EHR-MPC: Inference-Time Control for Sepsis Treatment with Generative Patient Digital Twins

Joshua Pickard JPICKARD@BROADINSTITUTE.ORG
Broad Institute of MIT and Harvard

Wei Qi
Broad Institute of MIT and Harvard

Na Li
Harvard University

Ann Woolley
Brigham and Women's Hospital

Lisa Cosimi
Brigham and Women's Hospital

Roy Kishony
Technion–Israel Institute of Technology

Deborah Hung
Broad Institute of Harvard and MIT

Abstract

Sepsis is a leading cause of mortality, yet optimal treatment policies remain contested. Existing reinforcement learning (RL) approaches learn fixed strategies for sepsis treatment, limiting adaptability to changing clinical objectives during inference. We propose EHR-MPC, a framework that decouples learning patient dynamics from optimizing treatment by training a patient digital twin in the form of a generative electronic health record (EHR) model. The digital twin predicts clinical trajectories under interventions and enables model predictive control (MPC) to optimize treatments via inference-time planning over simulations. We evaluate EHR-MPC on a multicenter ICU sepsis cohort spanning 8 hospitals in the Mass General Brigham health system using both off-policy importance sampling and on-policy simulation-based evaluation. Relative to RL baselines, EHR-MPC achieves comparable off-policy performance and improved simulation performance. Unlike RL, this work frames sepsis treatment optimization as inference-time control over learned patient dynamics, establishing a general framework for decision making with generative clinical models.

1. Introduction

Treatment of sepsis patients is a major healthcare challenge affecting more than 48 million people and resulting in 11 million deaths per year globally (World Health Organization, 2024). Extensive work has explored optimal sepsis treatment policies in both clinical trials (Venkatesh et al., 2018; Annane et al., 2018) and reinforcement learning (Raghu et al., 2017a,b, 2018; Komorowski et al., 2018; Huang et al., 2022). In particular, corticosteroid administration remains a longstanding clinical question, with more than 60 randomized controlled trials conducted (Schumer, 1976; Annane et al., 2025). Still, evidence guiding which patients should receive corticosteroids, as well as the appropriate timing and dosage, remains contested (Marik, 2018).

While both clinical trials and reinforcement learning (RL) aim to identify optimal policies for sepsis management, their recommendations remain largely disconnected. For instance, one RL study suggests that "optimal treatment [with corticosteroids] may be more restrictive than routine clinical practice" (Bologheanu et al., 2023), in contrast to randomized clinical trial evidence indicating that corticosteroids "probably reduce 28-day and hospital mortality among patients with sepsis" (Annane et al., 2025). This discrepancy highlights a limitation of RL approaches, as despite strong retrospective performance under off-policy evaluation, learned policies are often difficult to interpret, adapt, or validate in clinical practice (Zhang et al., 2022; Frommeyer et al., 2025). Optimizing a fixed reward function without explicitly modeling patient dynamics limits the ability of RL-based treatment strategies to accommodate individualized clinical goals or to adapt to competing objectives (Jayaraman et al., 2024). In practice, sepsis management requires balancing trade-offs such as short-term hemodynamic stabilization versus long-term organ-system preservation, informed by clinician judgment, evolving standards of care, and patient-specific context (Prescott et al., 2026). Because RL typically entangles patient dynamics and reward specification within a single model, adapting a learned policy to new clinical objectives requires retraining, limiting robustness to changing goals, deployment settings, and human-in-the-loop constraints.

The need to improve sepsis management, together with the limitations of prior RL approaches, motivates decoupling the learning of patient dynamics from the optimization of treatment decisions, so that clinical objectives are specified at inference time rather than embedded in a fixed policy. We propose EHR-MPC, a framework that learns a generative digital twin of patient trajectories from electronic health record (EHR) data and performs treatment optimization using model predictive control (MPC). The digital twin simulates counterfactual trajectories under candidate interventions, while an MPC controller evaluates and selects action sequences according to clinically specified objectives at inference time, forming a closed-loop system that can incorporate clinician feedback (Fig. 1). This framework enables treatment strategies to be optimized under new objectives without retraining, supporting flexible, objective-aware, and human-in-the-loop clinical decision-making.

Generalizable Insights about Machine Learning in Healthcare. This work suggests three potential advantages of decoupling dynamics from policy optimization. First, separating patient dynamics from intervention policies changes the structure of the learning problem, as modeling patient trajectories explicitly yields a digital twin that is reusable across objectives, whereas policy learning must entangle both dynamics and treatment objectives. Second, generative EHR models support decision-making algorithms, where pretrained patient digital twins enable downstream optimization via inference-time methods such as model predictive control (MPC). This suggests a broader class of clinical machine learning systems that extend beyond prediction to planning over learned patient simulators. Third, moving from training-time policy optimization to inference-time planning enables adaptation to changing clinical objectives without retraining, supporting more flexible and objective-aware decision-making. Together, these insights point toward a paradigm in which reusable models of patient dynamics replace fixed policies as the primary object of learning, with treatment decisions obtained through inference-time control over these models.

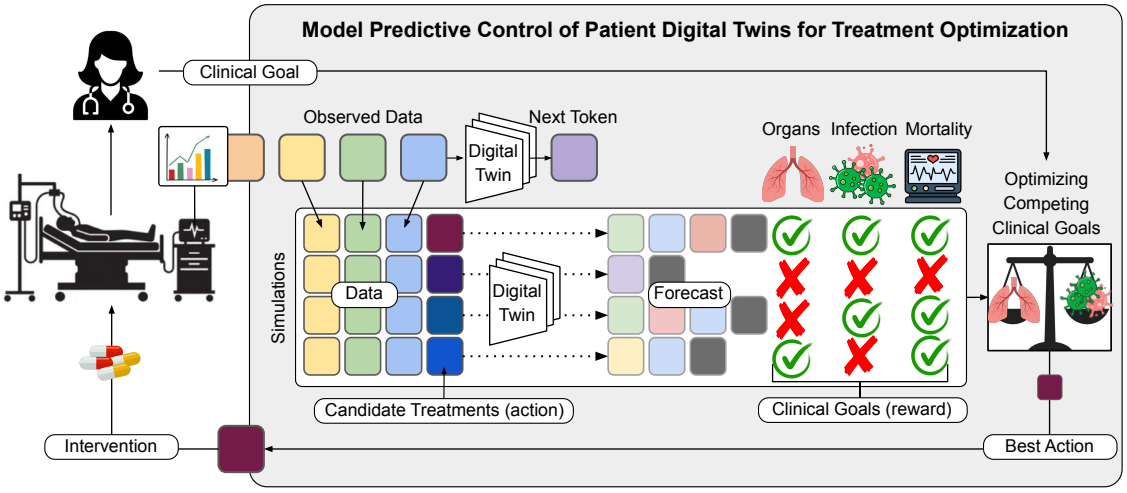

Figure 1: **Digital Twin MPC for Treatment Optimization.** (left) A patient generates data that is observed by a clinician, tokenized, and sent to the digital twin. (simulation) Using observed data, a digital twin performs forecasts on the state of the patient as if several candidate treatments were applied. Each forecast is evaluated according to several clinical outcomes. (optimization) Based upon clinical goals defined at inference time, the action token generating the best simulated outcome is translated to the proposed intervention.

2. Related Work

We compare two existing paradigms for resolving the question of steroid use in sepsis and introduce a third framework. Table 1 summarizes this section.

Optimal use of steroids in sepsis remains unresolved. The use of corticosteroids for sepsis and septic shock has been debated for decades, with trials dating back to the 1970s (Schumer, 1976). Despite substantial randomized controlled trials, no consensus has emerged on if, when, and for whom, steroids are beneficial (Venkatesh et al., 2018; Annane et al., 2018, 2025). This ambiguity reflects that steroid therapy involves a trade-off between suppressing harmful inflammation and impairing host immune response, and the balance of these effects varies across patients and over time (Wiersinga and Seymour, 2018). As a result, the optimal treatment strategy is dynamic and patient-specific, depending on the patient state and treatment goals. These challenges make steroid use in sepsis management a longstanding, open challenge and an example of sequential decision-making under uncertainty, motivating the use of machine learning to find improved treatment strategies.

Reinforcement learning for sepsis. The AI Clinician was an RL model trained to recommend vasopressor and fluid dosing for sepsis patients, reporting improved off-policy evaluation performance (Komorowski et al., 2018). Subsequent work expanded the action space to include steroids (Bologheanu et al., 2023), antibiotics (Futoma et al., 2018; Wang et al., 2024), and continuous treatment representations (Raghu et al., 2017b; Li et al., 2020), while also incorporating safety constraints and uncertainty modeling (Nanayakkara et al., 2022; Tu et al., 2025). A predominant RL paradigm has relied on Q-learning–based methods, with policy evaluation commonly performed using off-policy estimators such as

Table 1: **Comparison of paradigms for sepsis treatment optimization.** The EHR-MPC framework combines the interpretability of clinical trials with the data-driven approach of RL and enables inference-time adaptation. Characterizations vary by implementation.

	Clinical Trials	**RL (value-based)**	**EHR-MPC (ours)**
Data	Prospective	Retrospective	Retrospective
Interpretability	High	Limited	High
Policy	Fixed protocol	Trained	Inference-time
Dynamics	Not modeled	Implicit	Explicit
Adaptivity	Limited	Limited	High
Evaluation	Randomization	Off-policy	Simulation

importance sampling or doubly robust methods (Peng et al., 2018; Jia et al., 2020; Liu et al., 2021; Huang et al., 2022; Böck et al., 2022; Wu et al., 2023; Zhang et al., 2024; Choi et al., 2024; Drudi et al., 2024). Across these approaches, policies are learned from retrospective EHR data by optimizing value functions that implicitly encode both patient dynamics and clinical objectives. Despite extensive methodological development, these approaches have had limited translation into clinical practice. More broadly, many results are validated on a small number of benchmark datasets, such as MIMIC and eICU, raising concerns about generalizability (Pollard et al., 2018; Johnson et al., 2023). Collectively, this suggests that learning policies directly from observational data without explicitly modeling patient dynamics can yield brittle and potentially misaligned treatment recommendations.

Foundation models as patient digital twins for control. Large language models trained on clinical notes and structured EHR tokens have demonstrated strong performance on diagnostic and predictive tasks (Lee et al., 2020; Rasmy et al., 2021). More recently, trajectory-level foundation models have been shown to capture temporal physiological structure, enabling prediction of patient evolution over time (Renc et al., 2024, 2025; Makarov et al., 2025; Li et al., 2025). These models can be interpreted as *digital twins*, that is, data-driven simulators of patient dynamics that evolve under different clinical interventions. This framing enables decision-making through optimization over predicted patient trajectories (Alge et al., 2024; Pickard et al., 2025; Prunella et al., 2026). In parallel, classical model predictive control (MPC) has been applied in clinical settings such as drug delivery and glucose regulation, but relies on hand-crafted physiological models that limit flexibility and scalability (Ionescu et al., 2008; Bruttomesso et al., 2009; Naşcu et al., 2014). In contrast, learned digital twins provide a data-driven foundation for MPC, enabling inference-time optimization over treatments to improve patient outcomes.

3. EHR-MPC: Electronic Health Record Model Predictive Control

We introduce EHR-MPC, a framework that decouples (i) learning patient dynamics from real-time EHR data and (ii) optimizing clinical decisions. We formalize drug administration as a sequential decision-making problem, then present a two-stage approach of a digital twin patient model and an MPC controller that optimizes treatment actions during inference.

3.1. Problem Formulation

We model sepsis management as a state–action–reward process. At each decision time t, the *state* $x(t)$ consists of all clinical data available in the ICU (vitals, labs, interventions, demographics, and device settings). An *action* $a(t)$ includes any clinical intervention, such as the administration of corticosteroids. A *reward* function $R(x(t), a(t))$ encodes the clinical objective, such as the Sequential Organ Failure Assessment (SOFA) score or mortality.

This formulation is compatible with both previous RL studies and the proposed framework. Unlike RL, which explicitly trains a policy by entangling patient dynamics and clinical objectives during training, EHR-MPC trains only a generative model of patient dynamics and obtains treatment decisions entirely through inference-time planning.

3.2. Learning and Simulating Patient Digital Twins

We learn a model to forecast patient trajectories by training on tokenized EHR data. Data are pulled directly from an operational EHR database shared by several hospitals so that the model is designed for constraints that arise during the real-time deployment at a clinical setting.

Tokenized EHR. Each patient is represented as a sequence $x_{1:k} = (x_1, x_2, \ldots, x_k)$ where tokens x_i are ordered in time and correspond to clinical events such as measurements of vitals, labs, device settings, or drug administrations. At time t the state is $x(t) = x_{1:k_t}$. The state variable $x_{1:k}$ represents both historic patient data and clinical interventions. Continuous measurements are discretized into bins, and interventions (e.g., administering corticosteroids) are represented with action tokens. Additional special tokens to indicate start and end of the sequence, unknown clinical events, and the passing of four hour intervals, are also included. From time t to time $t + h$, the patient state updates as:

$$x_{1:k_{t+h}} \leftarrow \text{StreamUpdate}(x_{1:k_t}, t + h). \tag{1}$$

There are $k_{t+h} - k_t$ new tokens added to the patient state during this time interval. The duration h is a time horizon that represents intervals such as the time delay in streaming data from the EHR database or the forecasting horizon for predicting patient outcomes. Tokens are indexed by both order (k) and time (t) to accommodate irregular sampling schedules in EHR and event triggered control (Heemels et al., 2012). The StreamUpdate operator is modeled by the generative patient digital twin.

Simulated patient dynamics. Given the current patient history $x_{1:k_t}$, the digital twin defines the next-token distribution

$$p_\theta(x_{k_t+1} \mid x_{1:k_t}), \tag{2}$$

where θ are the parameters of the twin model. This distribution p_θ can be instantiated using sequence models such as transformers (Makarov et al., 2025; Li et al., 2025). We train a moderate-scale model on our cohort, though the formulation readily scales to larger pretrained models (Renc et al., 2024).

The digital twin p_θ is trained to forecast patient trajectories over a fixed horizon h (e.g., the next 24 hours). Beginning from time t, this corresponds to generating tokens up to a

horizon $t + h$, which are distributed according to

$$p_\theta(x_{k_t:k_{t+h}} \mid x_{k_t}) = \prod_{i=k_t}^{k_{t+h}} p_\theta(x_i \mid x_{<i}). \tag{3}$$

The number of tokens required to reach time $t + h$ depends on the realized sequence. For example, when $h = 24$ hours, generation proceeds until six `[4 hour]` tokens have been produced.

Simulation via Next-Token Prediction. To generate patient trajectories under a candidate treatment sequence $a_{k_t:k_{t+h}}$, we *force* the corresponding action tokens into the input and recursively generate future tokens using the autoregressive model:

$$p(x_{k_t:k_{t+h}} \mid x_{1:k_t}, a_{k_t:k_{t+h}}) = \prod_{i=k_t}^{k_{t+h}} p(x_i \mid x_{<i}, a_{<i}), \tag{4}$$

where a_{k_t} denotes the actions applied up to step i. The resulting model defines a controlled generative process (Plumerault et al., 2020). This procedure yields a counterfactual trajectory in which the generated tokens represent the model's predicted evolution of patient state (e.g., vitals, labs, and interventions) under the proposed treatment.

Eq. 4 enables the model p_θ to act as a *patient digital twin*: a data-driven, virtual replica that evolves in response to real-time EHR data and proposed interventions (Laubenbacher et al., 2024). The digital twin can be viewed as a simulator of patient trajectories, forecasting how physiology may evolve under alternative treatment strategies and enabling comparison of candidate interventions before action.

3.3. Model Predictive Control via Token-Forced Rollouts

To derive treatment recommendations, we apply model predictive control (MPC) to the learned patient digital twin p_θ at each decision time. Rather than learning a fixed policy, the controller performs explicit planning at inference time. MPC evaluates candidate intervention sequences by simulating their effect on future patient trajectories, scores the trajectories under a given objective, and selects the action sequence with the highest predicted utility. Because MPC operates entirely at inference time, the clinical objective (i.e., reward function) can be specified after training. This enables flexible optimization across different and potentially evolving clinical goals, and allows the controller to evaluate interventions that were not explicitly anticipated during training of the digital twin.

Token-forced rollouts. Given a candidate action sequence $a_{t:t+h}$ over a horizon h, we simulate a counterfactual trajectory by *forcing* the corresponding action tokens into the model input and generating future tokens by sampling from the distribution defined in Eq. 4. At each step k_t, the next token is sampled

$$\hat{x}_{k_{t+1}} \sim p(x_{k_{t+1}} \mid x_{1:k_t}, a_{k_t:k_{t+h}}), \tag{5}$$

yielding a trajectory $\hat{x}_{t:t+h}$ that represents the predicted evolution of patient state under the proposed interventions.

This procedure simulates "what would happen" under a treatment sequence. By explicitly modeling intervention effects, the framework yields (i) a mechanistic forecast of patient trajectory evolution, rather than only final outcomes, and (ii) a basis for evaluating and comparing outcomes across alternative treatment strategies.

Trajectory evaluation. Each simulated trajectory is evaluated according to a reward function. The purpose of these functions is to evaluate how well each patient trajectory fits defined clinical endpoints. Common endpoints considered in clinical trials for learning improved sepsis treatment policies include: patient mortality, length of stay in the hospital and ICU, and organ failure, which is commonly measured with the Sequential Organ Failure Assessment (SOFA) score (Vincent et al., 1996).

A reward function $R : x_{1:k_t} \to \mathbb{R}$ assigns a numeric score to each patient trajectory. Computationally, R acts as a decoding objective over trajectories, converting predicted clinical sequences into a scalar utility that can be optimized via search over intervention candidates. These reward functions can be specified in two ways. In an explicit formulation, specific tokens (e.g., `[Mortality]`) directly induce reward or penalty. In an implicit formulation, clinically relevant outcomes such as SOFA score or length of stay are estimated from the trajectory rather than directly observed in the token sequence. This distinction is necessary because several endpoints are not directly represented as tokens. For example, length of stay is only known after transfer or discharge, and SOFA components may be only partially and irregularly observed during an ICU stay. As a result, these quantities are partially observable and must be inferred from the evolving trajectory rather than read off the sequence directly. In both formulations, constructing R can be achieved for any clinical objective by constructing either explicit functions operating on the token sequence or with task-specific heads that predict clinical outcomes.

Evaluation is performed at inference time, enabling the same learned dynamics model to support optimization over multiple, potentially changing objectives without retraining p_θ. This flexibility is a consequence of decoupling patient dynamics from clinical objectives. Because R is never embedded in the model p_θ, new reward functions can be introduced or modified at deployment time without any retraining. This allows a single trained digital twin to simultaneously serve clinicians with different treatment priorities, adapting to evolving standards of care or patient-specific goals without additional model development.

Action selection and receding-horizon control. During inference, the controller selects a sequence of interventions by solving a planning problem over the digital twin:

$$a_t^* = \arg\max_{a_{k_t:k_{t+h}}} \mathbb{E}_{\hat{x}_{k_t:k_{t+h}} \sim p_\theta(\cdot|x_{1:k_t}, a_{k_t:k_{t+h}})} \left[R(\hat{x}_{k_t:k_{t+h}}) \right]. \tag{6}$$

In practice, this optimization is intractable to solve exactly and is approximated via simulation-based search (Garcia et al., 1989). Candidate action sequences are sampled (or constructed via heuristic exploration), their corresponding trajectories are generated by sampling from p_θ, and the resulting trajectories are scored using R. The action sequence with highest estimated value is selected. See Algorithm 1.

Only the first action a_t^* is executed, after which the system observes updated patient data and repeats the optimization. This receding-horizon procedure enables continual replanning as new information becomes available, analogous to clinical practice where treatment decisions are repeatedly updated in response to evolving patient state.

Methodological interpretation. EHR-MPC decouples learning of patient dynamics from decision-making by using a generative digital twin to explicitly simulate future trajectories under candidate intervention sequences. At inference time, decisions are obtained through simulation-based search, where each candidate intervention is evaluated by rolling out the learned model and scoring the resulting trajectory against a clinically specified objective. This shifts computation from offline policy fitting to online planning, enabling clinical objectives to be modified without retraining. This formulation enables three capabilities absent from standard RL approaches, namely (i) inference-time objective specification, (ii) direct inspection of counterfactual patient trajectories, and (iii) reuse of a single dynamics model across multiple clinical goals.

4. Cohort Selection

4.1. Clinical Setting and Data Source

We assembled a multi-site cohort of ICU patients from a large health system comprising eight hospitals. This included patient encounters at two academic medical centers (AMC) and six community hospitals (CH) observed from 2022 onward. Data originate from the institutional electronic health record system, which feeds into a relational database. Patient encounter information, including lab results, vital signs, locations, medication administrations, diagnoses, and other signals were extracted from the production EHR system.

4.2. Cohort Identification

We constructed the study cohort from all patient encounters with at least one ICU admission across a curated set of more than 30 critical care units, spanning medical, surgical, cardiac, and mixed ICUs across all participating sites. From this eligible population, we sampled 36,930 patients, including both sepsis and non-sepsis ICU admissions. The resulting cohort is temporally uniform over the collection dates and stratified across ICU locations in proportion to their underlying patient volumes. Including a broad ICU population was intentional, since sepsis may develop during an ICU stay and diagnostic labeling based on ICD codes is known to be imperfect and sensitive to evolving clinical definitions (Liu et al., 2022). This design reflects deployment conditions in which the model operates over general ICU admissions rather than a pre-filtered diagnostic cohort, and avoids coupling cohort construction to the same coding schemes used to define the target outcome. Cohort characteristics stratified by site are reported in Table 2.

4.3. Data Extraction

For each encounter, we assembled a complete longitudinal EHR record by joining across flowsheet, laboratory, medication, procedure, administrative, demographic, and diagnostic tables within the EHR relational database. We included available records from each patient admission as well as those occurring within two days before or after the patient encounter. Time-unrestricted historical data for comorbidities, prior diagnoses, mortality outcomes, and admission records to other hospitals were also included. This design ensures that each patient representation incorporates both proximal clinical dynamics and relevant longitudinal context.

Table 2: **Patient cohort.** 36k ICU patients from eight hospitals, collected between 2022 and 2026.

Location	Total	MGH	BWH	SLM	NWH	WDH	BWF	CDH	MVH
Hospital Type		AMC	AMC	CH	CH	CH	CH	CH	CH
Demographics									
Patients	36930 (100.0%)	13651	12291	3541	2256	1849	1465	1358	519
LOS (hrs)	66.2 (27.6–154.6)	83.0	74.3	54.7	49.0	38.5	44.9	40.8	7.8
Mortality	12657 (34.3%)	4285	4337	1370	864	561	494	629	117
Female (n)	16101 (43.6%)	5487	5322	1654	1146	841	746	641	254
Median Age (yrs)	67 (55–76)	66	66	68	72	67	67	68	74
Median BMI	26.9 (23.1–31.4)	27.0	26.8	27.0	25.7	27.8	26.4	27.1	26.0
SOFA Scores									
SOFA IQR	2–5	2–5	2–5	2–5	1–4	2–5	1–5	1–4	1–3
Days > 3 subscores (%)	78.8	79.5	81.3	77.3	70.6	74.9	75.5	64.9	57.0
Treatments									
Vasopressor	21097 (57.1%)	9594	7067	1448	1105	842	572	404	65
Antibiotic	22635 (61.3%)	7430	8260	2358	1386	1112	967	907	215
Corticosteroid	12616 (34.2%)	4633	4405	1155	722	551	458	567	125
Diagnoses									
Sepsis	4129 (11.2%)	1053	994	660	348	324	296	410	44
Septic shock	1495 (4.0%)	333	379	211	134	140	135	145	18
Respiratory failure	6741 (18.3%)	1809	1865	1056	353	433	463	667	95
Hepatic failure	831 (2.3%)	385	136	110	40	42	51	59	8
Heart failure	8435 (22.8%)	3000	2451	1069	534	469	409	384	119
Coagulopathy	1226 (3.3%)	406	468	98	81	51	56	43	23
Encephalopathy	1277 (3.5%)	275	222	238	74	130	110	211	17
Pneumonia	5026 (13.6%)	1468	1113	916	372	304	300	424	129
Shock (other)	3581 (9.7%)	1075	1396	236	182	246	261	177	8

Working with a production EHR database introduces several well-known data quality challenges. First, the database schema evolves over time as EHR systems are updated and hospital configurations change. Second, substantial heterogeneity exists across sites in how clinical variables are recorded, including differences in flowsheet structure, laboratory naming conventions, medication ordering systems, and unit-level documentation practices. Third, many clinical events have uncertain or delayed timestamps, particularly for diagnosis codes, laboratory processing times, and documentation-based observations, which can introduce temporal ambiguity in the recorded trajectories.

These challenges are standard in large-scale EHR analysis and are mitigated through careful manual curation of key mappings and alignment between backend data structures and their corresponding clinical semantics. For example, while the SOFA score is derived from more than 100 raw database fields, these map to a small number of underlying physiological measurements. In our framework, the use of a token-based, self-supervised sequence model provides additional robustness to these issues by learning directly from observed event streams without requiring perfect alignment of individual fields or strict synchronization of measurement times.

4.4. Tokenized EHR Representation

Following extraction and preprocessing, each patient trajectory was converted to a discrete token sequence suitable for modeling as described in Section 3. Following Renc et al. (2024), continuous measurements were discretized into bins and medications, procedures, and ad-

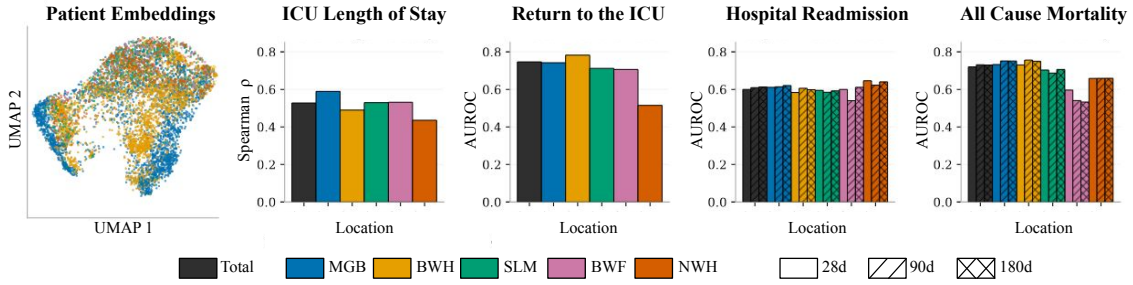

Figure 2: **Representations and Forecasting.** Prediction of outcomes from patient embeddings and random forest classifiers.

ministrative (admit, transfer, discharge, etc.) events were mapped to dedicated vocabulary entries. The tokenized dataset contained 469 million content tokens, 90.3% of which indicated observational data rather than interventions or other data. The median patient trajectory sequence length contained 20,232 tokens of which 1,058 are unique. Detailed token-type breakdowns and vocabulary statistics are provided in Section A.2.

5. Experiments

Following our decoupled formulation of learning and policy optimization, experiments are organized into two stages. First, validating the digital twin as a predictive and intervention-aware model (Section 5.1). Second, evaluating treatment policies derived from the model using both off-policy and simulation-based methods (Section 5.2) [1].

5.1. Digital Twin Construction and Forecasting Evaluation

We train a transformer to parameterize p_θ (Section 3) and evaluate it on the following tasks.

Outcome Prediction from Learned Patient Representations. We assess representation quality learned by the digital twin according to their ability to predict several standard clinical endpoints. We fit lightweight random forest models to predict the ICU length of stay (time-to-event), ICU readmission within the same encounter (binary), hospital readmission (binary), and all-cause mortality (binary) from the internal representations learned by the digital twin of each trajectory until the first ICU day (Fig. 2). Across hospital sites, the embeddings achieve consistent, nontrivial predictive performance, indicating that they capture meaningful signal. We also observe modest site-level structure in the embedding space, suggesting partial but not complete alignment across institutions.

Patient Forecasting. We evaluate the digital twin p_θ as a generative model of clinical time series using next-token prediction (NTP). Performance is assessed via next-m token set recall and token-level accuracy stratified by semantic type (Stein et al., 2023). For multi-step forecasting, we measure whether true future tokens appear within the top-i predictions over horizons $m \in \{5, 10\}$ (Table 3, left). We also report accuracy by token category

1. See the appendix for complete experimental details.

(Table 3, right), including medications, location/unit transitions, and laboratory, vital, and order events. Medication tokens are predicted most accurately, reflecting their temporal persistence and repeated administration. In contrast, laboratory, vital, and order tokens exhibit lower accuracy, consistent with their higher temporal variability and the stochastic nature of short-term physiologic dynamics.

Table 3: **Digital Twin Forecasting.** Next-token prediction (NTP) evaluation. Left: Overall performance (Set Recall, %). Right: Accuracy stratified by token type.

	Next-m tokens	
Top-i	$m = 5$	$m = 10$
10	86.2	82.8
20	89.4	85.9
50	92.9	89.9

Token Type	Top-1	Top-5	Top-10
Medications	54.8	74.0	82.7
Location / unit	31.4	58.7	66.0
Labs, vitals & orders	27.2	54.3	65.0

Dose-Response Sensitivity. We evaluate whether the digital twin exhibits sensitivity to pharmacologic intervention tokens in line with Eq. 4. For each high-acuity patient (SOFA ≥ 6), we construct counterfactual 24-hour trajectories by injecting repeated drug tokens into the observed context as if the drug were administered at a higher dose and then generate future tokens by sampling from the digital twin p_θ. The number of drug tokens is varied from 1 to 20, and estimated mortality risk is predicted using a task specific prediction head (see Table 4). Corticosteroid tokens induce a consistent, monotonic decrease in predicted mortality risk across increasing injection levels. Antibiotics and vasopressors exhibit weaker and less monotonic responses, with more variability across dose levels. These results suggest that the model does not respond uniformly to intervention tokens, but instead exhibits drug-specific variation.

The prediction and forecasting accuracy together with the intervention sensitivity satisfy empirical conditions for local controllability. This supports the use of MPC to steer the learned digital twin.

5.2. EHR-MPC Treatment Policy Optimization

Using the learned digital twin p_θ, we solve Eq. 6 to optimize treatment policies. Evaluating treatment policies learned from observational EHR data presents significant statistical challenges. Off-policy weighted importance sampling (WIS) estimates policy value from observed trajectories but exhibits high variance when the learned and observed policies differ substantially; WIS is included here to enable comparison with prior RL approaches for sepsis that rely on similar protocols. Simulation-based evaluation enables on-policy rollout under the learned dynamics but introduces dependence on model fidelity. Due to these fundamental limitations, we interpret results jointly across both evaluation frameworks, treating consistency between them as stronger evidence than either alone.

Off-Policy Evaluation via Importance Sampling. We evaluate EHR-MPC under two reward functions: (i) minimizing SOFA score and (ii) minimizing mortality risk, comparing against a Q-network baseline and the observed clinician policy using per-decision

Table 4: **Dose-response sensitivity in the digital twin.** Predicted mortality under counterfactual rollouts with increasing numbers of injected intervention tokens. Values correspond to $P(\text{mortality})$ from a frozen prediction head.

Dose tokens	1	2	4	6	8	10	12	15	18	20
Steroids	0.38	0.33	0.41	0.39	0.39	0.37	0.31	0.22	0.16	0.12
Vasopressor	0.49	0.56	0.58	0.66	0.68	0.73	0.67	0.62	0.54	0.52
Antibiotics	0.44	0.50	0.44	0.48	0.45	0.41	0.54	0.47	0.43	0.50

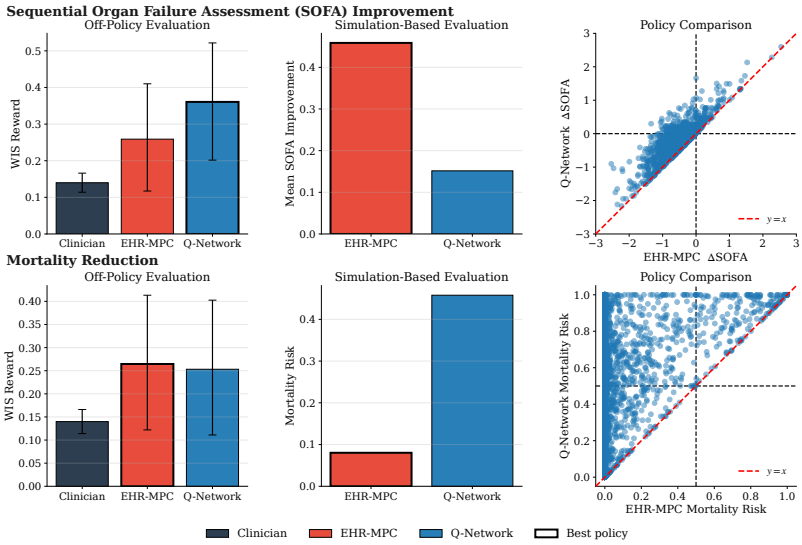

Figure 3: **Offline Evaluation of EHR-MPC Treatment Policies.** The learned EHR-MPC policy is evaluated with WIS (left) and simulated patient outcomes (middle, right). The improvement of patient outcomes according to EHR-MPC and Q-networks is shown per patient (right).

weighted importance sampling (WIS). Although p_θ is trained on the full ICU cohort, evaluation is restricted to sepsis patients. Across both objectives, EHR-MPC and the Q-network achieve higher estimated value than the clinician policy under WIS (Fig. 3, left). Point estimates between EHR-MPC and the Q-network are similar with overlapping confidence intervals.

On-Policy Evaluation via Digital Twin Simulator. We assess learned policies using the digital twin as a simulator. Patient trajectories are initialized from observed tokens up to the first ICU day, after which each policy (EHR-MPC and Q-networks) selects interventions at daily intervals as the simulated patient evolves over time. Unlike WIS, which evaluates fixed logged trajectories under distribution shift, simulation-based evaluation enables on-policy rollouts under the learned digital twin. Across simulations, EHR-MPC consistently outperforms Q-network policies aimed to improve daily SOFA scores and decrease mortality risk, both in terms of average effect (Fig. 3, center) and individual simulations (Fig. 3,

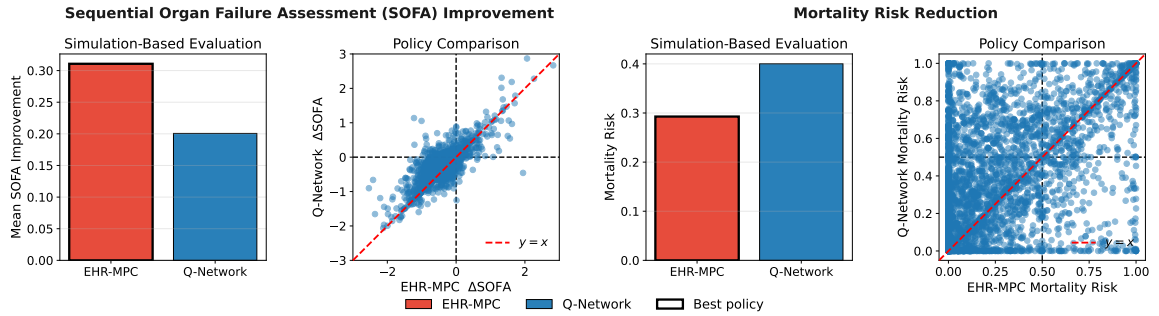

Figure 4: **MPC Robustness to Model Misalignment.** Performance of MPC and Q-networks are evaluated on the digital twin simulator when the MPC controller uses a less accurate version of the patient digital twin.

right). We examine the robustness of this result to model misspecification in the following experiment.

Robustness to Model Misspecification. A potential concern with simulator-based evaluation is that EHR-MPC may trivially optimize the evaluation metric by exploiting the same dynamics model it plans with. To address this, we decouple the *planning model* from the *evaluation model* using checkpoints from different training epochs: MPC planning uses an early-epoch checkpoint (40% of training epochs), while policy scoring and simulation use a later, more-converged checkpoint. The Q-network is also trained on the representations of the planning model used by MPC. This separates action selection from action evaluation, providing a more rigorous test of generalization under model mismatch. Under this protocol, EHR-MPC continues to outperform the clinician policy with only modest performance degradation relative to the same-model baseline (Fig. 4). Degradation is larger for mortality risk minimization than for SOFA, consistent with SOFA providing denser per-step rewards that are easier to optimize.

Treatment Policy Divergence. To complement the off-policy and simulation-based evaluations, we examine differences in action distributions between EHR-MPC, Q-networks, and the observed clinical policy (Fig. 5, top). The policies are stratified by location and recommendations of individual drugs. We further quantify policy differences using Jensen-Shannon divergence (JSD) over action distributions across hospital sites (Fig. 5, bottom). The Q-network policy remains close to the clinical policy across most sites, while EHR-MPC shows greater divergence from both. Finally, the clinical policy shows the greatest heterogeneity across sites, whereas both EHR-MPC and Q-network policies are more consistent across locations (Fig. 6). Analogous analyses for SOFA score reduction are provided in Figs. 7 and 8.

6. Discussion

EHR-MPC frames sepsis treatment optimization as an inference-time planning problem over a learned generative model of patient dynamics. The framework separates two distinct optimization problems of (i) modeling how patients evolve under interventions and (ii)

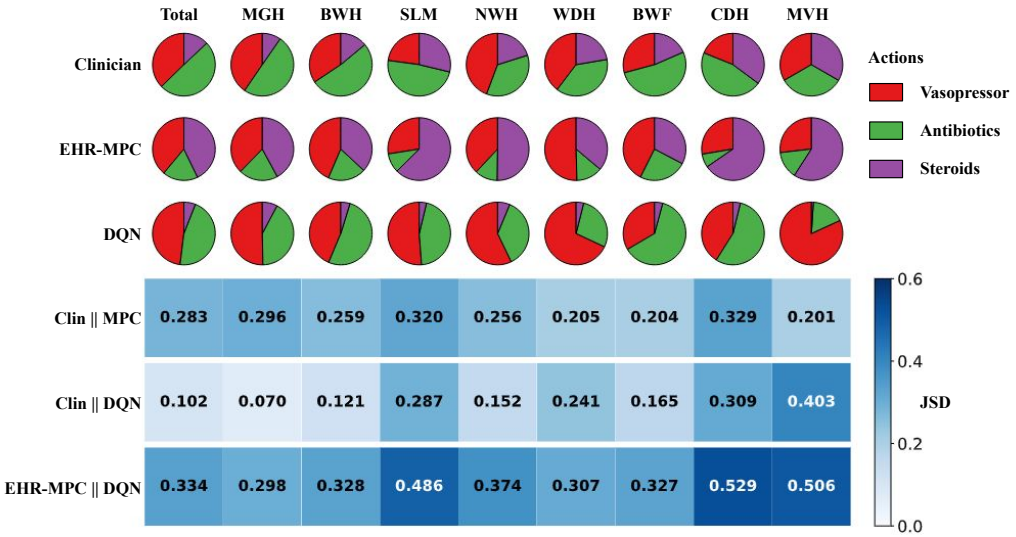

Figure 5: **Policy Divergence for Mortality Optimization.** (Top) The distribution of recommended drugs is shown for each of the three policies across locations. (Bottom) The JSD divergence between these distributions is shown.

selecting interventions that optimize clinical goals. Previous RL for sepsis addressed these tasks jointly to learn a fixed policy. Once p_θ is learned, EHR-MPC determines treatment strategies that are flexible to new clinical objectives without retraining, providing a robust and generalizable tool for decision-making at inference time.

Interpreting Policy Evaluations. Based on the WIS, the EHR-MPC's comparable performance relative to Q-networks demonstrates the decoupled framework for treatment optimization can achieve comparable performance to RL methods that study sepsis treatment optimization. Moreover, the increased divergence between EHR-MPC and the clinician while maintaining comparable performance with the Q-networks, which have policies more similar to the clinicians, indicate that EHR-MPC overcomes the penalty of WIS for diverging from observed data. This is an important property for searching and optimizing the treatment space. Based on the simulations, EHR-MPC consistently outperforms Q-networks on both SOFA improvement and mortality reduction, and this advantage persists under the model misspecification protocol, where the MPC controller planned over a poorly identified simulation model. This robustness rules out the concern that EHR-MPC is merely exploiting a shared dynamics model, and supports the interpretation that online planning provides a robust framework for treatment optimization. Across both evaluation schemes, which are standard in RL and MPC respectively, the results support the use of EHR-MPC as a framework for inference-time treatment optimization.

Examining the action distributions provides further context for these results. Regarding specific treatment recommendations, the Q-network reduces corticosteroid utilization relative to clinical practice, consistent with prior RL studies (Bologheanu et al., 2023). In contrast, EHR-MPC recommends increased steroid use when optimized for mortality reduction, more consistent with clinical trial evidence that corticosteroids reduce 28-day

mortality in sepsis patients (Annane et al., 2025). The site-level variation shows that both learned policies are more consistent than the clinical policy.

Taken together, EHR-MPC is supported both empirically—by strong simulation performance and off-policy estimates—and clinically, by producing treatment recommendations that better align with the broader evidence base than competing algorithms.

Capabilities enabled by inference-time control. By decoupling patient dynamics from treatment optimization, EHR-MPC supports several capabilities that are difficult to achieve with RL models. First, treatment strategies are computed via inference-time planning, allowing decisions to be adapted to patient state and evolving clinical objectives without retraining. Second, a single learned dynamics model can support multiple clinical objectives, enabling trade-offs between competing goals. Third, safety and feasibility constraints can be incorporated directly during planning, allowing candidate interventions to be filtered or penalized without requiring changes to the underlying model. Finally, the framework explicitly models and predicts patient trajectories under interventions, providing clinicians with interpretable forecasts of patient state rather than only scalar reward estimates. Together, these capabilities position EHR-MPC as a flexible decision-support framework that aligns with the requirements of clinical practice.

Limitations. First, the primary limitation of this work is the fidelity and validation of the learned digital twin. While p_θ is trained on intervention-aware patient trajectories, the training process does not provide guarantees of correct system identification. This reflects a well-known challenge in system identification and control (Ljung, 1998; Bemporad and Morari, 2007), and more broadly in off-policy evaluation for sequential decision-making in healthcare (Shalit et al., 2017; Oberst and Sontag, 2019). Improving the reliability of learned patient dynamics, as well as validating their behavior under interventions, remains an open and active problem in data-driven clinical modeling.

Second, offline policy evaluation remains a substantial challenge for validating this work along with other RL for sepsis optimization. Off-policy estimators such as WIS exhibit high variance and sensitivity to support mismatch between policies, specifically when evaluating policies that diverge from clinician behavior (Precup et al., 2000; Gottesman et al., 2018). In contrast, simulation-based evaluation alleviates these statistical issues by enabling rollout under the learned dynamics, but introduces dependence on model accuracy. Neither evaluation framework alone is sufficient, but their consistency provides stronger evidence than either in isolation. While EHR-MPC enables simulation-based, on-policy evaluation, substantial work remains to improve the fidelity, generalizability, and validation of such simulations.

From a computational standpoint, this work adopts standard architectures and evaluation schemes to isolate the effect of the proposed framework. However, each component, including the generative model, the MPC planner, and off-policy evaluation, could be further optimized. While the results demonstrate the effectiveness of EHR-MPC, performance may be improved through more advanced architectures and more extensive hyperparameter tuning. Nonetheless, these results establish the utility of EHR-MPC under standard modeling choices, suggesting opportunities for improvement through advances in model architecture and optimization that are orthogonal to the proposed framework.

Ultimately, the value of decoupling dynamics from policy optimization depends on how faithfully the digital twin captures patient physiology under intervention. Improving digital twin fidelity, validating learned dynamics prospectively, and developing principled evaluation frameworks for simulation-based policies are the central open problems for translating this class of methods into clinical practice.

References

Olivia P Alge, Joshua Pickard, Winston Zhang, Shuyang Cheng, Harm Derksen, Gilbert S Omenn, Jonathan Gryak, J Scott VanEpps, and Kayvan Najarian. Continuous sepsis trajectory prediction using tensor-reduced physiological signals. *Scientific Reports*, 14 (1):18155, 2024.

Djillali Annane, Alain Renault, Christian Brun-Buisson, Bruno Megarbane, Jean-Pierre Quenot, Shidasp Siami, Alain Cariou, Xavier Forceville, Carole Schwebel, Claude Martin, et al. Hydrocortisone plus fludrocortisone for adults with septic shock. *New England Journal of Medicine*, 378(9):809–818, 2018.

Djillali Annane, Josef Briegel, David Granton, Eric Bellissant, Didier Keh, Yizhak Kupfer, Romain Pirracchio, Bram Rochwerg, et al. Corticosteroids for treating sepsis in children and adults. *Cochrane Database of Systematic Reviews*, (6), 2025.

Alberto Bemporad and Manfred Morari. Robust model predictive control: A survey. In *Robustness in identification and control*, pages 207–226. Springer, 2007.

Markus Böck, Julien Malle, Daniel Pasterk, Hrvoje Kukina, Ramin Hasani, and Clemens Heitzinger. Superhuman performance on sepsis mimic-iii data by distributional reinforcement learning. *PLoS One*, 17(11):e0275358, 2022.

Razvan Bologheanu, Lorenz Kapral, Daniel Laxar, Mathias Maleczek, Christoph Dibiasi, Sebastian Zeiner, Asan Agibetov, Ari Ercole, Patrick Thoral, Paul Elbers, et al. Development of a reinforcement learning algorithm to optimize corticosteroid therapy in critically ill patients with sepsis. *Journal of Clinical Medicine*, 12(4):1513, 2023.

Daniela Bruttomesso, Anne Farret, Silvana Costa, Maria Cristina Marescotti, Monica Vettore, Angelo Avogaro, Antonio Tiengo, Chiara Dalla Man, Jerome Place, Andrea Facchinetti, et al. Closed-loop artificial pancreas using subcutaneous glucose sensing and insulin delivery and a model predictive control algorithm: preliminary studies in padova and montpellier, 2009.

Yunho Choi, Songmi Oh, Jin Won Huh, Ho-Taek Joo, Hosu Lee, Wonsang You, Cheng-mok Bae, Jae-Hun Choi, and Kyung-Joong Kim. Deep reinforcement learning extracts the optimal sepsis treatment policy from treatment records. *Communications medicine*, 4(1): 245, 2024.

Cristian Drudi, Maximiliano Mollura, H Lehman Li-wei, and Riccardo Barbieri. A reinforcement learning model for optimal treatment strategies in intensive care: assessment of the role of cardiorespiratory features. *IEEE Open Journal of Engineering in Medicine and Biology*, 5:806–815, 2024.

Timothy C Frommeyer, Michael M Gilbert, Reid M Fursmidt, Youngjun Park, John Paul Khouzam, Garrett V Brittain, Daniel P Frommeyer, Ean S Bett, and Trevor J Bihl. Reinforcement learning and its clinical applications within healthcare: A systematic review of precision medicine and dynamic treatment regimes. In *Healthcare*, volume 13, page 1752. MDPI, 2025.

Joseph Futoma, Anthony Lin, Mark Sendak, Armando Bedoya, Meredith Clement, Cara O'Brien, and Katherine Heller. Learning to treat sepsis with multi-output gaussian process deep recurrent q-networks. 2018.

Carlos E Garcia, David M Prett, and Manfred Morari. Model predictive control: Theory and practice—a survey. *Automatica*, 25(3):335–348, 1989.

Omer Gottesman, Fredrik Johansson, Joshua Meier, Jack Dent, Donghun Lee, Srivatsan Srinivasan, Linying Zhang, Yi Ding, David Wihl, Xuefeng Peng, et al. Evaluating reinforcement learning algorithms in observational health settings. *arXiv preprint arXiv:1805.12298*, 2018.

Wilhelmus PMH Heemels, Karl Henrik Johansson, and Paulo Tabuada. An introduction to event-triggered and self-triggered control. In *2012 ieee 51st ieee conference on decision and control (cdc)*, pages 3270–3285. IEEE, 2012.

Yong Huang, Rui Cao, and Amir Rahmani. Reinforcement learning for sepsis treatment: A continuous action space solution. In *Machine Learning for Healthcare Conference*, pages 631–647. PMLR, 2022.

Clara M Ionescu, Robin De Keyser, Bismark Claure Torrico, Tom De Smet, Michel MRF Struys, and Julio E Normey-Rico. Robust predictive control strategy applied for propofol dosing using bis as a controlled variable during anesthesia. *IEEE Transactions on biomedical engineering*, 55(9):2161–2170, 2008.

Pushkala Jayaraman, Jacob Desman, Moein Sabounchi, Girish N Nadkarni, and Ankit Sakhuja. A primer on reinforcement learning in medicine for clinicians. *NPJ digital medicine*, 7(1):337, 2024.

Yan Jia, John Burden, Tom Lawton, and Ibrahim Habli. Safe reinforcement learning for sepsis treatment. In *2020 IEEE International conference on healthcare informatics (ICHI)*, pages 1–7. IEEE, 2020.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018.

Reinhard Laubenbacher, Borna Mehrad, Ilya Shmulevich, and Natalia Trayanova. Digital twins in medicine. *Nature computational science*, 4(3):184–191, 2024.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Hao Li, Bowen Deng, Chang Xu, Zhiyuan Feng, Viktor Schlegel, Yu-Hao Huang, Yizheng Sun, Jingyuan Sun, Kailai Yang, Yiyao Yu, et al. Mira: Medical time series foundation model for real-world health data. *arXiv preprint arXiv:2506.07584*, 2025.

Luchen Li, Ignacio Albert-Smet, and Aldo A Faisal. Optimizing medical treatment for sepsis in intensive care: from reinforcement learning to pre-trial evaluation. *arXiv preprint arXiv:2003.06474*, 2020.

Bonnie Liu, Milena Hadzi-Tosev, Yang Liu, Kayla J Lucier, Anchit Garg, Sophie Li, Nancy M Heddle, Bram Rochwerg, and Shuoyan Ning. Accuracy of international classification of diseases, 10th revision codes for identifying sepsis: a systematic review and meta-analysis. *Critical care explorations*, 4(11):e0788, 2022.

Ran Liu, Joseph L Greenstein, James C Fackler, Jules Bergmann, Melania M Bembea, and Raimond L Winslow. Offline reinforcement learning with uncertainty for treatment strategies in sepsis. *arXiv preprint arXiv:2107.04491*, 2021.

Lennart Ljung. System identification. In *Signal analysis and prediction*, pages 163–173. Springer, 1998.

Nikita Makarov, Maria Bordukova, Papichaya Quengdaeng, Daniel Garger, Raul Rodriguez-Esteban, Fabian Schmich, and Michael P Menden. Large language models forecast patient health trajectories enabling digital twins. *npj Digital Medicine*, 8(1):588, 2025.

Paul E Marik. Steroids for sepsis: yes, no or maybe. *Journal of Thoracic Disease*, 10(Suppl 9):S1070, 2018.

Thesath Nanayakkara, Gilles Clermont, Christopher James Langmead, and David Swigon. Unifying cardiovascular modelling with deep reinforcement learning for uncertainty aware control of sepsis treatment. *PLOS Digital Health*, 1(2):e0000012, 2022.

Ioana Naşcu, Alexandra Krieger, Clara Mihaela Ionescu, and Efstratios N Pistikopoulos. Advanced model-based control studies for the induction and maintenance of intravenous anaesthesia. *IEEE Transactions on biomedical engineering*, 62(3):832–841, 2014.

Michael Oberst and David Sontag. Counterfactual off-policy evaluation with gumbel-max structural causal models. In *International Conference on Machine Learning*, pages 4881–4890. PMLR, 2019.

Xuefeng Peng, Yi Ding, David Wihl, Omer Gottesman, Matthieu Komorowski, Li-wei H Lehman, Andrew Ross, Aldo Faisal, and Finale Doshi-Velez. Improving sepsis treatment strategies by combining deep and kernel-based reinforcement learning. In *AMIA Annual Symposium Proceedings*, volume 2018, page 887, 2018.

Joshua Pickard, Cooper Stansbury, Amit Surana, Lindsey Muir, Anthony Bloch, and Indika Rajapakse. Dynamic sensor selection for biomarker discovery. *Proceedings of the National Academy of Sciences*, 122(41):e2501324122, 2025.

Antoine Plumerault, Hervé Le Borgne, and Céline Hudelot. Controlling generative models with continuous factors of variations. *arXiv preprint arXiv:2001.10238*, 2020.

Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):180178, 2018.

Doina Precup, Richard S Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. 2000.

Hallie C Prescott, Massimo Antonelli, Waleed Alhazzanic, Morten Hylander Møller, Fayez Alshamsi, Luciano CP Azevedo, Emilie Belley-Cote, Jan De Waele, Lennie Derde, Joanna C Dionnec, et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2026. *Intensive care medicine*, pages 1–74, 2026.

Michela Prunella, Chiara Romano, Alessandro Borri, Nicola Altini, Maria Domenica Di Benedetto, Pieter Annaert, Karel Allegaert, Anne Smits, and Vitoantonio Bevilacqua. Evolutionary digital twin framework for optimal aminoglycoside dosing in neonates with suspected sepsis. *npj Digital Medicine*, 2026.

Aniruddh Raghu, Matthieu Komorowski, Imran Ahmed, Leo Celi, Peter Szolovits, and Marzyeh Ghassemi. Deep reinforcement learning for sepsis treatment. *arXiv preprint arXiv:1711.09602*, 2017a.

Aniruddh Raghu, Matthieu Komorowski, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach. In *Machine learning for healthcare conference*, pages 147–163. PMLR, 2017b.

Aniruddh Raghu, Matthieu Komorowski, and Sumeetpal Singh. Model-based reinforcement learning for sepsis treatment. *arXiv preprint arXiv:1811.09602*, 2018.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86, 2021.

Pawel Renc, Yugang Jia, Anthony E Samir, Jaroslaw Was, Quanzheng Li, David W Bates, and Arkadiusz Sitek. Zero shot health trajectory prediction using transformer. *NPJ digital medicine*, 7(1):256, 2024.

Pawel Renc, Michal K Grzeszczyk, Nassim Oufattole, Deirdre Goode, Yugang Jia, Szymon Bieganski, Matthew BA McDermott, Jaroslaw Was, Anthony E Samir, Jonathan W Cunningham, et al. Foundation model of electronic medical records for adaptive risk estimation. *GigaScience*, 14:giaf107, 2025.

William Schumer. Steroids in the treatment of clinical septic shock. *Annals of surgery*, 184 (3):333–341, 1976.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pages 3076–3085. PMLR, 2017.

George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in Neural Information Processing Systems*, 36:3732–3784, 2023.

Rui Tu, Zhipeng Luo, Chuanliang Pan, Zhong Wang, Jie Su, Yu Zhang, and Yifan Wang. Offline safe reinforcement learning for sepsis treatment: Tackling variable-length episodes with sparse rewards. *Human-Centric Intelligent Systems*, 5(1):63–76, 2025.

Balasubramanian Venkatesh, Simon Finfer, Jeremy Cohen, Dorrilyn Rajbhandari, Yaseen Arabi, Rinaldo Bellomo, Laurent Billot, Maryam Correa, Parisa Glass, Meg Harward, et al. Adjunctive glucocorticoid therapy in patients with septic shock. *New England Journal of Medicine*, 378(9):797–808, 2018.

J-L Vincent, Rui Moreno, Jukka Takala, Sheila Willatts, Arnaldo De Mendonça, Hajo Bruining, C Kathryn Reinhart, PeterM Suter, and Lambertius G Thijs. The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure: On behalf of the working group on sepsis-related problems of the european society of intensive care medicine (see contributors to the project in the appendix). *Intensive care medicine*, 22 (7):707–710, 1996.

Yuan Wang, Anqi Liu, Jucheng Yang, Lin Wang, Ning Xiong, Yisong Cheng, and Qin Wu. Clinical knowledge-guided deep reinforcement learning for sepsis antibiotic dosing recommendations. *Artificial intelligence in medicine*, 150:102811, 2024.

W Joost Wiersinga and Christopher W Seymour. Handbook of sepsis. *Handbook of sepsis*, 2018.

World Health Organization. Sepsis. https://www.who.int/news-room/fact-sheets/detail/sepsis, May 2024. Accessed: 2026-04-03.

XiaoDan Wu, RuiChang Li, Zhen He, TianZhi Yu, and ChangQing Cheng. A value-based deep reinforcement learning model with human expertise in optimal treatment of sepsis. *NPJ Digital Medicine*, 6(1):15, 2023.

Kristine Zhang, Henry Wang, Jianzhun Du, Brian Chu, Aldo Robles Arévalo, Ryan Kindle, Leo Anthony Celi, and Finale Doshi-Velez. An interpretable rl framework for pre-deployment modeling in icu hypotension management. *npj Digital Medicine*, 5(1):173, 2022.

Tianyi Zhang, Yimeng Qu, Deyong Wang, Ming Zhong, Yunzhang Cheng, and Mingwei Zhang. Optimizing sepsis treatment strategies via a reinforcement learning model. *Biomedical Engineering Letters*, 14(2):279–289, 2024.

Appendix A. Cohort and Data Description

A.1. Patient Categorization and Clinical Rewards

Disease definitions. Sepsis and related comorbidities were identified using ICD-10 codes as listed in Table 5.

Diagnosis	ICD-10 Code
Sepsis	A40, A41, R65.2
Septic shock	R65.21
Respiratory failure	J96, J80
AKI	N17
CKD	N18
Hepatic failure	K72
Heart failure	I50
Coagulopathy / DIC	D65, D68
Encephalopathy	G93.4, F05
Pneumonia	J18, J15
Shock (other)	R57.0, R57.1, R57.8, R57.9

Table 5: ICD-10 codes used to define clinical conditions in the study cohort.

Mortality. In-hospital mortality was defined as a documented death during the index hospital encounter or within 72 hours of discharge. Mortality labels were extracted from the Epic EHR system's outcome and encounter tables.

SOFA score. The Sequential Organ Failure Assessment (SOFA) score was computed from raw EHR flowsheet and laboratory data across six organ systems (Table 6). Each patient's stay was partitioned into non-overlapping 24-hour windows anchored at first event; patients with fewer than 24 hours of data were excluded. Within each window, the worst (most abnormal) value was used for each component, except PaO_2 and platelets where the lowest observed value is the worst. Subscores follow standard 0–4 SOFA thresholds; the total is their sum over non-missing components.

Where arterial PaO_2 was unavailable, it was estimated from SpO_2 via a linear approximation of the oxygen–haemoglobin dissociation curve: $\widehat{PaO_2} = 3(SpO_2 - 90) + 60\,\text{mmHg}$, valid for $SpO_2 \in [75, 100]\%$. FiO_2 was resolved by hierarchy: (i) directly measured, (ii) estimated from O_2 flow rate as $20 + 4\times\dot{V}$ (L/min) capped at 60%, or (iii) assumed 21%. Mechanical ventilation was inferred from O_2-device flowsheet values containing "vent"; SOFA respiratory subscores of 3–4 require ventilation (else capped at 2). Vasopressor detection used a fixed-score of 2 regardless of dose, as reliable infusion-rate data were unavailable.

ΔSOFA was computed component-wise between consecutive windows and summed only over components non-missing in *both* windows, avoiding artifacts from missing data:

$$\Delta\text{SOFA}^{(t)} = \textstyle\sum_{k\in\mathcal{K}_t}\big(s_k^{(t+1)} - s_k^{(t)}\big), \tag{7}$$

where $\mathcal{K}_t$ is the set of components with non-missing scores in both window t and $t{+}1$.

Table 6: **SOFA computation rules per 24-hour window.** All EHR field patterns are case-insensitive regular expressions matched on flowsheet event names. "Agg." denotes within-window aggregation. †Scores 3–4 require concurrent mechanical ventilation; otherwise subscore capped at 2.

Component	EHR Field Pattern(s)	Agg.	Thresholds (Score: Range)	Implementation Notes

Respiratory PaO_2/FiO_2	`^PO2` (excl. venous)
	`^SpO2\s*|/|`
	`^FIO2\s*|/|`
	`O2 Flow Rate`
		Min/Max	**0:** $\geq$400
			1: [300, 400)
			2: [200, 300)
			3: [100, 200)†
			4: <100†
				If PaO_2 absent, impute from SpO_2: $\widehat{PaO_2} = 3(SpO_2 - 90) + 60$ mmHg. FiO_2 hierarchy: measured $\rightarrow$ flow estimate ($20 + 4\times$ L/min) $\rightarrow$ 21%.

Coagulation Platelets $10^3/\mu$L	`^PLT$`	Min	**0:** $\geq$150 **1:** [100, 150) **2:** [50, 100) **3:** [20, 50) **4:** <20	
Hepatic Bilirubin mg/dL	`^Total Bilirubin$` `^Bilirubin,\s*Total`	Max	**0:** $\leq$1.2 **1:** (1.2, 2] **2:** (2, 6] **3:** (6, 12] **4:** >12	
Cardio MAP mmHg	`art.*\bmap\b` Fallback: `^MAP\b` Medication names	Min	**0:** MAP $\geq$70 **1:** MAP <70 **2-4:** Dependent on vasoactive medications dose	ART preferred over NIBP
Neuro GCS	`Glasgow Coma Scale`	Min	**0:** 15 **1:** [13, 15) **2:** [10, 13) **3:** [6, 10) **4:** <6	Missing window yields NaN rather than imputing
Renal Creatinine mg/dL	`^Creatinine$` `^Creatinine\s*Whole Bld$`	Max	**0:** $\leq$1.2 **1:** (1.2, 2.0] **2:** (2.0, 3.5] **3:** (3.5, 5.0] **4:** >5.0	

A.2. Patient Tokenization

Tables 7 and 8 summarize the tokenization scheme.

Table 7: **Tokenization statistics.** Content tokens exclude special tokens: `[BOS]`, `[EOS]`, `[PAD]`, `[UNK]`, `[MASK]`, `[4 hours]`.

Statistic	Value	Unit
Vocabulary		
Patients	36,930	people
Vocabulary size	96,268	tokens
Content tokens	469,272,491	tokens
Token type breakdown (content tokens; cohort total)		
Observation (vital / lab)	423,557,340 (90.3%)	tokens
Medication	24,030,597 (5.1%)	tokens
Procedure / order	17,277,634 (3.7%)	tokens
Administrative	4,306,243 (0.9%)	tokens
Sequence statistics per patient (median, IQR)		
Sequence length	20232 (12007–28978)	tokens
Unique tokens	1058 (756–1533)	tokens
Per-patient content-token fractions (median, IQR)		
Observation fraction	88.6 (84.6–92.8)	%
Medication fraction	4.6 (2.9–6.5)	%
Procedure fraction	4.4 (2.9–6.5)	%
Administrative fraction	1.0 (0.4–2.5)	%
Complexity and quality		
Token entropy	3.86 (2.11–5.50)	nats
Vocab coverage (unique / length)	5.6 (4.0–8.7)	%
UNK rate	0.01 (0.00–0.03)	%

Table 8: **Per-hospital tokenization statistics.**

Statistic	Total	MGH	BWH	SLM	NWH	WDH	BWF	CDH	MVH
Sequence statistics (median)									
Patients, n	36,930	13,651	12,291	3,541	2,256	1,849	1,465	1,358	519
Sequence length (tokens)	20232	22878	20940	19568	19827	11155	20165	14535	16485
Unique tokens per patient	1058	1258	1191	840	853	771	820	772	490
Content-token fractions (%)									
Observation (vital / lab)	88.6	90.2	89.9	86.2	85.6	86.5	85.7	84.8	83.1
Medication	4.6	4.4	4.5	4.5	5.3	4.4	5.1	5.4	3.8
Procedure / order	4.4	3.6	4.0	5.4	6.0	6.2	6.1	6.5	6.8
Administrative	1.0	0.6	0.7	3.1	2.1	2.0	2.0	2.4	4.8
Complexity and quality (median)									
Token entropy (nats)	3.86	4.48	4.59	2.12	2.16	3.10	2.15	2.32	1.03
UNK rate (%)	0.01	0.01	0.01	0.02	0.02	0.02	0.02	0.03	0.03

Appendix B. Experimental Details

B.1. Digital Twin Architecture

The patient digital twin is a decoder-only transformer with the following architecture:

- **Model family:** GPT-2-style causal language model trained for next token prediction on the tokenized EHR data and finetuned task heads for predicting mortality risk and daily changes in SOFA.
- **Parameters:** 74.8M total (backbone).
- **Embedding dimension:** $d = 512$.
- **Vocabulary:** consists of discretized vitals, labs, medications, procedures, administrative events, demographics, and special tokens (`[BOS]`, `[EOS]`, `[PAD]`, `[UNK]`, `[4 hour]`).
- **Context length:** 512 tokens (block size).

The base transformer model was trained on 7 NVIDIA A6000 GPUs for approximately 18 hours.

Patient trajectories are long, exceeding the model's context window. To obtain patient trajectory-level representations for outcome prediction, we aggregate representations across sequential, non-overlapping windows spanning the trajectory. This design aligns with the approximately Markovian nature of short-horizon ICU dynamics, consistent with previous sepsis simulators (Böck et al., 2022).

Evaluating Next Token Prediction. Standard top-i accuracy measures whether the single true next token appears in the model's top-i predictions. Set Recall generalizes this to multi-step forecasting horizons.

Given a context prefix $x_{1:k}$, define the *ground-truth set* $\mathcal{T}_m = \{x_{k+1}, x_{k+2}, \ldots, x_{k+m}\}$ as the set of distinct token identities observed in the next m positions. Let $\hat{\mathcal{T}}_i$ be the set of token identities in the model's top-i predictions from the last-position logits:

$$\hat{\mathcal{T}}_i = \{\, \mathrm{id}_j \mid j \in \mathrm{argtop}_i\, p_\theta(x_{k+1} \mid x_{1:k}) \,\}\,. \tag{8}$$

Set Recall at budget i over horizon m is:

$$\mathrm{SR@}(i, m) = \frac{|\mathcal{T}_m \cap \hat{\mathcal{T}}_i|}{|\mathcal{T}_m|}. \tag{9}$$

For each patient in the held-out set, we sample random positions as context endpoints and collect both the immediate next token (for top-i accuracy) and the next-m token set (for Set Recall).

B.2. Outcome Prediction Heads

SOFA delta head. A two-layer MLP with architecture $512 \to 256 \to 7$, predicting 24-hour per-component SOFA deltas. Trained with mean squared error loss on frozen backbone representations. The seven outputs correspond to total SOFA and six organ-system components (respiratory, coagulation, liver, cardiovascular, CNS, renal).

Mortality prediction head. A three-layer MLP with architecture LayerNorm(512) $\rightarrow$ Linear$(512, 256)$ $\rightarrow$ GELU $\rightarrow$ Dropout(0.2) $\rightarrow$ Linear$(256, 128)$ $\rightarrow$ GELU $\rightarrow$ Dropout(0.2) $\rightarrow$ Linear$(128, 1)$, predicting log-odds of in-hospital mortality. Trained with binary cross-entropy loss with class-weight balancing (positive rate: 38.5%) for 100 epochs with early stopping (patience 20), learning rate 10^{-3}, batch size 256, using Adam optimizer. Achieves AUROC = 0.9352 and AUPRC = 0.9142 on held-out patients.

Separate heads are trained for different backbone checkpoints (early and late fine-tuning epochs) to support cross-epoch evaluation.

B.3. MPC Optimization Procedure

The MPC planing algorithm is outlined in Algorithm 1. At each decision point, four intervention options are considered: giving a combination of vasopressor, antibiotics, or steroids, or the option to not give one of these medications. Up to 5 representative drug tokens are considered per drug class. The final hidden state (last-token representation) of the transformer after rollout is extracted and passed to the outcome prediction head. The action minimizing the predicted outcome score is selected.

The autoregressive generation uses a greedy (argmax) decoding strategy, generating future tokens token-by-token until six `[4 hour]` tokens have been produced (approximately 24 hours of simulated patient state).

Algorithm 1 EHR-MPC Planning via Token-Forced Rollouts

Require: Current patient state $x_{1:k_t}$, digital twin p_θ, reward function R, planning horizon h, number of candidates N

for $i = 1$ to N **do**
  Sample candidate action sequence $a^{(i)}_{k_t:k_{t+h}}$
  Initialize trajectory $\hat{x}^{(i)}_{1:k_t} \leftarrow x_{1:k_t}$
  for $\tau = k_t$ to k_{t+h} **do**
    Force action token $a^{(i)}_\tau$ into model input
    Sample next token:

$$\hat{x}^{(i)}_{k_{\tau+1}} \sim p_\theta\big(x_{k_{\tau+1}} \mid \hat{x}^{(i)}_{1:k_\tau}, a^{(i)}_{k_t:k_\tau}\big)$$

  end for
  Compute trajectory reward:

$$J^{(i)} \leftarrow R\big(\hat{x}^{(i)}_{k_t:k_{t+h}}\big)$$

end for
Select best sequence:

$$i^* \leftarrow \arg\max_i J^{(i)}$$

return first action $a^{(i^*)}_t$

B.4. Per-Decision Weighted Importance Sampling (PDWIS)

The behavior policy $\mu(a \mid s)$ is estimated by fitting a multinomial logistic regression classifier on the concatenated last-token representations of all training-set patient decision steps, predicting which of the four action classes was taken by the clinician. At each evaluation step t for trajectory i, the per-step IS ratio is

$$\rho_t^{(i)} = \frac{\pi(a_t^{(i)} \mid s_t^{(i)})}{\mu(a_t^{(i)} \mid s_t^{(i)})}, \tag{10}$$

where π is the evaluation policy (softmax over predicted scores, temperature $\tau = 0.5$). Cumulative weights are clipped at $c_{\max} = 10$ to reduce variance:

$$w_t^{(i)} = \prod_{\tau \le t} \min\left(\rho_\tau^{(i)},\, c_{\max}\right). \tag{11}$$

The PDWIS estimate is

$$\hat{V}^\pi = \frac{\sum_i \sum_t w_t^{(i)}\, r_t^{(i)}}{\sum_i \sum_t w_t^{(i)}}, \tag{12}$$

where $r_t^{(i)} = -\Delta\text{SOFA}_t^{(i)}$ (positive when SOFA improves) is taken from the observed EHR data.

B.5. Reinforcement Learning Baselines

To situate our framework against existing methods, we compare against Q-network-based policies trained on the same frozen patient representations $\mathbf{h}_p$. A Q-network $Q_\psi : \mathbb{R}^{512} \times \mathcal{A} \to \mathbb{R}$ estimates the expected discounted return of each (state, action) pair under a fixed reward function r, and is trained to minimize the Bellman residual:

$$\mathcal{L}_{\text{Q}}(\psi) = \mathbb{E}\left[\left(r(p,a) + \gamma \max_{a'} Q_{\bar{\psi}}(\mathbf{h}_{p'}, a') - Q_\psi(\mathbf{h}_p, a)\right)^2\right], \tag{13}$$

where γ is the discount factor and $Q_{\bar{\psi}}$ is a periodically updated target network. Q-networks constitute the canonical fused policy-dynamics baseline: the value function implicitly encodes both patient dynamics and the clinical objective, and adapting to a new reward function requires full retraining of Q_ψ. We train separate Q-networks for each reward function considered in our experiments. Each Q-network was trained on a single A6000 GPU for up to 300 epochs with a learning rate of 10^{-3}.

B.6. Additional Results

Figures 6-8 illustrate policy distributions across mortality and SOFA rewards and locations.

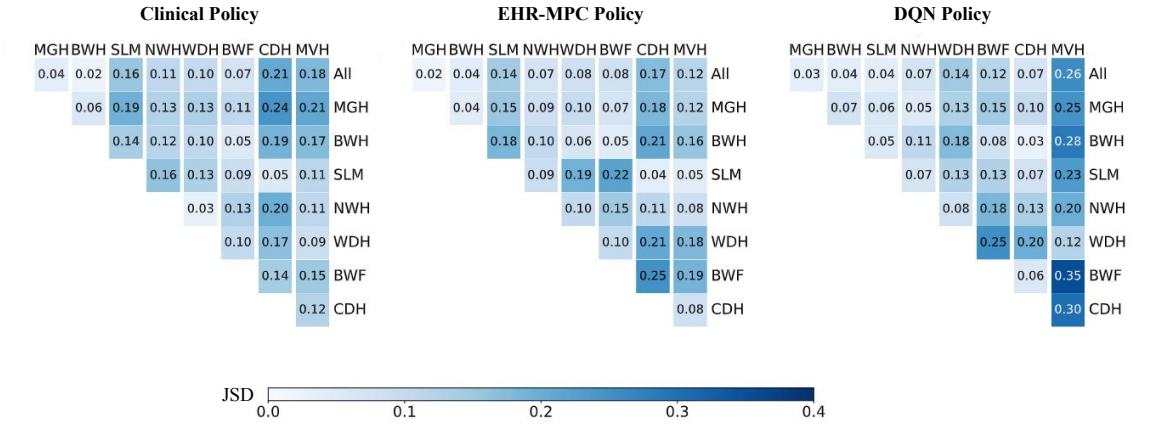

Figure 6: **Policy Variation for Mortality Risk Minimization.**

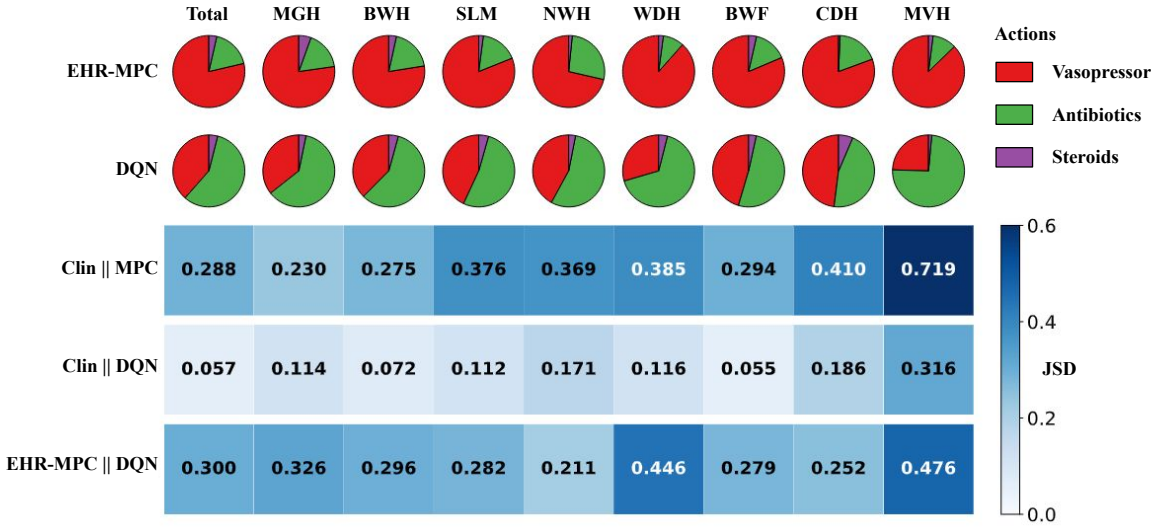

Figure 7: **Policy Distributions and Variation for SOFA Minimization.**

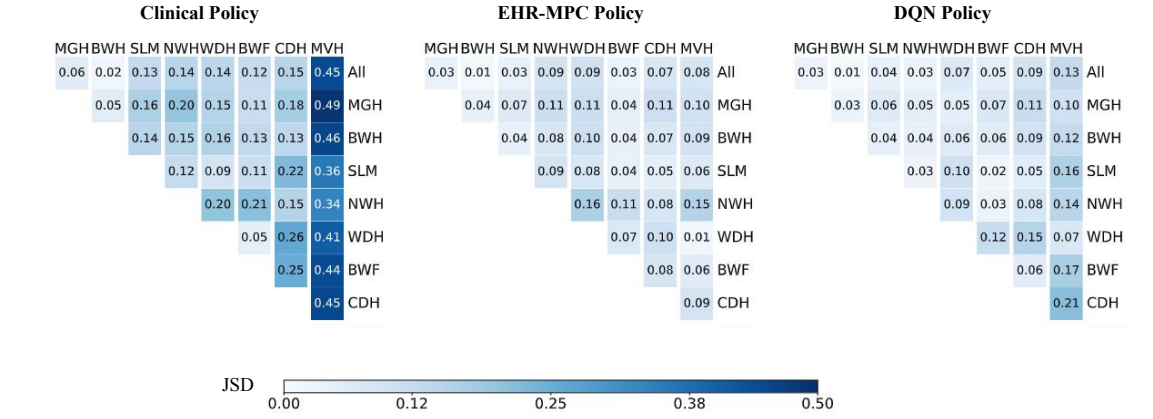

Figure 8: **Policy Variation for SOFA Minimization.**