LEARNING CONTINUOUS PATIENT TRAJECTORIES FROM ELECTRONIC HEALTH RECORDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Electronic health records provide irregular observations of latent patient states that evolve continuously over time. Recent autoregressive models condition on clinical histories to forecast future events as sequences of discrete observations. Conversely, multi-marginal flow matching provides a continuous-time formulation, but using multiple observations to supervise training paths does not itself give the learned dynamics access to preceding patient history. We introduce `EHRFlow`, a multi-marginal flow-matching framework that conditions on encoded patient history, thereby allowing future dynamics to depend on the patient's prior clinical trajectory. Our proposed framework accommodates irregular observation times and supports forecasting at arbitrary horizons. Across controlled synthetic benchmarks, `EHRFlow` improves clinical-code forecasting and latent-state recovery. On real-world clinical datasets comprising more than one million patients, including an independent external validation cohort, `EHRFlow` improves horizon-averaged top-5 clinical-code accuracy over autoregressive and history-independent flow-matching baselines. Finally, in a controlled counterfactual simulation, conditional guidance approximates the known effect of an antihypertensive intervention without training a task-specific outcome model.

1 INTRODUCTION

Forecasting the future health trajectories of individual patients would enable more personalised care by anticipating future patient states and medically relevant events. Recent advancements in autoregressive and generative machine learning have demonstrated the ability to learn from patient medical history using electronic health records (EHRs) to generate sequences of future diagnoses, treatments, and measurements (Choi et al., 2016b; Zhang et al., 2026; Shmatko et al., 2025).

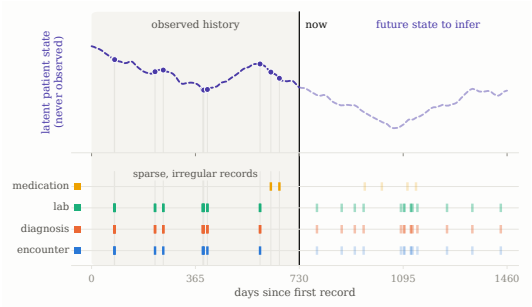


Figure 1: Schematic of an unobserved, continuously evolving patient state (top) and sparse EHR observations (bottom).

Despite these developments, trajectory forecasting remains an unresolved problem in clinical machine learning (Goldstein et al., 2017). Modeling patient trajectories is highly challenging as the data are longitudinal, high-dimensional, mixed-type, and irregularly sampled. Furthermore, patients differ in observation schedules and rates of clinical progression, without a common temporal reference point. Additionally, forecasting future patient states requires knowledge of their clinical history, as patients in similar current states may have different futures depending on their preceding disease course, treatment exposure, or rate of progression (Swain et al., 2003; Coresh et al., 2014). EHRs provide only sparse, partial observations of a patient state that continues to evolve between records. This motivates modeling continuous latent trajectories conditioned on clinical history to infer future states (Figure 1).

Several existing works have tried to address these challenges. Autoregressive models condition on patient history but predominantly represent clinical data as discrete events (Shmatko et al., 2025; Waxler et al., 2025; Kraljevic et al., 2024). As a result, these approaches do not explicitly repre-

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.

sent the continuously evolving patient state between recorded events, and forecasting at a distant horizon typically requires sequential rollout generation, which may accumulate prediction errors. Conversely, continuous-time models capture evolving patient dynamics. Neural ordinary differential equations (ODEs) parameterise dynamics with a neural vector field (Chen et al., 2018), and continuous normalizing flows use these dynamics to transport probability distributions over time (Grathwohl et al., 2019).

Flow matching proposes learning the vector field by regressing directly onto target velocities along prescribed probability paths (Lipman et al., 2023), avoiding numerical integration during training. However, flow matching methods typically rely on paired samples to construct a probability path and require specifying how observations across time are coupled. For instance, in many biological settings, observations are collected as unpaired population snapshots along a shared experimental timeline, thus requiring trajectory reconstruction such as through optimal transport (Schiebinger et al., 2019; Tong et al., 2020). EHR data, meanwhile, provide the opportunity to leverage multiple observations from the same patient when learning these dynamics. Recent multi-marginal and stream-level flow-matching approaches use multiple observed states to construct training paths (Rohbeck et al., 2025; Wei & Ma, 2025), with applications to longitudinal medical imaging (Islam et al., 2025). However, in these works, learned vector fields are parameterised by only the current patient state and temporal information, without access to the patient's preceding history. Related work by Zhang et al. (2024) demonstrated that conditioning the vector field on the most recent observations can improve performance. However, their work focused on low-dimensional clinical time series data and formulated the target flow using pairs rather than multiple trajectory observations, as in multi-marginal approaches.

We introduce `EHRFlow`, a history-conditioned multi-marginal flow-matching framework for continuous latent-state forecasting from longitudinal and irregularly sampled EHRs. Our framework uses multiple observations from each patient to supervise velocity during training, and conditions the vector field on encoded patient history and the sampled calendar duration, thus preserving relative observation times. At inference, the field is integrated from the encoded anchor state, conditioned on the preceding history, to forecast an arbitrary horizon.

Contributions. Our core contributions are threefold:

1. **Conceptually**, we formulate longitudinal EHR forecasting as learning continuous latent patient trajectories and provide a theoretical illustration of why matching population marginals alone does not necessarily recover individual trajectories.
2. **Methodologically**, we introduce `EHRFlow`, extending multi-marginal flow matching with explicit history conditioning and a temporal formulation that accommodates irregular observations and arbitrary forecast horizons.
3. **Empirically**, we demonstrate across both synthetic benchmarks and real-world EHR datasets, comprising more than one million patients and including an independent external validation cohort, that our method improves forecasting performance over autoregressive and continuous-time baselines. We further recover a known treatment effect through intervention-conditioned forecasting in a controlled simulation.

2 Related Work

Sequence models for clinical data. Early RNN-based approaches such as RETAIN (Choi et al., 2016a) and DoctorAI (Choi et al., 2016b) used patient histories to predict clinical outcomes. Transformers enabled richer longitudinal representations, including BEHRT (Li et al., 2020), MOTOR (Steinberg et al., 2024), and APOLLO (Zhang et al., 2026). Foresight (Kraljevic et al., 2024), ETHOS (Renc et al., 2024), DELPHI-2M (Shmatko et al., 2025), and COMET (Waxler et al., 2025) extend next-token prediction to generate clinical timelines, while SurvivEHR (Gadd et al., 2026) uses a competing-risks framework. These models accommodate irregular timing through temporal encodings or time-to-event predictions, but generally represent evolution through discrete events. `EHRFlow` instead uses longitudinal context to condition a continuously evolving latent patient state.

Continuous-time and state-space models. Neural ODEs (Chen et al., 2018) and Latent ODEs (Rubanova et al., 2019) learn continuous dynamics, while Neural CDEs (Kidger et al., 2020), GRU-ODE-Bayes (De Brouwer et al., 2019) and ContiFormer (Chen et al., 2023) incorporate irregular

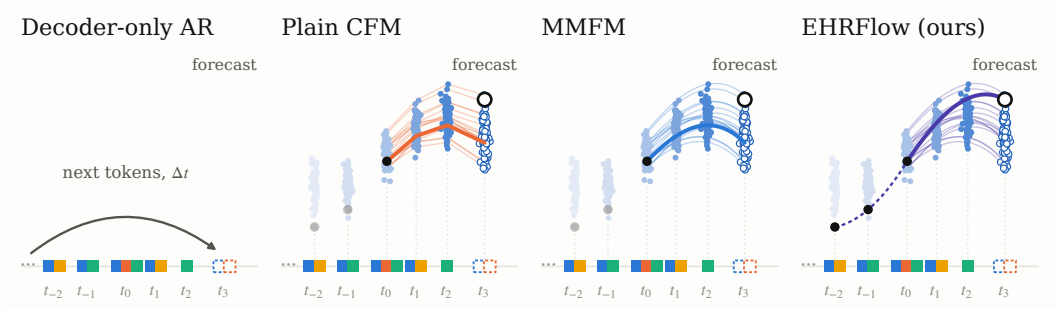


Figure 2: Autoregressive models condition on the observed sequence but predict future events discretely. Conditional flow matching (CFM) models have pairwise latent transitions. Multi-marginal flow matching (MMFM) uses multiple observations to define the training trajectory, and `EHRFlow` additionally conditions the learned vector field on the preceding patient history.

observations into evolving hidden states. For models trained through numerical integration, learning from long records across large patient populations can be computationally demanding. Attentive state-space models capture dependence on preceding latent states (Alaa & van der Schaar, 2019), while TrajSurv learns continuous latent trajectories for survival prediction (Zeng et al., 2025). `EHRFlow` instead learns from directly supervised velocities along paths constructed from multiple observations, avoiding trajectory integration during flow training.

Flow matching for continuous dynamics. Flow matching learns continuous vector fields through simulation-free regression onto conditional velocity targets (Lipman et al., 2023; Albergo & Vanden-Eijnden, 2023), with extensions incorporating data geometry (Kapusniak et al., 2024) and applications to probabilistic forecasting (Kollovieh et al., 2025). Trajectory Flow Matching (Zhang et al., 2024) conditions on only the most recent observations for clinical time-series modelling, using pairwise flow supervision. Multi-marginal (Rohbeck et al., 2025) and Gaussian-process (Wei & Ma, 2025) formulations instead use multiple observations to supervise training paths, but do not condition the learned field on the preceding patient history. `EHRFlow` builds on these multi-observation path constructions by explicitly conditioning the vector field on the patient history available at the forecast anchor. Its latent formulation supports forecasts of both continuous and discrete clinical tokens at arbitrary horizons. Future observations supervise the training path, while preceding observations provide inference-time context (see Figure 2).

3 Problem Formulation and Preliminaries

3.1 Observed and Latent Patient States

We assume that the clinical state of a patient evolves continuously over time and that EHRs provide sparse, irregular, and incomplete observations of this underlying process. For patient i, we observe a time-stamped sequence

$$\mathcal{X}_i = \{(\tau_{i,j}, x_{i,j})\}_{j=1}^{m_i}, \qquad \tau_{i,1} < \cdots < \tau_{i,m_i}, \tag{1}$$

where $x_{i,j} \in \mathbb{R}^{d_x}$ denotes the feature vector constructed from patient i's recorded clinical information at time $\tau_{i,j}$. Each feature vector aggregates all categorical event indicators, continuous-valued measurements, and other observation indicators possible to observe at a single time point. We encode each observation in the sequence into a lower-dimensional continuous-valued latent representation,

$$z_{i,j} = E_\psi(x_{i,j}) \in \mathbb{R}^{d_z}, \qquad \mathcal{Z}_i = \{(\tau_{i,j}, z_{i,j})\}_{j=1}^{m_i}, \tag{2}$$

where $\mathcal{Z}_i$ is the corresponding time-stamped latent sequence. A decoder D_ϕ maps latent states to predictions in the observation space.

For patient i, let $a \in \{1, \ldots, m_i\}$ index an anchor observation $x_{i,a}$ at time $\tau_{i,a}$. We denote the patient history available at and before the anchor as $\mathcal{H}_{i,a} = \{(\tau_{i,j}, x_{i,j})\}_{j=1}^{a}$. Given a forecast horizon $\Delta > 0$, our objective is to predict the future observed patient state at time $\tau_{i,a} + \Delta$, using the encoded anchor state $z_{i,a}$ and the observed history $\mathcal{H}_{i,a}$. We model this trajectory evolution in latent space using a neural vector field, which specifies the instantaneous rate of change of the latent

state. Integrating the field produces a latent forecast, which is then decoded into predictions of EHR observations. For notational simplicity, we now focus on a generic patient trajectory and omit the patient index i.

3.2 Trajectory modeling with conditional flow matching

We first consider two encoded observations z_{s_0} and z_{s_1}, which are sampled from the same patient at calendar time τ_{s_0}, τ_{s_1} and separated by $\Delta = \tau_{s_1} - \tau_{s_0}$ calendar days. Conditional flow matching (Lipman et al., 2023; Albergo & Vanden-Eijnden, 2023) constructs a conditional probability path between them over normalised flow time $s \in [0, 1]$. A common construction is

$$z_s = \mu_s(z_{s_0}, z_{s_1}) + \sigma_s \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \qquad s \sim \mathcal{U}(0, 1), \tag{3}$$

where μ_s and σ_s specify the mean interpolation and noise scale. Holding the sampled ϵ fixed, differentiation with respect to s gives the conditional velocity

$$u_s = \frac{\partial \mu_s(z_{s_0}, z_{s_1})}{\partial s} + \frac{\partial \sigma_s}{\partial s}\epsilon. \tag{4}$$

For example, the linear interpolant $\mu_s(z_{s_0}, z_{s_1}) = (1 - s) \cdot z_{s_0} + s \cdot z_{s_1}$ with $\sigma_s = 0$ yields $u_s = z_{s_1} - z_{s_0}$. To accommodate observation pairs separated by different calendar durations, we condition the vector field on $\lambda = \log(1 + \Delta)$ and train it to predict the conditional velocity:

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{(z_{s_0}, z_{s_1}, \Delta) \sim \pi, s, \epsilon} \left[\| v_\theta(z_s, s, \lambda) - u_s \|_2^2 \right], \tag{5}$$

where π is the empirical distribution over all pairs of consecutive observations belonging to the same patient, across all patients. Because longitudinal records provide a natural correspondence between observations, no optimal-transport coupling between patients is required.

Note that normalised flow time s denotes relative position within an observation window, which maps to calendar time $\tau := \tau_{s_0} + s \cdot \Delta$. The target u_s is a velocity with respect to normalised flow time. For an interpolated z, we have $\mathrm{d}z/\mathrm{d}s = \Delta \mathrm{d}z/\mathrm{d}\tau$, thereby motivating the need to condition the vector field on calendar distance Δ to distinguish different window durations with respect to calendar days.

3.3 Trajectory modeling with multi-marginal flow matching

Pairwise conditional flow matching constructs a training path using only two observed patient states. However, EHRs contain many intermediate observations that provide necessary information about the underlying clinical dynamics. We therefore adapt multi-marginal flow matching (Rohbeck et al., 2025) to construct the training path jointly from multiple, irregularly sampled observations. Let us consider $K \geq 2$ encoded observations and calendar times from a single patient's trajectory starting at τ_j as $\{(\tau_k, z_k)\}_{k=j}^{K+j-1}$, with $\tau_j < \cdots < \tau_{K+j-1}$. Let $\Delta = \tau_{K+j-1} - \tau_j$ be the window's span of total calendar days and $\lambda = \log(1 + \Delta)$. We first normalise their time-stamps into normalised flow time:

$$t_k = \frac{\tau_k - \tau_j}{\Delta}, \qquad 0 = t_j < \cdots < t_{K+j-1} = 1. \tag{6}$$

Let $\mathcal{W} = \{(t_k, z_k)\}_{k=j}^{K+j-1}$ denote the sampled window. To accommodate both deterministic and stochastic constructions of interpolants connecting the observed states, we write

$$z_s = I_s(\mathcal{W}, \xi), \qquad u_s = \frac{\partial I_s(\mathcal{W}, \xi)}{\partial s}, \tag{7}$$

where ξ denotes the sampled path randomness and is held fixed when differentiating, and the interpolant must pass through the observed states, i.e., $I_{t_k}(\mathcal{W}, \xi) = z_k$, for $k = j, \ldots, K + j - 1$. This yields the multi-marginal flow-matching objective

$$\mathcal{L}_{\text{MMFM}} = \mathbb{E}_{(\mathcal{W}, \Delta) \sim \pi, s, \xi} \left[\| v_\theta(z_s, s, \lambda) - u_s \|_2^2 \right], \tag{8}$$

where π denotes the empirical distribution over windows pooled across patients, each patient contributing up to a fixed number of windows drawn uniformly from its time-window candidates. Note that this formulation is agnostic to the interpolant family.

4 EHRFLOW

Although multi-marginal formulations incorporate multiple trajectory observations into training paths, this supervision does not itself provide the learned vector field with access to patient history (Rohbeck et al., 2025). Related work conditions flow dynamics on recent measurements (Zhang et al., 2024), but models only pairwise transitions in observation space and conditions on the most recent observations rather than the broader longitudinal history.

History conditioning. To address this gap, we consider the patient's observed history $\mathcal{H}_a = \{(\tau_l, x_l)\}_{l=1}^{a}$, where a indexes the first observation of the sampled window $\mathcal{W}$, so that $\tau_a = \tau_j$. Thus, the history $\mathcal{H}_a$ includes only observations up to and including the anchored observation. We propose to condition the vector field on a fixed-dimensional representation of the patient's anchored history:

$$h_a = g_\omega(\mathcal{H}_a) \in \mathbb{R}^{d_h}, \tag{9}$$

where g_ω is a learned history encoder. During flow training, g_ω is frozen and h_a is held fixed along the entire path, independently of s.

We provide theoretical motivation for history-conditioning in Appendix B. First, we show that conditioning on history cannot increase the optimal flow-matching loss, and that the improvement is strict whenever history changes the expected velocity given the current state (Theorem 1). Second, in a linear-Gaussian example with noisy history, both history-independent and history-conditioned fields recover the correct population marginals, but history-conditioning yields strictly lower individual forecast error. This error decreases as the history becomes more informative and vanishes in the noiseless limit (Theorem 2). Figure 3 illustrates Theorem 2, showing how history conditioning distinguishes trajectories that share the same current state.

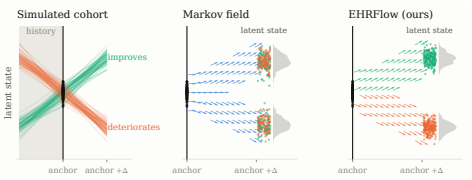


Figure 3: Patients with similar current states may have different future trajectories depending on how they reached those states.

Optimisation. We pretrain the encoder and decoder using a modified variational autoencoder objective (Kingma & Welling, 2013), where the reconstruction loss is adapted for both categorical and continuous variables. We separately pretrain the history encoder g_ω using a GRU (Chung et al., 2014) to predict the clinical event recorded at the patient's next observation from the preceding K observations. During flow training, the observation encoder E_ψ and history encoder g_ω are frozen, while the vector field v_θ and decoder D_ϕ are trained jointly. Training windows begin at a randomly sampled anchor observation and end at the first observation at least M days later, where M is drawn from a fixed set of forecast horizons. We minimise

$$\begin{aligned}\mathcal{L}_{\texttt{EHRFlow}} = \mathbb{E}_{(\mathcal{W},\Delta,\mathcal{H}_a)\sim\pi,\, s,\, \xi}\Big[&\|v_\theta(z_s, s, \lambda, h_a) - u_s\|_2^2 \\ &+ \ell_{\text{rec}}(D_\phi(z_s + (1-s)\cdot v_\theta(z_s, s, \lambda, h_a)),\, x_{K+j-1})\Big],\end{aligned} \tag{10}$$

where x_{K+j-1} is the final observation of the window $\mathcal{W}$ and $\hat{z}_1 = z_s + (1-s)\cdot v_\theta(z_s, s, \lambda, h_a)$ is a single-step approximation to the latent endpoint. The reconstruction loss ℓ_{rec} combines cross-entropy for categorical variables and squared error for continuous variables, and helps supervise the vector field and decoder through this endpoint surrogate without numerical integration.

Inference. At inference, no future observations are required. Given a patient's observable history $\mathcal{H}_a$ up until the most recent observation x_a and the requested forecast horizon Δ in calendar days, we compute the encoded history representation h_a and then integrate the learned history-conditioned vector field over the normalised flow time $s \in [0, 1]$:

$$\frac{\mathrm{d}\hat{z}_s}{\mathrm{d}s} = v_\theta(\hat{z}_s, s, \lambda, h_a)\,, \qquad \hat{z}_0 = z_a = E_\psi(x_a), \tag{11}$$

where $\lambda = \log(1+\Delta)$. The ODE solution map Φ_θ defines the predicted latent state at calendar time $\tau_a + \Delta$ and corresponding predicted observation state:

$$\hat{z}_1 = \Phi_\theta(z_a, h_a, \lambda), \qquad \hat{x}(\tau_a + \Delta) = D_\phi(\hat{z}_1). \tag{12}$$

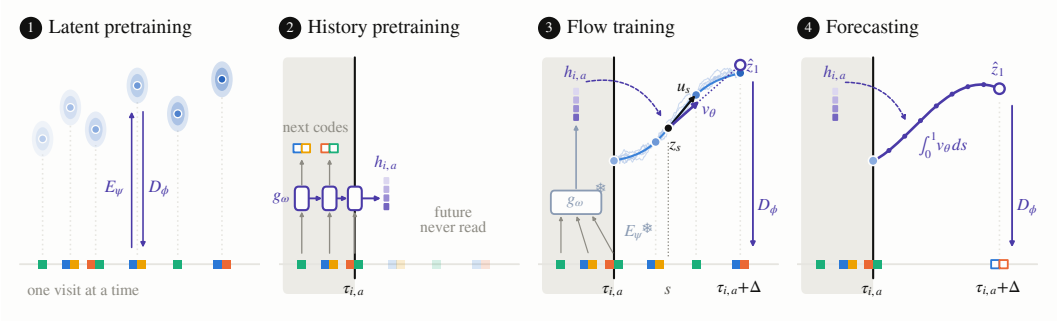


Figure 4: **Training and forecasting with `EHRFlow`.** (1) An encoder and decoder learn visit-level latent representations. (2) A history encoder is pretrained on observations up to the forecast anchor. (3) Both encoders are frozen. Future observations define paths and target velocities for training the history-conditioned field. (4) At inference, the field is integrated from the encoded anchor and decoded into clinical predictions.

We report `EHRFlow` (GP-CFM) for Gaussian-process interpolants (Wei & Ma, 2025) and `EHRFlow` (MMFM) for natural cubic-spline interpolants with bridge noise (Rohbeck et al., 2025). Full objectives, architectures, and optimisation details are provided in Appendix D.5. The training and inference procedure is summarized in Figure 4 .

5 Experiments

We evaluate `EHRFlow` across controlled synthetic benchmarks and large-scale longitudinal EHR datasets. Our evaluation is structured around four questions:

(i) **Trajectory forecasting:** How accurately can flow-based models forecast future clinical observations compared with autoregressive and continuous-time sequence models?

(ii) **Real-world generalisation:** Do the benefits extend to primary- and secondary-care EHRs, including an independent external cohort?

(iii) **History conditioning and optimisation**: Which aspects of historical conditioning and training contribute to forecasting performance?

(iv) **Intervention modeling:** Can the learned dynamics be conditionally guided to recover a known intervention effect in a controlled simulation?

5.1 Experimental setup

Datasets and tasks. We evaluate on eight datasets total, spanning from fully synthetic to real-world clinical observations. First, we validate on five controlled synthetic datasets, each with 20,000 simulated patients, where we vary the data-generating mechanism to model five distinct disease trajectories: *Anthracycline*, which models cumulative treatment exposure; *RA on DMARD*, which models delayed treatment response; *CKD*, which models patient-specific progression rates; *RRMS*, which models relapse history; and *Heart failure*, which models a Markovian worsening trajectory (see Appendix D.1). Second, we validate on 23,072 patients generated using the clinical simulator Synthea (Walonoski et al., 2018). Finally, we validate on two real-world EHR datasets each containing over 500 million clinical events. The AMC-EHR dataset contains 647,447 patients treated at a large academic medical center (AMC) in the United States. The second dataset contains two separate cohorts provided by the United Kingdom Clinical Practice Research Datalink (CPRD) (Herrett et al., 2015). The CPRD dataset provides a natural training and testing split, where the training "Aurum" cohort comprises 815,092 patients drawn from primary-care practices and the externally validated "Gold" cohort comprises 53,599 patients from a separate set of practices and EHR software. Full cohort definitions, preprocessing, and split procedures are provided in Appendix D. We evaluate forecasting at $\Delta \in \{90, 180, 365, 730, 1095\}$ days from the observed anchor. Our primary metric is top-5 clinical-code accuracy, which measures whether the recorded target code appears among the five highest-ranked predictions. This accommodates the possibility that patients have several concurrent conditions, only some of which are diagnosed or recorded at a given time. Top-1 accuracy and macro-AUROC across horizons are reported in Appendix C.

Table 1: Top-5 clinical-code accuracy across synthetic benchmarks, averaged over forecast horizons. Entries show mean ± standard deviation across runs.

	Anthracycline	RA on DMARD	CKD	RRMS	Heart failure	Synthea
Decoder-only	0.510 ± 0.017	0.821 ± 0.010	0.614 ± 0.008	0.600 ± 0.023	0.405 ± 0.019	0.432 ± 0.029
Encoder-decoder	0.507 ± 0.015	0.860 ± 0.007	0.668 ± 0.016	0.593 ± 0.013	0.403 ± 0.019	0.278 ± 0.121
GRU-ODE-Bayes	0.546 ± 0.022	0.829 ± 0.042	0.656 ± 0.022	0.524 ± 0.026	0.386 ± 0.064	0.439 ± 0.052
Neural CDE	0.411 ± 0.066	0.729 ± 0.061	0.577 ± 0.069	0.491 ± 0.044	0.305 ± 0.084	0.601 ± 0.027
CFM (adjacent)	0.574 ± 0.035	0.971 ± 0.003	0.812 ± 0.014	0.781 ± 0.024	0.602 ± 0.026	0.790 ± 0.014
Metric FM	0.597 ± 0.036	0.963 ± 0.004	0.789 ± 0.021	0.759 ± 0.062	0.627 ± 0.053	0.766 ± 0.008
GP-CFM	0.779 ± 0.004	0.977 ± 0.001	0.912 ± 0.003	0.895 ± 0.005	0.770 ± 0.005	$\mathbf{0.875} \pm 0.004$
MMFM	0.785 ± 0.003	0.980 ± 0.001	0.913 ± 0.004	0.896 ± 0.003	0.767 ± 0.004	0.874 ± 0.003
`EHRFlow` (GP-CFM)	0.863 ± 0.007	$\mathbf{0.986} \pm 0.002$	$\mathbf{0.941} \pm 0.001$	$\mathbf{0.944} \pm 0.005$	$\mathbf{0.787} \pm 0.004$	$\mathbf{0.875} \pm 0.003$
`EHRFlow` (MMFM)	$\mathbf{0.865} \pm 0.007$	0.985 ± 0.002	0.940 ± 0.004	0.942 ± 0.004	0.775 ± 0.009	$\mathbf{0.875} \pm 0.003$

Baselines. We compare against three families of clinical forecasting methods: first, two decoder-only and encoder–decoder autoregressive transformers (Shmatko et al., 2025; Zhang et al., 2026) that directly predict future event tokens; second, GRU-ODE-Bayes (De Brouwer et al., 2019) and Neural CDEs (Kidger et al., 2020) which model irregular observations through continuous hidden-state dynamics; and third, four flow-based baselines – conditional flow matching (CFM) (Lipman et al., 2023), Metric FM (Kapusniak et al., 2024), Gaussian-process stream-level flow matching (GP-CFM) (Wei & Ma, 2025), and multi-marginal flow matching (MMFM) (Rohbeck et al., 2025). The flow-based methods use a common observation representation and vector-field architecture, with an additional geopath network for Metric FM and additional history inputs for `EHRFlow`. Autoregressive baselines are closely matched in parameter count to the flow-based methods, with only small residual differences due to architectural design. Architectural details, training objectives, and inference procedures are provided in Appendix D.6.

5.2 Synthetic trajectory forecasting

We first evaluate across the six synthetic datasets modeling different clinical trajectories. Multi-marginal methods outperform adjacent CFM and Metric FM, with similar results across Gaussian-process and spline interpolants (Table 1). Our results highlight that history conditioning provides further gains, particularly for cumulative exposure and relapse history. In the aggregate top-5 comparison, both `EHRFlow` variants outperform all baselines under a paired, stratified sign-flip randomisation test with Benjamini–Hochberg correction ($q < 10^{-5}$). On Synthea, top-5 accuracy remains comparable to history-independent multi-marginal models, although history conditioning improves macro-AUROC at longer horizons (see Appendix C).

Furthermore, we found that the benefit of history conditioning depends on the evaluation metric. On the Markovian heart-failure dataset, history conditioning provides smaller and less consistent gains than on datasets with explicitly history-dependent dynamics, consistent with history primarily helping recover the underlying state from incomplete observations. We additionally evaluate latent-state recovery against the known simulated latent state at forecasted time-points unobserved in the training data (Appendix C). Predictions at the same time point obtained directly or within longer rollouts show similar accuracy and 93–94% mean top-5 overlap, supporting empirical trajectory consistency (see Table 7). `EHRFlow` achieves the lowest latent-state recovery error across synthetic benchmarks, while separate off-grid evaluations demonstrate generalisation to unseen forecast horizons (see Figure 12). Computational cost is also reduced: synthetic training and evaluation runs take 0.18–0.76 hours with `EHRFlow`, compared with 10.02–56.94 hours for autoregressive baselines (see Table 10).

Table 2: Top-5 clinical-code accuracy on real-world CPRD and AMC-EHR datasets, averaged over forecast horizons.

	CPRD	AMC-EHR
Decoder-only	0.368 ± 0.012	0.259 ± 0.008
Encoder-decoder	0.335 ± 0.001	0.232 ± 0.037
GP-CFM	0.554 ± 0.021	0.275 ± 0.004
MMFM	0.548 ± 0.014	0.279 ± 0.004
`EHRFlow` (GP-CFM)	$\mathbf{0.638} \pm 0.009$	0.283 ± 0.013
`EHRFlow` (MMFM)	0.621 ± 0.021	$\mathbf{0.302} \pm 0.016$

5.3 Real-world trajectory forecasting

Our results demonstrate that history conditioning also significantly improves horizon-averaged top-5 accuracy in real-world settings ($q < 0.05$, BH-corrected; Table 2). On the CPRD dataset, `EHRFlow` improves over its history-independent counterparts by 8.4 percentage points with GP-CFM interpolation and 7.3 points with MMFM. On the AMC-EHR dataset, the corresponding gains are 0.8

and 2.3 points. Meanwhile, the comparison with autoregressive methods depends on the metric and forecast horizon. Both EHRFlow variants outperform the autoregressive baselines on horizon-averaged top-5 accuracy. For short horizons, we observe autoregressive models achieve higher macro-AUROC on AMC-EHR, whereas EHRFlow (MMFM) performs best at longer horizons (Appendix C). This pattern likely reflects the error accumulation that occurs during autoregressive rollout. The gains on both primary- and secondary-care records indicate that the benefits of history conditioning extend across healthcare settings, including the independent CPRD Gold cohort used for external validation.

5.4 History-conditioning and optimisation ablations

We conduct targeted ablations to isolate the contribution of history conditioning, its incorporation, and key optimisation choices (see Table 3).

Parameter matching. Because conditioning the vector field on a history representation increases the input dimensionality of the MLP, EHRFlow contains slightly more parameters than the corresponding history-independent GP-CFM and MMFM baselines. We therefore reduce the width of the vector-field MLP to match the baselines' parameter count. The resulting performance is largely unchanged, indicating that our method's improvement does not arise simply from increased model capacity. As a complementary control, we retain the full EHRFlow architecture but replace the history representation with zeros. Performance then returns to that of the history-independent baselines.

How history is incorporated. We next show that using the history representation as the flow source, regardless of whether we additionally condition on history, performs worse than our choice of using the current patient state as the source while conditioning on history. We also show that training the history encoder jointly with the flow objective, instead of pretraining and freezing it, performs worse than using a pretrained history representation. Fully end-to-end optimisation of the latent encoder, decoder, history encoder, and vector field likewise reduces accuracy, supporting the use of pretrained representations under the evaluated optimisation settings. We also found that removing the auxiliary decoder loss, thereby freezing it, substantially reduces accuracy, highlighting the importance of observation-space supervision during flow training. Finally, adding history conditioning to adjacent CFM also improves forecasting, but remains below the multi-marginal models. These ablations indicate that the gains of EHRFlow arise from both explicit historical conditioning and multi-observation trajectory supervision. Dataset-level results are provided in Appendix C. We additionally vary the history-window length to assess sensitivity to the amount of preceding context.

Table 3: EHRFlow ablations: top-5 clinical-code accuracy averaged across synthetic datasets and forecast horizons.

	Average top-5 acc.	
	GP-CFM	MMFM
EHRFlow	**0.899** $\pm$ 0.001	**0.897** $\pm$ 0.002
EHRFlow matched	0.897 $\pm$ 0.002	**0.897** $\pm$ 0.002
EHRFlow, zero context	0.868 $\pm$ 0.002	0.869 $\pm$ 0.001
EHRFlow, end-to-end GRU	0.873 $\pm$ 0.004	0.878 $\pm$ 0.003
EHRFlow, fully end-to-end	0.876 $\pm$ 0.008	0.865 $\pm$ 0.004
EHRFlow, no decoder loss	0.599 $\pm$ 0.013	0.606 $\pm$ 0.011
History as source	0.865 $\pm$ 0.003	0.867 $\pm$ 0.003
History as source + context	0.864 $\pm$ 0.007	0.865 $\pm$ 0.003
No history	0.868 $\pm$ 0.001	0.869 $\pm$ 0.002
CFM (adjacent) + history	0.776 $\pm$ 0.005	
CFM (adjacent)	0.755 $\pm$ 0.007	
CFM (MMFM windows) + history	0.808 $\pm$ 0.007	

5.5 Guided intervention modelling

As an additional controlled experiment, we test whether EHRFlow captures intervention-dependent patient dynamics in a simulation with known ground truth.

Evaluation protocol. We construct a Synthea hypertension experiment in which the same patient population is simulated twice from identical initial random states, where one population has antihypertensive prescriptions suppressed. Treatment is defined as antihypertensive initiation and the outcome as systolic blood pressure, with simulator-defined treatment effect as $\theta = -20.48$ mmHg. We condition EHRFlow on treatment assignment using classifier-free guidance (Ho & Salimans, 2022) and evaluate both matched trajectories, where both treatment conditions are available for each patient, and a split-arm setting, where each patient contributes to only one condition. The guidance weight is selected using validation outcome-prediction error without access to θ.

Results. Guided `EHRFlow` approximates the simulator-defined effect with an average error of 0.07–2.07 mmHg across both models and evaluation settings, equivalent to a relative error of 0.3%–10.1% (Table 4). These results suggest that the learned latent trajectories can preserve clinically meaningful changes in individual physiological variables, as demonstrated here for blood pressure under a simulated intervention.

Table 4: Recovery of the simulator-defined hypertension effect with CFG.

Design	Observed	Guided	Error
EHRFlow (GP-CFM)			
Paired	-23.20 ± 0.59	-19.42 ± 2.33	1.06 ± 2.33
Split-arm	-22.58 ± 0.59	-18.41 ± 1.62	2.07 ± 1.62
EHRFlow (MMFM)			
Paired	-23.20 ± 0.59	-18.83 ± 2.27	1.65 ± 2.27
Split-arm	-22.58 ± 0.59	-20.41 ± 1.72	0.07 ± 1.72

6 Discussion

Generative models of patient trajectories have attracted growing interest (Liu & Altman, 2025), with language-inspired transformers demonstrating success in forecasting clinical events (Shmatko et al., 2025; Waxler et al., 2025). However, autoregressive forecasting typically requires sequential rollout and represents patient evolution through discrete observations. Encoder–decoder architectures such as APOLLO offer alternative ways to encode clinical history and predict future events, but inherently model trajectories as a sequence of discrete events (Zhang et al., 2026). Continuous-time models instead represent evolving latent states, although numerical integration during training can make learning computationally demanding. Flow matching offers a scalable route to learning these dynamics through simulation-free vector-field training.

Applying flow matching to longitudinal EHRs requires several adaptations. Pairwise conditional flow matching supervises paths between two observations, whereas multi-marginal methods incorporate multiple observations that guide the trajectory (Rohbeck et al., 2025; Wei & Ma, 2025). However, multi-observation supervision does not itself provide inference-time access to preceding patient history. Clinical forecasting also requires suitable latent representations and accommodation of irregular observations and varying forecast durations. `EHRFlow` combines pretrained observation and history encoders with history-conditioned dynamics and duration-dependent normalised flow time. Jointly fine-tuning the vector field and decoder with a single-step linear endpoint surrogate provides observation-space supervision without numerical integration during training.

The results support the value of combining continuous latent dynamics with historical context for individual patient forecasting. Historical context helps distinguish patients with similar current records but different subsequent trajectories, with benefits extending to large-scale primary- and secondary-care datasets. Furthermore, our results on synthetic state recovery and forecasting at unseen horizons demonstrate the quality of our model's learned patient representation. Our method's similar performance across rollout durations further supports forecast consistency, while similar performance across interpolant families suggests that the findings are not specific to path construction.

Limitations. Continuous latent modelling does not guarantee recovery of biological dynamics, as EHRs also reflect coding practices, missingness, and informative observation times driven by healthcare utilisation that our model does not explicitly capture. Our evaluation is limited to continuous and discrete clinical variables, although the encoder-based formulation supports incorporating histopathology, imaging, and genomic data; the effectiveness of these extensions remains to be demonstrated. Predictive gains also depend on the metric and horizon, with autoregressive models remaining competitive on macro-AUROC at shorter horizons. Our staged training procedure outperforms fully end-to-end optimisation under the evaluated settings, motivating future work on objectives that more effectively co-adapt latent representations and dynamics. Finally, the intervention experiment is a controlled proof of concept rather than evidence of causal treatment-effect estimation from observational data. Broader external validation, uncertainty assessment, and prospective evaluation remain necessary before clinical use.

Conclusion and clinical perspective. `EHRFlow` enables continuous, history-conditioned forecasting from population-scale clinical records, potentially supporting earlier intervention and personalised care. Future work should examine whether the evolving latent embeddings can also serve as reusable representations for downstream clinical tasks (Zhang et al., 2026). More broadly, this formulation offers a step towards patient digital twins that could help explore treatment strategies and optimise care, subject to further clinical validation.

REFERENCES

Ahmed M. Alaa and Mihaela van der Schaar. Attentive state-space modeling of disease progression. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/1d0932d7f57ce74d9d9931a2c6db8a06-Paper.pdf.

Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=li7qeBbCR1t.

Lazar Atanackovic, Xi Zhang, Brandon Amos, Mathieu Blanchette, Leo J Lee, Yoshua Bengio, Alexander Tong, and Kirill Neklyudov. Meta flow matching: Integrating vector fields on the wasserstein manifold. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=9SYczU3Qgm.

Ioana Bica, Ahmed M. Alaa, James Jordon, and Mihaela van der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. *The Eighth International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/2002.04083.

Ruben Branco, Marta Gromicho, Mamede de Carvalho, Piero Fariselli, and Sara C. Madeira. Patientflow: Learning to generate mixed-type longitudinal clinical data with flow matching. *Artificial Intelligence in Medicine*, 176:103392, 06 2026. ISSN 0933-3657. URL http://dx.doi.org/10.1016/j.artmed.2026.103392.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/69386f6bb1dfed68692a24c8686939b9-Paper.pdf.

Yuqi Chen, Kan Ren, Yansen Wang, Yuchen Fang, Weiwei Sun, and Dongsheng Li. Contiformer: Continuous-time transformer for irregular time series modeling. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=YJDz4F2AZu.

Edward Choi, Mohammad Taha Bahadori, Joshua A. Kulas, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. RETAIN: an interpretable predictive model for healthcare using reverse time attention mechanism. In *Proceedings of the 29th International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2016a. URL https://arxiv.org/abs/1608.05745.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. Doctor AI: Predicting clinical events via recurrent neural networks. In *Proceedings of the 1st Machine Learning for Healthcare Conference*, Proceedings of Machine Learning Research. PMLR, 2016b. URL https://proceedings.mlr.press/v56/Choi16.html.

Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. In *Proceedings of the 2nd Machine Learning for Healthcare Conference*, Proceedings of Machine Learning Research. PMLR, 2017. URL https://proceedings.mlr.press/v68/choi17a.html.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014. URL https://arxiv.org/abs/1412.3555.

Josef Coresh, Tanvir Chowdhury Turin, Kunihiro Matsushita, Yingying Sang, Shoshana H. Ballew, Lawrence J. Appel, Hisatomi Arima, Steven J. Chadban, Massimo Cirillo, Ognjenka Djurdjev, Jamie A. Green, Gunnar H. Heine, Lesley A. Inker, Fujiko Irie, Areef Ishani, Joachim H. Ix, Csaba P. Kovesdy, Angharad Marks, Takayoshi Ohkubo, Varda Shalev, Anoop Shankar, Chi Pang Wen, Paul E. de Jong, Kunitoshi Iseki, Benedicte Stengel, Ron T. Gansevoort, and Andrew S. Levey. Decline in estimated glomerular filtration rate and subsequent risk of end-stage renal

disease and mortality. *JAMA*, 311(24):2518–2531, 06 2014. ISSN 0098-7484. doi: 10.1001/jama.2014.6634. URL http://dx.doi.org/10.1001/jama.2014.6634.

Edward De Brouwer, Jaak Simm, Adam Arany, and Yves Moreau. GRU-ODE-Bayes: Continuous modeling of sporadically-observed time series. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/455cb2657aaa59e32fad80cb0b65b9dc-Paper.pdf.

Cristóbal Esteban, Stephanie L. Hyland, and Gunnar Rätsch. Real-valued (medical) time series generation with recurrent conditional GANs, 2017. URL https://arxiv.org/abs/1706.02633.

Charles Gadd, Krishna Gokhale, Aditya Acharya, Jennifer Cooper, Francesca Crowe, Leah Fitzsimmons, Thomas Jackson, Krishnarajah Nirantharakumar, Christopher Yau, Rebecca Birch, Marco Canducci, Dominic Danks, Alexander d'Elia, Alastair Denniston, Sarah Flanagan, Suzy Gallier, Naijie Guan, Xin Guan, Imane Guellil, Georgios Gkoutos, Shamil Haroon, Eleanor Hathaway, Louise Jackson, Janet Lord, Zeinab Majid, Tom Marshall, George Morris, Charlotte Owen, Elizabeth Sapey, Chris Sainsbury, Charlotte Spurway, Peter Tino, Steven Wambua, Amaya Azcoaga-Lorenzo, Colin McCowan, Luciana Rocha Pedro, Muhammad Usman, Natalia Hong, Sara Matijevic, Kaspar Martens, Tim Williams, and Puja Myles. SurvivEHR: a competing risks, time-to-event foundation model for multiple long-term conditions from primary care electronic health records. *npj Digital Medicine*, 9(1), May 2026. ISSN 2398-6352. URL http://dx.doi.org/10.1038/s41746-026-02709-z.

Benjamin A Goldstein, Ann Marie Navar, Michael J Pencina, and John P A Ioannidis. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 24(1):198–208, May 2017. ISSN 1527-974X. URL http://dx.doi.org/10.1093/jamia/ocw042.

Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. FFJORD: Free-form continuous dynamics for scalable reversible generative models. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJxgknCcK7.

Lin Lawrence Guo, Jason Fries, Ethan Steinberg, Scott Lanyon Fleming, Keith Morse, Catherine Aftandilian, Jose Posada, Nigam Shah, and Lillian Sung. A multi-center study on the adaptability of a shared foundation model for electronic health records. *npj Digital Medicine*, 7(1), 09 2024. ISSN 2398-6352. doi: 10.1038/s41746-024-01166-w. URL http://dx.doi.org/10.1038/s41746-024-01166-w.

Emily Herrett, Arlene M Gallagher, Krishnan Bhaskaran, Harriet Forbes, Rohini Mathur, Tjeerd van Staa, and Liam Smeeth. Data resource profile: Clinical practice research datalink (CPRD). *International Journal of Epidemiology*, 44(3):827–836, 09 2015. ISSN 1464-3685. doi: 10.1093/ije/dyv098. URL http://dx.doi.org/10.1093/ije/dyv098.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

Guillaume Huguet, Daniel Sumner Magruder, Alexander Tong, Oluwadamilola Fasina, Manik Kuchroo, Guy Wolf, and Smita Krishnaswamy. Manifold interpolating optimal-transport flows for trajectory inference. In *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=ahAEhOtVif.

Mohammad Mohaiminul Islam, Thijs P. Kuipers, Sharvaree Vadgama, Coen de Vente, Afsana Khan, Clara I. Sánchez, and Erik J. Bekkers. Longitudinal flow matching for trajectory modeling, 2025. URL https://arxiv.org/abs/2510.03569.

Kacper Kapusniak, Peter Potaptchik, Teodora Reu, Leo Zhang, Alexander Tong, Michael M. Bronstein, Joey Bose, and Francesco Di Giovanni. Metric flow matching for smooth interpolations on the data manifold. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=fE3RqiF4Nx.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series, 2020. URL https://arxiv.org/abs/2005.08926.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013. URL https://arxiv.org/abs/1312.6114.

Marcel Kollovieh, Marten Lienen, David Lüdke, Leo Schwinn, and Stephan Günnemann. Flow matching with gaussian process priors for probabilistic time series forecasting. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=uxVBbSlKQ4.

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, April 2024. ISSN 2589-7500. URL http://dx.doi.org/10.1016/S2589-7500(24)00025-6.

Rahul G. Krishnan, Uri Shalit, and David Sontag. Deep kalman filters, 2015. URL https://arxiv.org/abs/1511.05121.

Rahul G. Krishnan, Uri Shalit, and David Sontag. Structured inference networks for nonlinear state space models, 2016. URL https://arxiv.org/abs/1609.09869.

Jin Li, Benjamin J. Cairns, Jingsong Li, and Tingting Zhu. Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications. *npj Digital Medicine*, 6(1), May 2023. ISSN 2398-6352. URL http://dx.doi.org/10.1038/s41746-023-00834-7.

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), April 2020. ISSN 2045-2322. URL http://dx.doi.org/10.1038/s41598-020-62922-y.

Bryan Lim, Ahmed Alaa, and Mihaela van der Schaar. Forecasting treatment responses over time using recurrent marginal structural networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/56e6a93212e4482d99c84a639d254b67-Paper.pdf.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Kara Liu and Russ B Altman. Conditional generative models for synthetic tabular data: Applications for precision medicine and diverse representations. *Annual review of biomedical data science*, 8(1):21–49, 2025. URL https://doi.org/10.1146/annurev-biodatasci-103123-094844.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes, 2022. URL https://arxiv.org/abs/2204.07258.

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*, Proceedings of Machine Learning Research. PMLR, 04 Dec 2021. URL https://proceedings.mlr.press/v158/pang21a.html.

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*, Proceedings of Machine Learning Research. 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), 09 2024. ISSN 2398-6352. doi: 10.1038/s41746-024-01235-0. URL http://dx.doi.org/10.1038/s41746-024-01235-0.

Martin Rohbeck, Edward De Brouwer, Charlotte Bunne, Jan-Christian Huetter, Anne Biton, Kelvin Y. Chen, Aviv Regev, and Romain Lopez. Modeling complex system dynamics with flow matching across time and conditions. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=hwnObmOTrV.

Yulia Rubanova, Ricky T. Q. Chen, and David K Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/42a6845a557bef704ad8ac9cb4461d43-Paper.pdf.

Silas Ruhrberg Estévez, Nicolas Huynh, Tennison Liu, Roderik M. Kortlever, Gerard I. Evan, David L. Bentley, and Mihaela van der Schaar. CellBRIDGE: Learning cellular trajectories via interaction-aware alignment. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=W9EgZALquo.

Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, Lia Lee, Jenny Chen, Justin Brumbaugh, Philippe Rigollet, Konrad Hochedlinger, Rudolf Jaenisch, Aviv Regev, and Eric S. Lander. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943.e22, February 2019. ISSN 0092-8674. URL http://dx.doi.org/10.1016/j.cell.2019.01.006.

Lifeng Shen and James Kwok. Non-autoregressive conditional diffusion models for time series prediction. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR, 2023. URL https://proceedings.mlr.press/v202/shen23d.html.

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(8088):248–256, 09 2025. ISSN 1476-4687. URL http://dx.doi.org/10.1038/s41586-025-09529-3.

Ethan Steinberg, Ken Jung, Jason A. Fries, Conor K. Corbin, Stephen R. Pfohl, and Nigam H. Shah. Language models are an effective patient representation learning technique for electronic health record data, 2020. URL https://arxiv.org/abs/2001.05295.

Ethan Steinberg, Jason Alan Fries, Yizhe Xu, and Nigam Shah. MOTOR: A time-to-event foundation model for structured medical records. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NialiwI2V6.

Sandra M. Swain, Fredrick S. Whaley, and Michael S. Ewer. Congestive heart failure in patients treated with doxorubicin: A retrospective analysis of three trials. *Cancer*, 97:2869–2879, May 2003. ISSN 1097-0142. URL http://dx.doi.org/10.1002/cncr.11407.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. CSDI: Conditional score-based diffusion models for probabilistic time series imputation, 2021. URL https://arxiv.org/abs/2107.03502.

Brandon Theodorou, Cao Xiao, and Jimeng Sun. Synthesize high-dimensional longitudinal electronic health records via hierarchical autoregressive language model. *Nature Communications*, 14(1), August 2023. ISSN 2041-1723. URL http://dx.doi.org/10.1038/s41467-023-41093-0.

Alexander Tong, Jessie Huang, Guy Wolf, David van Dijk, and Smita Krishnaswamy. TrajectoryNet: A dynamic optimal transport network for modeling cellular dynamics, 2020. URL https://arxiv.org/abs/2002.04461.

Alexander Y. Tong, Nikolay Malkin, Kilian Fatras, Lazar Atanackovic, Yanlei Zhang, Guillaume Huguet, Guy Wolf, and Yoshua Bengio. Simulation-free Schrödinger bridges via score and flow matching. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*. PMLR, 2024. URL https://proceedings.mlr.press/v238/tong24a.html.

Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3):230–238, August 2018. ISSN 1527-974X. URL http://dx.doi.org/10.1093/jamia/ocx079.

Zifeng Wang and Jimeng Sun. PromptEHR: Conditional electronic healthcare records generation with prompt learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2022. URL https://aclanthology.org/2022.emnlp-main.185/.

Shane Waxler, Paul Blazek, Davis White, Daniel Sneider, Kevin Chung, Mani Nagarathnam, Patrick Williams, Hank Voeller, Karen Wong, Matthew Swanhorst, Sheng Zhang, Naoto Usuyama, Cliff Wong, Tristan Naumann, Hoifung Poon, Andrew Loza, Daniella Meeker, Seth Hain, and Rahul Shah. Generative medical event models improve with scale, 2025. URL https://arxiv.org/abs/2508.12104.

Ganchao Wei and Li Ma. Stream-level flow matching with gaussian processes. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=qg9p1I5lmp.

Sihang Zeng, Lucas Jing Liu, Jun Wen, Meliha Yetisgen, Ruth Etzioni, and Gang Luo. TrajSurv: Learning continuous latent trajectories from electronic health records for trustworthy survival prediction. In *Proceedings of the 10th Machine Learning for Healthcare Conference*, volume 298. PMLR, 2025. URL https://proceedings.mlr.press/v298/zeng25a.html.

Andrew Zhang, Tong Ding, Sophia J. Wagner, Caiwei Tian, Ming Y. Lu, Rowland Pettit, Joshua E. Lewis, Alexandre Misrahi, Dandan Mo, Long Phi Le, and Faisal Mahmood. A multimodal and temporal foundation model for virtual patient representations at healthcare system scale, 2026. URL https://arxiv.org/abs/2604.18570.

Xi Zhang, Yuan Pu, Yuki Kawamura, Andrew Loza, Yoshua Bengio, Dennis Shung, and Alexander Tong. Trajectory flow matching with applications to clinical time series modelling. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=fNakQltI1N.

SUPPLEMENTARY INFORMATION

A Extended related work

Patient representations and generative latent dynamics. Beyond the clinical sequence models discussed in the main text, pretrained models learn reusable representations from structured electronic health records for transfer to downstream clinical tasks (Steinberg et al., 2020; Guo et al., 2024; Pang et al., 2021). Earlier deep generative state-space models learned latent transition and observation distributions from longitudinal data, including applications to patient trajectories and treatment counterfactuals (Krishnan et al., 2015; 2016).

Synthetic longitudinal health records. A complementary literature develops generative models for synthetic health data. Early approaches such as medGAN generate discrete patient records using adversarial training (Choi et al., 2017), while recurrent conditional GANs generate multivariate medical time series (Esteban et al., 2017). Subsequent methods model longitudinal EHR structure, including PromptEHR (Wang & Sun, 2022) and the hierarchical autoregressive model HALO (Theodorou et al., 2023). Other approaches jointly generate continuous measurements and discrete clinical variables (Li et al., 2023), while PatientFlow applies flow matching to generate synthetic longitudinal clinical records (Branco et al., 2026).

Diffusion-based time-series modeling. Diffusion models provide another framework for generative modeling of temporal data. TimeGrad combines autoregressive forecasting with conditional denoising diffusion (Rasul et al., 2021), while CSDI conditions score-based diffusion on observed values for probabilistic imputation and also supports interpolation and forecasting (Tashiro et al., 2021). Non-autoregressive conditional diffusion models generate future sequences jointly, avoiding sequential sampling across forecast time steps (Shen & Kwok, 2023). These approaches offer alternatives for modeling distributions over temporal observations, whereas `EHRFlow` learns a vector field describing latent patient evolution over the forecast interval.

Biological trajectories and conditional transport. Related trajectory-inference problems arise in biology, where destructive measurements often prevent repeated observation of the same cell. Waddington-OT infers couplings between cell populations sampled at different times (Schiebinger et al., 2019), while TrajectoryNet learns continuous transport dynamics from such snapshots (Tong et al., 2020). Manifold-based approaches incorporate data geometry into trajectory reconstruction (Huguet et al., 2022), and CellBRIDGE incorporates interaction information when aligning cellular states across time (Ruhrberg Estévez et al., 2026). Meta Flow Matching conditions on population embeddings to generalise across initial distributions (Atanackovic et al., 2025), while related work combines score and flow matching to learn stochastic dynamics through Schrödinger bridges from unpaired samples (Tong et al., 2024). Longitudinal EHRs provide within-patient correspondence across observations, allowing us to construct training paths directly and condition forecasts on individual history.

Treatment-conditioned generation and counterfactual prediction. Classifier-free guidance combines conditional and unconditional model predictions to control generation without a separate classifier (Ho & Salimans, 2022). We use this mechanism to generate treatment-conditioned trajectories in a controlled simulation. A distinct literature targets counterfactual outcome prediction from observational data: recurrent marginal structural networks model treatment assignment and outcomes over time (Lim et al., 2018), while Counterfactual Recurrent Networks (Bica et al., 2020) and the Causal Transformer (Melnychuk et al., 2022) learn representations designed to address time-varying confounding. These methods rely on causal identification assumptions. Our intervention experiment instead evaluates recovery of a known simulator-defined treatment effect; conditional guidance alone does not establish causal identification from observational EHRs.

B Theoretical motivation for history conditioning

We establish that history conditioning cannot increase the optimal population flow-matching loss, and give a simple example separating population marginal matching from individual trajectory forecasting.

B.1 History conditioning reduces the optimal flow-matching loss

Theorem 1 (Optimality of history conditioning)**.** *Let*

$$X = (Z_s, s, \lambda),$$

let H denote the history representation, and let $U \in \mathbb{R}^{d_z}$ be the target velocity with $\mathbb{E}\|U\|_2^2 < \infty$. Define

$$R_0^* = \inf_v \mathbb{E}\|U - v(X)\|_2^2, \qquad R_H^* = \inf_w \mathbb{E}\|U - w(X,H)\|_2^2.$$

Then

$$R_0^* - R_H^* = \mathbb{E}\left\|\mathbb{E}[U \mid X, H] - \mathbb{E}[U \mid X]\right\|_2^2 \geq 0.$$

Equality holds if and only if

$$\mathbb{E}[U \mid X, H] = \mathbb{E}[U \mid X] \quad \textit{a.s.}$$

Proof. Squared-error risk is minimised by the conditional mean. Let

$$m_0 = \mathbb{E}[U \mid X], \qquad m_H = \mathbb{E}[U \mid X, H].$$

Then

$$R_0^* = \mathbb{E}\|U - m_0\|_2^2 = \mathbb{E}\|(U - m_H) + (m_H - m_0)\|_2^2.$$

Since

$$\mathbb{E}[U - m_H \mid X, H] = 0,$$

the cross term vanishes, giving

$$R_0^* = R_H^* + \mathbb{E}\|m_H - m_0\|_2^2.$$

□

Thus, the benefit of history is determined by the squared change it induces in the conditional mean velocity. If the expected dynamics are already determined by the current state and time, history provides no improvement.

B.2 Population marginals do not determine individual trajectories

We next give a simple Gaussian example showing that the same distinction also appears after integrating the learned field: two fields can generate the same population distribution while producing different forecasts for individual trajectories.

Consider

$$Z_t = A + tB, \qquad A, B \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \qquad t \in [-1, 1].$$

Here A determines the state at the forecast anchor and B determines the patient-specific rate of progression. The anchor is

$$Z_0 = A.$$

Suppose the previous state is observed with noise,

$$Y_{-1} = Z_{-1} + \eta, \qquad \eta \sim \mathcal{N}(0, \sigma^2),$$

independently of (A, B). We define the history representation

$$H = Z_0 - Y_{-1} = B - \eta.$$

Its reliability is

$$\rho^2 = \mathrm{Corr}(H, B)^2 = \frac{1}{1+\sigma^2}, \qquad \kappa = 1 - \rho^2.$$

Thus, ρ^2 increases as the history becomes more informative about the patient's progression rate, with $\rho = 1$ corresponding to noiseless history.

During training the model uses the path

$$Z_s = A + sB, \qquad U = B, \qquad s \sim \mathcal{U}(0,1).$$

Theorem 2 (Population marginals versus individual trajectories)**.** *For every $\rho \in (0, 1]$, the following hold:*

1. ***Optimal fields.*** *The population-optimal history-independent and history-conditioned fields are*
$$v^*(z, s) = \frac{s}{1 + s^2} z, \qquad w^*(z, s, h) = \rho^2 h + \frac{s\kappa}{1 + \kappa s^2} \left(z - s\rho^2 h\right).$$
2. ***Flow-matching risks.*** *The corresponding optimal risks are*
$$R_0^* = \frac{\pi}{4}, \qquad R_H^* = \sqrt{\kappa} \arctan \sqrt{\kappa}.$$
Hence $R_H^* < R_0^*$ for $\rho > 0$, and $R_H^* = 0$ when $\rho = 1$.
3. ***Integrated flows.*** *Starting from $Z_0 = A$, the resulting trajectories are*
$$\widehat{Z}_s^0 = A\sqrt{1 + s^2}, \qquad \widehat{Z}_s^H = s\rho^2 H + A\sqrt{1 + \kappa s^2}.$$
Both recover the correct population marginals:
$$\widehat{Z}_s^0 \overset{d}{=} \widehat{Z}_s^H \overset{d}{=} Z_s \sim \mathcal{N}(0, 1 + s^2),$$
and the history-conditioned flow additionally satisfies
$$\widehat{Z}_s^H \mid H \overset{d}{=} Z_s \mid H.$$
4. ***Individual forecast error.*** *For every $s \in (0, 1]$,*
$$\mathbb{E}\left[\left(\widehat{Z}_s^0 - Z_s\right)^2\right] = \left(\sqrt{1 + s^2} - 1\right)^2 + s^2,$$
whereas
$$\mathbb{E}\left[\left(\widehat{Z}_s^H - Z_s\right)^2\right] = \left(\sqrt{1 + \kappa s^2} - 1\right)^2 + \kappa s^2.$$
The history-conditioned error is strictly smaller for $\rho > 0$, decreases with ρ^2, and vanishes when $\rho = 1$.

Proof. Since (B, H) is jointly Gaussian,
$$B \mid H = h \sim \mathcal{N}(\rho^2 h, \kappa).$$
Conditionally on $H = h$,
$$\mathbb{E}[Z_s \mid h] = s\rho^2 h, \qquad \mathrm{Var}(Z_s \mid h) = 1 + \kappa s^2, \qquad \mathrm{Cov}(B, Z_s \mid h) = s\kappa.$$
Gaussian conditioning therefore gives
$$\mathbb{E}[B \mid Z_s = z, H = h] = \rho^2 h + \frac{s\kappa}{1 + \kappa s^2} \left(z - s\rho^2 h\right),$$
which is w^*. Without conditioning on history,
$$\mathbb{E}[B \mid Z_s = z] = \frac{s}{1 + s^2} z,$$
which gives v^*.

The corresponding conditional variances are
$$\mathrm{Var}(B \mid Z_s) = \frac{1}{1 + s^2}, \qquad \mathrm{Var}(B \mid Z_s, H) = \frac{\kappa}{1 + \kappa s^2}.$$
Integrating over $s \in [0, 1]$ yields
$$R_0^* = \int_0^1 \frac{1}{1 + s^2}\, ds = \frac{\pi}{4},$$
and
$$R_H^* = \int_0^1 \frac{\kappa}{1 + \kappa s^2}\, ds = \sqrt{\kappa} \arctan \sqrt{\kappa}.$$

Integrating the optimal fields from $\widehat{Z}_0 = A$ gives

$$\widehat{Z}_s^0 = A\sqrt{1+s^2}, \qquad \widehat{Z}_s^H = s\rho^2 H + A\sqrt{1+\kappa s^2}.$$

Since A is independent of H,

$$\widehat{Z}_s^H \mid H = h \sim \mathcal{N}\left(s\rho^2 h, 1+\kappa s^2\right),$$

which is exactly the conditional distribution of $Z_s \mid H = h$. Marginalising over H gives

$$\widehat{Z}_s^H \sim \mathcal{N}(0, 1+s^2),$$

and the same population marginal is obtained by $\widehat{Z}_s^0$.

Finally, write

$$B = \rho^2 H + R, \qquad R \sim \mathcal{N}(0, \kappa),$$

with R independent of (A, H). Then

$$\widehat{Z}_s^H - Z_s = A\left(\sqrt{1+\kappa s^2} - 1\right) - sR,$$

while

$$\widehat{Z}_s^0 - Z_s = A\left(\sqrt{1+s^2} - 1\right) - sB.$$

Independence of the two terms in each expression gives the stated mean-squared errors. The history-conditioned error is strictly increasing in $\kappa = 1-\rho^2$, and therefore decreases as the history becomes more reliable. For $\rho = 1$, we have $\kappa = 0$ and $H = B$, so

$$\widehat{Z}_s^H = A + sB = Z_s.$$

□

Figure 5 evaluates Theorem 2 empirically. We train both fields as small MLPs by flow-matching regression on samples from the linear-Gaussian model, integrate them from $Z_0 = A$, and evaluate on held-out samples for $\rho \in \{0.1, \ldots, 1\}$. The learned flows match the closed forms in all three respects: forecast error, optimal risk, and marginal and conditional variance.

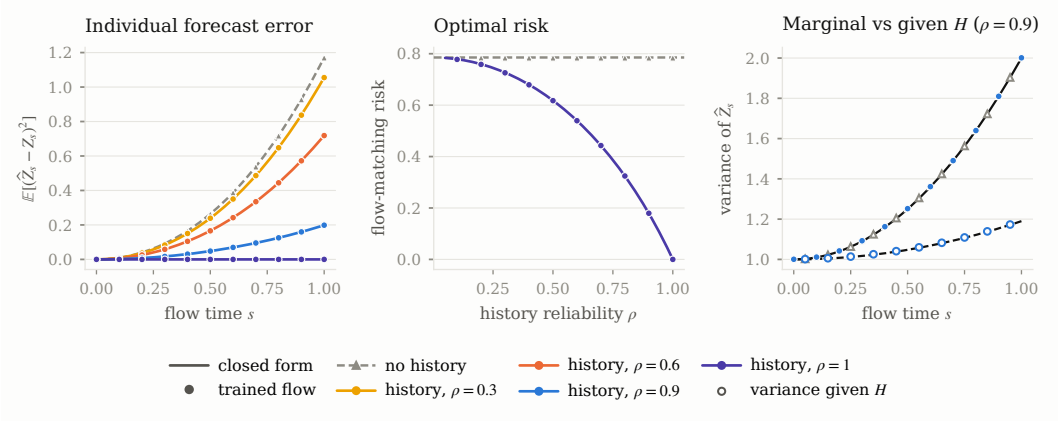


Figure 5: **Left:** individual forecast error against flow time. **Centre:** flow-matching risk against history reliability ρ. **Right:** variance of the generated state at $\rho = 0.9$; both flows recover the marginal $1 + s^2$, only the history-conditioned flow recovers the variance given H, $1 + \kappa s^2$.

C Extended results

C.1 The choice of the interpolant

Across both the synthetic and real-world datasets, we observed that the choice of multi-marginal interpolant had a substantially smaller effect on predictive performance than history conditioning. This suggests that, provided the interpolant captures the observed marginals, much of the variation between particular interpolation schemes may be attenuated when the corresponding vector field is learned by the neural network.

We illustrate this effect in Figure 6. For a fixed set of observed marginals, multiple valid interpolation paths exist that pass through the same knots. Although different interpolants induce substantially different conditional velocity targets, flow matching learns a marginal vector field by regressing over these conditional targets (Lipman et al., 2023; Albergo & Vanden-Eijnden, 2023). This regression can therefore attenuate interpolant-specific variability, providing a possible explanation for the substantially smaller differences observed between the fitted fields in Figure 6.

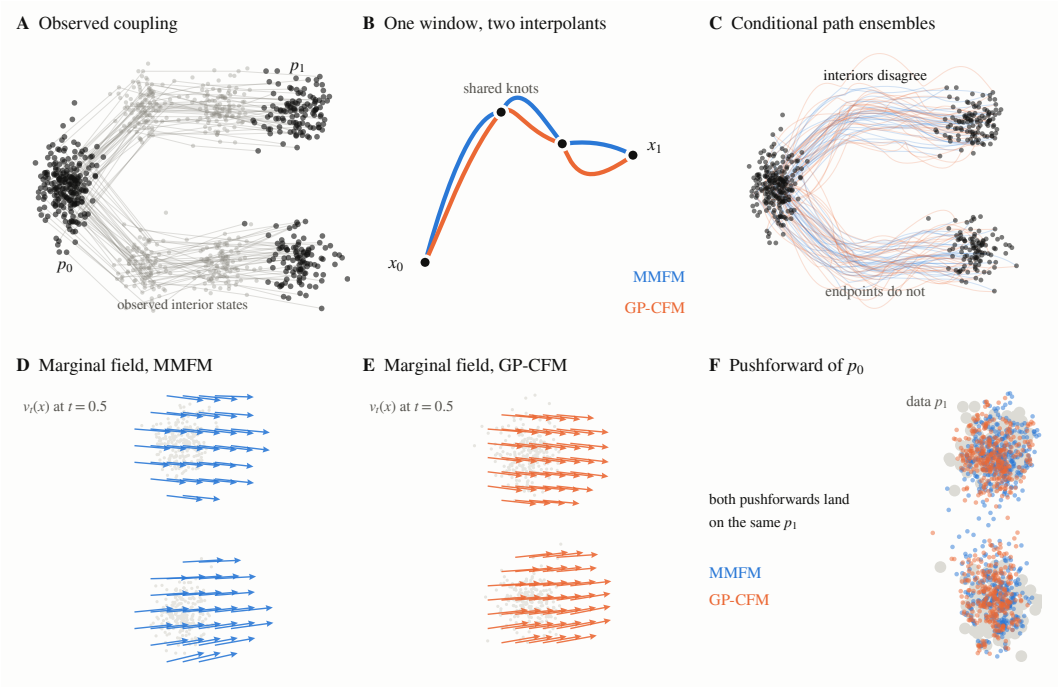


Figure 6: Smooth vector-field paths through the same knots on synthetic data.

We examine this phenomenon more systematically in Figure 7 using a two-dimensional synthetic example with six different knot-interpolating families. The conditional paths, and consequently their velocity targets, differ substantially across interpolants. However, these differences are reduced after fitting the same MLP architecture to the corresponding targets, consistent with the network averaging over variation in the conditional paths. Differences are reduced further after integrating the fitted vector fields: while small discrepancies remain at $t = 0.5$, the final pushforward distributions are nearly indistinguishable. Across ten seeds, pairwise Wasserstein-1 distances between the predicted terminal distributions are only 0.02–0.04.

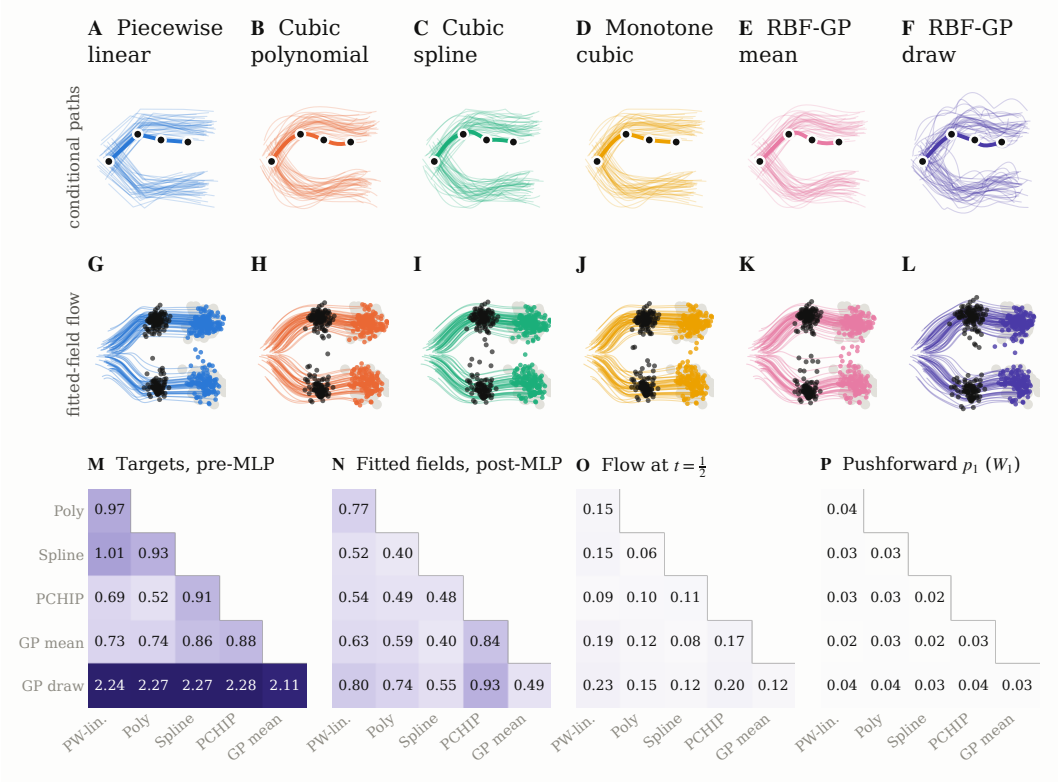


Figure 7: Six knot-interpolating families fitted with the same network architecture (10 seeds). Differences between conditional paths are reduced in the fitted flows, with pairwise endpoint Wasserstein-1 distances of 0.02–0.04.

C.2 Extended forecasting metrics

We extend the horizon-averaged results in the main text by reporting top-5 accuracy, top-1 accuracy, and macro-AUROC separately at each forecast horizon. These metrics assess complementary aspects of prediction: top-5 accuracy measures whether the recorded code appears among the five highest-ranked predictions, top-1 accuracy requires it to be the highest-ranked prediction, and macro-AUROC assesses discrimination across evaluated code classes.

Synthetic benchmarks. Figures 8, 9 and 10 report results across the five controlled synthetic datasets and Synthea. Multi-marginal methods generally outperform adjacent CFM and Metric FM, while history conditioning provides additional gains that depend on the data-generating mechanism. The clearest benefits occur in Anthracycline and RRMS, where preceding observations provide information about cumulative exposure and relapse history. Improvements across both top-1 accuracy and macro-AUROC indicate that these benefits extend beyond the top-5 metric used in the main text.

The additional metrics also reveal settings in which history conditioning provides limited benefit. On Heart failure, the multi-marginal variants have similar top-1 accuracy, as expected for the Markovian disease mechanism, despite modest differences in top-5 accuracy. On Synthea, history conditioning yields little improvement in top-5 accuracy but notably improves macro-AUROC at longer horizons. Performance generally declines with increasing forecast duration, although this pattern is not universal across datasets and metrics.

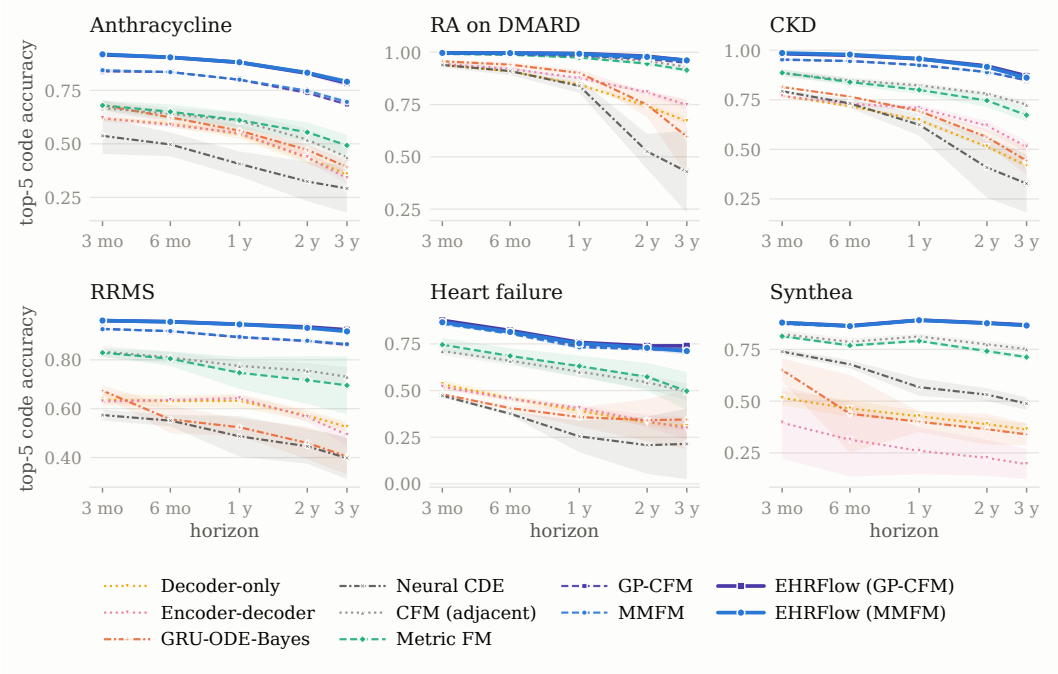


Figure 8: Clinical-code top-5 accuracy by forecast horizon across the five synthetic datasets and Synthea. Lines show the mean across random seeds; bands indicate ± one standard deviation.

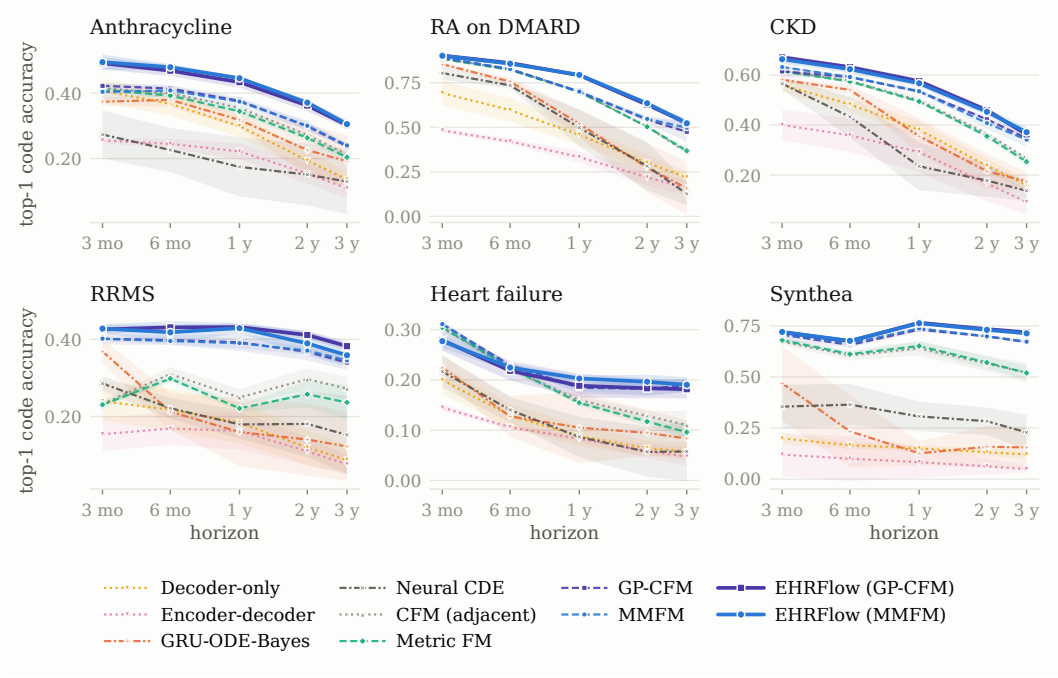


Figure 9: Clinical-code top-1 accuracy by forecast horizon across the five synthetic datasets and Synthea. Lines show the mean across random seeds; bands indicate ± one standard deviation.

Real-world datasets. Figure 11 reports all three metrics on CPRD and AMC-EHR. On CPRD, history conditioning produces clear improvements over the corresponding history-independent flow models in top-5 accuracy and macro-AUROC, while the differences in top-1 accuracy are smaller.

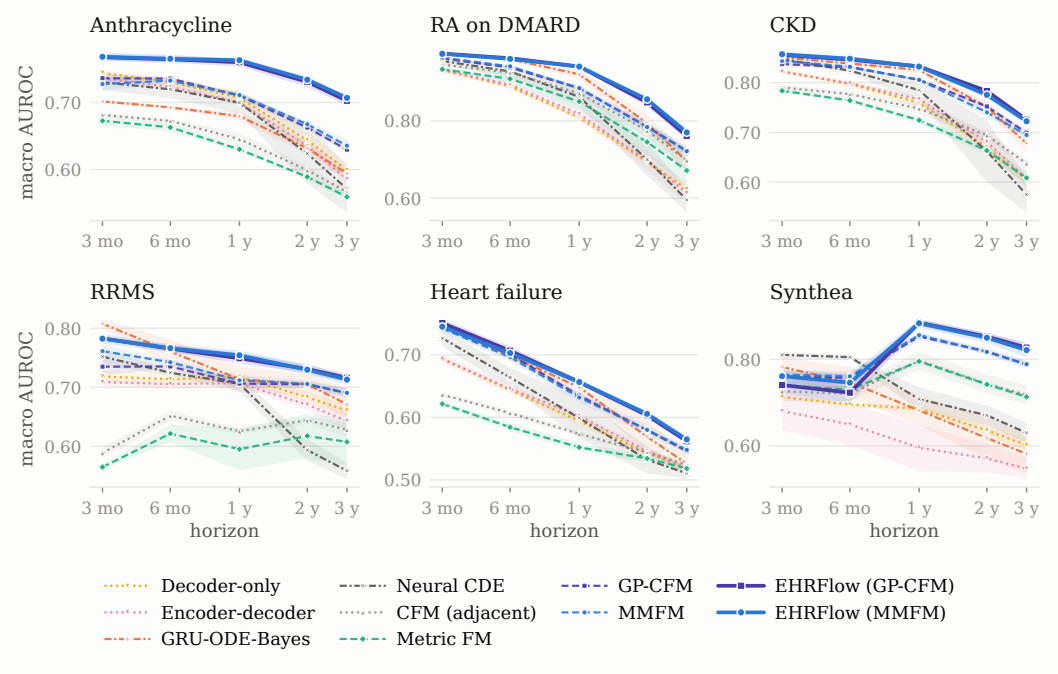


Figure 10: Clinical-code macro-AUROC by forecast horizon across the five synthetic datasets and Synthea. Lines show the mean across random seeds; bands indicate ± one standard deviation.

The autoregressive baselines deteriorate more strongly with increasing horizon, particularly for top-5 accuracy. On AMC-EHR, the relative ranking depends more on the metric: autoregressive models remain competitive at short horizons and achieve higher short-horizon macro-AUROC, whereas `EHRFlow` (MMFM) performs best at longer horizons.

C.3 Evidence for continuous latent dynamics

Accurate predictions at the training horizons alone do not establish that a model captures useful latent dynamics between them. We therefore examine three complementary properties: recovery of the underlying simulated state, agreement between forecasts of the same time point obtained using different rollout durations, and predictive performance at horizons outside the training grid.

Recovery of the simulated latent state. Table 5 evaluates whether forecasted latent representations retain information about the simulator's underlying disease-severity state. We read out severity using a linear map fitted on encoded anchor states and measure mean-squared error against the ground-truth severity at the forecast horizon. Results are averaged across the 90-, 365-, and 1,095-day horizons and summarised by the median and standard deviation across seeds.

History conditioning reduces recovery error relative to the corresponding history-independent multi-marginal models across all five controlled datasets. For example, on RRMS, error decreases from 2.23 to 1.58 for GP-CFM and from 2.35 to 1.67 for MMFM. On CKD, the corresponding reductions are from 0.99 to 0.78 and from 1.01 to 0.73. These results suggest that the forecasting improvements are accompanied by better recovery of the simulated state, rather than being confined to clinical-code prediction.

Interpolation and extrapolation across forecast horizons. Table 6 and Figure 12 evaluate forecasts at horizons not included in the training grid. We distinguish interior horizons of 270, 545, and 912 days from horizons outside the trained range: 45 days and 1,460 days. The latter horizons test predictions at shorter and longer durations than the nominal training horizons, respectively. Figure 12 places these evaluations alongside the trained horizons to show how performance changes with forecast duration.

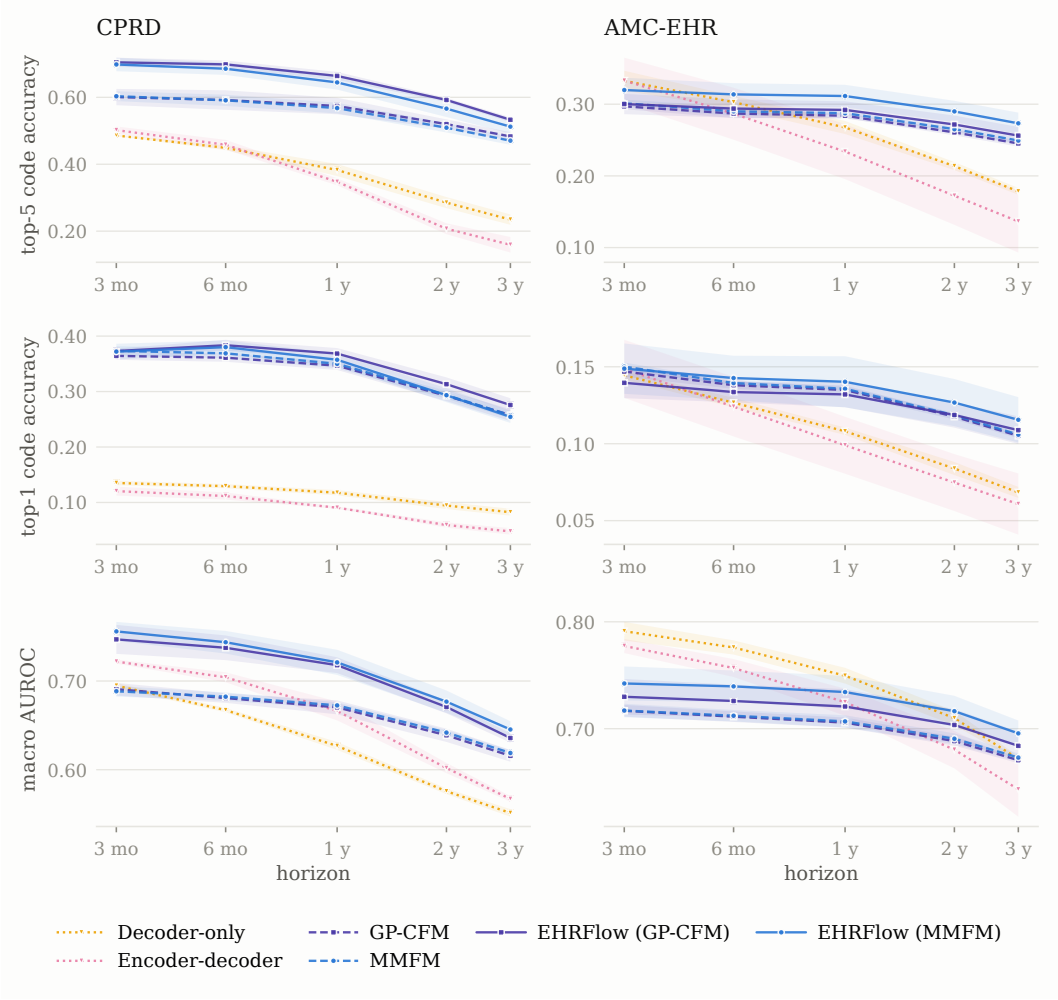


Figure 11: Clinical-code forecasting performance by horizon on CPRD (external Gold cohort) and AMC-EHR. Rows show top-5 accuracy, top-1 accuracy, and macro-AUROC. Lines show the mean across random seeds; shaded bands indicate $\pm$ one standard deviation.

At interior horizons, history conditioning retains clear benefits on Anthracycline, CKD, and RRMS. For example, Anthracycline top-5 accuracy increases from 0.767 to 0.851 for GP-CFM and from 0.774 to 0.853 for MMFM. These gains also persist in the aggregate evaluation outside the trained range. Improvements are less consistent on Heart failure and Synthea, and the history-conditioned variants do not outperform their counterparts in every setting.

Agreement across rollout durations. Because the vector field is conditioned on the requested duration, forecasts of the same calendar time do not necessarily agree when obtained from rollouts of different lengths. Table 7 examines this directly. For each nested pair $h < H$, we compare a forecast integrated to day h with the state at day h along a rollout targeting day H, using the same anchor and historical context. We report the top-5 accuracy of each forecast and the fraction of codes shared between their top-5 prediction sets.

Across the evaluated datasets and horizon pairs, mean top-5 overlap is 93.9% for `EHRFlow` (GP-CFM) and 93.1% for `EHRFlow` (MMFM). Direct and nested forecasts also have similar mean accuracy: 0.920 versus 0.921 for GP-CFM and 0.919 versus 0.921 for MMFM. Agreement varies across datasets and horizon pairs, with lower overlap for some longer rollouts.

Table 5: Median and s.d. of latent-state recovery on the synthetic benchmarks across seeds.

	Heart failure	RA on DMARD	CKD	RRMS	Anthracycline
CFM (adjacent)	1.48 ± 0.05	0.36 ± 0.04	1.14 ± 0.01	1.96 ± 0.06	1.07 ± 0.06
Metric FM	1.41 ± 0.08	0.34 ± 0.03	1.10 ± 0.06	2.06 ± 0.04	1.35 ± 0.23
GP-CFM	1.24 ± 0.06	0.26 ± 0.00	0.99 ± 0.05	2.23 ± 0.12	1.07 ± 0.05
MMFM	1.34 ± 0.02	0.27 ± 0.01	1.01 ± 0.06	2.35 ± 0.11	0.98 ± 0.01
`EHRFlow` (GP-CFM)	**1.18** ± 0.12	**0.23** ± 0.00	0.78 ± 0.03	**1.58** ± 0.06	**0.81** ± 0.01
`EHRFlow` (MMFM)	1.24 ± 0.12	0.25 ± 0.01	**0.73** ± 0.12	1.67 ± 0.08	0.94 ± 0.05

Table 6: Top-5 clinical-code accuracy at the off-grid horizons, averaged over the horizons of each block (mean ± std over seeds).

	Heart failure	RA on DMARD	CKD	RRMS	Anthracycline	Synthea
Interior off-grid horizons (270, 545, 912 d)						
GRU-ODE-Bayes	0.352 ± 0.091	0.813 ± 0.043	0.611 ± 0.025	0.495 ± 0.013	0.518 ± 0.020	0.302 ± 0.031
Neural CDE	0.267 ± 0.119	0.664 ± 0.066	0.492 ± 0.100	0.438 ± 0.055	0.386 ± 0.098	0.562 ± 0.032
CFM (adjacent)	0.573 ± 0.030	0.967 ± 0.003	0.797 ± 0.012	0.769 ± 0.029	0.557 ± 0.040	0.705 ± 0.012
Metric FM	0.599 ± 0.063	0.957 ± 0.006	0.769 ± 0.024	0.743 ± 0.079	0.583 ± 0.042	0.664 ± 0.018
GP-CFM	0.738 ± 0.006	0.974 ± 0.001	0.906 ± 0.004	0.887 ± 0.005	0.767 ± 0.003	0.813 ± 0.006
MMFM	0.735 ± 0.005	0.977 ± 0.002	0.905 ± 0.004	0.889 ± 0.004	0.774 ± 0.004	**0.814** ± 0.003
`EHRFlow` (GP-CFM)	**0.756** ± 0.006	**0.985** ± 0.002	**0.934** ± 0.003	**0.939** ± 0.005	0.851 ± 0.009	0.810 ± 0.005
`EHRFlow` (MMFM)	0.746 ± 0.008	0.984 ± 0.002	0.933 ± 0.005	0.937 ± 0.003	**0.853** ± 0.007	0.806 ± 0.006
Outside the trained range (45, 1460 d)						
GRU-ODE-Bayes	0.476 ± 0.100	0.708 ± 0.157	0.627 ± 0.031	0.579 ± 0.020	0.548 ± 0.003	0.521 ± 0.026
Neural CDE	0.407 ± 0.105	0.677 ± 0.131	0.585 ± 0.087	0.478 ± 0.047	0.414 ± 0.094	0.603 ± 0.039
CFM (adjacent)	0.586 ± 0.029	0.947 ± 0.008	0.772 ± 0.025	0.774 ± 0.027	0.534 ± 0.025	0.781 ± 0.010
Metric FM	0.600 ± 0.056	0.937 ± 0.009	0.730 ± 0.023	0.754 ± 0.070	0.578 ± 0.034	0.758 ± 0.008
GP-CFM	0.794 ± 0.004	0.963 ± 0.004	0.866 ± 0.008	0.879 ± 0.006	0.730 ± 0.013	0.855 ± 0.005
MMFM	0.793 ± 0.020	0.969 ± 0.003	0.872 ± 0.007	0.882 ± 0.007	0.744 ± 0.007	0.861 ± 0.008
`EHRFlow` (GP-CFM)	**0.817** ± 0.012	**0.969** ± 0.007	**0.889** ± 0.007	**0.934** ± 0.004	0.829 ± 0.010	**0.864** ± 0.004
`EHRFlow` (MMFM)	0.789 ± 0.024	0.968 ± 0.003	0.881 ± 0.014	0.932 ± 0.004	**0.837** ± 0.011	0.862 ± 0.004

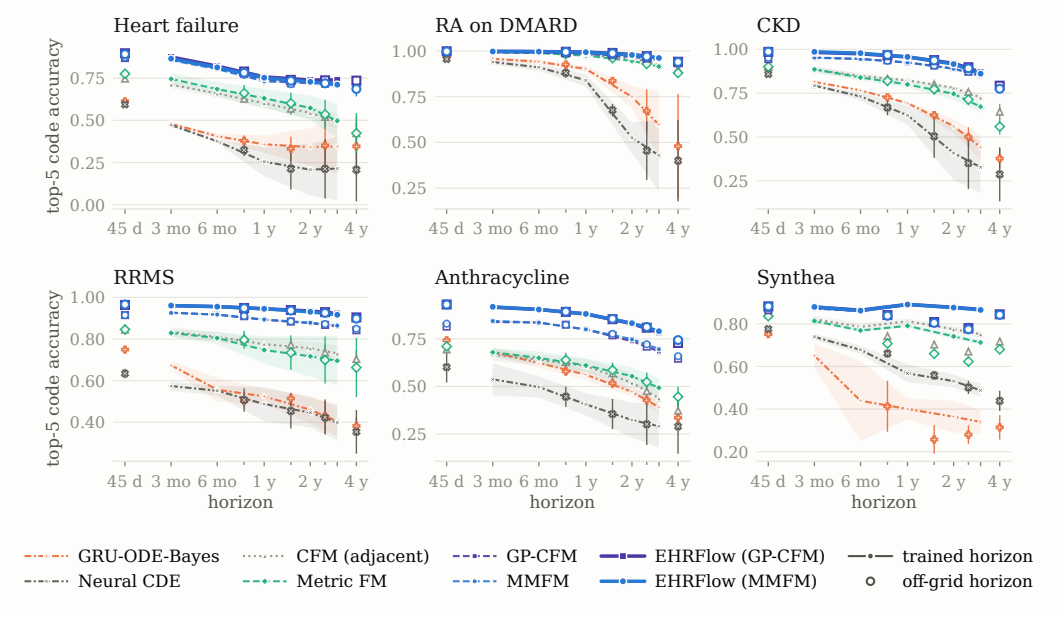


Figure 12: Clinical-code top-5 accuracy at trained and off-grid forecast horizons across the synthetic benchmarks. Off-grid evaluation tests interpolation between training horizons and extrapolation outside the training range. Results are averaged across seeds and shown with ± 1 s.d.

Table 7: Trajectory self-consistency per synthetic dataset and nested pair (h read off a rollout to H). Direct: top-5 code accuracy of a rollout to day h; Nested: the same day read off the rollout to H; Overlap: shared fraction of the two top-5 sets. Mean ± s.d. over seeds.

	`EHRFlow` (GP-CFM)			`EHRFlow` (MMFM)		
	Direct	Nested	Overlap	Direct	Nested	Overlap
Heart failure						
90 d in 180 d	0.872 ± 0.007	0.876 ± 0.008	0.952 ± 0.012	0.867 ± 0.012	0.870 ± 0.014	0.940 ± 0.015
90 d in 365 d	0.872 ± 0.007	0.874 ± 0.008	0.929 ± 0.017	0.867 ± 0.012	0.869 ± 0.009	0.913 ± 0.015
180 d in 365 d	0.811 ± 0.004	0.817 ± 0.005	0.936 ± 0.006	0.807 ± 0.020	0.813 ± 0.018	0.928 ± 0.018
180 d in 730 d	0.811 ± 0.004	0.820 ± 0.006	0.909 ± 0.015	0.807 ± 0.020	0.819 ± 0.006	0.896 ± 0.028
365 d in 730 d	0.743 ± 0.008	0.755 ± 0.009	0.915 ± 0.005	0.749 ± 0.019	0.759 ± 0.015	0.922 ± 0.013
365 d in 1095 d	0.743 ± 0.008	0.751 ± 0.011	0.876 ± 0.027	0.749 ± 0.019	0.760 ± 0.012	0.886 ± 0.027
Mean	0.809 ± 0.003	0.815 ± 0.006	0.920 ± 0.011	0.808 ± 0.017	0.815 ± 0.012	0.914 ± 0.018
RA on DMARD						
90 d in 180 d	0.998 ± 0.000	0.998 ± 0.000	0.962 ± 0.006	0.998 ± 0.001	0.998 ± 0.000	0.961 ± 0.010
90 d in 365 d	0.998 ± 0.000	0.997 ± 0.001	0.938 ± 0.010	0.998 ± 0.001	0.997 ± 0.001	0.935 ± 0.017
180 d in 365 d	0.997 ± 0.001	0.997 ± 0.000	0.955 ± 0.009	0.997 ± 0.001	0.996 ± 0.000	0.954 ± 0.015
180 d in 730 d	0.997 ± 0.001	0.996 ± 0.001	0.932 ± 0.011	0.997 ± 0.001	0.995 ± 0.001	0.930 ± 0.020
365 d in 730 d	0.994 ± 0.001	0.993 ± 0.001	0.951 ± 0.012	0.994 ± 0.001	0.992 ± 0.001	0.955 ± 0.012
365 d in 1095 d	0.994 ± 0.001	0.993 ± 0.000	0.940 ± 0.012	0.994 ± 0.001	0.991 ± 0.001	0.942 ± 0.015
Mean	0.996 ± 0.000	0.996 ± 0.000	0.946 ± 0.008	0.996 ± 0.001	0.995 ± 0.001	0.946 ± 0.014
CKD						
90 d in 180 d	0.984 ± 0.002	0.987 ± 0.001	0.960 ± 0.020	0.986 ± 0.002	0.988 ± 0.001	0.964 ± 0.005
90 d in 365 d	0.984 ± 0.002	0.983 ± 0.003	0.946 ± 0.028	0.986 ± 0.002	0.984 ± 0.001	0.946 ± 0.009
180 d in 365 d	0.976 ± 0.003	0.979 ± 0.001	0.964 ± 0.019	0.978 ± 0.004	0.980 ± 0.001	0.963 ± 0.006
180 d in 730 d	0.976 ± 0.003	0.976 ± 0.002	0.949 ± 0.029	0.978 ± 0.004	0.977 ± 0.002	0.944 ± 0.008
365 d in 730 d	0.957 ± 0.004	0.961 ± 0.002	0.961 ± 0.010	0.957 ± 0.003	0.962 ± 0.003	0.953 ± 0.011
365 d in 1095 d	0.957 ± 0.004	0.959 ± 0.002	0.948 ± 0.017	0.957 ± 0.003	0.960 ± 0.003	0.935 ± 0.013
Mean	0.972 ± 0.002	0.974 ± 0.001	0.955 ± 0.020	0.974 ± 0.003	0.975 ± 0.001	0.951 ± 0.006
RRMS						
90 d in 180 d	0.963 ± 0.003	0.968 ± 0.002	0.962 ± 0.010	0.961 ± 0.004	0.965 ± 0.004	0.947 ± 0.012
90 d in 365 d	0.963 ± 0.003	0.962 ± 0.005	0.942 ± 0.010	0.961 ± 0.004	0.962 ± 0.006	0.925 ± 0.015
180 d in 365 d	0.959 ± 0.002	0.962 ± 0.003	0.969 ± 0.003	0.956 ± 0.006	0.958 ± 0.004	0.958 ± 0.012
180 d in 730 d	0.959 ± 0.002	0.955 ± 0.003	0.945 ± 0.005	0.956 ± 0.006	0.956 ± 0.005	0.933 ± 0.022
365 d in 730 d	0.948 ± 0.001	0.948 ± 0.002	0.971 ± 0.004	0.942 ± 0.009	0.945 ± 0.005	0.956 ± 0.018
365 d in 1095 d	0.948 ± 0.001	0.945 ± 0.003	0.955 ± 0.007	0.942 ± 0.009	0.943 ± 0.007	0.938 ± 0.029
Mean	0.957 ± 0.001	0.957 ± 0.002	0.957 ± 0.005	0.953 ± 0.006	0.955 ± 0.004	0.943 ± 0.017
Anthracycline						
90 d in 180 d	0.915 ± 0.006	0.917 ± 0.006	0.940 ± 0.010	0.920 ± 0.005	0.925 ± 0.003	0.941 ± 0.008
90 d in 365 d	0.915 ± 0.006	0.902 ± 0.004	0.908 ± 0.014	0.920 ± 0.005	0.909 ± 0.005	0.906 ± 0.011
180 d in 365 d	0.904 ± 0.006	0.905 ± 0.007	0.937 ± 0.012	0.899 ± 0.007	0.906 ± 0.005	0.932 ± 0.012
180 d in 730 d	0.904 ± 0.006	0.890 ± 0.006	0.904 ± 0.017	0.899 ± 0.007	0.896 ± 0.007	0.897 ± 0.012
365 d in 730 d	0.879 ± 0.009	0.881 ± 0.012	0.933 ± 0.010	0.864 ± 0.009	0.879 ± 0.007	0.919 ± 0.010
365 d in 1095 d	0.879 ± 0.009	0.870 ± 0.011	0.909 ± 0.017	0.864 ± 0.009	0.873 ± 0.011	0.890 ± 0.011
Mean	0.900 ± 0.006	0.894 ± 0.007	0.922 ± 0.012	0.894 ± 0.007	0.898 ± 0.005	0.914 ± 0.010
Synthea						
90 d in 180 d	0.889 ± 0.004	0.888 ± 0.003	0.945 ± 0.020	0.886 ± 0.008	0.887 ± 0.006	0.935 ± 0.014
90 d in 365 d	0.889 ± 0.004	0.887 ± 0.003	0.925 ± 0.021	0.886 ± 0.008	0.888 ± 0.007	0.910 ± 0.021
180 d in 365 d	0.876 ± 0.003	0.878 ± 0.004	0.945 ± 0.017	0.873 ± 0.003	0.877 ± 0.004	0.935 ± 0.010
180 d in 730 d	0.876 ± 0.003	0.877 ± 0.003	0.923 ± 0.020	0.873 ± 0.003	0.877 ± 0.004	0.911 ± 0.018
365 d in 730 d	0.900 ± 0.004	0.900 ± 0.005	0.936 ± 0.015	0.900 ± 0.004	0.900 ± 0.004	0.921 ± 0.007
365 d in 1095 d	0.900 ± 0.004	0.900 ± 0.005	0.921 ± 0.013	0.900 ± 0.004	0.900 ± 0.005	0.905 ± 0.009
Mean	0.888 ± 0.004	0.888 ± 0.004	0.933 ± 0.017	0.886 ± 0.005	0.888 ± 0.005	0.920 ± 0.012
Average over datasets						
90 d in 180 d	0.937 ± 0.002	0.939 ± 0.001	0.954 ± 0.007	0.936 ± 0.003	0.939 ± 0.003	0.948 ± 0.004
90 d in 365 d	0.937 ± 0.002	0.934 ± 0.003	0.931 ± 0.008	0.936 ± 0.003	0.935 ± 0.002	0.922 ± 0.004
180 d in 365 d	0.921 ± 0.001	0.923 ± 0.001	0.951 ± 0.006	0.918 ± 0.004	0.922 ± 0.003	0.945 ± 0.004
180 d in 730 d	0.921 ± 0.001	0.919 ± 0.002	0.927 ± 0.007	0.918 ± 0.004	0.920 ± 0.002	0.919 ± 0.006
365 d in 730 d	0.904 ± 0.002	0.906 ± 0.002	0.945 ± 0.004	0.901 ± 0.004	0.906 ± 0.003	0.938 ± 0.004
365 d in 1095 d	0.904 ± 0.002	0.903 ± 0.002	0.925 ± 0.007	0.901 ± 0.004	0.904 ± 0.003	0.916 ± 0.007
Mean	0.920 ± 0.000	0.921 ± 0.001	0.939 ± 0.006	0.919 ± 0.003	0.921 ± 0.002	0.931 ± 0.005

C.4 HISTORY-CONDITIONING ABLATIONS

We extend the aggregate ablations in the main text with dataset-level results and an evaluation of how performance changes with the amount of historical context.

Dataset-level ablations. Table 8 separates the effects of model capacity, history availability, optimisation, and the way historical information enters the flow. Matching the vector-field parameter count largely preserves performance, whereas replacing the history representation with zeros returns results close to the history-independent models. This pattern is particularly clear on Anthracycline and RRMS, supporting the interpretation that the gains arise from informative historical context rather than additional vector-field capacity.

Using a learned history representation as the flow source performs less well than conditioning a flow initialised from the encoded anchor state. Jointly training the history encoder with the flow also generally underperforms pretraining and freezing it under the evaluated optimisation settings. Fully end-to-end training likewise underperforms the staged procedure across all datasets for both interpolant families, with increased variability on several benchmarks. Removing the auxiliary decoder loss substantially reduces accuracy across datasets. This ablation removes observation-space supervision and leaves the decoder frozen, so it establishes the importance of this combined training choice rather than isolating either component individually. Finally, adding history to adjacent CFM improves average performance but remains below the multi-marginal variants.

Table 8: History-conditioning ablations per dataset: top-5 clinical-code accuracy averaged over horizons (mean $\pm$ std over seeds).

	Anthracycline	RA on DMARD	CKD	RRMS	Heart failure	Synthea
MMFM path						
`EHRFlow`	**0.865** $\pm$ 0.007	0.985 $\pm$ 0.002	0.940 $\pm$ 0.004	0.942 $\pm$ 0.004	0.775 $\pm$ 0.009	0.875 $\pm$ 0.003
`EHRFlow` param. matched	0.858 $\pm$ 0.004	0.986 $\pm$ 0.001	0.938 $\pm$ 0.005	0.940 $\pm$ 0.006	0.784 $\pm$ 0.010	0.877 $\pm$ 0.004
`EHRFlow`, zero context	0.783 $\pm$ 0.005	0.979 $\pm$ 0.002	0.910 $\pm$ 0.004	0.898 $\pm$ 0.003	0.766 $\pm$ 0.011	0.876 $\pm$ 0.002
`EHRFlow`, end-to-end GRU	0.807 $\pm$ 0.014	0.982 $\pm$ 0.002	0.926 $\pm$ 0.007	0.919 $\pm$ 0.005	0.757 $\pm$ 0.010	0.874 $\pm$ 0.003
`EHRFlow`, fully end-to-end	0.767 $\pm$ 0.021	0.981 $\pm$ 0.002	0.919 $\pm$ 0.006	0.911 $\pm$ 0.011	0.748 $\pm$ 0.009	0.861 $\pm$ 0.008
`EHRFlow`, no decoder loss	0.460 $\pm$ 0.023	0.834 $\pm$ 0.015	0.691 $\pm$ 0.036	0.475 $\pm$ 0.027	0.474 $\pm$ 0.025	0.703 $\pm$ 0.012
History as source	0.780 $\pm$ 0.007	0.978 $\pm$ 0.002	0.910 $\pm$ 0.003	0.892 $\pm$ 0.009	0.763 $\pm$ 0.004	0.878 $\pm$ 0.003
History as source + context	0.773 $\pm$ 0.004	0.979 $\pm$ 0.002	0.905 $\pm$ 0.006	0.886 $\pm$ 0.008	0.770 $\pm$ 0.010	0.877 $\pm$ 0.004
No history	0.785 $\pm$ 0.003	0.980 $\pm$ 0.001	0.913 $\pm$ 0.004	0.896 $\pm$ 0.003	0.767 $\pm$ 0.004	0.874 $\pm$ 0.003
GP-CFM path						
`EHRFlow`	0.863 $\pm$ 0.007	**0.986** $\pm$ 0.002	0.941 $\pm$ 0.001	0.944 $\pm$ 0.005	**0.787** $\pm$ 0.004	0.875 $\pm$ 0.003
`EHRFlow` param. matched	0.863 $\pm$ 0.003	0.984 $\pm$ 0.002	**0.941** $\pm$ 0.003	**0.944** $\pm$ 0.000	0.772 $\pm$ 0.010	0.875 $\pm$ 0.002
`EHRFlow`, zero context	0.780 $\pm$ 0.012	0.980 $\pm$ 0.002	0.913 $\pm$ 0.002	0.894 $\pm$ 0.002	0.768 $\pm$ 0.007	0.873 $\pm$ 0.002
`EHRFlow`, end-to-end GRU	0.793 $\pm$ 0.019	0.979 $\pm$ 0.002	0.919 $\pm$ 0.011	0.913 $\pm$ 0.009	0.761 $\pm$ 0.009	0.871 $\pm$ 0.005
`EHRFlow`, fully end-to-end	0.805 $\pm$ 0.021	0.983 $\pm$ 0.003	0.926 $\pm$ 0.006	0.909 $\pm$ 0.021	0.759 $\pm$ 0.015	0.874 $\pm$ 0.003
`EHRFlow`, no decoder loss	0.492 $\pm$ 0.033	0.829 $\pm$ 0.006	0.666 $\pm$ 0.028	0.453 $\pm$ 0.042	0.465 $\pm$ 0.043	0.690 $\pm$ 0.028
History as source	0.775 $\pm$ 0.006	0.976 $\pm$ 0.001	0.908 $\pm$ 0.002	0.890 $\pm$ 0.009	0.764 $\pm$ 0.005	**0.879** $\pm$ 0.003
History as source + context	0.775 $\pm$ 0.014	0.975 $\pm$ 0.003	0.899 $\pm$ 0.017	0.886 $\pm$ 0.006	0.772 $\pm$ 0.011	0.877 $\pm$ 0.003
No history	0.779 $\pm$ 0.004	0.977 $\pm$ 0.001	0.912 $\pm$ 0.003	0.895 $\pm$ 0.005	0.770 $\pm$ 0.005	0.875 $\pm$ 0.004
CFM (linear)						
CFM (adjacent) + history	0.655 $\pm$ 0.013	0.957 $\pm$ 0.016	0.815 $\pm$ 0.026	0.834 $\pm$ 0.021	0.599 $\pm$ 0.040	0.799 $\pm$ 0.026
CFM (adjacent)	0.574 $\pm$ 0.035	0.971 $\pm$ 0.003	0.812 $\pm$ 0.014	0.781 $\pm$ 0.024	0.602 $\pm$ 0.026	0.790 $\pm$ 0.014
CFM (MMFM windows) + history	0.743 $\pm$ 0.020	0.963 $\pm$ 0.009	0.871 $\pm$ 0.021	0.860 $\pm$ 0.021	0.585 $\pm$ 0.045	0.830 $\pm$ 0.013

Sensitivity to history-window length. Figure 13 varies the number of day-states K available to the history encoder and reports top-5 accuracy averaged across forecast horizons. On Anthracycline and RRMS, performance improves as historical observations are added and then broadly plateaus. CKD shows an earlier improvement followed by relatively little sensitivity to longer windows. In contrast, RA on DMARD and Synthea vary little over the evaluated window lengths, while Heart failure shows modest, non-monotonic changes. The similar patterns across the two interpolant families suggest that sensitivity to context length is primarily associated with the available historical information rather than the particular path construction. The default window of $K = 32$ lies within the region of broadly stable performance across these benchmarks.

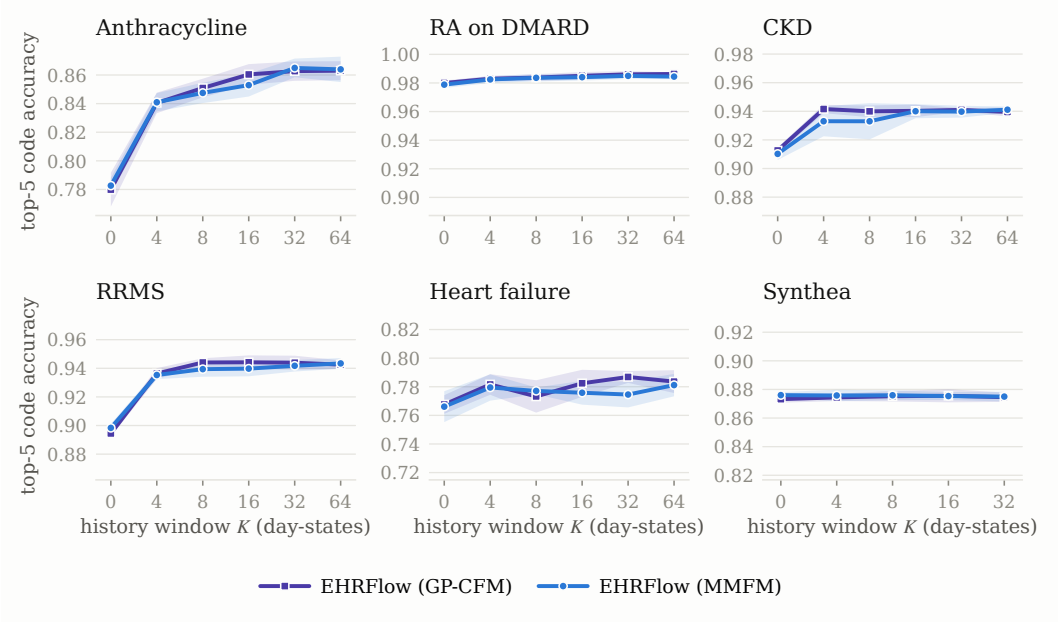


Figure 13: Top-5 clinical-code accuracy, averaged across forecast horizons, is shown for each dataset as the number of day-states K available to the history encoder varies. Lines show the mean across random seeds; shaded bands indicate $\pm$ one standard deviation.

D Experimental details

D.1 Synthetic datasets

To evaluate the benefits of history conditioning in a controlled environment with a ground truth latent trajectory, we constructed five datasets with different history mechanisms. Each dataset provides daily simulation of 20,000 patients for up to five years with no mortality events. Each patient i carries a two-dimensional latent state

$$z_{i,t} = \left(z_{i,t}^{(s)},\, z_{i,t}^{(d)}\right), \tag{13}$$

which defines a disease-severity axis and a disease-domain axis, together with an unobserved scalar baseline $b_i \sim \mathcal{N}(0,1)$ and initial state $z_{i,0}^{(s)} \sim \mathcal{N}(0.25\, b_i,\, 0.35^2)$, $z_{i,0}^{(d)} \sim \mathcal{N}(0,\, 0.25^2)$. To simulate patient evolution, we perform an Euler–Maruyama step

$$z_{i,t+1} = \mathrm{clip}\Big(z_{i,t} + \kappa\, f(z_{i,t}, a_{i,t}, b_i)\,\Delta t + \eta\, \sigma(z_{i,t})\,\xi_{i,t}\sqrt{\Delta t},\ -3,\ 6\Big), \qquad \xi_{i,t} \sim \mathcal{N}(0, I_2), \tag{14}$$

with $\Delta t = 1/365.25$ years, treatment indicator $a_{i,t} \in \{0,1\}$, and dataset-level multipliers κ and η on the whole drift and diffusion (default 1). The shared drift and diffusion are

$$f^{(s)} = r_i\,(0.35 + 0.45\, b_i)\, s_i + 0.25\, z^{(d)} - \tau\, a, \tag{15}$$

$$f^{(d)} = -0.25\, z^{(d)} + 0.12\, z^{(s)} + 0.05\, b_i, \tag{16}$$

$$\sigma(z) = \mathrm{diag}\big(0.08 + 0.03\max(z^{(s)}, 0),\ 0.06 + 0.02\,|z^{(d)}|\big), \tag{17}$$

where r_i is the patient-specific progression rate (1 except in CKD), s_i is a fixed per-patient progression sign drawn once at baseline ($s_i = +1$ with probability 0.35, progressing; $s_i = -0.6$ otherwise, remitting), and $\tau = 0.25$ is the instantaneous treatment effect.

Observation process. We model visits with probability $\min(0.95,\, 10\,\Delta t)$. A visit emits an outpatient encounter, a diagnosis code and two laboratory values. The diagnosis is read off a 5×3 grid over the latent state: the severity bin is the number of edges in $\{-0.5, 0.5, 1.5, 2.5\}$ that lie at or below $z^{(s)}$ and the domain $\in \{\text{renal}, \text{cardiac}, \text{metabolic}\}$ is read off $z^{(d)}$ with edges $\{-0.3, 0.3\}$,

giving 15 diagnosis codes. The laboratory values are $50 + 35\, z^{(s)} + \mathcal{N}(0, 8^2)$ (severity marker) and $1 + 0.25\, z^{(d)} + 0.15\, a + \mathcal{N}(0, 0.08^2)$ (renal marker). Treatment status is redrawn every 30 days as $a \sim \text{Bernoulli}\left(\text{sigmoid}(-2 + 0.25\, z^{(s)} + 1.5\, a + \delta \mathbb{1}[a = 0])\right)$, where δ is a dataset-level start-logit shift (0 unless stated), so patients both start and stop treatment; each start and stop is written as a medication event.

Below, we describe the mechanisms used to introduce dependence on patient history. These synthetic datasets are inspired by broad clinical patterns, such as cumulative treatment exposure and delayed treatment response, but serve as simplified illustrations of history-dependent dynamics rather than validated mechanistic models of disease progression or treatment effects in real patients.

Heart failure: Markov dynamics. Severity follows a five-level ladder with set-points $\mu_k = -1 + k$, $k \in \{0, \ldots, 4\}$; the initial level is read off $z_0^{(s)}$ with the diagnosis-grid edges. On each day, with probability $1/100$, the level moves to an adjacent level, upward with probability 0.7 and downward with probability 0.3, reflecting at the boundaries.

RA on DMARD: delayed treatment effect. Treatment initiation at t_s is observed immediately (medication event) and is absorbing, but the latent responds only inside a lagged window,

$$a_t^{\text{eff}} = \mathbb{1}[t_s + 270 \le t < t_s + 670]\,, \tag{18}$$

during which the treatment term in $f^{(s)}$ is $-3.5\, a_t^{\text{eff}}$; outside it the treatment has no effect on the latent (the renal marker still carries the $0.15\, a$ term). This dataset uses $\kappa = 0.3$, $\eta = 0.7$ and a start-logit shift $\delta = -1.5$.

CKD: patient-specific progression rate. Each patient has a fixed unobserved rate

$$r_i = \text{clip}\big(\text{LogNormal}\big(-\tfrac{1}{2}\, 0.9^2,\, 0.9\big)\,,\ \tfrac{1}{4},\ 4\big)\,, \tag{19}$$

which multiplies only the progression term $(0.35 + 0.45\, b_i)\, s_i$ of $f^{(s)}$, not the coupling or treatment terms. This dataset uses $\kappa = 0.9$, $\eta = 0.7$.

RRMS: relapse history. Patients sit at a baseline set-point $\mu_i^{\text{base}} \in \{-1, 0, 1\}$ (chosen from the initial state) and experience 90-day flares during which the set-point rises by 2 severity bins. The first onset is uniform on the first 270 days; subsequent onsets form a renewal process with

$$T_{k+1} - T_k \;=\; \max\big(91,\ \text{round}(270 \cdot G_k)\big) \text{ days}, \qquad G_k \sim \Gamma(\text{shape} = 64,\ \text{scale} = 1/64), \tag{20}$$

i.e. mean 270 days and coefficient of variation $1/\sqrt{64} = 0.125$.

Anthracycline: cumulative exposure. Thirty-day treatment episodes start with a per-patient daily hazard $h_i = \rho_i/180$, $\rho_i = \text{clip}(\text{LogNormal}(-\frac{1}{2}0.7^2, 0.7), \frac{1}{4}, 4)$, while no episode is active, and each start and stop is written as a medication alongside the shared treatment review above. Exposure accumulates as

$$E_t = \sum_{u \le t} \mathbb{1}[\text{exposed at } u], \tag{21}$$

and once $E_t \ge 180$ days the set-point rises permanently by 2 severity bins,

$$\mu_t = \mu_i^{\text{base}} + 2 \cdot \mathbb{1}[E_t \ge 180], \tag{22}$$

with $\mu_i^{\text{base}} \in \{-1, 0, 1\}$ as in RRMS.

D.2 SYNTHEA.

For a more realistic patient simulator we also generated a cohort of 23,072 patients with Synthea v3.4.0 (Walonoski et al., 2018). Unlike the controlled datasets, Synthea is produced by rule-based clinical state machines rather than our latent simulator. We retain conditions, hospitalisations (inpatient, emergency and urgent-care encounters), observations (laboratory results, vital signs and survey items), medication starts and stops, and deaths; all other encounter classes (ambulatory, wellness, outpatient, home, virtual) are dropped. A summary of the different datasets is provided in Table 9.

Table 9: Synthetic datasets and the source of history dependence. Event and code counts are those of the processed data the models are trained on.

Dataset	Patients	Events	Codes	Predictive information
Heart failure	20,000	5,311,935	20	Current state only
RA on DMARD	20,000	5,013,692	18	Time since treatment initiation
CKD	20,000	5,234,333	20	Historical progression rate
RRMS	20,000	5,264,932	20	Time since previous relapse
Anthracycline	20,000	5,587,408	22	Cumulative treatment exposure
Synthea	23,072	20,586,237	1,893	Rule-based longitudinal disease history

D.3 Real-world datasets

CPRD. We use a heart-failure cohort from the UK Clinical Practice Research Datalink (Herrett et al., 2015) covering both the Aurum and Gold primary-care databases: 868,691 patients (Aurum 815,092, Gold 53,599) with heart-failure index dates between 2006 and 2023, and 650,147,233 events. Event sources are multimorbidity onsets, laboratory tests, medications (drug and drug-class codes), BMI, blood pressure, smoking and alcohol status, frailty scores, NYHA class, heart-failure phenotype, A&E visits, hospital admissions and admission diagnoses, and mortality, plus three static demographic rows per patient (sex, ethnicity, deprivation). We train on Aurum, hold out a random 15% of Aurum patients for validation and epoch selection, and report model performance on all 53,599 Gold patients (35,758,467 events) as an external test set. The study was approved by the relevant CPRD Ethics Board.

AMC-EHR. The dataset is derived from an OMOP CDM instance of a tertiary hospital in the USA (outpatient visits, hospital admissions, emergency visits, observation-stay visits, conditions, drugs, measurements and mortality). We only included patients with at least five events to construct a full trajectory, yielding a dataset of 529,125,023 events over 647,447 patients. Splits are patient-level (15% validation, 15% test). The dataset was exempt from IRB review at the host institution because it was de-identified.

D.4 Event representation

Every dataset is converted to one normalised event table (patient, date, source, code, optional value). The flow and continuous-time models consume a fixed per-event feature vector: a source one-hot, a code one-hot over the 511 most frequent codes plus an unknown token, the standardised value and a has-value flag, four standardised time columns (age, time since registration, time since index diagnosis, days since the previous event), two indicator flags (action event, terminal event) and a constant bias column. Days are binned so a patient trajectory is a sequence of day-states.

D.5 EHRFlow implementation

Latent autoencoder. A variational autoencoder maps each day-state to a latent $z \in \mathbb{R}^d$. Encoder and decoder each have one hidden layer with SiLU activation and dropout 0.05; the encoder outputs mean and log-variance (clamped to $[-10, 10]$), the decoder emits source logits, code logits and the continuous block. Reconstruction loss is cross-entropy on source and code plus $0.1\times$ MSE on the continuous block; the KL term has weight 10^{-3}. The autoencoder is pretrained for 3 epochs with AdamW; the encoder is then frozen and the latent mean is used at evaluation. The decoder keeps training during the flow stage.

Vector field. The velocity network is an MLP with three hidden layers (SiLU, dropout 0.05) over $(z_t,\ s,\ \log(1+\Delta t),\ c)$, where $s \in [0,1]$ is path time, Δt the calendar gap of the pair in days and c the history context (absent on the no-history arms). Parameters are as follows: hidden 488 / latent 64 on the synthetic datasets, 567 / 256 on CPRD, 614 / 64 on AMC-EHR.

History encoder. A single-layer GRU (hidden 128) reads the $K = 32$ day-states ending at the anchor (the anchor and its 31 predecessors) and a linear layer projects its final state to a 32-dimensional

context. The GRU is pretrained for 2 epochs on cross-entropy and then frozen, so the flow stage receives a fixed history summary.

Training. Pairs are anchor/target day-states from the same patient. Flow models sample 4 knots per path at least 2% of the window apart. The GP path draws the interpolant between knots from a Gaussian process with a linear endpoint mean and an RBF kernel on the residual. The spline path (MMFM) is a natural cubic spline through the same knots with the bridge standard deviation of Rohbeck et al. (2025), $\sigma_t = \sigma_{\text{base}}\, 4(t_{k+1} - t)(t - t_k)/\Delta_k^2$, $\sigma_{\text{base}} = 0.1$. The training loss is the MSE between the predicted and target path velocities (with the target detached), plus, with weight 1, the observation-space reconstruction loss of the decoded one-step latent endpoint estimate $D_\phi(z_t + (1-t)\, v_\theta(z_t, t, \lambda, h_a))$ against the final observation of the window.

Optimisation and inference. We used AdamW with learning rate 10^{-3}, weight decay 10^{-6} and gradient-norm clip 1.0. On the synthetic datasets, we used batch size 512 and 16 epochs; on CPRD and AMC-EHR, we used batch size 4,096 and 5 epochs. The epoch with the best validation loss is selected for all models. Trajectories are integrated with 16 Euler-midpoint steps in path time from the anchor latent to the horizon and decoded to source and code logits.

D.6 Baseline implementations

We compare against models that learn longitudinal patient trajectories and support forecasting across multiple horizons. Task-specific predictors for a fixed clinical endpoint may achieve strong endpoint accuracy, but do not provide intermediate patient states, arbitrary-horizon forecasts, or trajectory rollout, and therefore do not assess the same modelling objective. Clinical-code prediction nevertheless provides a common endpoint metric for comparing forecasting accuracy across the trajectory models considered here.

CFM, MMFM, GP-CFM and Metric FM. All flow baselines are the identical network of Appendix D.5 without the history encoder; they differ only in the loss and pair sampler. CFM (adjacent) pairs consecutive day-states with a straight-line path (Lipman et al., 2023). Metric FM learns a geodesic-path network (2 layers of width 32, trained for 5 epochs and frozen) under a LAND diagonal metric $1/(\sum_j w_j\, dx_j^2 + \rho)^\alpha$ with $\rho = 0.01$, $\alpha = 1$ and bandwidth set to the median pairwise latent distance (Kapusniak et al., 2024).

Continuous-time latent-state baselines (GRU-ODE-Bayes, Neural CDE). Both consume the same day-state feature vectors as the flow models and share an observation head: a categorical distribution over the source block, a categorical over the code block (soft targets, since day-states are mean-aggregated) and a diagonal Gaussian over the continuous columns. GRU-ODE-Bayes (De Brouwer et al., 2019) integrates the full GRU-ODE $\mathrm{d}h/\mathrm{d}t = (1 - z(h))(g(h) - h)$ between observations with Euler steps of 7 days and applies the GRU-Bayes update at each observation, with the paper's pre-jump likelihood plus a 10^{-4}-weighted post-jump term. The Neural CDE (Kidger et al., 2020) follows the linearly interpolated day-state path (time as the first channel) with a fixed-step RK4 solver and scores the next observation by extrapolating the state over the gap with the features held fixed; the vector field is a 2-layer MLP of width 256. Both are trained on the observation negative log-likelihood of 64-day-state windows. Training uses AdamW with the flow arms' learning rate and weight decay, a cosine schedule, and 16 epochs.

Decoder-only. A causal Transformer over event tokens: token embedding plus a continuous-time encoding of $(\text{age}/100, \log \text{gap}/10)$ through a two-layer MLP, added to the token embedding; pre-norm encoder layers with GELU and $4\times$ feed-forward width; a causal mask in which same-day tokens cannot attend to each other and the diagonal is always open. Following Delphi-2M there is no separate time head: each token logit is the log-rate of an independent exponential risk, the total rate is $\exp \mathrm{logsumexp}$ over non-padding tokens, and the waiting-time NLL is weighted 1.0 against the next-token cross-entropy (Shmatko et al., 2025). One `<no_event>` filler token is inserted every five years of follow-up from the patient's first event; fillers are trained on but never scored. Vocabulary caps are 32 sources / 512 events. Widths: hidden 116 with 4 layers and 4 heads on the synthetic datasets (144 on Synthea), 190 with 4 layers and 2 heads on CPRD, 172 with 4 layers on AMC-EHR.

Encoder–decoder. Source and code embeddings (source width $\max(8, \text{hidden}/4)$, code width the remainder) are concatenated to the hidden width; the eight numeric columns of the event feature vector pass through a two-layer MLP and are added, followed by a learned positional embedding and layer normalisation (Zhang et al., 2026). Transformer encoder layers process the window and a single learned query is decoded by cross-attention into one state vector that feeds the next-event head and a separate exponential log-rate head for the gap, trained with the same waiting-time NLL with one filler token per five years as above. Widths: hidden 106 with 2 encoder and 2 decoder layers and 2 heads on the synthetic datasets (132 on Synthea), 176 on CPRD, 160 on AMC-EHR.

Endpoints evaluation. For each horizon $H \in \{90, 180, 365, 730, 1095\}$ days an anchor day t is paired with the first day at or after $t + H$ carrying a label-eligible event, provided it lies within $\min(0.25H, 180)$ days of $t + H$. On the synthetic and Synthea datasets every event day of an evaluation patient is an anchor and pairs are capped at 250,000 per horizon with at least one pair per patient. On CPRD and AMC-EHR a single landmark per patient is chosen once for all horizons: the first event with at least three (CPRD) or 32 (AMC-EHR) prior events and, on CPRD, at least 180 days after the heart-failure index date, so that every horizon scores the same anchors. We test each (`EHRFlow` variant, baseline) comparison with a one-sided stratified sign-flip permutation test (H_1: `EHRFlow` higher), datasets as strata so that each dataset carries equal weight regardless of its scale, the statistic being the mean over strata of the mean paired difference. Table cells report mean $\pm$ one standard deviation over seeds unless otherwise specified.

D.7 Ablations implementation

All ablation arms use the synthetic recipe (16 epochs, hidden 488 / latent 64 / 3 hidden layers, MMFM sampler with 4 knots, 5 seeds) and differ from `EHRFlow` in one factor:

- **No history**: the same network without the history encoder (plain GP-CFM or MMFM).
- **`EHRFlow`, end-to-end GRU**: the GRU is not pretrained; it is trained jointly with the vector field through the flow loss.
- **`EHRFlow`, fully end-to-end**: the latent encoder, decoder, history encoder, and vector field are all trained jointly during flow training, rather than freezing the latent and history encoders as in the standard `EHRFlow` training procedure.
- **`EHRFlow`, zero context**: the GRU is instantiated, so the field input width and parameter count are identical to `EHRFlow`, but its output is set to zero in both training and evaluation.
- **`EHRFlow` matched**: hidden width 481, so that the field MLP has the reference parameter count despite the wider input.
- **`EHRFlow`, no decoder loss**: the decoder weight is set to $\beta = 0$, so the flow stage minimises the flow-matching loss alone. The decoder receives no gradient and, like the encoder, stays fixed at its weights from latent pretraining; everything else is `EHRFlow`.
- **History as source**: a variational history encoder maps the last $K = 32$ day-states to a diagonal-Gaussian prior $q_\phi(z_0 \mid \mathcal{H}_a)$, KL-regularised with the same weight as the latent VAE, and a sample from it replaces the encoded anchor as the starting point of the flow. The vector field is the unconditioned GP-CFM / MMFM field.
- **History as source + context**: the same prior is the source, and the same history summary additionally conditions the field as in `EHRFlow`.
- **CFM (adjacent)**: conditional flow matching with the linear interpolant on pairs of consecutive observations, with no multi-observation path construction.
- **CFM (adjacent) + history**: the same model conditioned on the frozen, pretrained history representation h_a, exactly as in `EHRFlow`.
- **CFM (MMFM windows) + history**: the same history-conditioned model trained on the anchor–target pairs used by MMFM, but with a linear interpolant that ignores intermediate observations. This is closely related to the history-conditioned pairwise formulation of trajectory flow matching (Zhang et al., 2024), but omits its uncertainty-prediction component, whose interpretation in observation space would require propagating latent uncertainty through the decoder.

Table 10: Wall-clock time in hours for each of the synthetic datasets (mean ± std over seeds).

Wall-clock (h)	Anthracycline	RA on DMARD	CKD	RRMS	Heart failure	Synthea
Decoder-only	56.94 ± 26.02	47.02 ± 20.78	51.38 ± 26.94	52.84 ± 24.04	50.75 ± 22.82	23.36 ± 11.54
Encoder-decoder	11.41 ± 4.18	12.81 ± 4.68	13.82 ± 4.66	10.02 ± 4.52	15.24 ± 5.64	14.34 ± 4.52
GRU-ODE-Bayes	15.95 ± 2.70	9.78 ± 1.12	12.40 ± 1.53	12.86 ± 2.06	13.00 ± 1.76	13.78 ± 1.81
Neural CDE	4.32 ± 0.92	3.70 ± 0.39	4.37 ± 0.51	4.37 ± 0.48	4.39 ± 0.43	12.72 ± 3.78
CFM (adjacent)	1.15 ± 0.28	0.27 ± 0.00	0.29 ± 0.01	0.52 ± 0.05	0.30 ± 0.01	1.57 ± 1.92
Metric FM	1.31 ± 0.20	0.87 ± 0.19	1.13 ± 0.05	0.57 ± 0.09	1.11 ± 0.15	0.63 ± 0.02
GP-CFM	0.88 ± 0.39	0.48 ± 0.03	0.51 ± 0.03	0.80 ± 0.32	0.88 ± 0.42	0.73 ± 0.03
MMFM	0.55 ± 0.40	0.21 ± 0.02	0.22 ± 0.04	0.54 ± 0.34	0.63 ± 0.42	0.45 ± 0.03
`EHRFlow` (GP-CFM)	0.66 ± 0.24	0.40 ± 0.01	0.43 ± 0.02	0.61 ± 0.19	0.68 ± 0.28	0.76 ± 0.02
`EHRFlow` (MMFM)	0.36 ± 0.20	0.18 ± 0.02	0.19 ± 0.04	0.35 ± 0.17	0.41 ± 0.29	0.50 ± 0.01

D.8 Treatment-effect experiment

Cohort. A Synthea population of 20,000 living patients aged 40–90 is generated once with the stock modules (treatment arm) and once with antihypertensive prescriptions suppressed (control arm), yielding 26,390 and 26,879 patient records including deceased patients (66.6M and 64.0M events). Because Synthea prescribes antihypertensives to every hypertensive patient, we construct the control arm by suppressing prescriptions, obtaining paired simulated trajectories under both treatment conditions.

Estimand and evaluation designs. Treatment is defined as the first antihypertensive prescription. The target outcome is the last systolic blood-pressure reading within 1,825 days of treatment initiation, requiring at least two outcome readings. The simulator-defined ground-truth effect is -20.48 mmHg (sd 14.45; $n = 19{,}618$ paired patients). We evaluate two designs: *Paired*, in which both treatment conditions are available for each patient and effects are averaged over treated patients; and *Split-arm*, in which each patient contributes only one treatment condition, assigned by a hash of the patient identifier.

Guided `EHRFlow`. The treatment arm enters the vector field through a 32-dimensional embedding, replaced by a learned null token with probability 0.1 during training for classifier-free guidance (Ho & Salimans, 2022). At inference, the guided field $v_\varnothing + w\,(v_a - v_\varnothing)$ is integrated with 16 steps. The guidance weight is selected from 14 values in $[0.5, 3.0]$ by minimising validation mean squared error between the rolled-out outcome and each patient's observed factual reading, without access to the counterfactual ground truth.

D.9 Computational resources

All main experiments, including the synthetic benchmarks, ablations, and treatment-effect experiments, were run on a machine with a single NVIDIA RTX 6000 Ada GPU (48 GB) and two AMD EPYC 7713 64-core processors, inside a container limited to 32 CPU threads and a 120 GiB memory cgroup (Ubuntu 22.04, Python 3.13, PyTorch 2.7.1, CUDA 12.6). Table 10 reports wall-clock time per run for the synthetic datasets (training and evaluation). The CPRD and AMC-EHR experiments were run on separate specialised compute environments that complied with the respective data-access and security requirements of the data providers. Unlike the synthetic datasets, these datasets did not fit entirely in memory, so their runtimes were substantially influenced by I/O bottlenecks, with individual runs taking several days. All synthetic experiments, including Synthea and the ablations, were repeated with five random seeds; CPRD and AMC-EHR experiments used three and five seeds, respectively, and treatment-effect recovery experiments used ten seeds.

E AI use statement

We used large language models (LLMs) to assist with improving the clarity of the manuscript, writing code for the experiments and refining the formatting of tables and figures. LLMs were not used for research ideation or any other substantive contributions that would merit authorship. The authors take responsibility for the final content of this work.

F Reproducibility

Implementation details for `EHRFlow` and all baselines, together with the synthetic data-generation procedures, are provided in Appendix D. The Synthea simulator is publicly available at `https://github.com/synthetichealth/synthea`. The CPRD and AMC-EHR datasets cannot be released publicly owing to data-access restrictions. Qualified researchers may apply for access to CPRD data through `https://www.cprd.com/`. Code will be released upon acceptance of the manuscript.