encoders provide alternative inductive biases for modeling those changes. Probabilistic variants add event-time covariates, Gaussian residual distributions, B-spline residual flows, and covariance-coupled residual sampling. Gaussian heads model heteroscedastic uncertainty, while B-spline flows represent flexible state-dependent densities and trajectory envelopes. In the main comparison, the sampled B-spline model improves on G-Transformer in VRMSE_{37} and CRPS at both horizons and is close to the continuous-flow model.

All-hours RMSE combines alive-in-ICU physiology with post-release and post-death padding, and oracle-state decompositions show that release/death timing is a large component of the score. Future simulators should model stay-in-ICU, ICU release, and death as a patient-specific competing-risk transition process and should select models by validation rollout RMSE together with status likelihood. Treatment-response evidence is also necessarily split by data source: MIMIC evaluates factual rollout under the observed care process, whereas CVSim evaluates known-effect counterfactual response under a mechanistic simulator. The MIMIC-seeded CVSim results show partial transfer but attenuated response in difficult regimes, so counterfactual validity in real EHR data remains unresolved.

The predictive check uses MIMIC-IV v3.1, whereas G-Transformer reports MIMIC-IV v1.0 results [Johnson et al., 2021]; this source-version difference changes cohort details but does not change the qualitative comparison. Within the matched evaluation, the B-spline model improves factual prediction, persistence baselines reveal the limits of RMSE as a simulator metric, and learned residual distributions expose uncertainty structure that static residual noise cannot represent. Efficient inference and distribution visualization are important directions for extending these models.

This study uses existing de-identified MIMIC-IV and eICU records; we collected no new patient data. Access to both datasets requires credentialing and a data-use agreement [Johnson et al., 2024, Pollard et al., 2019]. For MIMIC-IV, the source institution reviewed data collection and waived informed consent [Johnson et al., 2024].

A reliable ICU trajectory simulator could support policy stress tests, synthetic cohort generation, model-based sensitivity analyses, and preclinical evaluation of treatment-regime hypotheses. The present results identify the requirements for that use: accurate factual rollout, explicit patient-existence state, treatment-policy realism, calibrated uncertainty, and known-effect validation in semi-synthetic, mechanistic, and non-public clinical settings. The next technical steps are competing-risk status transitions, treatment occurrence-plus-dose models, event-level data beyond hourly bins when available, LLM or state-space simulator backbones with clinical adapters, and synthetic-to-real response training that preserves known treatment effects without degrading factual MIMIC fit.

Acknowledgments and Disclosure of Funding

This research work was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) [Grant reference number EP/Y035216/1] Centre for Doctoral Training in Data-Driven Health (DRIVE-Health) at King's College London. Computational resources were provided by King's Computational Research, Engineering and Technology Environment (CREATE) at King's College London [King's College London, 2026].

The authors declare no competing interests.

References

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, pages 1171–1179, 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/hash/e995f98d56967d946471af29d7bf99f1-Abstract.html.

Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific Reports*, 8(1):6085, 2018. doi: 10.1038/s41598-018-24271-9.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016. doi: 10.1145/2939672.2939785. URL https://arxiv.org/abs/1603.02754.

Carl de Boor. *A Practical Guide to Splines*. Springer, revised edition, 2001. ISBN 9780387953663. URL https://link.springer.com/book/9780387953663.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. In *Advances in Neural Information Processing Systems*, volume 32, pages 7511–7522, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/7ac71d433f282034e088473244df8c02-Abstract.html.

Thomas Frost, Hrisheekesh Vaidya, and Steve Harris. The hidden risks of temporal resampling in clinical reinforcement learning. *arXiv preprint arXiv:2602.06603v3*, 2026. URL https://arxiv.org/abs/2602.06603v3.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. doi: 10.1161/01.CIR.101.23.e215. URL https://pubmed.ncbi.nlm.nih.gov/10851218/.

Thomas Heldt, Ramakrishna Mukkamala, George B. Moody, and Roger G. Mark. CVSim: An open-source cardiovascular simulator for teaching and research. *The Open Pacing, Electrophysiology and Therapy Journal*, 3:45–54, 2010. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC3178445/.

Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman and Hall/CRC, Boca Raton, 2020. URL https://miguelhernan.org/whatifbook.

George Hripcsak, David J. Albers, and Adler Perotte. Parameterizing time in electronic health record studies. *Journal of the American Medical Informatics Association*, 22(4):794–804, 2015. doi: 10.1093/jamia/ocu051. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC6169471/.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35 (1):73–101, 1964. doi: 10.1214/aoms/1177703732.

Alexander Immer, Emanuele Palumbo, Alexander Marx, and Julia Vogt. Effective Bayesian heteroscedastic regression with deep neural networks. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 53996–54019. Curran Associates, Inc., 2023. doi: 10.52202/075280-2349. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a901d5540789a086ee0881a82211b63d-Paper-Conference.pdf.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV. *PhysioNet*, 2021. doi: 10.13026/s6n6-xd98. URL https://physionet.org/content/mimiciv/1.0/. Version 1.0.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV clinical database demo. *PhysioNet*, 2023a. doi: 10.13026/dp1f-ex47. URL https://physionet.org/content/mimic-iv-demo/2.2/. Version 2.2; Open Data Commons Open Database License v1.0.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV. *PhysioNet*, October 2024. doi: 10.13026/kpb9-mt58. URL https://physionet.org/content/mimiciv/3.1/. Version 3.1.

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):1, January 2023b. ISSN 2052-4463. doi: 10.1038/s41597-022-01899-x.

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5574–5584. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/2650d6089a6d640c5e85b2b88265dc2b-Paper.pdf.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. In *Advances in Neural Information Processing Systems*, volume 33, pages 6696–6707, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4a5876b450b45371f6cfe5047ac8cd45-Abstract.html.

King's College London. King's Computational Research, Engineering and Technology Environment (CREATE), 2026. URL https://doi.org/10.18742/rnvf-m076. Accessed 24 October 2026.

Rui Li, Stephanie Hu, Mingyu Lu, Yuria Utsumi, Prithwish Chakraborty, Daby M. Sow, Piyush Madan, Jun Li, Mohamed Ghalwash, Zach Shahn, and Li-wei H. Lehman. G-Net: a recurrent network approach to g-computation for counterfactual prediction under a dynamic treatment regime. In *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pages 282–299. PMLR, 2021. URL https://proceedings.mlr.press/v158/li21a.html.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal Transformer for estimating counterfactual outcomes. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 15293–15329. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/melnychuk22a.html.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021. URL https://jmlr.org/papers/v22/19-1028.html.

Tom Pollard, Alistair Johnson, Jesse Raffa, Leo Anthony Celi, Omar Badawi, and Roger Mark. eICU Collaborative Research Database. PhysioNet, 2019. URL https://physionet.org/content/eicu-crd/2.0/. Version 2.0.

Tom Pollard, Benjamin E. Moody, Li-wei H. Lehman, Brian J. Gow, Chrystinne Fernandes, Chen Xie, Alistair Johnson, Roger G. Mark, and Thomas Heldt. PhysioNet as a global platform for biomedical research. *Nature Health*, 1(8):792–795, 2026. doi: 10.1038/s44360-026-00096-z.

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.

James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7 (9–12):1393–1512, 1986. ISSN 0270-0255. doi: 10.1016/0270-0255(86)90088-6.

Mervyn Singer, Clifford S. Deutschman, Christopher W. Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R. Bernard, Jean-Daniel Chiche, Craig M. Coopersmith, Richard S. Hotchkiss, Mitchell M. Levy, John C. Marshall, Greg S. Martin, Steven M. Opal, Gordon D. Rubenfeld, Tom van der Poll, Jean-Louis Vincent, and Derek C. Angus. The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA*, 315(8):801–810, 2016. doi: 10.1001/jama.2016.0287. URL `https://pubmed.ncbi.nlm.nih.gov/26903338/`.

Andrew Stirn, Harm Wessels, Megan Schertzer, Laura Pereira, Neville Sanjana, and David Knowles. Faithful heteroscedastic regression with neural networks. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 5593–5613. PMLR, 25–27 Apr 2023. URL `https://proceedings.mlr.press/v206/stirn23a.html`.

Hong Xiong, Feng Wu, Leon Deng, Megan Su, and Li-wei H. Lehman. G-Transformer: Counterfactual outcome prediction under dynamic and time-varying treatment regimes. In Kaivalya Deshpande, Madalina Fiterau, Shalmali Joshi, Zachary Lipton, Rajesh Ranganath, and Iñigo Urteaga, editors, *Proceedings of the 9th Machine Learning for Healthcare Conference*, volume 252 of *Proceedings of Machine Learning Research*. PMLR, 16–17 Aug 2024. URL `https://proceedings.mlr.press/v252/xiong24a.html`.

A Appendix

A.1 Cohort construction

Table A.1: MIMIC-IV v3.1 sepsis cohort construction. The main experiments use a subject-level split with one ICU stay per subject. The G-Transformer paper used MIMIC-IV v1.0; version and source-definition differences are therefore reported explicitly.

Step	Count	Notes
Raw ICU stays	94,458	From `mimiciv_icu.icustays`.
Unique subjects	65,366	From `mimiciv_icu.icustays`.
Sepsis-3 eligible stays	12,369	Cohort table after sepsis and surgical-exclusion logic.
Current cohort stays with ICU LOS < 24h	1,316	Short-stay/release trajectories remain in the source cohort and are padded after release.
Impossible timestamps	0	No impossible timestamps detected in the final cohort table.
Final eligible stays	12,369	Postgres cohort count; cache count may be subset.
Final unique subjects	10,845	Postgres cohort count.
One-stay-per-subject cohort	10,845	First eligible stay per subject.
Primary train/val/test split	7,147 / 893 / 894	Subject-level split matching the G-Transformer reported split size; remaining eligible subjects are reserved for auxiliary analyses.

A.2 G-Transformer target scale and metric definitions

The G-Transformer MIMIC predictive-check table reports Causal Transformer 14.37/15.73, G-Net 7.017/7.670, and G-Transformer 6.993/7.643 for $k = 2/k = 6$. A reconstruction using 36 variables and log-scale scoring gives 7.653/6.971 for G-Transformer at $k = 2/k = 6$. If the published hour labels are reversed, the reported G-Transformer scores become 7.643/6.993, close to this reconstruction. This interpretation assumes a reversal and the reconstruction was selected from several scoring choices. Our main predictive check reports vector-normalized RMSE over the 37 continuous or treated-continuous variables in Table A.2. Scalar normalized individual RMSE averages squared error over both timesteps and variables before taking the square root. Vector-normalized individual RMSE sums squared error across the 37-variable state vector at each future timestep before the square root. If variable-level errors are similar on the normalized scale, the vector metric is approximately $\sqrt{37}$ times the scalar metric; this explains why scalar values near 1 correspond to vector values near 6–8.

A.3 Scored continuous variable panel

Table A.2: Scored continuous and treated-continuous variables. All RMSE metrics normalize each variable by the training-set standard deviation before aggregation.

Variable	Group	Unit or role
Heart Rate	Vital sign	beats/min
Diastolic Blood Pressure	Vital sign	mmHg
Systolic Blood Pressure	Vital sign	mmHg
Mean Blood Pressure	Vital sign	mmHg
Minimum Mean Blood Pressure	MAP-derived	mmHg
Minimum Change in Mean Blood Pressure from Baseline	MAP-derived	mmHg
Minimum Mean Blood Pressure from Baseline	MAP-derived	mmHg relative to baseline definition
Minimum Change in Mean Blood Pressure from Previous	MAP-derived	mmHg
Temperature	Vital sign	degrees Celsius
SOFA Score	Severity score	treated as continuous
Change in SOFA Score from Baseline	Severity score	score difference
Change in SOFA Score from Previous	Severity score	score difference

Variable	Group	Unit or role
Platelet	Laboratory	counts/10^9 L
Hemoglobin	Laboratory	g/dL
Calcium	Laboratory	mg/dL
BUN	Laboratory	table unit mmol/L; raw MIMIC source audited for mg/dL conversion
Creatinine	Laboratory	mg/dL
Bicarbonate	Laboratory	mmol/L
Lactate	Laboratory	mmol/L
O2 Requirement Level	Oxygen support	ordinal level treated as continuous
Change in O2 from Baseline	Oxygen support	level difference
Change in O2 from Previous	Oxygen support	level difference
pO2	Arterial blood gas	mmHg
sO2	Arterial blood gas	percent
spO2	Vital sign	percent
pCO2	Arterial blood gas	mmHg
Total CO2	Laboratory	mEq/L
pH	Arterial blood gas	numerical pH
Base excess	Arterial blood gas	mmol/L
Weight	Body measure	kg
Change in Weight	Body measure	kg
Respiratory Rate	Vital sign	breaths/min
Fluid Volume	Fluids	maintenance fluid amount, mL
Urine Output	Fluids	mL
Bolus Volume	Treatment amount	administered bolus amount, mL
Cumulative Fluids	Fluids	mL
Fluid Balance in ICU	Fluids	mL

A.4 Architecture and scenario sweep

Some population statistics predate the split, and the earlier G-Transformer implementation reads padded histories.

Table A.3: Architecture and temporal-encoding comparison. Rows use the 37-variable vector-normalized RMSE definition; historical preparation and rollout protocols differ from the main predictive check. Observation-probability heads are retained as ablations because they did not consistently improve probabilistic-head rollout.

Scenario	Head/temporal setup	$k = 2$	$k = 6$
G-Transformer baseline, per-covariate MSE	conditional mean	7.180	7.124
G-Transformer baseline, per-covariate MSE	empirical residual	8.529	8.525
MSE grouped encoders	direct MSE	7.271	7.245
Event-time+Δt+p_{obs} B-spline grouped	spline mean	7.679	6.635
NLL Gaussian grouped	model mean	7.794	7.788
Event-time+Δt NLL Gaussian grouped	model mean	7.880	6.777
Event-time NLL Gaussian grouped	model mean	7.928	7.838
Event-time+Δt+p_{obs} Gaussian grouped	model mean	8.101	6.966
B-spline grouped, no temporal heads	spline mean	8.300	8.301
Causal Transformer	empirical residual	13.391	12.484
MSE shared encoder	direct MSE	17.526	14.909
XGBoost autoregressive baseline	empirical residual	18.873	20.192
XGBoost with static covariates, validation-tuned	empirical residual	18.955	20.242
NLL Gaussian per covariate	model mean	21.741	21.281
Event-time NLL Gaussian per covariate	model mean	23.520	24.247

Table A.4: Scalar continuous-normalizing-flow (CNF) residual comparison. The CNF head was tested as a smoother residual-flow alternative to B-splines. This earlier four-basis flow did not improve point rollout or distributional scoring. The eight-basis flow in the main comparison is a separate fit.

Head	Rule	k	$\mathrm{VRMSE}_{37} \downarrow$	One-step NLL $\downarrow$	Readout
B-spline residual flow	mean	6	6.635	–	retained expressive flow
Residual CNF	sample MC	6	7.089	3.566	slower, not competitive
Residual CNF	mean	6	8.217	3.566	worse point rollout
Gaussian NLL	mean	6	6.777	–	calibrated baseline

Table A.5: Paired-bootstrap confidence intervals for an earlier MIMIC comparison. The bootstrap unit is the patient-level 37-variable vector-normalized individual RMSE, with 5,000 resamples. Lower values are better.

Model	k	VRMSE_{37} [95% CI]	Difference from carry-forward [95% CI]
Carry-forward	2	7.980 [7.357,8.710]	0.000 [0.000,0.000]
Train population mean	2	6.529 [5.883,7.266]	-1.450 [-1.584,-1.323]
Causal Transformer	2	13.391 [12.862,13.990]	5.412 [5.114,5.710]
XGBoost autoregression	2	18.955 [18.370,19.575]	10.997 [10.704,11.286]
G-Transformer	2	7.180 [6.532,7.867]	-0.799 [-0.961,-0.637]
PULSE, reference	2	6.921 [6.293,7.617]	-1.058 [-1.212,-0.908]
Anchored residual empirical state+treatment rollout	2	6.758 [6.131,7.474]	-1.221 [-1.375,-1.073]
Carry-forward	6	6.633 [5.940,7.405]	0.000 [0.000,0.000]
Train population mean	6	6.484 [5.815,7.250]	-0.148 [-0.267,-0.030]
Causal Transformer	6	12.484 [11.906,13.147]	5.851 [5.568,6.149]
XGBoost autoregression	6	20.242 [19.617,20.916]	13.630 [13.335,13.925]
G-Transformer	6	7.124 [6.451,7.905]	0.491 [0.334,0.648]
PULSE, reference	6	5.969 [5.285,6.727]	-0.664 [-0.767,-0.564]
Anchored residual empirical state+treatment rollout	6	5.871 [5.195,6.622]	-0.762 [-0.867,-0.665]

The paired comparison between reference PULSE and G-Transformer is also significant under the same patient bootstrap: PULSE reduces VRMSE_{37} by 0.259 at $k = 2$ (95% CI 0.140–0.383) and by 1.155 at $k = 6$ (95% CI 1.020–1.293).

A.5 eICU validation panels

We report three earlier eICU validation panels. The seven-vital core scores heart rate, systolic blood pressure, diastolic blood pressure, mean blood pressure, temperature, respiratory rate, and oxygen saturation. The 22-variable extended panel adds broader eICU physiologic and laboratory variables. The seven-vital and 22-variable panels report raw-scale eICU RMSE and coverage; their aggregation differs across tables, and their values are not numerically comparable to the MIMIC 37-variable vector-normalized metric. Tables A.7 and A.8 add grouped probabilistic follow-ups on the same 22-variable eICU panel, including B-spline heads. The 22-variable grouped models use recorded targets and factual future treatments in both tables. The narrower shared HR/MAP panel in Table A.9 is retained only because it is the common output space across MIMIC, eICU, and CVSim.

Table A.6: eICU-local raw-scale rollout panels. The seven-vital core uses grouped models and MC32; the 22-variable extended panel uses per-covariate models and MC16. RMSE and MAE pool recorded targets. The 22-variable carry-forward uses the training mean for variables unobserved in the prefix; the per-covariate model rows retain an evaluation that omitted observation masks from their histories.

Panel	Model	Test stays	MC	RMSE ↓	MAE ↓	cov95
Seven-vital core	Carry-forward	4,332	1	13.987	8.632	–
Seven-vital core	Grouped MSE + static MC	4,332	32	11.923	7.832	0.928
Seven-vital core	Grouped NLL	4,332	32	11.754	7.655	0.921
Seven-vital core	Grouped NLL + sigma reg.	4,332	32	**11.636**	7.900	0.945
22-variable extended	Carry-forward	2,552	1	212.991	46.377	–
22-variable extended	Per-covariate MSE	2,552	16	153.438	41.224	0.892
22-variable extended	Per-covariate NLL	2,552	16	**151.046**	35.181	0.871

Table A.7: eICU 22-variable grouped probabilistic evaluation at $k = 6$. Each model uses 100 trajectories for 2,552 test stays, observed-history masks, static covariates and factual future treatments. Scores use recorded targets; VRMSE_{22} and raw RMSE average per-stay errors, unlike the pooled error in Table A.6. Carry-forward uses the training mean before a variable's first measurement. CVSim30 adds partially labelled synthetic data.

Training data	Model	Rule	VRMSE_{22} ↓	Raw RMSE ↓	Carry-forward	Train-hour mean	Skill vs CF ↑
eICU	MoG2	sample MC100	4.434	173.583	2.616	2.603	-0.695
eICU	B-spline	sample MC100	4.577	235.611	2.616	2.603	-0.750
MIMIC+eICU	MoG2	sample MC100	2.839	194.149	2.616	2.603	-0.085
MIMIC+eICU	B-spline	sample MC100	3.209	172.618	2.616	2.603	-0.227
M+E+CVSim30	MoG2	sample MC100	3.083	185.390	2.616	2.603	-0.178
M+E+CVSim30	B-spline	sample MC100	3.428	162.797	2.616	2.603	-0.310

Table A.8: Distribution scores for the models in Table A.7, using the same 100 trajectories and recorded targets. CRPS and weighted interval score (WIS) use the training-standardized scale and pool recorded variable–hour targets. Cov90 is coverage of the central 90% interval, with linearly interpolated sample quantiles. The probability integral transform (PIT) Kolmogorov–Smirnov (KS) statistic compares empirical sample ranks with a uniform distribution. Width CV is the coefficient of variation of the standardized 90% interval width across targets.

Training data	Model	Rule	CRPS ↓	WIS ↓	PIT KS ↓	cov90	Width CV
eICU	MoG2	sample MC100	0.869	3.012	0.158	0.962	0.517
eICU	B-spline	sample MC100	0.900	3.157	0.191	0.912	0.456
MIMIC+eICU	MoG2	sample MC100	0.661	2.457	0.180	0.980	0.663
MIMIC+eICU	B-spline	sample MC100	0.648	2.265	0.094	0.964	0.423
M+E+CVSim30	MoG2	sample MC100	0.710	2.636	0.220	0.981	0.587
M+E+CVSim30	B-spline	sample MC100	0.676	2.311	0.148	0.952	0.544

Table A.9: eICU shared HR/MAP normalized transfer panel at $k = 6$. This two-variable panel is retained for comparability with CVSim, which maps cleanly to heart rate and mean arterial pressure. All rows evaluate 300 eICU test stays.

Training data	Model	scalar ↓	VRMSE_2 ↓	Carry-forward	Pop. mean	Skill vs CF ↑
MIMIC	MoG4 event-time NLL	0.721	1.020	1.141	1.335	0.106
MIMIC	B-spline event-time flow	0.762	1.078	1.141	1.335	0.055
eICU	MoG4 event-time NLL	0.718	1.015	1.138	1.323	0.107
eICU	B-spline event-time flow	0.736	1.041	1.138	1.323	0.085
MIMIC+eICU	MoG4 event-time NLL	0.735	1.040	1.137	1.324	0.086
MIMIC+eICU	B-spline event-time flow	0.767	1.084	1.137	1.324	0.047
MIMIC+eICU+CVSim	MoG4 event-time NLL	**0.599**	**0.847**	0.928	1.365	0.087
MIMIC+eICU+CVSim	B-spline event-time flow	0.633	0.895	0.928	1.365	0.036

A.6 Cross-dataset interleaving on a shared panel

The cross-dataset runs use a shared heart-rate/mean-arterial-pressure panel so that MIMIC, eICU, and CVSim can be trained and evaluated with a common output space. This panel is intentionally narrower than the MIMIC 37-variable predictive check, but it tests source transfer and synthetic augmentation under matched variables. Table A.10 reports the best large-tier $k = 6$ row for each training mix and evaluation source, so B-spline rows appear only when they are selected by that criterion. Table A.11 fixes the training source to MIMIC+eICU+CVSim and compares model families directly.

Table A.10: Large-tier $k = 6$ shared HR/MAP transfer results. Each row selects the best model family for that training mix and evaluation source from the retained grouped MSE, per-covariate MSE, MoG2, MoG4, and B-spline runs. VRMSE_2 is the two-variable vector-normalized score.

Eval source	Training data	Best model family	$\text{VRMSE}_2 \downarrow$	Carry-forward	Pop. mean	Skill vs CF ↑
MIMIC	MIMIC	MoG4 event-time NLL	1.134	1.245	1.407	0.089
MIMIC	eICU	MoG4 event-time NLL	1.136	1.242	1.413	0.086
MIMIC	MIMIC+eICU	MoG2 event-time NLL	1.132	1.242	1.405	0.089
MIMIC	MIMIC+eICU+CVSim	MoG4 event-time NLL	**0.988**	1.056	1.444	0.064
eICU	MIMIC	MoG4 event-time NLL	1.020	1.141	1.335	0.106
eICU	eICU	MoG4 event-time NLL	1.015	1.138	1.323	0.107
eICU	MIMIC+eICU	MoG2 event-time NLL	1.034	1.137	1.324	0.091
eICU	MIMIC+eICU+CVSim	per-covariate MSE	**0.834**	0.928	1.365	0.101

Table A.11: Large-tier $k = 6$ shared-panel family comparison with training fixed to MIMIC+eICU+CVSim. CVSim augmentation reduces the shared-panel baseline difficulty and improves both MIMIC and eICU factual scores, but the best factual family is not the best known-effect family.

Eval source	Training data	Model family	$\text{VRMSE}_2 \downarrow$	raw RMSE ↓	Skill vs CF ↑	Skill vs pop. ↑
eICU	MIMIC+eICU+CVSim	per-covariate MSE	**0.834**	11.667	0.101	0.389
eICU	MIMIC+eICU+CVSim	MoG4 event-time NLL	0.847	11.819	0.087	0.379
eICU	MIMIC+eICU+CVSim	grouped MSE	0.857	11.998	0.077	0.372
eICU	MIMIC+eICU+CVSim	MoG2 event-time NLL	0.891	12.699	0.040	0.347
eICU	MIMIC+eICU+CVSim	B-spline event-time flow	0.895	12.787	0.036	0.344
MIMIC	MIMIC+eICU+CVSim	MoG4 event-time NLL	**0.988**	13.536	0.064	0.316
MIMIC	MIMIC+eICU+CVSim	per-covariate MSE	0.989	13.590	0.063	0.315
MIMIC	MIMIC+eICU+CVSim	MoG2 event-time NLL	1.003	13.829	0.050	0.305
MIMIC	MIMIC+eICU+CVSim	B-spline event-time flow	1.020	14.012	0.034	0.294
MIMIC	MIMIC+eICU+CVSim	grouped MSE	1.030	14.057	0.025	0.287

Table A.12: Large-tier held-out CVSim known-effect summary on the same shared panel. Values average the held-out GC1s, GC2f, and GC2v regimes. Lower effect RMSE is better; sign accuracy measures agreement in treatment-effect direction.

Training data	Model family	Effect RMSE ↓	Trajectory RMSE ↓	Sign accuracy ↑
eICU	MoG4 event-time NLL	**1.533**	12.260	0.408
MIMIC+eICU+CVSim	B-spline event-time flow	1.538	7.781	**0.673**
MIMIC+eICU+CVSim	MoG4 event-time NLL	1.667	**3.111**	0.509
MIMIC+eICU+CVSim	MoG2 event-time NLL	1.668	3.040	0.424
MIMIC	MoG2 event-time NLL	1.670	21.194	0.262
MIMIC+eICU	B-spline event-time flow	1.674	12.186	0.479
MIMIC+eICU+CVSim	grouped MSE	1.729	4.026	0.390
eICU	MoG2 event-time NLL	1.731	11.429	0.413

A.7 Compute cost

Table A.13: Recorded training runtimes. Epoch counts reflect completed epochs, including early stopping. The G-Transformer row is the main-table fit on an NVIDIA A30; the other rows use earlier implementations and different training configurations, and their hardware is not consistently recorded. Times include training, validation, and setup; seconds per epoch are total time divided by completed epochs. Spline likelihood uses numerical inversion.

Run	Setup	Total runtime	Seconds/epoch
Causal Transformer	50 epochs, MSE, direct level	163s	3.3
G-Transformer	31 epochs, NVIDIA A30	7873s	254
Gaussian, grouped	20 epochs, batch 32, NLL	307s	15.3
Gaussian, per variable	12 epochs, batch 32, NLL	667s	55.6
Continuous-time Gaussian, grouped	25 epochs, batch 128, NLL	39s	1.6
Event-time Gaussian, grouped	24 epochs, batch 128, NLL	57s	2.4
B-spline residual flow	25 epochs, B-spline head without event-time inputs	573s	23
Event-time B-spline residual flow	25 epochs, event-time+Δt+p_{obs}, residual B-spline	959s	38
PULSE-CovSpline	20 epochs, staged standardized B-spline, latent severity, correlated residuals	3518s	176

The earlier PULSE-CovSpline implementation is therefore not compute-optimized. The principal engineering targets are vectorized rollout, fewer small device kernels, avoiding exact spline means or quadrature where unnecessary, and profiling the spline inverse/log-Jacobian path. These changes should reduce runtime without changing the statistical model.

A.8 Covariance-coupled residual analyses

Table A.14: Earlier PULSE-CovSpline analyses. The subset rows use a 350/75/75 subject split and the full 37-variable panel. The full-cohort rows show that generated covariance-coupled B-spline rollout does not beat persistence, while true death/release sharply improves the same model.

Model	Rule	k	MC	$\text{VRMSE}_{37}\downarrow$	CRPS $\downarrow$	skill vs CF $\uparrow$
PULSE-Spline	mean	2	8	8.202	240.08	-0.009
PULSE-CovSpline	mean	2	8	8.174	234.88	-0.006
PULSE-Spline	mean	6	8	5.139	115.54	0.051
PULSE-CovSpline	mean	6	8	4.901	112.35	0.095
PULSE-CovSpline	sample	6	100	5.327	103.42	0.017
Residual-correlation grouped	mean	6	8	4.984	115.20	0.080
Residual-correlation grouped	sample	6	100	5.310	105.58	0.020
PULSE-CovSpline, full cohort	sample	6	100	7.463	78.44	-0.125
PULSE-CovSpline, full cohort	mean	6	20	7.937	–	-0.197
PULSE-CovSpline, true death/release	mean	6	20	4.672	–	0.296

PULSE-CovSpline adds latent severity pooling, train-only covariance-based continuous-variable groups, and group-level correlated residual noise. It improves deterministic $k = 6$ rollout on the subset, while sampled rollout at MC=100 improves distributional scores relative to the deterministic rule but remains weaker for point RMSE. On the full cohort, however, the same covariance-coupled design is worse than carry-forward when future status is generated by the model. The true-death/release row is not deployable; it is included to show that the full-cohort failure is dominated by absorbing-state timing and padding. The residual-correlation variant clusters variables using train-only residuals $y_{t+1} - \hat{\mu}_t$ from a previously trained mean model. This sensitivity analysis adds a separate mean-model dependency and does not dominate the simpler covariance-grouped architecture in the first subset experiment.

Table A.15: MC100 calibration for the covariance-coupled subset experiments. PIT standard deviation is compared with the uniform value $1/\sqrt{12} \approx 0.289$; interval coverage is best when close to the nominal level.

Model	k	VRMSE_{37}	norm CRPS	PIT mean	PIT std	PIT KS	cov80	cov95
PULSE-CovSpline, sampled	2	8.098	0.622	0.523	0.321	0.093	0.806	0.909
PULSE-CovSpline, sampled	6	5.327	0.545	0.579	0.309	0.119	0.854	0.939
Residual-correlation grouped, sampled	6	5.310	0.552	0.586	0.307	0.125	0.859	0.938

Table A.16: Residual-correlation groups learned from train-only model residuals. Bolus volume is kept as a treatment amount outside the physiologic outcome groups.

Group	Variables
Weight	Change in Weight; Weight
Laboratory/ABG	BUN; Base excess; Bicarbonate; Calcium; Creatinine; Hemoglobin; Lactate; Platelet; Total CO2; pCO2; pH; pO2
sO2	sO2
Fluids	Cumulative Fluids; Fluid Balance in ICU; Fluid Volume; Urine Output
Physiology/severity	Change in O2 from Baseline; Change in O2 from Previous; Change in SOFA Score from Baseline; Change in SOFA Score from Previous; Diastolic Blood Pressure; Heart Rate; Mean Blood Pressure; Minimum Change in Mean Blood Pressure from Baseline; Minimum Change in Mean Blood Pressure from Previous; Minimum Mean Blood Pressure; Minimum Mean Blood Pressure from Baseline; O2 Requirement Level; Respiratory Rate; SOFA Score; Systolic Blood Pressure; spO2
Temperature	Temperature
Treatment amount	Bolus Volume

Table A.17: Largest latent-severity correlations in the PULSE-CovSpline subset experiment. The latent vector is learned only through rollout training; correlations are reported to check whether the pooled state aligns with future clinical status.

Latent dimension	Future variable	Pearson r	N
z_0	next mechanical ventilation	-0.642	1,800
z_9	future ICU release	-0.530	1,800
z_{14}	future ICU release	0.527	1,800
z_{10}	next mechanical ventilation	-0.512	1,800
z_3	next mechanical ventilation	-0.490	1,800
z_6	future ICU release	-0.485	1,800

A.9 Additional trajectory and density figures

Table A.18: Absorbing-state timing decomposition on the 37-variable MIMIC panel. Rows with true death/release timing use future status information and are error decompositions, not deployable simulators.

Protocol	k	VRMSE_{37}	alive-only	padded-only
Carry-forward	2	7.980	6.620	23.730
Carry-forward, true death/release	2	6.456	6.620	0.000
PULSE, generated death/release	2	7.578	6.187	23.734
PULSE, true death/release	2	5.116	5.242	0.000
PULSE, true treatments and binaries	2	4.975	5.101	0.000
Carry-forward	6	6.633	5.141	23.145
Carry-forward, true death/release	6	4.987	5.141	0.000
PULSE, generated death/release	6	6.700	5.184	22.379
PULSE, true death/release	6	4.169	4.273	0.000
PULSE, true treatments and binaries	6	4.082	4.188	0.000

A.10 Treatment-policy analyses

The treatment-policy analyses separate bolus volume from vasopressor status. Bolus is an amount administered within the hour; forward-filling it treats a past bolus as a persistent state and inflates future dose. Train-hour means beat carry-forward for bolus amount RMSE. Vasopressor is closer to a persistent binary process; carry-forward has better event discrimination and F1 than the fitted grouped neural policy, although train-rate baselines can have lower Brier by predicting the majority class. These results motivate occurrence-plus-amount modeling for bolus and start/continue/stop transitions for vasopressors.

A.11 Derived-variable consistency

Several scored variables are algebraically derived from primitives: MAP changes from baseline and previous hour, SOFA changes, oxygen changes, cumulative fluids, and fluid balance. A recomputation mode replaces independently predicted derived variables with values recomputed from predicted primitives. This improves semantic consistency and raises all-37 vector RMSE by approximately 0.90 at $k = 2$ and 0.72 at $k = 6$ when future variables are generated by the model. The all-37 metric

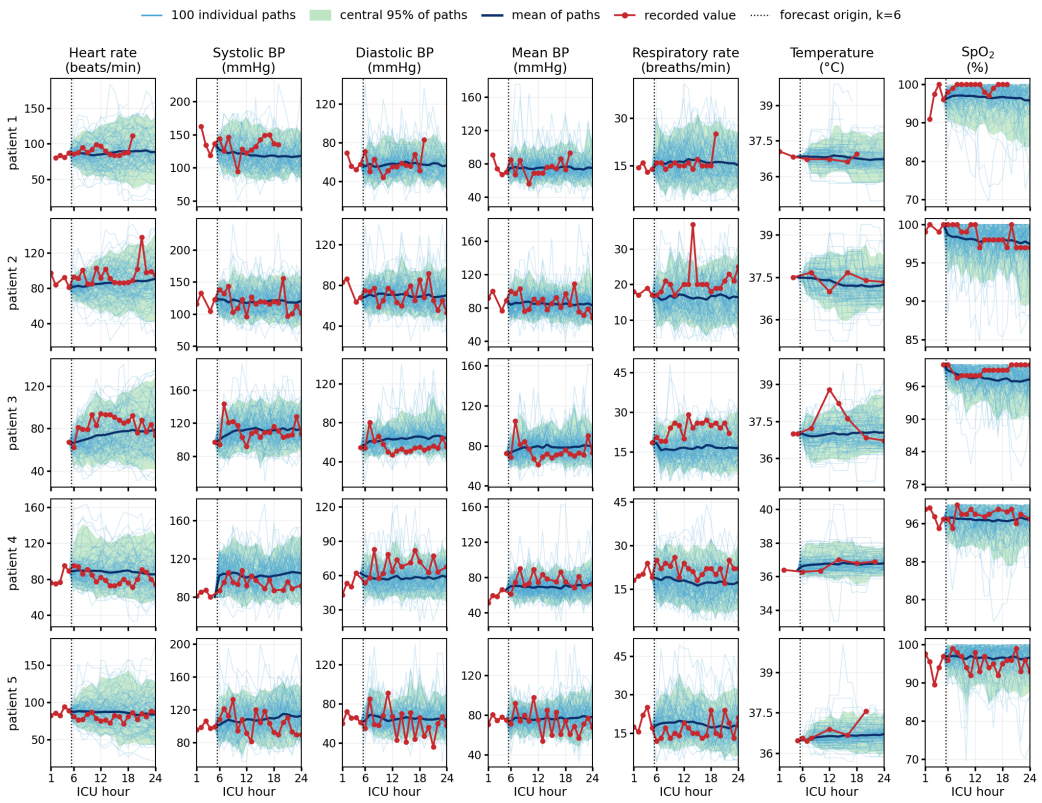


Figure A.1: PULSE-Spline MC100 sampled rollout for five public MIMIC-IV demo patients and seven vital variables. Faint blue lines are individual sampled trajectories; the green band is the empirical 95% interval and the dark line is the MC100 mean.

Table A.19: Treatment-policy examples. Oracle rows use test-hour summaries and are leakage controls.

Variable	Predictor	Split	RMSE/Brier	AUC	F1
Bolus Volume	carry-forward	test	345.7	–	–
Bolus Volume	train-hour mean	test	248.8	–	–
Bolus Volume	oracle test-hour mean	test	248.6	–	–
Vasopressor	carry-forward	test	0.136	0.740	0.557
Vasopressor	train-hour rate	test	0.129	0.548	0.000
Vasopressor	grouped neural policy	validation	–	0.589	0.245

can reward independent derived-channel prediction even when simulator algebra is inconsistent; we treat this as a predictive-check limitation.

A.12 CVSim known-effect summary

The CVSim tables summarize controlled continuations of the grouped anchored simulator. One set emphasizes the first intervention family, one emphasizes the second intervention family, and one increases the weight of paired counterfactual treatment-regime examples during training. These rows test how response supervision changes factual and known-effect performance under a mechanistic simulator with paired counterfactual trajectories.

The repaired grouped NLL anchor has the best factual population RMSE, and the GC2-aware continuation has the best average counterfactual population RMSE among the summarized clean-CVSim rows. This illustrates the expected tension in a semi-synthetic teacher experiment: more counterfactual supervision can improve response under a targeted intervention while degrading factual fit or other regimes.

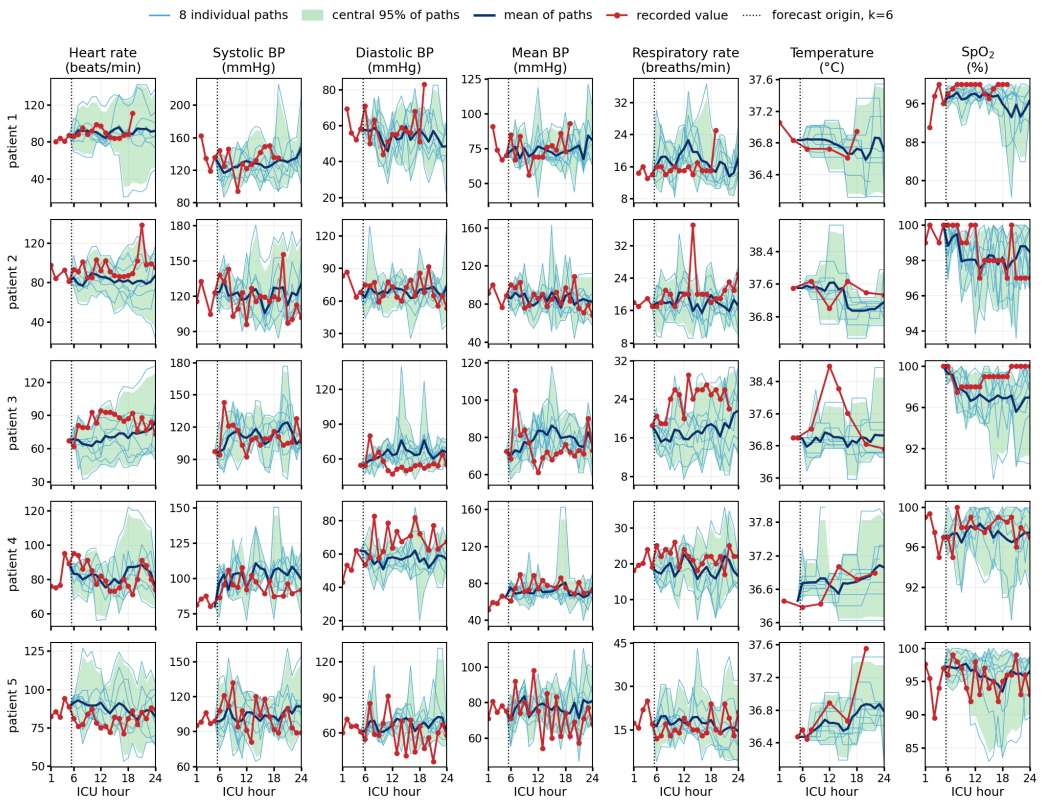


Figure A.2: PULSE-Spline MC8 sampled rollout for the same public-demo patients. With eight trajectories the individual sampled paths remain visually separable, showing how much of the uncertainty band comes from the recursive sampled simulation.

Table A.20: Clean CVSim known-effect summary. These rows use synthetic known-effect trajectories; MIMIC intervention rollouts are evaluated separately.

Model	factual pop.	cf. pop.	factual indiv.	cf. indiv.
Grouped NLL, continued on GC2 regime	4.671	27.649	24.686	38.680
Grouped NLL, counterfactual-heavy mixture	3.955	28.505	24.580	38.967
Grouped NLL, continued on GC1 regime	3.417	28.959	24.120	38.923
Grouped NLL, repaired anchored baseline	2.822	29.050	24.564	39.356
Grouped MSE, repaired anchored baseline	3.835	29.232	24.333	39.807

The corrupted-CVSim table probes the transition from clean mechanistic simulation to observation processes closer to routine EHR data. The intended deployment setting contains irregular measurement, measurement noise, treatment confounding, and censoring. Preserving known counterfactual labels while adding those observation processes tests whether the learned treatment-response mechanism remains usable after the simulator is made less idealized. NLL training improves CRPS in several heteroscedastic and strong-intervention rows; the mixed results support synthetic augmentation as a candidate regularizer requiring dedicated validation.

The MIMIC-seeded experiment separates factual rollout on the shared HR/MAP panel from known-effect response recovery. The response-loss model improves the known-effect metrics while losing part of the factual shared-panel gain, which supports reporting factual fit and response recovery as separate outcomes.

Mapped CVSim known-effect response for the PULSE-CovSpline variant: the model has low RMSE for mild mapped regimes but poor sign accuracy in several regimes, so factual MIMIC rollout and known-effect treatment response are reported as separate axes of evidence.

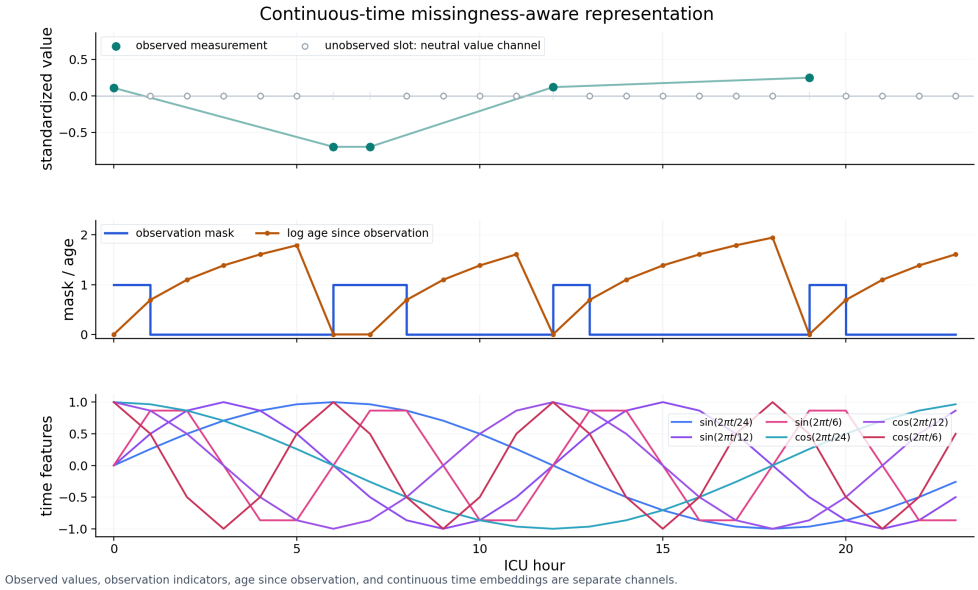


Figure A.3: Event-time input representation for one variable. The MIMIC feature table used here is hourly binned, but the model receives the value, observation indicator, time since the most recent measurement, and continuous-valued clock features separately. This distinguishes a newly observed state from a carried state without treating the hourly index as the only source of time information.

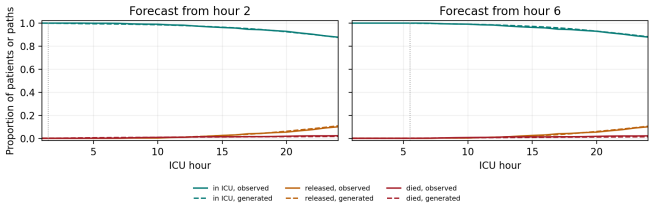
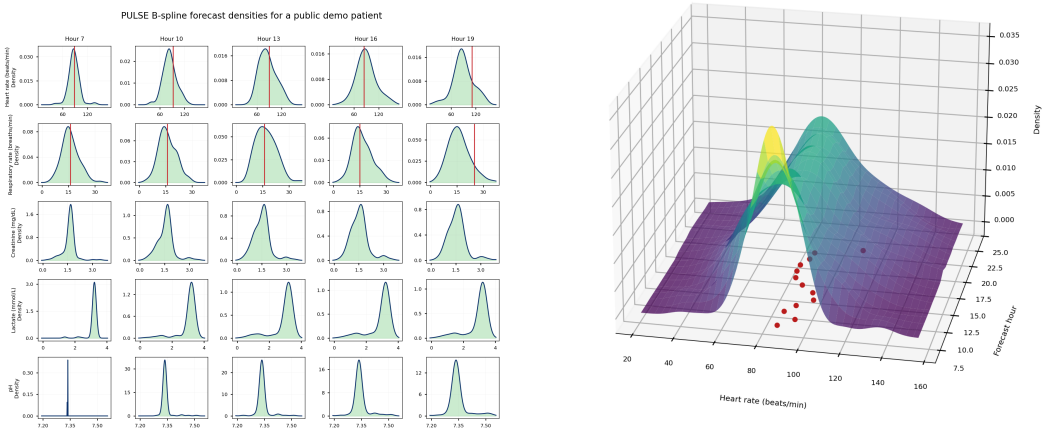


Figure A.4: True and predicted absorbing-state curves. All-hours RMSE penalizes a physiologically plausible rollout if the simulator releases or kills the patient at the wrong time.

A.13 Reproducibility record

Each experimental report records command lines, git identifiers, model identifiers, configuration hashes, random seeds, split hashes, covariate-list hashes, normalization-statistics hashes, database/source metadata, Monte Carlo counts, and metric variants. These records are intended to make the cohort, preprocessing, and rollout definitions auditable without exposing protected source data. The primary PULSE-Spline row uses a covariance-grouped model selected by validation likelihood and evaluated with a train-fitted treatment policy and MC=100.



(a) Forecast density gallery.

(b) Heart-rate forecast densities.

Figure A.5: B-spline forecasts for a public MIMIC-IV demo patient. The gallery and heart-rate manifold summarize 100 sampled paths; dots indicate available demo measurements.

Table A.21: Corrupted CVSim calibration examples. These rows use synthetic known-effect trajectories under EHR-like corruption.

Synthetic scenario	objective	normalized CRPS	cov95
gc2 balance, regime gc1	MSE	0.447	0.957
gc2 balance, regime gc1	NLL	0.459	0.958
gc2 balance, regime gc2v	MSE	0.366	0.992
gc2 balance, regime gc2v	NLL	0.351	0.993
heteroskedastic pilot, regime gc1	MSE	0.420	0.957
heteroskedastic pilot, regime gc1	NLL	0.377	0.963

Table A.22: 2k MIMIC-seeded CVSim response experiment on a shared HR/MAP panel. The paired response fine-tune sacrifices part of the factual shared-panel gain but improves known-effect treatment-response recovery. Lower RMSE is better.

Model	MIMIC $k = 2$	MIMIC $k = 6$	MIMIC-seeded effect	native CVSim effect
MIMIC-only	2.637	1.861	1.736	1.894
MIMIC/CVSim mixed	2.394	1.770	1.754	1.879
MIMIC/CVSim mixed + response loss	2.507	1.814	1.565	1.701

Table A.23: MIMIC-seeded CVSim response experiment. These trajectories initialize a mechanistic simulator from rough MIMIC baseline states and then apply treatment perturbations. The deltas are relative to the observed-policy simulation. Carry-forward has no transition response to altered actions, so it acts as a useful negative control for treatment-response plots.

Regime	Δ MAP	Δ HR	Δ CVP	Δ total blood volume
Withdrawal	-0.198	0.092	-0.555	-73.509
Mild escalation	0.381	-0.151	1.115	148.272
Extreme escalation	9.128	-1.847	11.808	1588.223

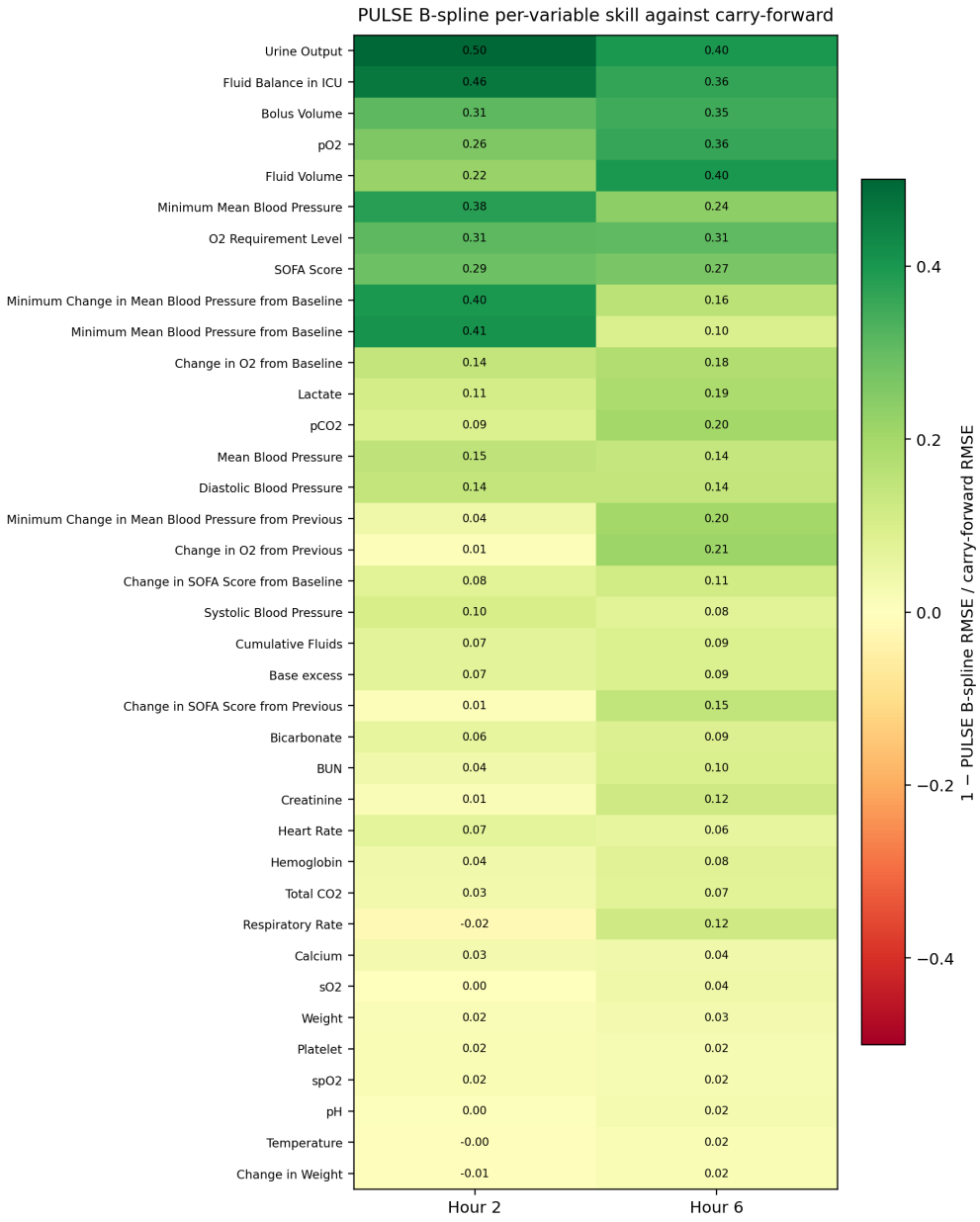


Figure A.6: Per-variable skill heatmap for the PULSE-Spline MC=100 rollout. The all-37 aggregate hides variables where carry-forward remains highly competitive.

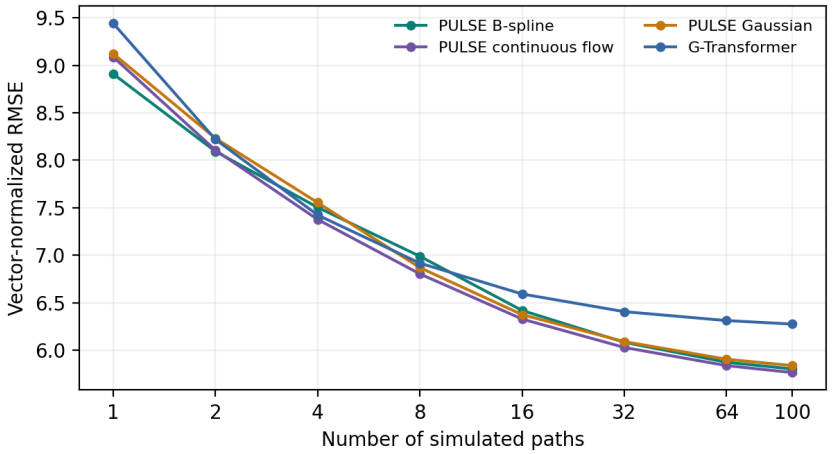


Figure A.7: Effect of Monte Carlo sample count on VRMSE_{37} at $k = 6$ for 891 held-out patients. Each point averages the first M of 100 saved trajectories from the fitted model, using the shared treatment policy and the all-hours scoring rule.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction describe PULSE, the MIMIC predictive check, the factual rollout results, and the CVSim known-effect counterfactual validation scope.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: Limitations are discussed in Section 6, including cohort mismatch, the error from the timing of death and ICU release, and the distinction between factual MIMIC rollout and CVSim known-effect counterfactual validation.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete proof?

   Answer: [N/A].

   Justification: The paper is empirical and methodological; it defines metrics and model parameterizations but does not claim new theoretical results.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all information needed to reproduce the main experimental results to the extent that it affects the claims?

   Answer: [Yes].

   Justification: Section 4 and Appendix A.13 describe the cohort, variables, metrics, rollout definitions, and provenance records. The code describes how to prepare the cohort, train the models, and evaluate their predictions using credentialed-access MIMIC-IV data.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results?

   Answer: [Yes].

   Justification: The code is publicly available at `https://github.com/manscrober/pulse-artifact`. MIMIC-IV and eICU require credentialed access and data-use agreements; the code builds patient caches from the downloadable MIMIC-IV files after import into DuckDB. Trained weights are not included.

6. **Experimental setting/details**

   Question: Does the paper specify all training and test details necessary to understand the results?

   Answer: [Yes].

   Justification: The paper specifies the split, variable panel, horizons, metrics, Monte Carlo aggregation, baseline families, and status/treatment rollout modes.

7. **Experiment statistical significance**

   Question: Does the paper report error bars or other appropriate information about statistical significance?

   Answer: [Yes].

   Justification: Table 2 and its discussion report patient-level paired bootstrap comparisons; Table A.5 gives intervals for the historical evaluation. Figure A.7 shows how error changes with the number of simulated trajectories. The bootstrap intervals describe patient variation for the fitted models.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on compute resources needed to reproduce the experiments?

   Answer: [Yes].

   Justification: Table A.13 reports completed epochs, training runtimes, and average time per epoch for selected models, with hardware specified where recorded.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform with the NeurIPS Code of Ethics?

   Answer: [Yes].

   Justification: The work uses de-identified credentialed clinical data and synthetic simulators for research evaluation; it does not propose deployment or clinical decision use.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive and negative societal impacts?

    Answer: [Yes].

    Justification: The discussion emphasizes that clinical rollout simulators can be misleading if evaluated incorrectly and argues against unsupported counterfactual claims from factual EHR data.

11. **Safeguards**

    Question: Does the paper describe safeguards for responsible release of high-risk data or models?

    Answer: [N/A].

    Justification: The paper does not introduce a deployable clinical model or release patient data; MIMIC-IV remains governed by its existing access controls.

12. **Licenses for existing assets**

    Question: Are existing assets properly credited and are license/terms respected?

    Answer: [Yes].

    Justification: The paper cites MIMIC-IV, eICU, and the sources of the baseline methods. Access to the clinical datasets is governed by PhysioNet credentialing and data-use agreements.

13. **New assets**

    Question: Are new assets introduced in the paper well documented?

    Answer: [Yes].

    Justification: The public code repository includes a README describing the model, training commands, and data access, together with a license.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include relevant details?

    Answer: [N/A].

    Justification: No crowdsourcing or prospective human-subject experiment is conducted.

15. **Institutional review board approvals or equivalent**

    Question: Does the paper describe risks, disclosures, and IRB approvals or equivalent review for human-subject research?

    Answer: [N/A].

    Justification: The empirical work uses an existing de-identified public critical-care dataset under its access requirements; no prospective data collection is performed.

16. **Declaration of LLM usage**

Question: Does the paper describe LLM usage if it is an important, original, or non-standard component of the core methods?

Answer: [N/A].

Justification: LLMs are not part of the model, evaluator, or experimental method described in the paper.