MedGym: A UNIFIED BENCHMARK FOR DYNAMIC MEDICAL TREATMENT REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical treatment recommendation poses several challenges to reinforcement learning (RL): patient physiology evolves in continuous time, measurements and interventions are performed at irregular intervals, and treatment effects vary substantially across individuals. Existing RL formulations and simulated environments, however, are based on discrete-time MDPs with fixed decision intervals. Thus, it remains difficult to evaluate whether RL methods can handle time-interval-dependent disease progression, personalized treatment response, and safety between consecutive measurement points. To address this gap, we introduce *MedGym*, a benchmark environment for dynamic treatment recommendation. *MedGym* models longitudinal patient evolution in a continuous-time framework and constructs a configurable medical RL benchmark from clinical data by using Physics-Informed Neural Networks. The resulting benchmark enables direct comparison between discrete-time and continuous-time methods under irregular treatment timing and patient-specific dynamics. Furthermore, *MedGym* supports evaluation from clinically important perspectives, such as personalization and trajectory-level safety. By providing a standardized and configurable benchmark for continuous-time dynamic treatment, *MedGym* enables more realistic and informative evaluation of medical RL methods.

1 INTRODUCTION

Medical treatment recommendation can naturally be formulated as a reinforcement learning (RL) problem, as treatment decisions must be dynamically adapted to a patient's clinical state, including vital signs, laboratory test results, symptoms, comorbidities, and prior treatment responses (Chakraborty & Murphy, 2014; Yu et al., 2021). Compared with standard RL benchmark settings, RL for medical treatment recommendation has two distinctive properties. First, a patient's clinical state evolves continuously, while observations are collected and treatments are updated at irregular, decision-dependent times. In contrast, standard RL benchmarks model the environment as a discrete-time Markov decision process (MDP) with uniformly spaced decision steps. Second, even under similar observed conditions and identical interventions, different patients may exhibit distinct latent clinical-state trajectories (Kravitz et al., 2004). In contrast, standard RL benchmarks typically learn a single population-level policy, without explicitly accounting for patient-specific latent dynamics.

Table 1: Comparison of *MedGym* with representative medical RL benchmarks. *MedGym* jointly exposes continuous-time patient evolution, adaptive intervention timing, personalization at multiple levels, online and offline policy learning, and trajectory-level safety under a common evaluation framework. (✓ explicitly supported as a benchmark dimension, ✗ not supported, △ partially supported or a related property is modeled but not evaluated as a controlled benchmark dimension.)

Benchmark	RL for Treatment	Continuous-Time Dynamics	Adaptive Timing	Population / Cluster / Individual Evaluation	Online + Offline Policy Learning	Trajectory-Level Safety
MedAgentGym (et al, 2026)	✗	✗	✗	✗	✗	✗
EpiCare (Hargrave et al., 2024)	✓	✗	✗	△	△	△
ICU-Sepsis (Choudhary et al., 2024)	✓	✗	✗	△	✗	✗
DTR-Bench (Luo et al., 2024)	✓	✗	✗	△	✗	✗
MedGym (ours)	✓	✓	✓	✓	✓	✓

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.

Despite these characteristics, existing medical RL benchmarks largely simplify dynamic treatment through discrete-time formulations with fixed or predefined decision intervals and primarily evaluate population-level policies. MedAgentGym (et al, 2026) focuses on biomedical coding and reasoning rather than treatment-policy learning, whereas treatment-oriented benchmarks, such as EpiCare (Hargrave et al., 2024), ICU-Sepsis (Choudhary et al., 2024), and DTR-Bench (Luo et al., 2024), retain discrete-time formulations. To the best of our knowledge, no existing benchmark jointly supports continuous-time patient evolution, irregular intervention timing, and explicit comparison between population-level and individualized treatment policies. Table 1 summarizes this gap. Consequently, existing benchmarks cannot reliably determine whether an RL method addresses the defining challenges of dynamic treatment or merely performs well under a simplified discrete-time approximation (Gottesman et al., 2018; 2019).

To address this benchmark gap, we introduce *MedGym*, a unified evaluation platform for dynamic medical treatment RL with two complementary layers. The first is a configurable environment-construction layer: a physiology-inspired ODE is combined with a learned neural residual, and a physics-informed neural network (PINN) approximates the resulting solution as a function of the current state, treatment action, and elapsed time, while clinical trajectories are used to construct patient-, cluster-, and population-level dynamics models. The second is a standardized evaluation layer that uses this environment family to make clinically relevant benchmark dimensions directly comparable, such as fixed versus adaptive intervention timing, population versus increasingly personalized policies, and treatment effectiveness versus trajectory-level safety. Thus, the continuous-time patient simulator is an enabling component of *MedGym*, while the benchmark novelty lies in the controlled evaluation capabilities supported by the complete platform.

Our contributions are threefold: (1) We develop a modular pipeline that converts irregular clinical trajectories from different patients into individualized continuous-time treatment simulators. (2) We establish a unified protocol for comparing fixed-interval and time-adaptive methods, population-level, cluster-level, and patient-specific (individual) policies, and offline and online RL. Population- and cluster-level policies are evaluated on held-out patient simulators, while all learned policies are assessed through matched closed-loop rollouts. (3) We instantiate *MedGym* in two clinically distinct treatment settings: sepsis using MIMIC-III, with six physiological state variables, three continuous treatment controls, a SOFA-based objective, and lactate-based trajectory safety; and acute hypotension using the Synthetic Acute Hypotension Dataset, with a MAP-centered treatment objective and distinct physiological safety constraints. Experiments show that intervention timing and the degree of personalization materially affect treatment performance across these settings, demonstrating the importance of evaluating these dimensions explicitly in dynamic treatment recommendation.

2 Dynamic Medical Treatment and Prior Works

Let $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^{d_s}$ and $\mathbf{u} \in \mathcal{U} \subseteq \mathbb{R}^{d_a}$ denote the patient state and treatment action, respectively, where both $\mathcal{X}$ and $\mathcal{U}$ are continuous spaces. The patient state evolves in continuous time according to $\mathbf{x}_t = \mathbf{x}_0 + \int_0^t f_\xi(\mathbf{x}_\tau, \mathbf{u}_\tau, \mathbf{w}_\tau)\mathsf{d}\tau$. Here, $\mathbf{w}_t \in \mathcal{W}$ denotes exogenous uncertainty, accounting for unpredictable factors such as latent physiological variation, heterogeneous treatment sensitivity, and clinical noise. Moreover, ξ characterizes patient-specific features, so that the state transition model may differ across individuals. To assess treatment quality, a bounded reward function $b : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ is defined from clinical health indicators, measuring the utility of treatment decisions over a finite and clinically meaningful range. The goal of dynamic medical treatment is to find a feedback policy $\pi : \mathcal{X} \to \mathcal{U}$ that maximizes the cumulative reward $b(\mathbf{x}_t, \mathbf{u}_t)$ over a fixed horizon $\mathcal{T} = [0, T]$, that is,

$$\max_{\pi \in \Pi} \mathbb{E}\left[\int_{t \in \mathcal{T}} b(\mathbf{x}_t, \pi(\mathbf{x}_t))\mathsf{d}t\right] \quad \text{s.t.} \quad \mathbf{x}_t = \mathbf{x}_0 + \int_0^t f_\xi(\mathbf{x}_\tau, \mathbf{u}_\tau, \mathbf{w}_\tau)\mathsf{d}\tau. \tag{$\mathcal{P}_\xi$}$$

Here, Π denotes the policy class. Patient states are observed, and treatment actions are updated at discrete decision times, which are irregular and depend on clinical judgment (Chakraborty & Murphy, 2014; Shortreed et al., 2011; Yu et al., 2021). Dynamic medical treatment exhibits three key characteristics: (a) Patient states' evolution admits a *continuous-time Markov Decision Process* (CTMDP). (b) Clinical measurements and treatment actions are *generated at irregular time intervals*, as their timing depends on clinicians' judgments of the current patient condition rather than on a fixed decision schedule. (c) Patient responses are heterogeneous, so each individual may follow a distinct dynamics and thus *requires personalized prediction and treatment* (Kravitz et al., 2004).

Recent reinforcement learning approaches to dynamic treatment commonly cast the problem as a discrete-time Markov Decision Process (MDP) with fixed or pre-specified decision intervals (Komorowski et al., 2018; Yan* et al., 2025; Peine et al., 2021). Accordingly, existing simulated benchmarks are also built on the discrete-time MDP formulation (Hargrave et al., 2024). In this setting, $\mathbf{s}_k = \mathbf{x}_{t_k}$ and $\mathbf{a}_k = \mathbf{u}_{t_k}$ denote the measured state and implemented treatment at the k-th decision time, respectively. The subsequent state $\mathbf{s}_{k+1} = \mathbf{x}_{t_{k+1}}$ ($t_{k+1} > t_k$) is assumed to follow the transition dynamics $\mathcal{T}(\mathbf{s}_{k+1} \mid \mathbf{s}_k, \mathbf{a}_k)$, which characterizes the distribution of the next state $\mathbf{s}_{k+1}$ conditioned on the current state $\mathbf{s}_k$ and action $\mathbf{a}_k$. The reward function $r : \mathcal{X} \times \mathcal{U} \to [0, r_{\max}]$ is defined by $r(\mathbf{s}_k, \mathbf{a}_k) := b(\mathbf{s}_k, \mathbf{a}_k)$, that is, it coincides with the instantaneous reward evaluated at time t_k. Then, dynamic medical treatment is represented by the discrete-time MDP $\mathcal{M} := \langle \mathcal{X}, \mathcal{U}, \mathcal{T}, \gamma, \rho_0 \rangle$, where $\gamma \in (0, 1]$ is the discount factor, and ρ_0 denotes the probability density of the initial patient state $\mathbf{s}_0$, typically reflecting the diversity of patient conditions at ICU admission or treatment initiation. A treatment policy is modeled as a stochastic mapping from each state to a probability density over admissible actions. Let $\tau := \{\mathbf{s}_0, \mathbf{a}_0, \ldots, \mathbf{s}_k, \mathbf{a}_k, \ldots\}$ denote a trajectory generated by a policy $\pi \in \Pi$. The value function under policy π and transition model $\mathcal{T}$ is defined as $V^{\pi}_{r,\mathcal{T}}(\mathbf{s}) = \mathbb{E}_{\pi}[\sum_{k=0}^{\infty} \gamma^k r(\mathbf{s}_k, \mathbf{a}_k) \mid \mathbf{s}_0 = \mathbf{s}]$. The corresponding population-level performance is given by $V^{\pi}_{r,\mathcal{T}}(\rho_0) := \mathbb{E}_{\mathbf{s}\sim\rho_0}[V^{\pi}_{r,\mathcal{T}}(\mathbf{s})]$. The resulting RL problem is formulated as

$$\max_{\pi \in \Pi} V^{\pi}_{r,\mathcal{T}}(\rho_0) \tag{DTRL}$$

Existing simulated environments for medical RL are built in discrete time. For example, EpiCare in (Hargrave et al., 2024) simulates longitudinal treatment through timestep-wise state transitions, where the patient evolves from one decision step to the next until remission or termination.

Gaps. By comparing the above two formulations, two important gaps emerge. First, the continuous-time treatment process in $\mathcal{P}_{\xi}$ is reduced to the discrete-time formulation in DTRL. Because clinical decision times t_k and t_{k+1} are generally irregular and condition-dependent (Chakraborty & Murphy, 2014; Murphy, 2005; Kidwell & Almirall, 2023), the evolution from $\mathbf{s}_k = \mathbf{x}_{t_k}$ to $\mathbf{s}_{k+1} = \mathbf{x}_{t_{k+1}}$ should depend on $t_{k+1} - t_k$. However, the transition model $\mathcal{T}(\mathbf{s}_{k+1} \mid \mathbf{s}_k, \mathbf{a}_k)$ omits this dependence and therefore cannot distinguish transitions induced by the same state–action pair over different durations, limiting its fidelity under sparse or unevenly spaced observations (Gottesman et al., 2018; 2019). Second, DTRL insufficiently represents patient heterogeneity. The patient-specific parameter ξ induces dynamics f_{ξ} under which the same intervention may produce different responses across patients (Kravitz et al., 2004). Yet existing benchmarks generally lack individual patient models, preventing patient-level policy evaluation, cross-patient comparison, and assessment of whether population-average optimality translates into optimality for each patient. For a patient model characterized by ξ, the individualized optimal policy can be denoted by $\pi^{\star}_{\xi}$. In contrast, solving DTRL yields a shared policy $\pi(\mathbf{s})$ that assigns the same treatment to various patients with identical observed states, regardless of ξ. Overall, evaluation settings that jointly capture continuous-time disease progression, irregular decision timing, and patient-specific treatment dynamics remain lacking.

3 ENVIRONMENT

Benchmarking pipeline. Figure 1 summarizes how to use *MedGym* as a benchmarking platform. Clinical trajectories, instantiated here by MIMIC-III, are first converted into the continuous-time simulator described later; the same construction pipeline can be re-instantiated with other clinical datasets. The environment supports two policy-learning routes. Online RL methods learn through direct interaction with *MedGym*, whereas offline RL methods learn from fixed trajectory datasets generated from the environment; when desired, a policy obtained by online RL can be used as the behavior policy for constructing such offline datasets. The resulting policies are then assessed through closed-loop rollouts in *MedGym* under a common evaluation interface. Importantly, the evaluation stage is organized around the benchmark dimensions rather than a single aggregate performance score. Policies can be compared under fixed or adaptive intervention timing and across population-, cluster-, and patient-level training settings, while treatment effectiveness and trajectory safety are reported separately through SOFA- and lactate-based metrics, respectively. This separation of environment construction, policy learning, and controlled rollout evaluation allows the same benchmark instance to support consistent comparisons across different RL paradigms and treatment-policy settings.

PINN-based framework. *MedGym* approximates the patient CTMDP using a Physics-Informed Neural Networks (PINN)-based framework as illustrated in Figure 2. The patient evolution is modeled

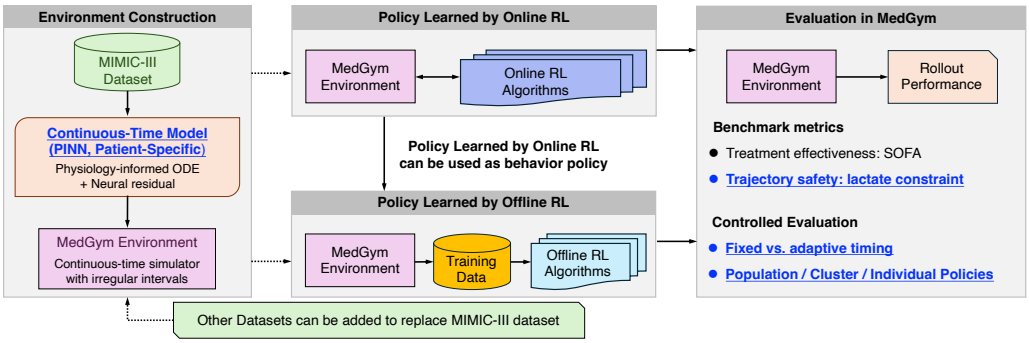


Figure 1: Benchmarking pipeline of *MedGym*. Clinical trajectories are used to construct a PINN-based continuous-time treatment environment, which supports both online RL through environment interaction and offline RL through generated training data. Learned policies are evaluated in *MedGym* under controlled comparisons of fixed versus adaptive timing and population-, cluster-, and individual-level policies, using SOFA-based treatment effectiveness and lactate-based trajectory safety metrics.

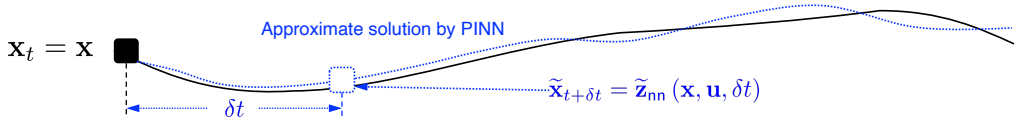


Figure 2: Schematic illustration of the PINN-based state transition in *MedGym*. Starting from the current state $\mathbf{x}_t = \mathbf{x}$, the patient state evolves over an elapsed interval δt under treatment action $\mathbf{u}$, and the future state at $t + \delta t$ is approximated by $\widetilde{\mathbf{x}}_{t+\delta t} = \widetilde{\mathbf{z}}_{\mathsf{nn}}(\mathbf{x}, \mathbf{u}, \delta t)$ from equation 1.

as a stochastic Ordinary Differential Equation (ODE) together with treatment interventions. We use a PINN to yield the following approximate solution:

$$\widetilde{\mathbf{z}}_{\mathsf{nn}}(\mathbf{x}, \mathbf{u}, \delta t) = \mathbf{F}_{\xi}^{\mathsf{nn}}(\mathbf{x}, \mathbf{u}, \delta t) + \mathbf{W}_{\mathsf{c}}, \ \mathbf{W}_{\mathsf{c}} \sim \widetilde{p}(\cdot|\mathbf{x}, \delta t)\,. \tag{1}$$

For any given current state $\mathbf{x}_t = \mathbf{x}$ at t, the estimated state $\widetilde{\mathbf{x}}_{t+\delta t}$ is given by $\widetilde{\mathbf{x}}_{t+\delta t} = \widetilde{\mathbf{z}}_{\mathsf{nn}}(\mathbf{x}, \mathbf{u}, \delta t)$. Here, ξ characterizes patient-specific factors, so that the deterministic component $\mathbf{F}_{\xi}^{\mathsf{nn}}(\mathbf{x}, \mathbf{u}, \delta t)$ captures individualized disease progression and treatment effects. This approximation consists of a deterministic component $\mathbf{F}_{\xi}^{\mathsf{nn}}(\mathbf{x}, \mathbf{u}, \delta t)$ learned by PINN and a stochastic component $\mathbf{W}_{\mathsf{c}} \sim \widetilde{p}(\cdot|\mathbf{x}, \delta t)$, whose conditional distribution depends on both the current state and the elapsed time interval δt. In our implementation, the residual term $\widetilde{p}(\cdot \mid \mathbf{x}_t, \delta t)$ is instantiated as an isotropic Gaussian distribution, i.e., $\mathbf{W}_{\mathsf{c}} \sim \mathcal{N}(\mathbf{0}, \delta t \cdot \sigma^2 \boldsymbol{I})$, providing a stochastic perturbation to the deterministic PINN dynamics. It is reasonable to incorporate δt into the variance scale, since a longer interaction interval may lead to greater uncertainty. The details of training $\mathbf{F}_{\xi}^{\mathsf{nn}}(\mathbf{x}, \mathbf{u}, \delta t)$ are presented in Appendix B.3.

Patient-specific models. Patient specificity is implemented at the dynamics-model level. For patient i, we train $\mathbf{F}_{\xi_i}^{\mathsf{nn}}$ using only that patient's longitudinal trajectory, where ξ_i serves as an abstract index of the resulting individualized dynamics rather than an explicitly inferred patient representation. Thus, different realizations of ξ are operationalized in the current benchmark through different patient-specific datasets and learned dynamics models. This construction is sufficient for the benchmark objective of examining how patient-specific dynamics affect prediction and decision making, while separating it from the distinct problem of identifying patient-specific factors from clinical information. Learning an explicit and transferable representation of ξ is therefore left as an open problem, and future personalization methods can be evaluated in *MedGym* by how well their inferred patient representations recover individualized dynamics and treatment performance.

Reason for using PINN. We adopt a PINN to exploit available physiological structure without assuming that the simplified ODE is exact. The physiology-inspired component provides a clinically meaningful inductive bias, while the neural residual absorbs model mismatch and patient-specific effects from data, offering a practical balance between purely mechanistic and fully data-driven dynamics models. Moreover, learning the resulting solution map as a function of $(\mathbf{x}, \mathbf{u}, \delta t)$ enables

direct evaluation at arbitrary elapsed times without repeatedly solving the ODE during RL rollouts, which is particularly suitable for the irregular-time and simulation-intensive setting of *MedGym*.

Sepsis instantiation. The reward is defined from a smooth SOFA-based surrogate with a penalty on vasopressor usage, while blood lactate is monitored separately as a trajectory-level safety quantity; details are given in Appendix B.2. These components are application-specific rather than intrinsic to *MedGym*: for another disease, the state variables, treatment actions, reward, and safety quantities can be replaced while retaining the same continuous-time environment interface.

Online policy learning algorithms. To evaluate online reinforcement learning methods in *MedGym*, we consider both standard discrete-time RL algorithms and recent continuous-time RL approaches. For the discrete-time setting, representative policy-learning baselines include Soft Actor-Critic (SAC), Proximal Policy Optimization (PPO) (Schulman et al., 2017) and Trust Region Policy Optimization (TRPO), together with safety-aware variants such as PPO Lagrangian (PPO Lag) and TRPO Lagrangian (TRPO Lag). These methods provide suitable baselines for benchmarking treatment performance under the *MedGym* environment. For the continuous-time setting, we follow the bounded-interaction formulation of TACOS (Treven et al., 2024), in which the policy jointly selects a treatment action and the duration for which that action is applied. TACOS reformulates the continuous-time control problem as an equivalent extended discrete-time MDP by accumulating the running reward over each selected physical-time interval. This formulation is particularly suitable for medical treatment, where measurements and treatment updates occur irregularly and the number of interventions is limited. The corresponding reformulation used in *MedGym* is detailed in Appendix C.1. Furthermore, as discussed in (Yan* et al., 2025), online policy learning methods can be naturally adapted to offline policy learning once a predictive environment model is available, thereby connecting online RL and model-based offline RL within the same benchmark framework.

Offline policy learning algorithms. To evaluate offline reinforcement learning methods in *MedGym*, we consider several representative value-based algorithms, including Deep Q-Network (DQN) (Mnih et al., 2015; Raghu et al., 2017; Komorowski et al., 2018), Conservative Q-Learning (CQL) (Kumar et al., 2020), and Guarded CQL (Yan* et al., 2025; Tumay et al., 2025). These methods learn treatment policies from pre-collected trajectories without requiring additional interaction with the environment during training, which is particularly relevant in medical settings where data are often retrospective and online exploration is limited (Levine et al., 2020; Thomas & Brunskill, 2016; Voloshin et al., 2021; Prudencio et al., 2023; Uehara et al., 2022). DQN provides a standard value-based baseline for policy learning from offline data. CQL mitigates extrapolation error by constraining learned actions to remain close to those supported by the dataset, making it particularly suitable for offline decision-making. Guarded CQL further improves robustness by penalizing overestimation on states and actions insufficiently supported by the data. Together, these methods cover a representative spectrum of commonly used offline RL approaches and enable a meaningful comparison of policy learning performance in longitudinal medical treatment settings.

Patient-model split. Let N_{psm} and n_{clu} denote the number of patient-specific models and the desired number of clusters, respectively. As detailed in Appendix B.3, patient models with low mutual estimation error are assigned to the same cluster. For two patient models M_i and M_j, this error evaluates M_j on transitions generated by M_i and M_i on transitions generated by M_j, providing a symmetric measure of their predictive compatibility. The cluster number n_{clu} is configurable, allowing the benchmark to control the trade-off between data sharing and personalization. Each cluster $\mathcal{C}_k, k = 1, ..., N_{\mathsf{psm}}/n_{\mathsf{clu}}$ is partitioned into two disjoint subsets: the training patient set $\mathcal{C}_k^{\mathsf{train}}$ and the test patient set $\mathcal{C}_k^{\mathsf{test}}$, satisfying $\mathcal{C}_k^{\mathsf{train}} \bigcap \mathcal{C}_k^{\mathsf{test}} = \emptyset$ and $\mathcal{C}_k^{\mathsf{train}} \bigcup \mathcal{C}_k^{\mathsf{test}} = \mathcal{C}_k$. All test sets $\mathcal{C}_k^{\mathsf{test}}$ contain the same number of patient models.

Cluster- and population-level PINN models. For any given patient model $M_{i,k} \in \mathcal{C}_k^{\mathsf{train}}$, the training dataset $D_{i,k}^{\mathsf{tp}}$ used for PINN regression is generated through system identification. Starting states, admissible treatment actions, and elapsed time intervals are sampled, after which the patient model is simulated under each treatment action over the corresponding interval. Each generated transition records the current state, applied action, elapsed interval, and resulting successor state. We then form the $k-$th cluster training dataset $\mathcal{D}_{k,\mathsf{clu}}^{\mathsf{tp}} = \bigcup D_{i,k}^{\mathsf{tp}}$ by aggregating the datasets of all patient models in $\mathcal{C}_k^{\mathsf{train}}$. Similarly, the population training dataset is defined by $\mathcal{D}_{\mathsf{pop}}^{\mathsf{tp}} = \bigcup_k \mathcal{D}_{k,\mathsf{clu}}^{\mathsf{tp}}$. The PINNs trained on $\mathcal{D}_{k,\mathsf{clu}}^{\mathsf{tp}}$ and $\mathcal{D}_{\mathsf{pop}}^{\mathsf{tp}}$ are referred to as the k-th cluster PINN model and the population PINN model, respectively. Note that the test patients are not used for population or cluster model construction.

Individual-, Cluster-, and population-level policies. A patient-specific policy $\pi_{i,\mathsf{ind}}$ is trained either online through direct interaction with a given patient model M_i or offline using the fixed dataset $D^{\mathsf{pl}}_{i,k}$. The latter is collected by rolling out a behavior policy on the patient model M_i. The k-th cluster-level policy $\pi_{k,\mathsf{clu}}$ is trained online by interacting with the k-th cluster PINN, and the population policy π_{pop} by interacting with the population PINN, so that each policy is optimized against the dynamics model of its own personalization level. For offline learning, $\pi_{k,\mathsf{clu}}$ is trained on the aggregated dataset $\mathcal{D}^{\mathsf{pl}}_k = \bigcup_i D^{\mathsf{pl}}_{i,k}$ and π_{pop} on the population-level dataset $\mathcal{D}^{\mathsf{pl}}_{\mathsf{pop}} = \bigcup_k \mathcal{D}^{\mathsf{pl}}_k$. For offline learning, π_{pop} is trained on the population-level dataset $\mathcal{D}^{\mathsf{pl}}_{\mathsf{pop}} = \bigcup_k \mathcal{D}^{\mathsf{pl}}_k$.

4 Evaluation Experiments

Evaluated Policies. We compare SAC (Haarnoja et al., 2018), TRPO (Schulman et al., 2015), and TRPOLag (Pasula, 2020) under fixed and adaptive interaction schedules. The adaptive interaction mechanism follows (Treven et al., 2024). We denote each method by ALG-T-G, where ALG $\in$ SAC, TRPO, TRPOLag denotes the RL algorithm, T $\in$ F, A denotes fixed or adaptive timing, and G $\in$ Pop, Clu, Ind denotes the population, cluster, or individual model or policy.

Model and Dataset. The settings used to construct the individual-, cluster-, and population-level PINN model and policies are summarized as follows: The number of cohorts is set as $N_{\mathsf{coh}} = 7$; In each cohort, the total number of patient-specific models is $N_{\mathsf{psm}} = 110$; Clusters per cohort is $N_{\mathsf{clu}} = 10$; Number of training models per cluster is $|\mathcal{C}^{\mathsf{train}}_k| = 10$; Number of test models per cluster is $|\mathcal{C}^{\mathsf{test}}_k| = 1$; Number of training models per cohort is $\sum_{k=1}^{N_{\mathsf{clu}}} |\mathcal{C}^{\mathsf{train}}_k| = 100$; Number of test models per cohort is $\sum_{k=1}^{N_{\mathsf{clu}}} |\mathcal{C}^{\mathsf{test}}_k| = 10$. Across the $N_{\mathsf{coh}} = 7$ cohorts this gives 770 patient-specific models and 70 held-out test patients, which is the evaluation set used in Table 2. For the offline setting, we adopt the same construction with a single cohort ($N_{\mathsf{coh}} = 1$), while all other settings remain unchanged. Complete optimization hyperparameters are provided in Appendix Table 7.

Benchmark questions. The evaluation is organized around four questions corresponding to the main benchmark dimensions of *MedGym*. **(Q1) Intervention timing:** Does allowing the policy to adapt the intervention interval improve treatment performance relative to a fixed interaction schedule? **(Q2) Personalization:** How does policy performance change from population-level to cluster-level and patient-level treatment, and how much of the individual-oracle benefit can be recovered by cluster-based policies? **(Q3) Learning setting:** Do the conclusions observed under online policy learning remain valid when policies must instead be learned from fixed offline datasets? **(Q4) Treatment effectiveness versus safety:** Do improvements in treatment effectiveness also translate into safer physiological trajectories, or can the two evaluation criteria lead to different conclusions?

4.1 Results and Discussions

Figures 3 and 4 jointly illustrate how the degree of personalization in dynamics modeling and policy learning affects prediction and control for the same test patient. The patient-specific setting serves as an oracle reference: its PINN attains the highest prediction fidelity and its policy the most favorable control outcomes, and both degrade as the level of personalization coarsens from cluster-level to population-level. The degradation in control is not a fidelity artefact, since all three policies are evaluated on the same patient-specific simulator and only the training regime changes. This consistent ordering indicates that patient heterogeneity should be considered in both environment construction and policy learning. An important direction for future medical RL is therefore to infer explicit and transferable patient representations, such as ξ, from limited clinical histories and use them for patient-adaptive policy optimization on unseen patients. *MedGym* provides a unified continuous-time benchmark for evaluating how closely such methods recover individualized dynamics and treatment performance relative to cluster-level, population-level, and individual-oracle references.

We next examine whether the patient-level observations persist across the held-out cohort. The normalized-error distributions in Figure 5 show that the patient-specific PINNs generally provide more accurate rollouts than the population-level PINN. The cohort-level results in Figure 6 and Table 2 are consistent with the preceding patient-level case study. Under both timing settings, the mean trajectory-wide and final SOFA scores follow the ordering Ind $<$ Clu $<$ Pop. The cluster policy outperforms its matched population policy for 71.4% and 65.7% of patients under fixed and

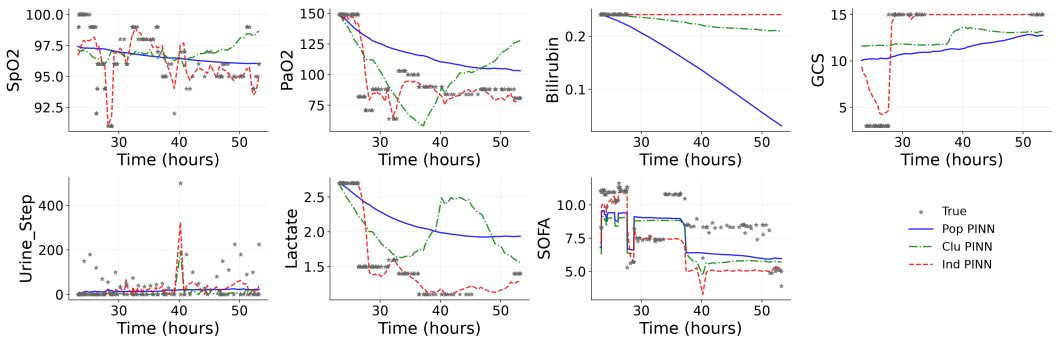


Figure 3: PINN rollout comparison for a representative held-out test patient ($T = 130$ h). The observed clinical trajectory is compared with full-horizon deterministic rollouts generated by the patient-specific (Ind), cluster-level (Clu), and population-level (Pop) PINNs across six physiological variables and the smoothed SOFA score. The test patient is excluded from the construction of the cluster- and population-level PINNs, while the patient-specific PINN serves as an oracle reference. NMSE values are 0.314, 0.784, and 0.846 for Ind, Clu, and Pop, respectively. The Ind PINN is fitted to this patient's full trajectory and is therefore evaluated in-sample. As an additional check on generalization, we retrain the Ind PINN for the same patient using only the first 70% of the trajectory and evaluate it on the remaining unseen 30%; the held-out rollout still captures the main trends (Appendix 7). A separate temporal split is unnecessary for the Clu and Pop models, since the test patient is already excluded entirely from their construction and their rollouts are therefore out-of-patient evaluations.

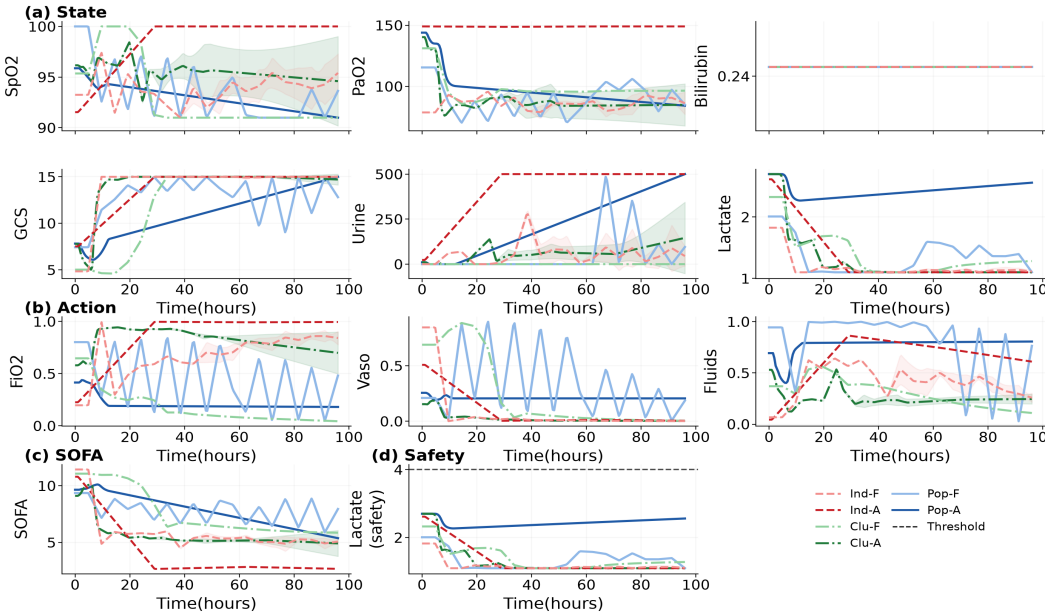
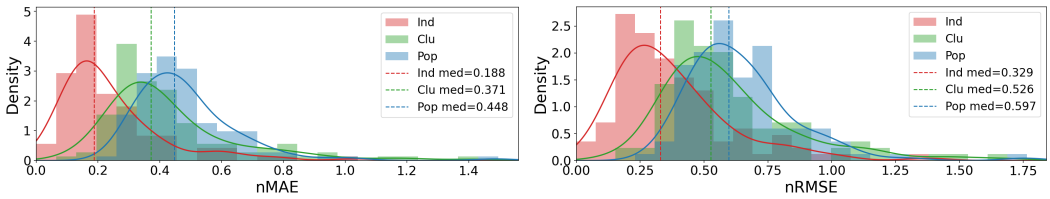


Figure 4: Closed-loop evaluation of Ind, Clu, and Pop policies on the patient-specific environment of the same representative test patient, using TRPOLag with $K = 20$. Fixed and adaptive interaction timing are compared in terms of (a) physiological state trajectories, (b) treatment actions, (c) lactate trajectories and the safety threshold of 4 mmol/L, and (d) SOFA trajectories. Under adaptive timing, the mean lactate values are 1.289, 1.338, and 1.357 mmol/L, and the mean SOFA scores are 3.082, 5.563, and 8.560 for Ind, Clu, and Pop, respectively. Results are averaged over five evaluation runs.

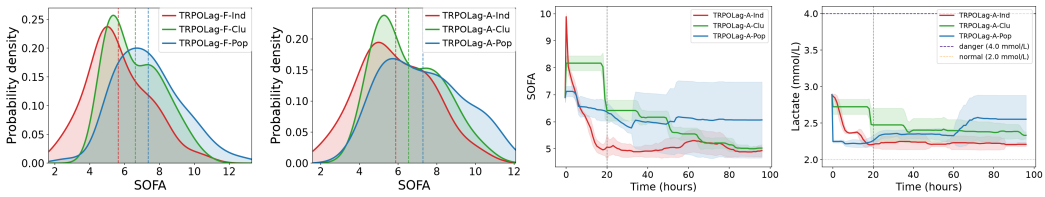
adaptive timing, respectively, showing that cluster-based training recovers part of the individual-oracle benefit. Adaptive timing reduces the trajectory-wide SOFA score for the population, cluster, and individual policy classes, although its effects on final SOFA and lactate safety are not uniform

(Table 2). Additional results for SAC, PPO, TRPO, and PPOLag are presented in Appendix D.1 and exhibit the same overall personalization trend.



(a) nMAE (b) nRMSE

Figure 5: Simulator fidelity evaluation. Distributions of (a) normalized MAE and (b) normalized RMSE for the individual, cluster, and population PINNs.



(a) Fixed (b) Adaptive (c) SOFA (d) Lactate

Figure 6: Cohort- and trajectory-level online evaluation of the individual-oracle (Ind), cluster-level (Clu), and population-level (Pop) TRPOLag policies. (a,b) Distributions of final SOFA scores at 96 hours across 70 held-out patients. Each patient-level score is averaged over ten noise realizations. (c,d) Representative SOFA and lactate trajectories (adaptive timing).

Policy	SOFA(all) (↓)	Lower SOFA vs. Pop (%)	SOFA(96h) (↓)	SafeTime% (↑)
TRPOLag-F-Ind	6.41 ± 1.72	91.4	$\mathbf{5.63 \pm 1.76}$	$\mathbf{97.46 \pm 12.65}$
TRPOLag-A-Ind	$\mathbf{6.20 \pm 1.87}$	81.4	5.87 ± 1.86	95.99 ± 13.76
TRPOLag-F-Clu	7.23 ± 1.59	71.4	6.61 ± 1.54	91.45 ± 23.62
TRPOLag-A-Clu	7.16 ± 1.68	65.7	6.55 ± 1.67	88.37 ± 25.61
TRPOLag-F-Pop	7.96 ± 1.90	–	7.34 ± 1.87	89.15 ± 29.45
TRPOLag-A-Pop	7.45 ± 1.99	–	7.28 ± 2.03	89.40 ± 24.14

Table 2: Online evaluation of TRPOLag on patient-specific simulators corresponding to 70 held-out patients. SOFA(all) and SOFA(96h) denote the trajectory-wide and final SOFA scores, respectively. Results are reported as mean and standard deviation across patients, each policy being trained with a single seed. Lower SOFA vs. Pop denotes the percentage of matched patients for which the individual or cluster policy achieves a lower final SOFA score than the population policy under the same timing setting. SafeTime% is the fraction of physical time with lactate below 4.0 mmol/L. F and A denote fixed and adaptive timing. Boldface indicates the best mean in each SOFA or safety column and does not imply statistical significance.

Besides, we examine whether the personalization benefits also exist in the offline setting, where policies must be learned from fixed behavior data. Table 3 evaluates two representative offline RL methods, DQN and Guarded Conservative Q-Learning (GCQL), and compares their individual-, cluster- and population-level policies under the same patient-specific evaluation environments. The online TRPOLag policies are evaluated on the same simulators as a reference; the individual one is also the behavior policy that generated the offline datasets. Both offline methods improve on their population-level counterparts, but the level of personalization at which the gain appears differs: GCQL benefits only from full individualization, whereas for DQN the cluster policy already recovers the gain. GCQL also remains closest to the behavior policy while staying nearer to the support of the behavior data. The advantage is not automatic, however: for CQL the cluster policy falls behind the population policy rather than lying between it and the individual one. These results show that access to patient-specific data alone is not sufficient to guarantee improved offline policies; the level

of personalization at which the benefit appears depends on the learning algorithm's ability to control distribution shift from the fixed dataset. This distinction illustrates an important role of *MedGym*: the benchmark can reveal when clinically relevant structure such as patient heterogeneity is successfully exploited by an RL method and when it is not. The CQL results, the alternative patient grouping, and the support-distance analysis are reported in Appendix D.4.

Policy	SOFA(all) (↓)	Lower SOFA vs. Pop (%)	SOFA(96h) (↓)	SafeTime% (↑)
TRPOLag-F-Ind (behavior)	5.89 ± 0.66	90.0	5.16 ± 0.85	100.00 ± 0.00
DQN-F-Ind	7.88 ± 1.02	90.0	7.81 ± 1.01	83.70 ± 33.63
DQN-F-Clu	7.84 ± 0.93	100.0	7.74 ± 1.01	89.06 ± 29.24
DQN-F-Pop	9.20 ± 0.97	–	9.34 ± 0.92	78.80 ± 38.91
GCQL-F-Ind	$\mathbf{6.27 \pm 0.72}$	**80.0**	$\mathbf{5.82 \pm 1.02}$	$\mathbf{99.94 \pm 0.18}$
GCQL-F-Clu	7.15 ± 0.75	40.0	7.00 ± 0.79	89.68 ± 20.67
GCQL-F-Pop	7.21 ± 0.88	–	7.06 ± 0.97	89.59 ± 21.85

Table 3: Offline-policy results for DQN, and GCQL on 10 held-out test patients. Metrics and abbreviations follow Table 2; dashes indicate that the population policy is not compared with itself.

Policy	MAP(all) [mmHg]	Higher Safe% vs. Pop (%) (↑)	MAP(48h) [mmHg]	SafeTime% (↑)	Lac(all) [mmol/L] (↓)
TRPOLag-F-Ind	69.47 ± 5.52	50.0	68.90 ± 4.97	77.8 ± 26.8	1.948 ± 1.346
TRPOLag-A-Ind	$\mathbf{72.89 \pm 3.52}$	**90.0**	$\mathbf{72.53 \pm 4.21}$	$\mathbf{96.3 \pm 3.4}$	$\mathbf{1.932 \pm 1.788}$
TRPOLag-F-Clu	81.01 ± 20.75	50.0	79.77 ± 24.86	76.6 ± 36.7	3.227 ± 2.073
TRPOLag-A-Clu	73.25 ± 18.03	60.0	72.27 ± 19.78	61.8 ± 37.1	3.954 ± 2.853
TRPOLag-F-Pop	77.99 ± 21.74	–	79.09 ± 29.14	52.7 ± 34.7	3.648 ± 2.068
TRPOLag-A-Pop	74.98 ± 25.06	–	76.84 ± 28.81	53.6 ± 38.8	3.565 ± 2.465
GCQL-F-Ind	$\mathbf{70.46 \pm 5.32}$	**70.0**	$\mathbf{70.97 \pm 5.52}$	$\mathbf{86.1 \pm 19.5}$	$\mathbf{2.005 \pm 1.338}$
GCQL-F-Clu	72.52 ± 9.23	40.0	82.17 ± 27.92	61.5 ± 31.3	3.298 ± 1.907
GCQL-F-Pop	77.56 ± 20.69	–	78.18 ± 26.76	65.1 ± 30.8	3.587 ± 2.025

Table 4: Evaluation of TRPOLag and GCQL policies on the acute hypotension cohort, reported as mean $\pm$ standard deviation across the 10 held-out test patients. MAP(all) and MAP(48h) are the trajectory-wide and the final mean arterial pressure, with clinical target zone $[65, 80]$ mmHg. SafeTime% is the fraction of physical time with MAP ≥ 65 mmHg, that is, it counts only the lower bound of the target zone, and Higher Safe% vs. Pop the percentage of patients for which Ind or Clu exceeds the population policy. Lac(all) is mean blood lactate, safe below 2.0 mmol/L.

For a more comprehensive analysis of the trajectory-level safety, additional schedule-independent trajectory-level safety analyses for the online policies are reported in Appendix D.3. To assess whether the benchmark construction and its principal evaluation dimensions are specific to MIMIC-III, we additionally instantiate *MedGym* using *the Synthetic Acute Hypotension Dataset*. The same environment-construction and evaluation protocol is applied with dataset-specific physiological variables and treatment definitions in Appendix B.4. The resulting evaluation again exhibits meaningful differences across fixed and adaptive intervention timing and across population-, cluster-, and individual-level policies, indicating that these benchmark dimensions are not artifacts of the MIMIC-III instantiation. Table 4 summarizes the representative results. Overall, *MedGym* enables controlled evaluation of medical RL algorithms in terms of patient personalization, trajectory-level safety, and continuous-time decision-making. The results consistently show that these dimensions affect policy performance and should be considered explicitly.

5 Conclusion

We introduced *MedGym*, a unified benchmark for dynamic medical treatment reinforcement learning. Built from clinical data, *MedGym* provides continuous-time patient environments in which elapsed treatment intervals and patient-specific dynamics explicitly affect state evolution. On top of this environment family, the benchmark supports controlled evaluation across fixed and adaptive intervention timing, different levels of personalization, online and offline policy learning, and treatment effectiveness and trajectory-level safety. Our experiments demonstrate that these evaluation dimensions materially affect the conclusions drawn about treatment policies. In particular, policy performance depends on intervention timing and the degree of personalization, while treatment effectiveness and safety need not improve simultaneously. These results highlight the importance of evaluating dynamic medical treatment RL beyond a single fixed-step, population-level setting.

ETHICS STATEMENT

This work is computational and uses only retrospectively collected, deidentified data from the credentialed-access MIMIC-III database; no new human- or animal-subject experiments were conducted. MIMIC-III was released with Institutional Review Board approval and a waiver of individual informed consent, and the data were accessed in accordance with the PhysioNet Data Use Agreement. All policies were trained and evaluated in learned patient simulators. The resulting policies are intended solely for benchmarking and must not be treated as clinical recommendations without rigorous prospective validation and clinical oversight.

REPRODUCIBILITY STATEMENT

The continuous-time transition model, the construction of patient-specific, cluster-level, and population-level PINNs, and the corresponding policy-learning and evaluation protocols are detailed in Appendix B and Appendix C. This appendix also specifies the fixed and adaptive timing settings, evaluated algorithms, held-out patient evaluation procedure, and performance and safety metrics. The remaining appendices document the MIMIC-III preprocessing procedure, physiological model, PINN architecture, patient-model clustering method, training hyperparameters, and additional experimental results. Source code for data preprocessing, simulator construction, policy training, evaluation, and figure generation is included in the supplementary material and will be made publicly available upon acceptance. Because MIMIC-III cannot be redistributed under the PhysioNet Data Use Agreement, credentialed users must obtain the data separately from PhysioNet and reconstruct the experimental inputs using the provided preprocessing scripts.

AI USE STATEMENT

In this work, we used generative AI tools as coding and writing assistants. AI assistance was used to help implement, review, and debug portions of the PINN training, reinforcement learning evaluation, and to draft and copy-edit portions of the manuscript. The final research problem, modeling choices, experimental design, and conclusions were determined by the authors. All AI-assisted code and text were reviewed by the authors, and every reported result was generated by the experimental code and checked against the corresponding outputs. No raw MIMIC-III records were provided to generative AI tools, which were not used to generate or modify experimental data or results. We take responsibility for the final content of this work.

REFERENCES

Bibhas Chakraborty and Susan A. Murphy. Dynamic treatment regimes. *Annual Review of Statistics and Its Application*, 1:447–464, 2014.

Kartik Choudhary, Dhawal Gupta, and Philip S. Thomas. Icu-sepsis: A benchmark mdp built from real medical data. *arXiv preprint arXiv:2406.05646*, 2024. doi: 10.48550/arXiv.2406.05646. URL https://arxiv.org/abs/2406.05646.

Ran Xu et al. Medagentgym: A scalable agentic training environment for code-centric reasoning in biomedical data science. *Proceedings of The Fourteenth International Conference on Learning Representations*, 2026.

Omer Gottesman, Fredrik Johansson, Joshua Meier, David Dent, Donghun Lee, Srivatsan Srinivasan, Linying Zhang, Yi Ding, David Wihl, Xiaoxiao Peng, Jiayu Yao, Isaac Lage, Constantin Mosch, Li-wei H. Lehman, Matthieu Komorowski, Aldo Faisal, Leo Anthony Celi, David Sontag, and Finale Doshi-Velez. Evaluating reinforcement learning algorithms in observational health settings. *arXiv preprint arXiv:1805.12298*, 2018.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature Medicine*, 25:16–18, 2019.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1861–1870. PMLR, 10–15 Jul 2018.

Mason Hargrave, Alex Spaeth, and Logan Grosenick. Epicare: A reinforcement learning benchmark for dynamic treatment regimes. *Proceedings of the 36th Advances in Neural Information Processing Systems*, 2024.

A.E.W. Johnson, T.J. Pollard, L. Shen, L.H. Hehman, M. Feng, M.Ghassemi, B. Moody, P. Szolovits, L.A. Celi, and R.G. Mark. Mimic-iii, a freely accessible critical care database. 3 (may. 2016), 2016. URL https://doi.org/10.1038/sdata.2016.35.

Kelley M. Kidwell and Daniel Almirall. Sequential, multiple assignment, randomized trial designs. *JAMA*, 329:336–337, 2023.

M. Komorowski, L.A. Celi, O. Badawi, A.C. Gordon, and A.A. Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24: 1716–1720, 2018.

Richard L. Kravitz, Naihua Duan, and Joel Braslow. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *The Milbank Quarterly*, 82:661–687, 2004.

A. Kumar, A. Zhou, G. Tucker, and S. Levine. Conservative q-learning for offline reinforcement learning. *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 33:1179–1191, 2020.

Nicholas Kuo, Simon Finfer, Louisa Jorm, and Sebastiano Barbieri. Synthetic acute hypotension and sepsis datasets based on MIMIC-III and published as part of the Health Gym project (version 1.0.0). *PhysioNet*, 2022.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Zhiyao Luo, Mingcheng Zhu, Fenglin Liu, Jiali Li, Yangchen Pan, Jiandong Zhou, and Tingting Zhu. Dtr-bench: An in silico environment and benchmark platform for reinforcement learning based dynamic treatment regime. *arXiv preprint arXiv:2405.18610*, 2024. doi: 10.48550/arXiv.2405.18610. URL https://arxiv.org/abs/2405.18610.

V. Mnih, K. Kavukcuoglu, D. Silver, and et al. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Susan A. Murphy. An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine*, 24:1455–1481, 2005.

Pranay Pasula. Lagrangian duality in reinforcement learning. *arXiv preprint arXiv:2007.09998*, 2020. doi: 10.48550/arXiv.2007.09998. URL https://arxiv.org/abs/2007.09998.

Alexander Peine, Ahmed Hallawa, Johannes Bickenbach, Gerrit Dartmann, Lejla B. Fazlic, Alexander Schmeink, Gerd Ascheid, Christian Thiemermann, Andreas Schuppert, Richard Kindle, Leo Celi, Gernot Marx, and Lukas Martin. Development and validation of a reinforcement learning algorithm to dynamically optimize mechanical ventilation in critical care. *npj Digital Medicine*, 4:1–12, 2021.

Rafael F. Prudencio, Marcos R. O. A. Maximo, and Esther L. Colombini. A survey on offline reinforcement learning: Taxonomy, review, and open problems. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–20, 2023.

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 *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pp. 147–163. PMLR, 2017.

M. Raissi, P. Perdikaris, and G.E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. doi: https://doi.org/10.1016/j.jcp.2018.10.045.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1889–1897. PMLR, 07–09 Jul 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. doi: 10.48550/arXiv.1707.06347. URL https://arxiv.org/abs/1707.06347.

Susan M. Shortreed, Eric Laber, Daniel J. Lizotte, T. Scott Stroup, Joelle Pineau, and Susan A. Murphy. Informing sequential clinical decision-making through reinforcement learning: An empirical study. *Machine Learning*, 84:109–136, 2011.

Philip S. Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. *arXiv preprint arXiv:1604.00923*, 2016.

Lenart Treven, Bhavya Sukhija, Yarden As, Florian Dörfler, and Andreas Krause. When to sense and control? a time-adaptive approach for continuous-time rl. *Proceedings of the 36th Advances in Neural Information Processing Systems*, 2024.

Aysın Tumay, Sophia Sun, Sonia Fereidooni, Aaron Dumas, Elise Jortberg, and Rose Yu. Guardian-regularized safe offline reinforcement learning for smart weaning of mechanical circulatory devices. *arXiv preprint arXiv:2511.06111*, 2025. doi: 10.48550/arXiv.2511.06111.

Masatoshi Uehara, Chengchun Shi, and Nathan Kallus. A review of off-policy evaluation in reinforcement learning, 2022. Review paper.

Christoph Voloshin, Hoang M. Le, Nan Jiang, and Yisong Yue. Empirical study of off-policy policy evaluation for reinforcement learning. *arXiv preprint arXiv:1911.06854*, 2021.

Runze Yan*, Xun Shen*, Akifumi Wachi, Sebastien Gros, Anni Zhao, and Xiao Hu. Offline guarded safe reinforcement learning for medical treatment optimization strategies. *Proceedings of the 39th Advances in Neural Information Processing Systems*, 2025.

Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys*, 55:5:1–5:36, 2021.

A LIMITATIONS

MedGym has several limitations that should be taken into account when interpreting the benchmark results.

First, the individual- and cluster-level personalization settings in the current benchmark should be interpreted as retrospective reference settings rather than deployable personalization procedures for previously unseen patients. The individual setting has access to a patient-specific dynamics model learned from that patient's trajectory and therefore serves as an oracle upper bound on the potential benefit of personalization. Likewise, the current cluster assignment is determined from similarities between learned patient-specific dynamics, which requires substantially more patient-specific information than would generally be available at the beginning of treatment. The present experiments therefore evaluate how policy performance changes when increasingly patient-specific dynamics are available, but they do not solve the separate problem of inferring an individual representation or assigning a new patient to an appropriate cluster from a short partial history. Learning such transferable representations and online assignment mechanisms is an important direction for future work.

Second, while *MedGym* supports trajectory-level safety evaluation in continuous time, the current benchmark does not include a dedicated continuous-time safe RL method. As a result, the present study mainly evaluates safety through policy outcomes and safety-aware reward or constraint formulations rather than through algorithms specifically designed for continuous-time safe exploration or safe policy improvement. Future work can extend *MedGym* by incorporating such methods once they become available, which would further strengthen the benchmark for safety-critical treatment recommendations.

Third, the current experiments do not include POMDP-specific algorithms, even though partial observability is highly relevant in medical treatment. In practice, the true patient condition is only indirectly revealed through clinical measurements, and therefore many treatment problems are naturally partially observable. Because *MedGym* is built on a continuous-time latent state evolution model, extending the benchmark to a POMDP interface is conceptually straightforward: one can define an observation mapping from the latent physiological state to a possibly noisy and incomplete observation vector, and then let policies act on the observation history or belief state rather than on the full latent state. This extension is left for future work.

Fourth, *MedGym* is constructed from observational clinical trajectories and therefore cannot establish the correctness of arbitrary counterfactual treatment responses. Each patient contributes observations under the treatment sequence that was actually administered, while the outcomes that would have occurred under alternative interventions are unobserved. Consequently, agreement with observed trajectories evaluates factual predictive fidelity but does not by itself validate simulator behavior under all state–action combinations that may be visited by a learned RL policy. This limitation cannot be eliminated by additional validation on another observational dataset, since such data likewise do not provide patient-level counterfactual ground truth. Establishing stronger counterfactual validity would require additional interventional evidence, richer treatment coverage together with appropriate causal assumptions, or independent physiological validation. Accordingly, *MedGym* should be interpreted as a clinically grounded research benchmark for controlled algorithm comparison rather than as a validated digital twin for predicting real-world treatment effects.

Finally, the current paper primarily focuses on the sepsis cohort as the main clinically motivated instantiation, while also evaluating a second clinical setting based on the Synthetic Acute Hypotension Dataset, together with a limited set of representative RL methods. While this is sufficient to demonstrate the key benchmark dimensions of continuous-time evolution, irregular intervention timing, personalization, and safety, broader validation across additional diseases, datasets, and algorithm classes would further improve the scope and robustness of the benchmark.

B Details of Environment Construction

B.1 MIMIC-III Dataset and Cohort Construction

MIMIC-III (Medical Information Mart for Intensive Care) is a large-scale, publicly available clinical database containing detailed records from over $40{,}000$ intensive care admissions (Johnson et al., 2016). The database includes longitudinal information such as demographics, vital signs, laboratory tests, medications, procedures, and other time-stamped clinical events, thereby providing a rich foundation for modeling dynamic treatment processes in critical care. For *MedGym*, MIMIC-III is particularly valuable because it reflects the key characteristics emphasized in this work: patient conditions evolve continuously over time, while measurements and interventions are collected at irregular intervals determined by clinical practice rather than by a fixed sampling schedule. This temporal structure makes the dataset well suited for learning continuous-time patient evolution models and for constructing benchmark environments in which the elapsed time between observations and treatments meaningfully affects state transitions. In addition, MIMIC-III contains substantial heterogeneity across patients in terms of disease severity, physiological trajectories, intervention histories, and outcomes. Such variability is essential for building a benchmark that supports the evaluation of personalized treatment policies rather than only population-level decision rules. By grounding *MedGym* in this real-world clinical diversity, we aim to construct an environment in which learned policies must account for both irregular treatment timing and individualized treatment response. Moreover, the widespread use of MIMIC-III in healthcare machine learning research, together with its extensive documentation and established preprocessing practices, facilitates reproducibility and comparability with prior studies. These properties make MIMIC-III an appropriate data foundation for *MedGym* and for the evaluation of reinforcement learning methods in continuous-time dynamic treatment recommendation.

From this database we construct the sepsis cohort used throughout the paper, yielding 110 patients in each of seven disjoint cohorts and hence the 770 patient-specific models used in Section 4. Because measurements arrive at clinically determined times rather than on a fixed schedule, the six state variables of Table 5 are resampled onto a common grid and the remaining gaps are filled by forward filling. This matters most for the laboratory variables, which are measured far less frequently than the vital signs; bilirubin in particular is near-constant within many stays, as the trajectories in Figure 4 show, so the hepatic component of the reward varies little for those patients. Treatment actions are continuous and are clipped to fixed admissible ranges, both when offline datasets are generated and when policies are evaluated, and all state variables are standardised to zero mean and unit variance using statistics computed from each patient's own record, while the action ranges are fixed across the cohort; both are reused unchanged for every policy and evaluation rollout. The numerical values of the extraction window, the resampling interval, the action bounds and the standardisation statistics are given in the supplementary material.

B.2 Sepsis Treatment: State, Reward, Safety, and Evaluation Quantities

Overview. For the sepsis instantiation of *MedGym*, we construct the treatment environment from the MIMIC-III sepsis cohort by extracting patient trajectories over a clinically relevant window around sepsis onset. Our goal is not to reproduce the full ICU record space, but to define a compact continuous-time treatment model that remains clinically meaningful, computationally tractable, and suitable for benchmarking both offline and online RL methods. To this end, we select a subset of physiological variables that are strongly associated with organ dysfunction and are routinely monitored in sepsis management. Treatment actions are modeled as continuous-valued clinical interventions, and the reward is designed from a smooth surrogate of organ failure severity. A key feature of *MedGym* is personalization: each patient trajectory is associated with an individual-specific variable ξ, which captures heterogeneity in disease progression and treatment response. Therefore, the environment does not represent a single generic sepsis process, but rather a family of related patient-specific treatment dynamics. Table 5 summarizes the state variables, interventions, and reward components used in this sepsis setting.

Reward Function. We construct the reward from the Sequential Organ Failure Assessment (SOFA) score, since SOFA provides a clinically interpretable summary of multi-organ dysfunction and is more suitable for sequential treatment evaluation than a purely terminal outcome.

Table 5: Summary of State Variables, Treatment Actions, and Reward Design in the Sepsis Setting of *MedGym*

Category	Variables	Description
Dynamic Features	SpO_2, PaO_2, Total bilirubin, GCS, Urine output, Lactate	Physiological variables describing respiratory, hepatic, neurological, renal, and metabolic status, selected to reflect major dimensions of organ dysfunction in sepsis.
Actions	FiO_2, Vasopressor rate, Intravenous fluids	Continuous treatment controls corresponding to oxygen support, vasopressor administration, and fluid resuscitation in ICU care.
Reward	SOFA-based surrogate with vasopressor penalty	A smooth approximation of organ failure severity used as dense feedback, together with a regularization term discouraging unnecessarily aggressive vasopressor usage.

The original SOFA scoring rule is threshold-based and assigns discrete integer levels to each physiological component:

$$\mathrm{SOFA}(\mathbf{x}, \mathbf{u}) = \sum_i \sum_k w_{i,k}\, \mathbf{1}(d_i\,(v_i - \tau_{i,k}) \geq 0)\,, \tag{2}$$

where $\mathbf{1}(\cdot)$ is the indicator function, i indexes the components, k indexes the thresholds $\tau_{i,k}$ of component i, and $w_{i,k}$ is the weight of the corresponding level. The direction coefficient is $d_i = -1$ for components scored when they fall below a threshold ($\mathrm{SpO_2}$, GCS, Urine) and $d_i = +1$ for those scored when they exceed one (Bili, Vaso), and all v_i are evaluated in raw clinical units. In *MedGym* each indicator is replaced by a smooth sigmoid step so that the reward remains informative between thresholds. Following the clinical SOFA definition, the cardiovascular component is scored through vasopressor dose, so the score depends on the treatment action as well as on the patient state. In our implementation, these thresholds are specified separately for each component according to the available variables in the dataset. For SpO_2, the cutoffs are 94, 90, 85, and 80 in percent. For bilirubin, the cutoffs are 1.2, 2.0, 6.0, and 12.0 mg/dL. For GCS, the cutoffs are 15, 13, 10, and 6. For urine output, the cutoffs are 500 and 200 mL/day. For vasopressor dosage, the cutoffs are 0.0, 0.05, 0.1, and 0.25 μg/kg/min. These choices follow the standard SOFA structure where possible, with modest adjustments to match the variables consistently available in our data.

In *MedGym*, this reward construction is based on the subset of organ systems that can be modeled reliably from the available data, namely respiratory, hepatic, neurological, renal, and cardiovascular status, corresponding to SpO_2, bilirubin, GCS, urine output, and vasopressor usage. The coagulation component is not included because of substantial missingness and inconsistent temporal coverage. Thus, the adopted score preserves the additive threshold-based logic of SOFA while adapting it to the continuous-time benchmark setting and the practical constraints of the dataset. We define the continuous-time running reward as

$$b(\mathbf{x}_t, \mathbf{u}_t) = -\,\mathrm{SOFA}(\mathbf{x}_t, \mathbf{u}_t) - \lambda_{\mathrm{v}}\,\mathrm{Vaso}(\mathbf{u}_t), \tag{3}$$

which measures treatment utility per unit of physical time. For a treatment action applied over an interval of duration δt, the reward used by the RL algorithm is obtained by integrating this running reward over the corresponding interval, following the TACOS reformulation described in Appendix C.1.

State Space. The state variables are selected to balance clinical relevance and modeling tractability. Rather than retaining a very high-dimensional ICU representation, we focus on variables that directly reflect major aspects of sepsis-related organ dysfunction and are closely aligned with the reward definition. The resulting state includes respiratory, hepatic, neurological, renal, and metabolic indicators that are routinely monitored in critical care and sufficiently informative for treatment

optimization. This choice yields a compact state representation that remains clinically meaningful while supporting stable learning and continuous-time trajectory modeling. In addition, *MedGym* incorporates personalization through ξ, so the same observed state may evolve differently across patients under the same intervention, thereby enabling evaluation of individualized treatment policies rather than only population-level strategies.

Explicit Clinical Safety Constraints. In addition to reward-based optimization, *MedGym* includes an explicit safety cost built from a small set of clinical violation criteria, with blood lactate as the primary one. Lactate is widely used as an indicator of impaired perfusion and physiological instability in sepsis, and sustained elevation is strongly associated with clinical deterioration. For this reason, it provides an interpretable and operationally meaningful quantity for monitoring safety in the benchmark. The role of the safety constraint is to prevent policies from driving the patient into hazardous regions that may not be sufficiently penalized by the reward alone. This is particularly important in *MedGym*, because treatments are applied over irregular intervals and unsafe evolution may arise between two consecutive decision times. By incorporating lactate-based safety monitoring, the benchmark supports evaluation of not only treatment effectiveness but also whether a policy maintains clinically acceptable continuous-time trajectories. Concretely, the safety cost has the same threshold form as equation 2, with a single threshold per criterion and unit weights,

$$c(\mathbf{x}, \mathbf{u}) = \sum_{j \in \mathcal{J}} \mathbf{1}(d_j \, (v_j - \theta_j) \geq 0) \,, \tag{4}$$

where $\mathcal{J}$ comprises vasopressor dose, lactate and oxygen saturation, with $d_j = +1$, $\theta_j = \bar{v}$ for $\mathrm{Vaso}(\mathbf{u})$; $d_j = +1$, $\theta_j = 4.0$ mmol/L for lactate; and $d_j = -1$, $\theta_j = 88\%$ for SpO_2. Thus $c \in \{0, 1, 2, 3\}$ counts the clinical criteria simultaneously violated at a decision point, weighted equally by default.cost limit

Evaluation Quantities. Treatment effectiveness is reported by two quantities computed from the same rollout. The trajectory-wide score is the physical-time average

$$\mathrm{SOFA}(\mathrm{all}) = \frac{1}{T} \int_0^T \mathrm{SOFA}(\mathbf{x}_t, \mathbf{u}_t) \, \mathrm{d}t, \tag{5}$$

and $\mathrm{SOFA}(96\mathrm{h}) = \mathrm{SOFA}(\mathbf{x}_T, \mathbf{u}_T)$ is the score at the end of the horizon. Both are evaluated on the same physical-time grid as the safety metrics of Appendix D.3, so neither depends on the number or the spacing of intervention times, and fixed- and adaptive-timing policies are compared on identical quantities.

B.3 Training Deterministic Part by PINN

To evaluate reinforcement learning methods in a clinically grounded continuous-time benchmark, *MedGym* constructs a family of patient-specific simulation environments from learned physiological dynamics. Each virtual patient is characterized by a personalization variable ξ, introduced in the main text, which modulates disease progression and treatment response. Consequently, *MedGym* defines not a single simulator but a collection of related environments indexed by ξ, enabling systematic evaluation of personalization in dynamic treatment recommendation. The environment is based on a continuous-time state evolution model of the form

$$\mathbf{x}_{t+\delta t} = \mathbf{F}_\xi(\mathbf{x}_t, \mathbf{u}_t, \delta t) + \mathbf{w}_{t,\delta t,\xi}, \tag{6}$$

where $\mathbf{x}_t$ is the patient state, $\mathbf{u}_t$ is the treatment action, δt is the elapsed interaction interval, and $\mathbf{w}_{t,\delta t,\xi}$ denotes residual uncertainty. The dependence on δt allows the transition model to capture irregular treatment and measurement intervals, while the dependence on ξ allows the dynamics to vary across patients. To obtain a data-driven simulator, we approximate the deterministic part of the dynamics by a Physics-Informed Neural Network (PINN) (Raissi et al., 2019), yielding

$$\widetilde{\mathbf{x}}_{t+\delta t} = \widetilde{\mathbf{z}}_{\mathsf{nn}}(\mathbf{x}_t, \mathbf{u}_t, \delta t, \xi) = \mathbf{F}_\xi^{\mathsf{nn}}(\mathbf{x}_t, \mathbf{u}_t, \delta t) + \mathbf{W}_{\mathsf{c}}, \qquad \mathbf{W}_{\mathsf{c}} \sim \widetilde{p}(\cdot \mid \mathbf{x}_t, \xi, \delta t). \tag{7}$$

Here, $\mathbf{F}_\xi^{\mathsf{nn}}(\mathbf{x}_t, \mathbf{u}_t, \delta t)$ is the personalized PINN approximation of the deterministic dynamics, while $\widetilde{p}(\cdot \mid \mathbf{x}_t, \xi, \delta t)$ models the residual stochasticity conditioned on the current state, the personalization variable, and the elapsed interval. Therefore, the transition mechanism in *MedGym* is not tied to a fixed discrete step, but adapts to irregular interaction intervals and individualized patient dynamics in

a principled manner. In *MedGym*, each episode is generated by sampling an initial patient condition together with a corresponding personalization variable ξ, and then rolling out the learned continuous-time dynamics under the chosen treatment policy. By varying both ξ and the stochastic realization of $\mathbf{W}_c$, the benchmark can generate a diverse collection of patient trajectories that reflect heterogeneous disease progression, treatment sensitivity, and uncertainty accumulation over irregular decision intervals. This construction is particularly important for the intended benchmark usage of *MedGym*. First, it preserves the continuous-time nature of patient evolution and therefore avoids the coarse approximation induced by fixed time discretization. Second, it explicitly incorporates personalization, allowing the benchmark to evaluate whether a method can adapt to patient-specific dynamics rather than relying only on population-level regularities. Third, because the same environment family can be used both to generate offline datasets and to evaluate learned policies through online rollouts, *MedGym* provides a unified platform for comparing offline and online reinforcement learning methods under realistic dynamic treatment settings.

Physical Model for PINN. We describe the patient dynamics by a continuous-time ODE that captures coarse interactions among respiratory, metabolic, neurological, renal, and hepatic functions. The purpose of this model is not to provide a fully mechanistic physiological simulator, but to encode a clinically interpretable structural prior that guides learning toward plausible patient evolution. Accordingly, *MedGym* combines a simplified physiology-inspired component with a flexible residual term learned from data. Personalization is introduced through ξ, so that both the disease progression law and treatment sensitivity are allowed to vary across patients. The physiological state is $\mathbf{x} = [\mathrm{SpO_2}, \mathrm{PaO_2}, \mathrm{Bili}, \mathrm{GCS}, \mathrm{Urine}, \mathrm{Lac}]^\top$ and the treatment action is $\mathbf{u} = [\mathrm{FiO_2}, \mathrm{Vaso}, \mathrm{Fluid}]^\top$, where $\mathrm{SpO_2}$ is peripheral oxygen saturation, $\mathrm{PaO_2}$ arterial oxygen tension, Bili serum bilirubin, GCS the Glasgow Coma Scale, Urine urine output, Lac blood lactate, $\mathrm{FiO_2}$ the fraction of inspired oxygen, Vaso vasopressor dose, and Fluid intravenous fluid rate. All variables are standardised to zero mean and unit variance; $[\cdot]^+ = \max(\cdot, 0)$ denotes rectification. The ODE is therefore defined in a normalized state space, where variables should be interpreted as relative deviations from nominal levels rather than absolute physiological quantities. For this reason, the terms below should be viewed as encoding directional and structural dependencies among variables, rather than exact physical laws. The oxygenation block describes a first-order tendency toward an $\mathrm{FiO_2}$-dependent equilibrium, together with modulation by metabolic stress and fluid administration:

$$\frac{d\,\mathrm{PaO_2}}{dt} = k_1(\xi)(\mathrm{FiO_2} - \mathrm{PaO_2}) - k_2(\xi)\,[\mathrm{Lac}]^+ + k_3(\xi)\,\mathrm{Fluid}, \tag{8}$$

$$\frac{d\,\mathrm{SpO_2}}{dt} = k_4(\xi)(\mathrm{PaO_2} - \mathrm{SpO_2}). \tag{9}$$

The first equation models how inspired oxygen and supportive treatment affect arterial oxygenation, while the second links peripheral saturation to arterial oxygen tension. Neurological status is modeled as deteriorating under metabolic burden and hypoxic stress:

$$\frac{d\,\mathrm{GCS}}{dt} = -k_5(\xi)\,[\mathrm{Lac}]^+ - k_6(\xi)\,h(\mathrm{SpO_2}), \tag{10}$$

where $h(\mathrm{SpO_2})$ is a nonnegative function that increases when oxygen saturation falls below normal levels. In normalized coordinates, this function captures relative hypoxic deviation rather than a hard clinical threshold. Lactate is driven upward by vasopressor-related hypoperfusion and oxygen deficit, and downward through clearance and dilution effects:

$$\frac{d\,\mathrm{Lac}}{dt} = k_7(\xi)\,\mathrm{Vaso} + k_8(\xi)\,h(\mathrm{SpO_2}) - k_9(\xi)\,[\mathrm{Lac}]^+ - k_{10}(\xi)\,\mathrm{Fluid}. \tag{11}$$

Urine output reflects renal response to fluid input and is suppressed by metabolic and vasomotor stress:

$$\frac{d\,\mathrm{Urine}}{dt} = k_{11}(\xi)\,\mathrm{Fluid} - k_{12}(\xi)\,[\mathrm{Lac}]^+ - k_{13}(\xi)\,\mathrm{Vaso}. \tag{12}$$

Bilirubin is used to represent slowly varying hepatic dysfunction:

$$\frac{d\,\mathrm{Bili}}{dt} = k_{14}(\xi)\,[\mathrm{Lac}]^+ - k_{15}(\xi)\,[\mathrm{Bili}]^+. \tag{13}$$

The coefficients $\{k_i(\xi)\}$ determine the strength of organ-level couplings and treatment effects for each patient. They should be understood as patient-dependent effective parameters rather than directly

measurable physiological constants. In the current implementation of *MedGym*, personalization is realized by indexing patient-specific models, so that ξ serves as an abstract label of individualized dynamics rather than an explicitly inferred latent representation. Thus, the present benchmark evaluates whether patient-specific dynamics improve prediction and control, while leaving the separate problem of learning an explicit transferable representation of ξ to future work. Since the true physiological dynamics are not fully known, the mechanistic component above is treated as an inductive bias rather than a complete model. There may exist substantial mismatch between the real dynamics and the simplified relations in equation 8–equation 13. To capture this discrepancy, we use the following hybrid model:

$$\frac{d\mathbf{x}}{dt} = \mathsf{MedODE}_{\xi}(\mathbf{x}, \mathbf{u}) + \mathsf{NN}_{\xi}(\mathbf{x}, \mathbf{u}). \tag{14}$$

Here ξ collects all patient-specific learnable quantities, namely the coefficients of MedODE_{ξ} and the weights of NN_{ξ}; a separate model is fitted to each patient and no parameter is shared across patients. Rather than integrating equation 14 numerically at every interaction, we jointly train a network that predicts the state derivative directly from $(\mathbf{x}, \mathbf{u})$, supervised both by finite differences of the recorded trajectory and by the right-hand side of equation 14. The personalized transition model $\mathbf{F}_{\xi}^{\mathsf{nn}}(\mathbf{x}, \mathbf{u}, \delta t)$ used in equation 7 advances this field over the elapsed time δt with a fixed number of equal sub-steps, so that an interaction of arbitrary duration is produced without an external ODE solver.

Training objective and optimization The loss is a weighted sum of four terms: a data term matching the predicted derivative to finite differences of consecutive observations, a multi-step term that unrolls the predicted field, a smoothness term on consecutive predicted derivatives, and a physics term matching the predicted field to the right-hand side of equation 14. Their weights follow a fixed three-stage schedule, from $20/40/10/0.1$ to $5/20/20/0.02$, which shifts emphasis from trajectory fitting to physical consistency; the physics term is evaluated at the recorded observation times, and intervals shorter than 0.05 hours are excluded because their finite differences are dominated by measurement noise. The derivative network uses three hidden layers of width 128 and the neural residual two hidden layers of width 128, about 5×10^4 parameters per patient in total, trained jointly with Adam at learning rates 10^{-4} and 5×10^{-5} for at most 10^4 epochs on the patient's full trajectory. The simulator also provides an observation noise model, which is disabled in all reported experiments.

Overall, the PINN construction in *MedGym* should be interpreted as a clinically motivated continuous-time structural prior together with a data-driven correction term. This design preserves continuous-time disease evolution, accommodates irregular interaction intervals, and supports personalized dynamics through ξ, while remaining sufficiently simple and modular for standardized benchmark evaluation.

Scope of the current physical model. The current implementation of *MedGym* uses MIMIC-III sepsis trajectories as a concrete use case, and the physiology-inspired ODE in Appendix B is therefore designed to reflect major organ interactions relevant to critical-care sepsis management. However, this does not mean that *MedGym* is restricted to sepsis as a benchmark concept. The main contribution of *MedGym* is the environment construction pipeline itself, namely the combination of a continuous-time structured dynamics model with data-driven residual learning for building personalized simulators from clinical data. This pipeline is modular and can be re-instantiated with different state variables, treatment actions, reward definitions, and physiology-inspired priors for other medical domains. For example, the same framework can in principle be extended to continuous-time treatment problems such as glucose management in diabetes or medication scheduling in oncology. Thus, the current sepsis setting should be understood as an initial benchmark instantiation rather than as a restriction on the broader applicability of the *MedGym* framework.

Interpretation of the individualized model. The individualized model (*Ind*) in *MedGym* should not be interpreted as a practically deployable personalized learning method for unseen patients. Rather, it serves as an *upper-bound evaluation setting* that measures the performance achievable when patient-specific dynamics are fully available. In this sense, the individualized model plays the role of an oracle benchmark for personalization, allowing us to quantify how much performance can potentially be gained when the underlying patient dynamics are accurately identified. The separate problem of inferring patient-specific representations such as ξ from a patient's partial history, and then using them to generalize to unseen patients, remains an important open research challenge. We

view *MedGym* as a benchmark that can support future evaluation of such representation-learning and patient-adaptation methods once they become available.

Patient clustering method. Let N_{psm} and N_{clu} denote the number of patient-specific models and the desired number of clusters, respectively. For patient model M_i, we represent its transition dynamics by the pretrained patient-specific PINN $\mathbf{F}^{\mathsf{nn}}_{\xi_i}(\mathbf{x}, \mathbf{u}, \delta t)$. Rather than clustering raw clinical variables, the proposed method compares the predictive behavior of patient-specific PINNs when they are evaluated under the same patient trajectories.

For each transition of patient j, we first use its own PINN to generate a one-step teacher target:

$$\widetilde{\mathbf{x}}_{j,t+1} = \mathbf{F}^{\mathsf{nn}}_{\xi_j}\left(\mathbf{x}_{j,t}, \mathbf{u}_{j,t}, \delta t_{j,t}\right). \tag{15}$$

Each prediction is re-anchored at the observed state $\mathbf{x}_{j,t}$ rather than recursively initialized from the preceding model prediction, thereby avoiding the accumulation of long-horizon rollout errors.

The PINN of another patient i is then evaluated using the same state, treatment action, and elapsed time:

$$\widehat{\mathbf{x}}_{i\to j,t+1} = \mathbf{F}^{\mathsf{nn}}_{\xi_i}\left(\mathbf{x}_{j,t}, \mathbf{u}_{j,t}, \delta t_{j,t}\right). \tag{16}$$

The transfer error from patient model M_i to patient model M_j is defined as

$$E_{ij} = \frac{1}{d_{\mathsf{x}} T_j} \sum_{m=1}^{d_{\mathsf{x}}} \sum_{t=1}^{T_j} \left(\frac{\widehat{x}^{(m)}_{i\to j,t+1} - \widetilde{x}^{(m)}_{j,t+1}}{s_m} \right)^2, \tag{17}$$

where d_{x} is the dimension of the physiological state, T_j is the number of valid transitions for patient j, and s_m is the population-level interquartile range of the m-th state variable. A smaller E_{ij} indicates that the two patient-specific PINNs produce more similar dynamics under the same clinical inputs.

Because the transfer-error scale may differ across target patients, each column of the error matrix is robustly normalized:

$$Z_{ij} = \frac{E_{ij} - \underset{r\neq j}{\mathrm{median}}(E_{rj})}{\max\left\{ \underset{r\neq j}{\mathrm{IQR}}(E_{rj}), \epsilon \right\}}, \tag{18}$$

where $\epsilon > 0$ prevents numerical instability when the interquartile range is close to zero. The i-th row of $\mathbf{Z}$ forms the transfer profile of patient model M_i.

The symmetric distance between patient models M_i and M_ℓ is defined as the root-mean-square difference between their transfer profiles:

$$\Delta_{i\ell} = \sqrt{\frac{1}{N_{\mathsf{psm}} - 2} \sum_{j\notin\{i,\ell\}} \left(Z_{ij} - Z_{\ell j}\right)^2}. \tag{19}$$

The columns corresponding to patients i and ℓ are excluded so that their diagonal self-errors do not affect the comparison. A small $\Delta_{i\ell}$ indicates that the two patient-specific PINNs respond similarly across other patient trajectories and therefore represent similar disease-progression and treatment-response dynamics.

The resulting distance matrix is used for capacity-constrained medoid clustering. For cluster $\mathcal{C}_k$, its medoid is defined as

$$m_k = \underset{u\in\mathcal{C}_k}{\arg\min} \sum_{v\in\mathcal{C}_k} \Delta_{uv}. \tag{20}$$

The clustering objective is

$$\min_{\{\mathcal{C}_k\}_{k=1}^{N_{\mathsf{clu}}}} \sum_{k=1}^{N_{\mathsf{clu}}} \sum_{i\in\mathcal{C}_k} \Delta_{i,m_k}, \tag{21}$$

subject to each patient model belonging to exactly one cluster and each cluster satisfying its prescribed capacity.

After clustering, each $\mathcal{C}_k$ is partitioned into the disjoint sets $\mathcal{C}_k^{\mathsf{train}}$ and $\mathcal{C}_k^{\mathsf{test}}$. Thus, the cluster membership of each test patient is determined by the PINN transfer-based clustering procedure before

policy evaluation. A test patient assigned to $\mathcal{C}_k^{\text{test}}$ is evaluated under the cluster-level policy of cluster k, but always on its own patient-specific simulator, and it is excluded from the construction and training of the cluster- and population-level PINNs and policies. The assignment therefore depends on similarities in learned patient dynamics rather than the SOFA score or reward subsequently obtained during policy evaluation.

Held-out evaluation of the patient-specific PINN. The patient-specific PINN in Figure 3 is fitted to the full trajectory of the test patient and therefore serves as an oracle reference, quantifying the accuracy attainable when the individual dynamics are fully available rather than the accuracy of predicting unseen dynamics. To separate these two cases, we train the individual PINN under a temporal split: the first 70% of the trajectory is used for training the PINN while the remaining 30% is used to evaluate the errors.

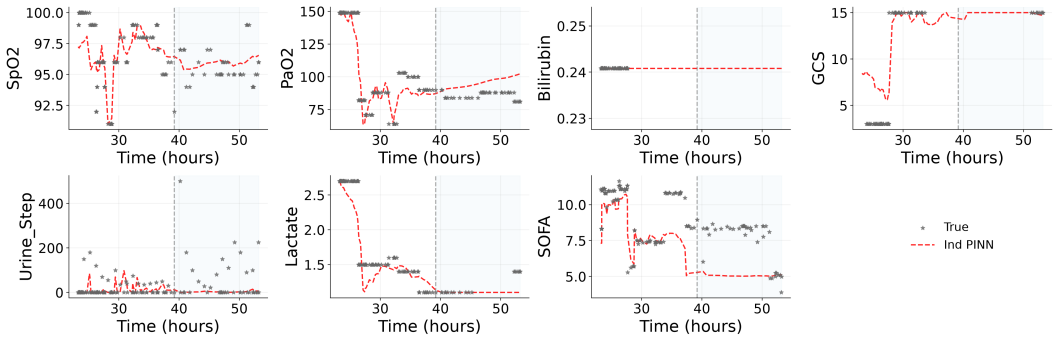


Figure 7: Held-out counterpart of Figure 3 for the same representative test patient. The patient-specific PINN is trained only on the first 70% of the trajectory, indicated by the shaded region, and all rollouts are evaluated on the remaining 30%.

B.4 Overview of the Synthetic Acute Hypotension Dataset

To evaluate whether *MedGym* generalizes beyond sepsis management, we instantiate a second environment from the Synthetic Acute Hypotension Dataset (Kuo et al., 2022), a GAN-generated corpus derived from MIMIC-III that contains $3{,}910$ ICU stays with $187{,}680$ hourly state–action pairs over 48 hours. Acute hypotension (mean arterial pressure below 65 mmHg) is a hemodynamic emergency with physiological dynamics and clinical objectives distinct from the multi-organ pathophysiology of sepsis: decisions are made hourly rather than on the irregular multi-hour scale of the sepsis cohort, the reward is a continuous MAP-based signal rather than a discrete SOFA surrogate, and the safety constraints differ. The same continuous-time formulation and personalization mechanism are applied throughout, with hyperparameters adjusted to the hourly resolution of this cohort.

State Space. The state space comprises ten variables that evolve continuously in time: eight observed physiological measurements and two latent quantities derived from the medication record, as listed in Table 6. The two latent variables, the effective circulating volume index V and the vasopressor effect-compartment concentration C_e, are obtained by forward integration of equation 25 along the recorded fluid and vasopressor history, and give the simulator an explicit representation of volume status and of the delay between an administered dose and its pressor effect. Key differences from sepsis include the explicit separation of MAP and pulse pressure as primary hemodynamic indicators, the hepatic marker ALT, and lactate as a direct indicator of tissue perfusion adequacy.

Action Space. The action space is three-dimensional, comprising the fluid rate, the vasopressor rate, and the inspired oxygen fraction. While fluid and vasopressor administration are originally categorical in the dataset, we treat them as continuous to maintain consistency with the sepsis experiments and to enable fine-grained policy optimization over arbitrary treatment intervals δt.

Reward Function. The clinical objective is to keep MAP within the target zone of 65 to 80 mmHg while avoiding vasopressor overuse, which causes gut ischaemia and lactate elevation. We therefore define the per-step reward as

Table 6: State variables, treatment actions, and reward design in the acute hypotension setting of *MedGym*. Variables marked $(*)$ are latent rather than observed.

Category	Variables	Description
Differential States	$V^{(*)}$, $C_e^{(*)}$, MAP, PP, UO, Lac, Cr, Hep, PaO_2, GCS	Effective circulating volume index, vasopressor effect-compartment concentration, mean arterial pressure, pulse pressure, urine output, lactate, creatinine, hepatic transaminase, arterial oxygen partial pressure, and Glasgow Coma Scale.
Actions	u_{fluid} (mL/kg/h) u_{vaso} (μg/kg/min) FiO_2	Fluid resuscitation, vasopressor administration in noradrenaline equivalents, and supplemental oxygen.
Reward	r_{MAP}, r_{vaso}, r_{lac}	Flat-top MAP reward on $[65, 80]$ mmHg with quadratic vasopressor and linear lactate penalties.

$$b(\mathbf{x}_t, \mathbf{u}_t) = r_{\text{MAP}}(\text{MAP}_t) + r_{\text{vaso}}(\tilde{u}_{\text{vaso},t}) + r_{\text{lac}}(\text{Lac}_t), \tag{22}$$

where $\tilde{u}_{\text{vaso}} \in [0, 1]$ is the vasopressor dose normalized by its maximum. The MAP term is flat inside the target zone and decays quadratically outside it,

$$r_{\text{MAP}}(\text{MAP}) = \begin{cases} +1, & 65 \le \text{MAP} \le 80, \\ \max\Big(-2,\ 1 - \big(\frac{\text{MAP}-65}{12.5}\big)^2\Big), & \text{MAP} < 65, \\ \max\Big(-2,\ 1 - \big(\frac{\text{MAP}-80}{12.5}\big)^2\Big), & \text{MAP} > 80, \end{cases} \tag{23}$$

and the penalty terms are

$$r_{\text{vaso}}(\tilde{u}_{\text{vaso}}) = -0.8\,\tilde{u}_{\text{vaso}}^2, \qquad r_{\text{lac}}(\text{Lac}) = -0.5\max(0,\ \text{Lac} - 2.0). \tag{24}$$

The flat top removes the maximum-dose optimum that a monotone reward would admit, the quadratic vasopressor penalty reflects the nonlinear toxicity of pressors, and the lactate penalty applies only above the clinical threshold of 2.0 mmol/L. This continuous MAP-centerd design contrasts with the discrete multi-organ SOFA score used in sepsis, and tests *MedGym* under a fundamentally different reward structure.

Dynamics. As in equation 14, the mechanistic component is combined with a learned residual and the two are trained jointly on each patient's record. Volume status enters through a Frank–Starling relation and the vasopressor through a first-order effect compartment, so that the two latent states follow

$$\frac{dV}{dt} = \frac{u_{\text{fluid}} + \text{UO}_0 - \text{UO}}{V_{\text{ref}}} - \frac{(V-1)_+}{\tau_{\text{leak}}}, \qquad \frac{dC_e}{dt} = \frac{u_{\text{vaso}} - C_e}{\tau_{c_e}}, \tag{25}$$

where $(\cdot)_+ = \max(\cdot, 0)$, and u_{fluid} and u_{vaso} are the fluid and vasopressor components of $\mathbf{u}$. The eight observed states are driven by these two: mean arterial pressure and pulse pressure relax towards equilibria set by the Frank–Starling stroke volume and by a saturating pressor effect that is attenuated under acidosis; urine output relaxes towards a value set by autoregulated glomerular filtration, itself reduced by venous congestion and by vasopressor-induced renal vasoconstriction; oxygen tension relaxes towards the alveolar gas value corrected for a congestion-dependent shunt; and lactate, creatinine, ALT and GCS follow production–clearance balances driven by oxygen debt, perfusion deficit and gut ischemia. The complete system and the reference values of its parameters are provided in the supplementary material, and each rollout starts from the patient's first recorded observation.

Explicit Clinical Safety Constraints. We additionally monitor four physiological risks that reward optimization alone cannot balance: tissue hypoperfusion indicated by elevated lactate, acute kidney injury indicated by oliguria, hypoxaemia indicated by low PaO_2, and iatrogenic hypertension caused by vasopressor overuse. The per-step cost is the sum of four graded components,

$$c(\mathbf{x}_t) = c_{\text{hyper}} + c_{\text{lac}} + c_{\text{UO}} + c_{\text{PaO}_2}, \qquad c(\mathbf{x}_t) \in [0,\, 5], \tag{26}$$

where

$$c_{\text{hyper}} = \min\big(2,\, \max(0,\, (\text{MAP} - 85)/10)\big), \tag{27}$$

$$c_{\text{lac}} = \begin{cases} 0, & \text{Lac} < 2.0, \\ \min(2,\, (\text{Lac} - 2.0)/2.0), & \text{Lac} \geq 2.0, \end{cases} \tag{28}$$

$$c_{\text{UO}} = \min\big(1,\, 1.5\max(0,\, (0.5 - \text{UO})/0.5)\big), \tag{29}$$

$$c_{\text{PaO}_2} = \min\big(1,\, \max(0,\, (60 - \text{PaO}_2)/60)\big). \tag{30}$$

The urine threshold of 0.5 mL/kg/h is the standard criterion for adequate renal perfusion, and the lactate threshold of 2.0 mmol/L is the established cutoff for tissue hypoperfusion. The Lagrangian-TRPO algorithm enforces $\mathbb{E}[c(\mathbf{x}_t)] \leq d$ through an adaptive multiplier λ, with the budget set to $d = 0.8$ in our experiments. Lactate therefore acts both as a reward penalty and as a constraint, mirroring the sepsis setting, whereas the hypertension term is unique to this cohort and prevents the multiplier from vanishing once MAP exceeds the upper bound of the target zone.

C Details of Policy Training

C.1 Online Policy Learning

This section describes the online policy training protocol used in *MedGym*. Online reinforcement learning interacts directly with the corresponding simulators during training. Policies are optimized by repeatedly collecting rollouts from the simulator and updating the policy in a closed loop manner. We consider the major online reinforcement learning methods, such as Soft Actor Critic (SAC), Proximal Policy Optimization (PPO), and Trust Region Policy Optimization (TRPO), as well as their safety constrained variants, Lagrangian PPO and Lagrangian TRPO. Following the benchmarking pipeline in Figure 1, each policy is trained through direct interaction with its corresponding MedGym environment.

Policy learning framework We formulate online policy learning as a continuous-time reinforcement learning problem on the MedGym environments. At each interaction step k, the policy π observes the current physiological state x_k and selects an action. For fixed-interval policies, the action corresponds to the treatment vector u_k, while for adaptive-time policies, the action is augmented as $\bar{u}_k = (u_k, \delta t_k)$, where δt_k denotes the next intervention interval.

The state consists of the physiological variables listed in Table 5, together with two additional components: the remaining time horizon and the remaining intervention budget. These variables provide temporal context and allow the policy to adapt its strategy based on the remaining decision horizon.

For fixed-interval policies, the action corresponds to the treatment vector u_k defined in Table 5. For adaptive-time policies, the action is augmented as $\bar{u}_k = (u_k, \delta t_k)$, where δt_k denotes the next intervention interval.

Given $(x_k, \bar{u}_k)$, the patient state evolves according to the PINN-based transition model. For both fixed- and adaptive-timing policies, the reward associated with an interaction is accumulated over the corresponding physical-time interval. The precise interval-reward definition and its relation to the continuous-time objective are given below for the adaptive-timing formulation. Policies are optimized from closed-loop trajectories collected through direct interaction with the simulator, in contrast to offline learning from a fixed dataset.

Time-adaptive policy learning framework. We follow the bounded-number-of-interactions formulation of TACOS (Treven et al., 2024). Let

$$0 = t_0 < t_1 < \cdots < t_K = T$$

denote the K intervention times over the fixed treatment horizon $[0, T]$, and define

$$\delta t_k := t_{k+1} - t_k.$$

At interaction k, the treatment action $\mathbf{u}_k$ is applied and held fixed over the interval $[t_k, t_{k+1})$.

Following TACOS, we denote the continuous-time state flow in *MedGym* by

$$\mathbf{x}_{k+1} = \Xi_\xi \left(\mathbf{x}_k, \mathbf{u}_k, \delta t_k \right), \tag{31}$$

where $\mathbf{x}_k = \mathbf{x}_{t_k}$ and Ξ_ξ is induced by the PINN-based continuous-time transition model in equation 1.

The corresponding reward flow is defined by accumulating the continuous-time running reward over the same physical-time interval:

$$\Xi_b \left(\mathbf{x}_k, \mathbf{u}_k, \delta t_k \right) := \int_0^{\delta t_k} b \left(\Xi_\xi(\mathbf{x}_k, \mathbf{u}_k, \tau), \mathbf{u}_k \right) \mathrm{d}\tau. \tag{32}$$

We therefore define the discrete reward at interaction k as

$$r_k = \Xi_b \left(\mathbf{x}_k, \mathbf{u}_k, \delta t_k \right). \tag{33}$$

For the bounded-interaction setting, the policy uses the current patient state, the remaining physical time, and the interaction index. Accordingly, following TACOS, we consider

$$\pi : \mathcal{X} \times \mathcal{T} \times \{0, \ldots, K-1\} \to \mathcal{U} \times [\delta t_{\min}, \delta t_{\max}], \tag{34}$$

where

$$\pi(\mathbf{x}_k, T - t_k, k) = (\mathbf{u}_k, \delta t_k). \tag{35}$$

Thus, the adaptive policy determines both the treatment action and the duration for which it is applied.

The resulting finite-horizon extended MDP optimizes

$$\max_\pi \ \mathbb{E} \left[\sum_{k=0}^{K-1} r_k \right], \qquad t_K = T. \tag{36}$$

Since the intervals $[t_k, t_{k+1})$ form a partition of the complete treatment horizon,

$$\begin{aligned} \sum_{k=0}^{K-1} r_k &= \sum_{k=0}^{K-1} \int_{t_k}^{t_{k+1}} b(\mathbf{x}_t, \mathbf{u}_k) \mathrm{d}t \\ &= \int_0^T b(\mathbf{x}_t, \mathbf{u}_t) \mathrm{d}t. \end{aligned} \tag{37}$$

Hence, the extended discrete-time objective equation 36 is the TACOS reformulation of the continuous-time treatment objective in equation $\mathcal{P}_\xi$. In particular, increasing δt_k does not reduce the accumulated SOFA-based penalty, because treatment utility continues to accumulate throughout the entire interval. Equation equation 37 gives the exact continuous-time-to-extended-MDP equivalence of the bounded-interaction TACOS formulation when the finite-horizon return is undiscounted, i.e., $\gamma = 1$ (Treven et al., 2024). In this case, changing the intervention intervals only changes how the fixed physical horizon is partitioned and does not change the total amount of time over which treatment utility is accumulated.

Lagrangian constrained RL We further consider Lagrangian variants of PPO and TRPO to incorporate safety constraints during online policy learning. The unconstrained PPO and TRPO policies optimize the expected discounted return using the reward advantage A_k^r estimated from on-policy rollouts. PPO performs this update through a clipped objective, while constrains by a KL-divergence trust region.

Safety is enforced through an explicit per-interaction cost. Let $c(\mathbf{x}, \mathbf{u})$ denote the instantaneous cost of the instantiation at hand, given in equation 4 for sepsis and in equation 26 for acute hypotension. The cost of interaction k is $C_k = c(\mathbf{x}_k, \mathbf{u}_k)$, evaluated at the decision state, and the constraint is imposed on the average of C_k over the K interactions of a rollout. Because the interaction budget K and the horizon T are identical in the fixed and the adaptive setting, both are subject to the

same constraint budget, and the reported safety quantities are in addition evaluated on a common physical-time grid over $[0, T]$ rather than at decision times, so the comparison between timing settings does not depend on where the interactions are placed. The reward and cost advantages A_k^r and A_k^c are computed from r_k and C_k, respectively.

Given the Lagrange multiplier $\lambda \geq 0$, we define

$$A_k^\lambda = \frac{A_k^r - \lambda A_k^c}{1 + \lambda}.$$

Thus, actions with high reward advantage are encouraged, while actions that increase the expected safety cost are penalized.

For Lagrangian PPO, we use A_k^λ in the standard PPO clipped surrogate objective:

$$L^{\mathrm{LagPPO}}(\theta) = \mathbb{E}_k \left[\min \left(\rho_k(\theta) A_k^\lambda, \mathrm{clip} \left(\rho_k(\theta), 1 - \epsilon, 1 + \epsilon \right) A_k^\lambda \right) \right],$$

where

$$\rho_k(\theta) = \frac{\pi_\theta(a_k \mid x_k)}{\pi_{\theta_{\mathrm{old}}}(a_k \mid x_k)}.$$

For Lagrangian TRPO, we use the same penalized advantage in the TRPO trust-region update:

$$\max_\theta \mathbb{E}_k \left[\rho_k(\theta) A_k^\lambda \right] \quad \text{s.t.} \quad \mathbb{E}_k \left[D_{\mathrm{KL}} \left(\pi_{\theta_{\mathrm{old}}}(\cdot \mid x_k) \,\|\, \pi_\theta(\cdot \mid x_k) \right) \right] \leq \delta.$$

The multiplier is updated after each rollout batch according to the observed constraint violation:

$$\lambda \leftarrow \left[\lambda + \eta_\lambda \left(\bar{c} - c_{\mathrm{limit}} \right) \right]_+,$$

where $\bar{c}$ is the mean rollout cost and c_{limit} is the cost threshold.

Detailed learning settings All policies are trained over the same fixed physical-time horizon $T = 96$ hours with a prescribed interaction budget K. For adaptive-time policies, the policy outputs

$$\bar{\mathbf{u}}_k = (\mathbf{u}_k, \delta t_k),$$

where $\delta t_k \in [\delta t_{\min}, \delta t_{\max}]$. Following the bounded-interaction formulation of TACOS (Treven et al., 2024), the selected durations satisfy

$$\sum_{k=0}^{K-1} \delta t_k = T, \qquad t_K = T. \tag{38}$$

This constraint is enforced by the environment: each proposed duration is clipped to $[\delta t_{\min}, \delta t_{\max}]$ and to the remaining time $T - t_k$, the last interaction takes the remainder, and an episode ends at $t_k = T$ rather than when the budget is exhausted, so both timing settings traverse the same physical horizon. Hence, the adaptive policy allocates the available interactions across the complete treatment horizon rather than changing the physical duration of the episode. For the fixed-timing baseline, the same horizon and interaction budget are used, with equidistant intervals $\delta t_k = T/K$.

The main difference among the algorithms is how rollout data are collected and reused. SAC is trained as an off-policy method. At each interaction step, the current policy generates a transition

$$(x_k, a_k, r_k, x_{k+1}, \mathrm{done}_k),$$

which is stored in a replay buffer. After the warm-up period, actor and critic networks are updated from mini-batches sampled from the replay buffer. Thus, SAC can reuse past transitions collected by earlier policies.

In contrast, PPO and TRPO are trained as on-policy methods. For each update, we first collect a rollout of length L using the current policy. The collected trajectory is then used to estimate returns and generalized advantages. PPO updates the policy by optimizing the clipped surrogate objective over multiple epochs and mini-batches, whereas TRPO updates the policy by solving a trust-region step with a KL-divergence constraint. After the update, the rollout data are discarded and new trajectories are collected with the updated policy. The Lagrangian variants follow the same data-collection procedure as their base algorithms, but replace the reward advantage with the penalized advantage defined above and update the Lagrange multiplier using the rollout-level cost violation. Each online policy is trained on a single simulator: an individual-level policy on the patient-specific PINN of its patient, a cluster-level policy on the cluster PINN of its cluster, and a population-level policy on the population PINN. The environment-step budget in Table 7 is therefore both the total and the per-simulator budget, and it is identical across the three personalization levels, so the comparison across levels is not confounded by differences in training exposure.

Table 7: RL training hyperparameters.

Component	Hyperparameter	Value
TACOS	$\delta t_{\min}$	0.5 h
	$\delta t_{\max}$	36.0 h
	Interaction budget K	20
	Fixed-timing interval T/K	4.8 h
	Total horizon T	96 h
Shared (online)	Hidden layer width	256
	Discount factor γ	1
SAC	Total environment steps	100,000
	Replay buffer capacity / warm-up	200,000 / 20,000
	Batch size	512
	Learning rate (actor / critic / α)	1×10^{-4}
	Soft-update rate τ	0.002
	Entropy temperature α	auto-tuned
	Target entropy	$-\dim(\mathcal{A})$
PPO / PPO-Lagrangian	Total environment steps	200,000
	Rollout length	2,048
	Update epochs per rollout	10
	Mini-batch size	64
	Actor / critic learning rate	3×10^{-4} / 1×10^{-3}
	GAE λ	0.95
	Clip range ϵ	0.2
	Entropy coefficient	0.01
	Cost limit c_{limit} (mean C_k) / lr λ	0.1 / 5×10^{-2}
TRPO / TRPO-Lagrangian	Total environment steps	200,000
	Rollout length	2,048
	Trust-region radius δ	0.01
	Critic learning rate	1×10^{-3}
	Value-function update steps	5
	Conjugate gradient iterations	10
	Line search (max back-tracks, coeff.)	10, 0.5
	GAE λ	0.95
	Cost limit c_{limit} (mean C_k) / lr λ	0.1 / 5×10^{-2}
Offline data	Behavior policy	TRPO-Lagrangian (fixed timing)
	Rollout episodes per patient S	5
	Action perturbation $\sigma_{\mathbf{u}}$	0.05
DQN / CQL / GCQL	Gradient steps	4,000
	Batch size	256
	Actor / critic learning rate	3×10^{-5}
	Soft-update rate τ	0.001
	Discount factor γ	0.99
CQL / GCQL	Conservative coefficient α	5.0
	Candidate actions / sampling s.d.	12 / 0.1
GCQL	Support radius multiplier	0.5
	Guard weight β (critic / actor)	0.5 / 2.0
	Guard sigmoid sharpness	20

C.2 Offline Policy Learning

This section describes the offline policy learning protocol used in MedGym. Unlike the online policy learning methods in Appendix C.1, offline reinforcement learning does not interact with the simulator during training. Instead, the policy is trained from a fixed dataset of transition tuples generated before policy optimization. After training, the learned offline policy is deployed in MedGym and evaluated by closed-loop rollouts. This follows the benchmark pipeline in Figure 2: MedGym is first used to

generate offline training data, the offline RL algorithm learns a policy from the fixed dataset, and the learned policy is then evaluated online in the same continuous-time patient simulators.

Throughout the offline evaluation, both the state space $\mathcal{X} \subseteq \mathbb{R}^{d_s}$ and the treatment action space $\mathcal{U} \subseteq \mathbb{R}^{d_a}$ are continuous, consistent with the environment definition in Section 2. Accordingly, action maximization and conservative regularization below are defined over the continuous action space; in implementation, the corresponding continuous-action operations are evaluated numerically using sampled candidate actions from the admissible action domain.

MedGym environments. As stated in the main text, the offline setting uses a single cohort of 110 patient-specific PINN simulators constructed in Appendix B, partitioned into 10 dynamics-based clusters. For each cluster $k \in \{1, \dots, 10\}$, one patient, denoted by j_k^{test}, is held out for testing, and the remaining ten patients form the training subset $\mathcal{C}_k^{\text{train}}$, giving 100 training patients and 10 held-out test patients. The patient-specific MedGym environments remain the basic environments for generating offline trajectories and evaluating learned policies. Cluster-level and population-level offline datasets are constructed only from the datasets of the training patients.

Patient-specific offline dataset. For each patient $i \in \{1, \dots, n\}$, we train a fixed-interval Lagrangian TRPO behavior policy $\pi^{(i)}_{\text{LagrangianTRPO}}$ on the corresponding PINN $F^{\text{nn}}_{\xi_i}(\mathbf{x}, \mathbf{u}, \delta t)$, using the online training protocol in Appendix C.1 with $\delta t_k = T/K$. Still on the same PINN $F^{\text{nn}}_{\xi_i}(\mathbf{x}, \mathbf{u}, \delta t)$, we generate an offline dataset, consisting of S rollout episodes,

$$\mathcal{D}^{(i)}_{\text{LagrangianTRPO}} := \left(\tau_s^{(i)}\right)_{s=1}^{S}, \quad \text{where} \quad \tau_s^{(i)} := \left(\mathbf{x}^{(i)}_{s,0}, \mathbf{u}^{(i)}_{s,0}, r^{(i)}_{s,0}, \mathbf{x}^{(i)}_{s,1}, \dots, \mathbf{x}^{(i)}_{s,H_i^{(s)}}\right).$$

To increase the diversity of the offline data and avoid generating a nearly deterministic dataset, we add small Gaussian perturbations to the behavior action and project the result back onto the admissible action set, via

$$\mathbf{u}^{(i)}_{s,k} = \Pi_{\mathcal{U}}\left(\mathbf{u}^{(i)}_k + \epsilon^{(i)}_{s,\mathbf{u}}\right), \qquad \epsilon^{(i)}_{s,\mathbf{u}} \sim \mathcal{N}(0, \sigma^2_{\mathbf{u}} I), \qquad \mathbf{u}^{(i)}_k = \pi^{(i)}_{\text{LagrangianTRPO}}(\mathbf{x}^{(i)}_{s,k}), \tag{39}$$

where $\Pi_{\mathcal{U}}$ denotes the projection onto the admissible action set and $\sigma_{\mathbf{u}} = 0.05$ in the normalized action space. Every rollout starts from the patient's recorded initial condition, which is not perturbed. Only the patient-specific datasets associated with $i \in \mathcal{I}_{\text{train}}$ are used to construct the cluster-level and population-level offline training datasets. The dataset of each held-out test patient is used only to train the corresponding individual oracle policy and is not included in cluster-level or population-level policy training.

Deep Q-Network. The *Deep Q-Network (DQN)* (Mnih et al., 2013) approximates the action-value function by $Q_\theta : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$. For the continuous treatment action space, we write the Bellman regression objective as

$$\min_{\theta \in \Theta} J_{\text{DQN}}(\theta) := \frac{1}{2}\mathbb{E}_{(\mathbf{x},\mathbf{u},r,\mathbf{x}') \sim \mathcal{D}}\left[r + \gamma \sup_{\mathbf{u}' \in \mathcal{U}} Q_\theta(\mathbf{x}', \mathbf{u}') - Q_\theta(\mathbf{x}, \mathbf{u})\right]^2. \tag{40}$$

Conservative Q-Learning. Directly applying value-based learning to a fixed offline dataset may assign erroneously high values to treatment actions insufficiently supported by the behavior data. CQL (Kumar et al., 2020) mitigates this extrapolation error by penalizing high values over the continuous action domain:

$$\min_{\theta \in \Theta} J_{\text{CQL}}(\theta) := J_{\text{DQN}}(\theta) + \alpha\left(\mathbb{E}_{\mathbf{x} \sim \mathcal{D}}\left[\log \int_{\mathcal{U}} \exp(Q_\theta(\mathbf{x}, \mathbf{u}))\, \mathrm{d}\nu(\mathbf{u})\right] - \mathbb{E}_{(\mathbf{x},\mathbf{u}) \sim \mathcal{D}}\left[Q_\theta(\mathbf{x}, \mathbf{u})\right]\right), \tag{41}$$

where ν denotes a reference distribution over the admissible continuous action space. The integral is approximated by Monte Carlo sampling in implementation.

Guarded CQL. Although CQL penalizes actions that may induce out-of-distribution (OOD) behavior, the learned policy may still assign high values to state–action pairs that are weakly

supported by the offline dataset. Guarded CQL (GCQL) (Yan* et al., 2025; Tumay et al., 2025) addresses this issue by introducing an additional support-aware regularization mechanism.

Rather than explicitly estimating the density of the offline data, we use the distance of a state–action pair from the center of the dataset as a computationally simple and conceptually effective surrogate for data support. This approximation does not reproduce the original density-based criterion exactly, but retains its main purpose: state–action pairs located farther from the region occupied by the behavior data are treated as less supported and are therefore penalized more strongly. Let

$$\mathbf{z} = (\mathbf{x}, \mathbf{u}) \in \mathcal{X} \times \mathcal{U},$$

and define the dataset centroid as

$$\overline{\mathbf{z}} = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{z} \in \mathcal{D}} \mathbf{z},$$

together with the support threshold

$$d_{\lim} = \tau \max_{\mathbf{z} \in \mathcal{D}} \|\mathbf{z} - \overline{\mathbf{z}}\|_2 ,$$

where $\tau = 0.5$ in our experiments. We define the corresponding guard indicator as

$$g(\mathbf{z}) = \mathbf{1} \{ \|\mathbf{z} - \overline{\mathbf{z}}\|_2 > d_{\lim} \},$$

which identifies state–action pairs lying outside the prescribed data-support region. In implementation the indicator is replaced by a sigmoid relaxation of the same threshold with sharpness 20, and the guard is applied to the critic and the actor objectives with separate weights, $\beta = 0.5$ and $\beta = 2.0$. The resulting guarded objective takes the form

$$\min_{\theta \in \Theta} J_{\mathrm{GCQL}}(\theta) = J_{\mathrm{CQL}}(\theta) + \beta \, \mathcal{R}_{\mathrm{guard}}(\theta). \tag{42}$$

We define the guarded regularizer as

$$\mathcal{R}_{\mathrm{guard}}(\theta) = \mathbb{E}_{\mathbf{x} \sim \mathcal{D},\, \mathbf{u} \sim \nu} \left[g(\mathbf{x}, \mathbf{u}) Q_\theta(\mathbf{x}, \mathbf{u}) \right], \tag{43}$$

where ν is the same reference distribution over admissible continuous actions used in the CQL regularizer, and $g(\mathbf{x}, \mathbf{u})$ is the guard indicator defined above. Thus, state–action pairs identified as weakly supported receive an additional Q_θ-dependent conservative penalty.

Patient-specific training. Given the fixed dataset $\mathcal{D}^{(i)}_{\text{LagrangianTRPO}}$, we train patient-specific DQN, CQL, and GCQL policies $\pi^{(i)}_{\text{DQN}}$, $\pi^{(i)}_{\text{CQL}}$, $\pi^{(i)}_{\text{GCQL}}$, respectively.

Patient-specific evaluation. After training, each $\pi^{(i)}_{\text{DQN}}$, $\pi^{(i)}_{\text{CQL}}$, and $\pi^{(i)}_{\text{GCQL}}$ is evaluated by closed-loop rollouts on the same PINN $F^{\text{nn}}_{\xi_i}(\mathbf{x}, \mathbf{u}, \delta t)$ that they were trained on. Similarly to equation 39, we produce the datasets $\mathcal{D}^{(i)}_{\text{DQN}}$, $\mathcal{D}^{(i)}_{\text{CQL}}$, and $\mathcal{D}^{(i)}_{\text{GCQL}}$. For each patient i, we compute treatment effectiveness and safety metrics. The main effectiveness metrics include cumulative reward, mean SOFA, and SOFA at the terminal horizon. The main safety metrics include mean lactate, lactate trajectory, and the lactate safety rate after the critical time threshold. This gives one patient-specific DQN, CQL, and GCQL policy for each of the 10 held-out test patients $\{j^{\text{test}}_k\}_{k=1}^{10}$, each trained on that patient's own offline dataset and evaluated on that patient's own simulator $F^{\text{nn}}_{\xi_{j^{\text{test}}_k}}$

Cluster-level offline dataset, training and evaluation. For each cluster k, we construct a cluster-level offline dataset by aggregating the patient-specific datasets of its training patients,

$$\mathcal{D}^{(k)}_{\text{clu}} = \bigcup_{i \in \mathcal{C}^{\text{train}}_k} \mathcal{D}^{(i)}_{\text{LagrangianTRPO}}.$$

Using $\mathcal{D}^{(k)}_{\text{clu}}$, we train cluster-level policies $\pi^{(k)}_{\text{DQN}}$, $\pi^{(k)}_{\text{CQL}}$ and $\pi^{(k)}_{\text{GCQL}}$ shared across the patients of cluster k. Each cluster-level policy is then evaluated on the simulator of the held-out test patient of the same cluster, i.e. on $F^{\text{nn}}_{\xi_{j^{\text{test}}_k}}$ — the same environment on which the individual- and population-level policies of that cluster are evaluated — producing the datasets $\mathcal{D}^{(k)}_{\text{DQN}}$, $\mathcal{D}^{(k)}_{\text{CQL}}$ and $\mathcal{D}^{(k)}_{\text{GCQL}}$.

Population wide offline dataset, training and evaluation. In addition to patient-specific offline learning, we construct a population-level offline dataset by aggregating all patient-specific datasets,

$$\mathcal{D}^{\text{pop}} = \bigcup_{i=1}^{100} \mathcal{D}^{(i)}_{\text{LagrangianTRPO}}.$$

Using $\mathcal{D}^{\text{pop}}$, we train single population-level policies $\pi^{\text{pop}}_{\text{DQN}}$, $\pi^{\text{pop}}_{\text{CQL}}$ and $\pi^{\text{pop}}_{\text{GCQL}}$ shared across all training patients. These are then evaluated separately on each of the 10 held-out patient-specific environments $\{F^{\text{nn}}_{\xi_{j^{\text{test}}_k}}\}_{k=1}^{10}$, i.e. on exactly the same environments used to evaluate the individual- and cluster-level policies, producing the datasets $\mathcal{D}^{\text{pop}}_{\text{DQN}}$, $\mathcal{D}^{\text{pop}}_{\text{CQL}}$ and $\mathcal{D}^{\text{pop}}_{\text{GCQL}}$.

Evaluation question. This offline protocol is designed to answer the following question: how does offline policy performance change as the training data become increasingly patient-specific? For each held-out test patient j^{test}_k, the three policy levels $\pi^{(j^{\text{test}}_k)}_{\diamond}$, $\pi^{(k)}_{\diamond}$ and $\pi^{\text{pop}}_{\diamond}$ are compared on the same simulator $F^{\text{nn}}_{\xi_{j^{\text{test}}_k}}$, where $\diamond$ can be DQN, CQL, or GCQL. They differ only in the offline dataset they are trained on: that patient's own trajectories, the trajectories of its cluster's training patients, or the trajectories of all training patients. If $\pi^{(j^{\text{test}}_k)}_{\diamond}$ consistently outperforms $\pi^{(k)}_{\diamond}$ and $\pi^{\text{pop}}_{\diamond}$, this indicates that offline policy learning benefits from patient-specific behavior data. If the cluster-level policy recovers part of that gap, the dynamics-based grouping transfers useful treatment structure across similar patients. If $\pi^{\text{pop}}_{\diamond}$ performs comparably, the aggregated offline dataset already captures enough shared treatment structure to support population-level generalization.

D Details of Evaluation Results

D.1 Online Policy Methods Evaluations

This subsection complements the main-text evaluation of TRPOLag by reporting additional online results for SAC, PPO, TRPO, and PPOLag. Figure 8 presents the corresponding final SOFA distributions and representative physiological trajectories, while Table 8 summarizes the quantitative results. All methods use the same 70 held-out patients, individual/cluster/population policy classes, and fixed/adaptive timing settings as the main-text evaluation. The results exhibit the same overall personalization pattern: the individual oracle generally achieves the lowest SOFA scores, while the cluster policies recover part of this benefit relative to the population policies.

D.2 Evaluation Under an Alternative Clustering Scheme

Model and Dataset Construction. To evaluate the sensitivity of the results to the patient-grouping strategy, we additionally consider an alternative setting with $N_{\mathsf{clu}} = 15$ patient groups. Each group contains one representative training patient and one matched held-out test patient, such that $|\mathcal{C}^{\mathsf{train}}_k| = 1$ and $|\mathcal{C}^{\mathsf{test}}_k| = 1$. The cluster-level model and policy for each group are constructed using its representative training patient. The population-level model and policy aggregate all training patients, giving $\sum_{k=1}^{N_{\mathsf{clu}}} |\mathcal{C}^{\mathsf{train}}_k| = 15$ training patients in total. Evaluation is conducted on the corresponding $\sum_{k=1}^{N_{\mathsf{clu}}} |\mathcal{C}^{\mathsf{test}}_k| = 15$ held-out test patients. These test patients are excluded from the construction and training of the cluster- and population-level models and policies. The evaluated RL algorithms, interaction-timing settings, and evaluation metrics remain the same as those used in the primary evaluation. Based on this alternative patient-grouping setting, Figure 9 and Table 9 report the results for SAC, PPO, TRPO, PPOLag, and TRPOLag under both fixed and adaptive timing. The figure presents the final SOFA distributions and representative physiological trajectories, while the table summarizes the corresponding treatment-effectiveness and safety metrics.

The results are consistent with those obtained under the primary 70-patient partition. The individual oracle generally achieves the lowest SOFA scores, while the cluster policies recover part of this personalization benefit relative to the population policies. Under the alternative partition, the cluster policy achieves a lower mean final SOFA score than the population policy in nine of the ten algorithm–timing combinations, with TRPO-A being the only exception. This result suggests that the benefit of cluster-based transfer persists across different patient-grouping strategies, although its magnitude depends on the selected partition.

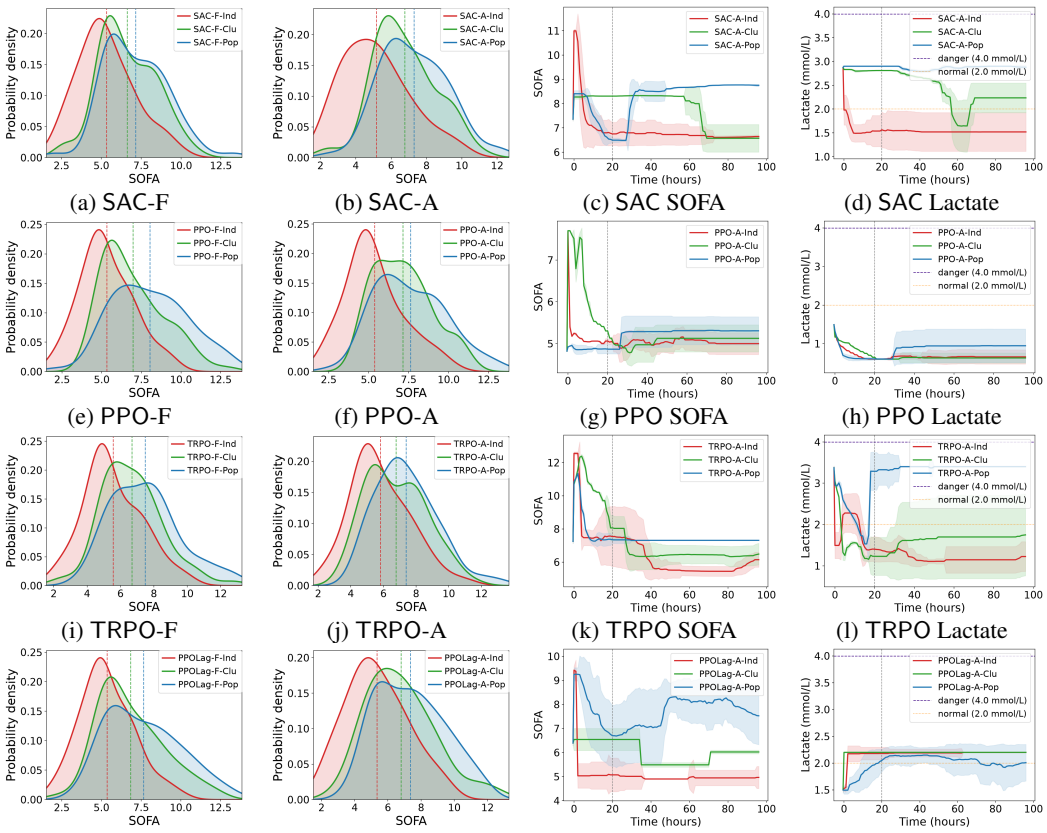


Figure 8: Additional online evaluations of SAC, PPO, TRPO, and PPOLag. Each row corresponds to one algorithm; the columns show the final SOFA distributions under fixed and adaptive timing, followed by representative adaptive-timing SOFA and Lactate trajectories.

Policy	SOFA(all) (↓)	Lower SOFA vs. Pop (%)	SOFA(96h) (↓)	SafeTime% (↑)
SAC-F-Ind	5.97 ± 1.67	97.1	5.32 ± 1.77	$\mathbf{91.95 \pm 22.83}$
SAC-A-Ind	$\mathbf{5.39 \pm 1.89}$	100.0	$\mathbf{5.16 \pm 1.90}$	90.34 ± 24.14
SAC-F-Clu	7.05 ± 1.74	70.0	6.62 ± 1.75	90.87 ± 24.90
SAC-A-Clu	6.86 ± 1.62	72.9	6.76 ± 1.73	89.98 ± 23.86
SAC-F-Pop	7.31 ± 1.87	–	7.15 ± 1.92	90.48 ± 26.80
SAC-A-Pop	7.20 ± 1.81	–	7.29 ± 1.86	90.71 ± 22.54
PPO-F-Ind	5.75 ± 1.66	100.0	$\mathbf{5.32 \pm 1.67}$	91.08 ± 23.76
PPO-A-Ind	$\mathbf{5.45 \pm 1.80}$	95.7	5.36 ± 1.80	91.35 ± 23.90
PPO-F-Clu	7.85 ± 1.85	72.9	6.97 ± 1.92	89.04 ± 28.76
PPO-A-Clu	7.68 ± 1.80	62.9	7.13 ± 1.77	$\mathbf{91.46 \pm 21.70}$
PPO-F-Pop	8.33 ± 2.20	–	8.05 ± 2.34	89.62 ± 28.46
PPO-A-Pop	8.00 ± 2.12	–	7.61 ± 2.19	89.06 ± 27.34
TRPO-F-Ind	6.46 ± 1.78	94.3	$\mathbf{5.55 \pm 1.73}$	$\mathbf{93.19 \pm 20.03}$
TRPO-A-Ind	$\mathbf{6.27 \pm 1.79}$	88.6	5.82 ± 1.80	90.49 ± 23.49
TRPO-F-Clu	7.73 ± 1.68	67.1	6.68 ± 1.81	92.23 ± 23.03
TRPO-A-Clu	7.62 ± 1.71	60.0	6.77 ± 1.80	91.27 ± 20.82
TRPO-F-Pop	8.48 ± 1.85	–	7.48 ± 2.05	91.17 ± 24.31
TRPO-A-Pop	7.96 ± 1.73	–	7.39 ± 1.84	93.09 ± 19.72
PPOLag-F-Ind	5.74 ± 1.65	98.6	5.35 ± 1.67	$\mathbf{97.72 \pm 12.05}$
PPOLag-A-Ind	$\mathbf{5.55 \pm 1.87}$	94.3	$\mathbf{5.34 \pm 1.83}$	96.60 ± 13.25
PPOLag-F-Clu	7.43 ± 1.96	68.6	6.84 ± 2.04	89.39 ± 27.51
PPOLag-A-Clu	7.23 ± 1.89	65.7	6.78 ± 2.05	89.27 ± 25.85
PPOLag-F-Pop	7.96 ± 2.10	–	7.65 ± 2.28	90.76 ± 27.12
PPOLag-A-Pop	7.62 ± 1.96	–	7.36 ± 2.00	90.24 ± 25.01

Table 8: Additional online-policy results on 70 held-out test patients. Metrics, comparisons, abbreviations, and boldface conventions follow Table 2; dashes indicate that the population policy is not compared with itself. The corresponding trajectory-level safety metrics are reported in Table 10.

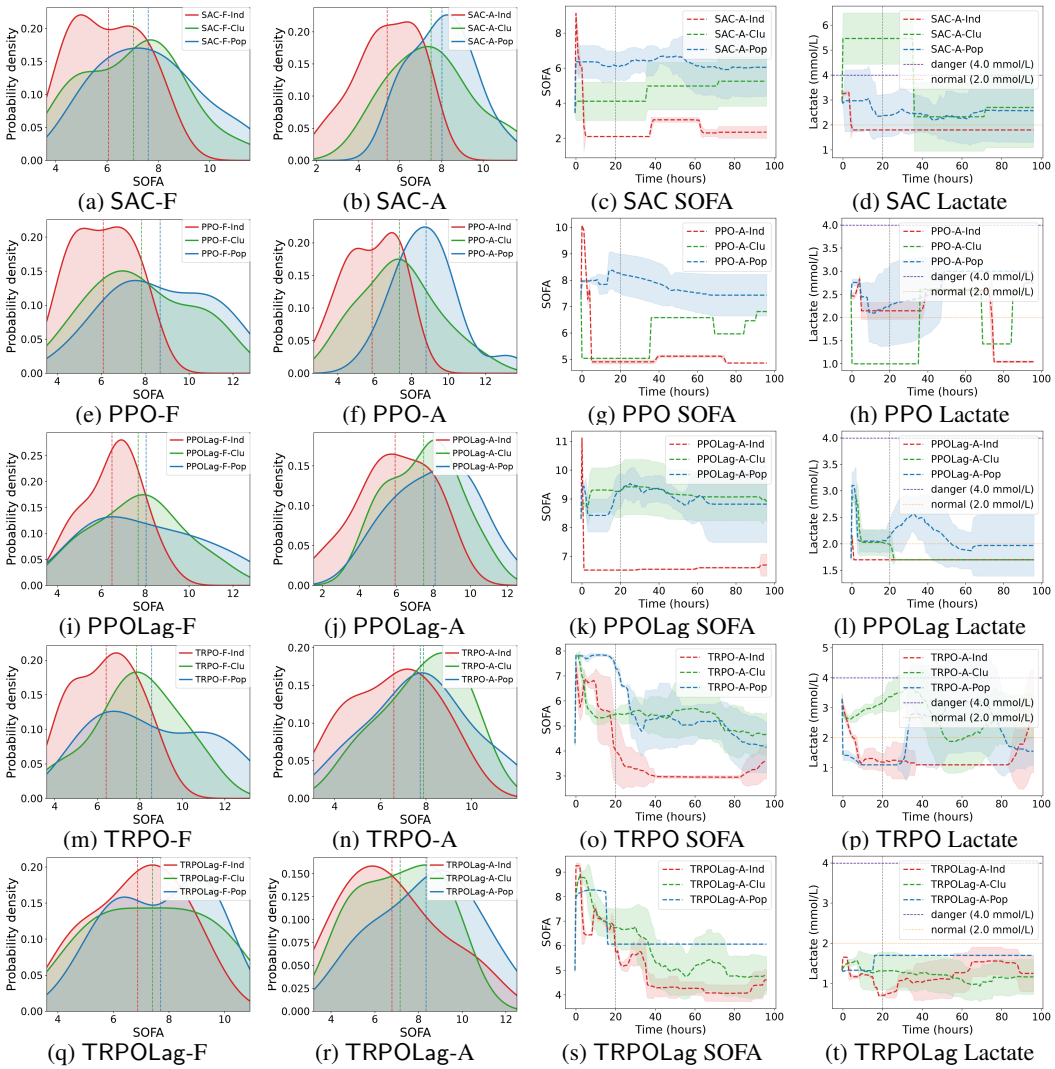


Figure 9: Online evaluation under the alternative 15-cluster partition. Each row corresponds to one RL algorithm. From left to right, the columns show the distributions of final SOFA scores under fixed and adaptive timing, followed by adaptive-timing SOFA and Lactate trajectories on a representative test patient. The distributional results are computed over 15 held-out patients, with each patient-level score averaged over ten noise realizations.

Policy	SOFA(all) (↓)	Lower SOFA vs. Pop (%)	SOFA(96h) (↓)	SafeTime% (↑)
SAC-F-Ind	6.73 ± 1.35	100.0	6.06 ± 1.33	$\mathbf{76.03 \pm 38.46}$
SAC-A-Ind	$\mathbf{5.79 \pm 1.46}$	100.0	$\mathbf{5.39 \pm 1.46}$	70.52 ± 41.71
SAC-F-Clu	7.28 ± 1.70	73.3	7.02 ± 1.86	74.02 ± 37.24
SAC-A-Clu	7.51 ± 1.84	80.0	7.50 ± 1.91	70.10 ± 34.31
SAC-F-Pop	7.80 ± 1.81	–	7.60 ± 1.89	70.25 ± 41.38
SAC-A-Pop	7.89 ± 1.49	–	8.02 ± 1.45	70.94 ± 34.44
PPO-F-Ind	6.56 ± 1.39	100.0	6.08 ± 1.31	78.64 ± 38.06
PPO-A-Ind	$\mathbf{5.92 \pm 1.42}$	100.0	$\mathbf{5.85 \pm 1.42}$	65.15 ± 41.45
PPO-F-Clu	8.11 ± 1.90	60.0	7.83 ± 2.18	78.47 ± 34.36
PPO-A-Clu	7.48 ± 1.88	86.7	7.33 ± 2.01	69.87 ± 37.24
PPO-F-Pop	8.88 ± 2.03	–	8.68 ± 2.24	73.44 ± 37.28
PPO-A-Pop	9.05 ± 1.72	–	8.78 ± 1.67	$\mathbf{85.89 \pm 16.47}$
PPOLag-F-Ind	6.82 ± 1.32	86.7	6.48 ± 1.26	$\mathbf{92.64 \pm 20.40}$
PPOLag-A-Ind	$\mathbf{6.23 \pm 1.98}$	100.0	$\mathbf{5.93 \pm 1.90}$	78.41 ± 29.20
PPOLag-F-Clu	7.80 ± 1.70	46.7	7.67 ± 1.95	78.93 ± 35.80
PPOLag-A-Clu	7.60 ± 1.79	73.3	7.47 ± 1.79	63.76 ± 36.93
PPOLag-F-Pop	8.14 ± 2.31	–	8.04 ± 2.42	65.51 ± 42.73
PPOLag-A-Pop	8.63 ± 1.60	–	8.09 ± 2.09	76.86 ± 30.76
TRPO-F-Ind	7.36 ± 1.51	100.0	$\mathbf{6.42 \pm 1.45}$	74.34 ± 38.30
TRPO-A-Ind	$\mathbf{7.00 \pm 1.87}$	66.7	6.58 ± 1.80	74.05 ± 37.81
TRPO-F-Clu	8.28 ± 1.54	80.0	7.84 ± 1.93	$\mathbf{79.61 \pm 39.03}$
TRPO-A-Clu	8.35 ± 1.61	46.7	7.88 ± 1.72	70.57 ± 31.74
TRPO-F-Pop	8.82 ± 2.29	–	8.55 ± 2.52	75.65 ± 34.35
TRPO-A-Pop	8.23 ± 1.62	–	7.74 ± 2.09	76.88 ± 33.03
TRPOLag-F-Ind	7.59 ± 1.41	93.3	6.87 ± 1.59	$\mathbf{91.82 \pm 21.52}$
TRPOLag-A-Ind	$\mathbf{7.09 \pm 2.05}$	73.3	$\mathbf{6.78 \pm 2.16}$	82.79 ± 25.84
TRPOLag-F-Clu	7.70 ± 1.79	73.3	7.40 ± 1.97	77.96 ± 34.42
TRPOLag-A-Clu	7.53 ± 1.74	73.3	7.16 ± 1.83	71.67 ± 36.45
TRPOLag-F-Pop	8.46 ± 1.35	–	7.70 ± 1.68	78.37 ± 32.61
TRPOLag-A-Pop	8.57 ± 1.75	–	8.34 ± 2.15	76.50 ± 34.60

Table 9: Online-policy results under the alternative clustering scheme on 15 held-out test patients. Metrics, comparisons, abbreviations, and boldface conventions follow Table 2; dashes indicate that the population policy is not compared with itself.

D.3 Trajectory-Level Safety Evaluation for Online RL

To evaluate safety independently of the number and spacing of intervention times, we assess all policies on a common physical-time grid over the complete treatment horizon $[0, T]$. Let $\mathrm{Lac}(t)$ denote the simulated lactate trajectory and let $\ell_{\mathrm{safe}} = 4.0$ mmol/L denote the safety threshold. We report four complementary metrics: *SafeTime%*, the fraction of physical time for which $\mathrm{Lac}(t) < \ell_{\mathrm{safe}}$; *ViolDur*, the total duration for which $\mathrm{Lac}(t) \geq \ell_{\mathrm{safe}}$; *PeakLac*, the maximum lactate value over the trajectory; and *ExcessAUC*, the time integral of the threshold exceedance $[\mathrm{Lac}(t) - \ell_{\mathrm{safe}}]_+$. These metrics respectively capture the amount of safe physical time, duration of unsafe exposure, worst-case excursion, and the combined duration and severity of threshold violation.

$$\mathrm{SafeTime}\% = \frac{100}{T} \int_0^T \mathbf{1}\{\mathrm{Lac}(t) < \ell_{\mathrm{safe}}\} \, \mathrm{d}t, \tag{44}$$

$$\mathrm{ViolDur} = \int_0^T \mathbf{1}\{\mathrm{Lac}(t) \geq \ell_{\mathrm{safe}}\} \, \mathrm{d}t, \tag{45}$$

$$\mathrm{PeakLac} = \max_{t \in [0,T]} \mathrm{Lac}(t), \tag{46}$$

$$\mathrm{ExcessAUC} = \int_0^T [\mathrm{Lac}(t) - \ell_{\mathrm{safe}}]_+ \, \mathrm{d}t. \tag{47}$$

Table 10 reports these metrics for all online policies. The Lagrangian individual policies are the safest overall, whereas the Lagrangian cluster and population policies show no consistent safety advantage over their unconstrained counterparts. This indicates that treatment effectiveness and trajectory-level safety can lead to different rankings of the same policies.

D.4 Offline Policy Methods Evaluations

This section reports two sets of offline results that complement Table 3 in the main text: the CQL results under the primary grouping, and the results of all three offline learners under the alternative patient grouping of Appendix D.2. Following the offline learning protocol in Appendix C.2, we

Policy	SafeTime% (↑)	ViolDur (h) (↓)	PeakLac (↓)	ExcessAUC (↓)
SAC-F-Ind	91.95 ± 22.83	7.71 ± 21.92	2.83 ± 2.57	12.88 ± 48.72
SAC-A-Ind	90.34 ± 24.14	9.26 ± 23.20	2.89 ± 2.08	17.26 ± 72.36
SAC-F-Clu	90.87 ± 24.90	8.77 ± 23.92	2.80 ± 2.36	25.84 ± 143.59
SAC-A-Clu	89.98 ± 23.86	9.61 ± 22.90	2.95 ± 2.34	24.35 ± 101.26
SAC-F-Pop	90.48 ± 26.80	9.14 ± 25.73	2.87 ± 2.62	36.29 ± 189.17
SAC-A-Pop	90.71 ± 22.54	8.90 ± 21.61	2.99 ± 2.53	26.96 ± 151.06
PPO-F-Ind	91.08 ± 23.76	8.55 ± 22.80	2.82 ± 2.46	12.73 ± 40.91
PPO-A-Ind	91.35 ± 23.90	8.27 ± 22.95	2.97 ± 2.01	13.48 ± 50.35
PPO-F-Clu	89.04 ± 28.76	10.53 ± 27.63	2.80 ± 2.63	31.65 ± 176.28
PPO-A-Clu	91.46 ± 21.70	8.19 ± 20.83	2.99 ± 2.59	27.23 ± 145.41
PPO-F-Pop	89.62 ± 28.46	9.96 ± 27.33	2.84 ± 2.54	16.24 ± 69.55
PPO-A-Pop	89.06 ± 27.34	10.48 ± 26.26	2.99 ± 2.28	21.35 ± 83.77
TRPO-F-Ind	93.19 ± 20.03	6.51 ± 19.19	2.84 ± 2.41	9.02 ± 35.38
TRPO-A-Ind	90.49 ± 23.49	9.11 ± 22.55	3.03 ± 2.36	13.19 ± 37.85
TRPO-F-Clu	92.23 ± 23.03	7.45 ± 22.10	2.82 ± 2.54	15.54 ± 58.15
TRPO-A-Clu	91.27 ± 20.82	8.37 ± 19.99	3.04 ± 2.63	25.96 ± 149.66
TRPO-F-Pop	91.17 ± 24.31	8.47 ± 23.38	2.83 ± 2.45	16.32 ± 71.68
TRPO-A-Pop	93.09 ± 19.72	6.60 ± 18.91	3.04 ± 2.60	12.20 ± 39.79
PPOLag-F-Ind	97.72 ± 12.05	2.15 ± 11.55	2.72 ± 2.14	3.50 ± 14.92
PPOLag-A-Ind	96.60 ± 13.25	3.24 ± 12.70	2.72 ± 1.37	3.61 ± 11.81
PPOLag-F-Clu	89.39 ± 27.51	10.20 ± 26.43	2.85 ± 2.64	35.60 ± 186.08
PPOLag-A-Clu	89.27 ± 25.85	10.29 ± 24.81	2.89 ± 2.52	34.23 ± 175.37
PPOLag-F-Pop	90.76 ± 27.12	8.86 ± 26.05	2.84 ± 2.53	14.47 ± 55.30
PPOLag-A-Pop	90.24 ± 25.01	9.36 ± 24.01	2.76 ± 1.71	14.50 ± 51.28
TRPOLag-F-Ind	97.46 ± 12.65	2.41 ± 12.12	2.66 ± 2.40	4.42 ± 23.64
TRPOLag-A-Ind	95.99 ± 13.76	3.84 ± 13.21	2.54 ± 1.10	4.02 ± 12.12
TRPOLag-F-Clu	91.45 ± 23.62	8.22 ± 22.71	2.79 ± 2.21	25.86 ± 127.96
TRPOLag-A-Clu	88.37 ± 25.61	11.16 ± 24.60	2.85 ± 2.12	23.64 ± 97.95
TRPOLag-F-Pop	89.15 ± 29.45	10.42 ± 28.31	2.76 ± 2.37	31.88 ± 150.16
TRPOLag-A-Pop	89.40 ± 24.14	10.18 ± 23.17	2.93 ± 2.55	33.57 ± 161.28

Table 10: Schedule-independent trajectory-level safety evaluation for all online RL policies. All metrics are computed over the common physical-time horizon $[0, T]$, independently of the number and spacing of intervention times. SafeTime% denotes the fraction of physical time for which lactate remains below 4.0 mmol/L; ViolDur denotes the total duration of lactate violation; PeakLac denotes the maximum lactate value over the trajectory; and ExcessAUC denotes the time integral of lactate above the 4.0 mmol/L threshold. Results are reported as mean and standard deviation across held-out patients. The large standard deviations of ExcessAUC reflect a zero-inflated distribution: the majority of patients maintain lactate below the threshold throughout and thus contribute zero excess, while a small number of patients with high initial lactate dominate the variance.

first generate fixed-dt offline datasets from the Lagrangian TRPO behavior policy and then train DQN, CQL, and GCQL policies without further simulator interaction during policy optimization. After training, all offline policies are deployed in the same MedGym environments and evaluated by closed-loop rollouts. This evaluation is intended to assess whether offline learning can recover useful patient-specific treatment policies from fixed behavior data, and whether the resulting individual policies improve over their corresponding population-level policies.

CQL under the primary grouping. CQL is reported here rather than in the main text because it behaves differently from the other two offline learners. Its individual policy attains the lowest trajectory-wide SOFA score of the three levels, but its cluster policy falls behind the population policy rather than sitting between the two. Since all three algorithms are trained on the same datasets and evaluated on the same simulators, the difference lies in how each controls extrapolation away from the behavior data.

Policy	SOFA(all) (↓)	Lower SOFA vs. Pop (%)	SOFA(96h) (↓)	SafeTime% (↑)
CQL-F-Ind	6.83 ± 0.61	50.0	6.55 ± 0.90	98.95 ± 2.99
CQL-F-Clu	7.59 ± 1.65	30.0	7.57 ± 1.74	87.34 ± 30.17
CQL-F-Pop	7.05 ± 1.15	–	6.98 ± 1.18	84.14 ± 32.87

Table 11: Offline-policy results for CQL under the primary grouping, evaluated on the same 10 held-out test patients and with the same metrics as Table 3.

Alternative patient grouping. Table 12 repeats the offline evaluation under the alternative grouping of Appendix D.2, in which each group contains one training and one test patient. GCQL again shows the clearest individual-over-population advantage, winning on 93.3% of matched patients, whereas for DQN the individual policy attains a slightly lower mean score but wins on fewer than half of the patients.

Policy	SOFA(all) (↓)	Lower SOFA vs. Pop (%)	SOFA(96h) (↓)	SafeTime% (↑)
TRPOLag-F-Ind (behavior)	7.53 ± 1.38	93.3	6.85 ± 1.56	92.13 ± 20.99
DQN-F-Ind	9.75 ± 1.54	46.7	9.72 ± 1.82	57.63 ± 38.95
DQN-F-Clu	9.01 ± 1.98	66.7	9.10 ± 2.27	81.61 ± 31.05
DQN-F-Pop	9.80 ± 1.09	–	9.77 ± 1.13	69.22 ± 36.92
CQL-F-Ind	8.78 ± 1.86	66.7	8.65 ± 1.98	86.60 ± 26.70
CQL-F-Clu	8.84 ± 1.52	53.3	8.82 ± 1.60	80.37 ± 36.15
CQL-F-Pop	9.18 ± 1.25	–	9.08 ± 1.06	80.02 ± 32.80
GCQL-F-Ind	$\mathbf{7.98 \pm 1.57}$	**93.3**	$\mathbf{7.73 \pm 1.74}$	$\mathbf{91.21 \pm 19.38}$
GCQL-F-Clu	8.36 ± 1.55	66.7	8.37 ± 1.61	80.71 ± 33.13
GCQL-F-Pop	8.80 ± 1.46	–	8.87 ± 1.56	75.38 ± 39.61

Table 12: Offline-policy results for DQN, CQL, and GCQL under the alternative grouping of Appendix D.2, evaluated on the patient-specific simulators of the corresponding 15 held-out test patients. Metrics and abbreviations follow Table 2; dashes indicate that the population policy is not compared with itself.

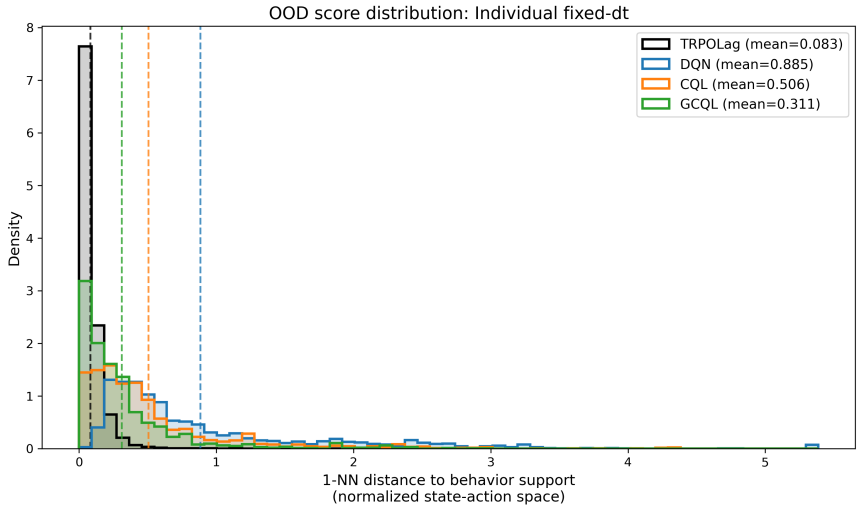


Figure 10: OOD support-distance distribution for fixed-dt individual offline policies. For each evaluated state-action pair, we compute its 1-nearest-neighbor distance to the TRPOLag behavior dataset in normalized state-action space. Smaller values indicate closer adherence to the behavior support. GCQL stays closer to the behavior support than DQN and CQL, while DQN exhibits the largest OOD deviation.

To further examine why GCQL yields a clearer personalization pattern, we quantify the degree to which each evaluated offline policy departs from the behavior support. For each patient i, let $\mathcal{D}^{(i)}_{\text{LagrangianTRPO}}$ denote the offline dataset generated by the TRPOLag behavior policy. We split it into a reference subset $\mathcal{D}^{(i)}_{\text{ref}}$ and a held-out query subset $\mathcal{D}^{(i)}_{\text{qry}}$. For a state-action pair $(\mathbf{x}, \mathbf{u})$ generated during evaluation, we define its nearest-neighbor support distance as

$$d^{(i)}_{\text{OOD}}(\mathbf{x}, \mathbf{u}) = \min_{(\mathbf{x}', \mathbf{u}') \in \mathcal{D}^{(i)}_{\text{ref}}} \left\| \begin{bmatrix} \bar{\mathbf{x}} \\ \bar{\mathbf{u}} \end{bmatrix} - \begin{bmatrix} \bar{\mathbf{x}}' \\ \bar{\mathbf{u}}' \end{bmatrix} \right\|_2, \tag{48}$$

where $(\bar{\mathbf{x}}, \bar{\mathbf{u}})$ denotes the coordinate-wise normalized state-action vector, with normalization statistics estimated from the offline behavior dataset. For an evaluation dataset $\mathcal{D}^{(i)}_{\pi}$ generated by policy π, we plot the empirical distribution of $d^{(i)}_{\text{OOD}}(\mathbf{x}, \mathbf{u})$ over all state-action pairs $(\mathbf{x}, \mathbf{u}) \in \mathcal{D}^{(i)}_{\pi}$ and all patients. For the TRPOLag curve, we compute the same distance using the held-out query subset $\mathcal{D}^{(i)}_{\text{qry}}$ rather than the reference subset itself, which avoids the trivial zero-distance comparison. Smaller values indicate that the evaluated policy remains closer to the support of the behavior data.

Figure 10 relates the preceding offline evaluation results to how far each method departs from the behavior-data support. DQN has the largest mean support distance and a long right tail, indicating that it frequently visits state-action regions poorly supported by the behavior dataset. CQL reduces this

deviation relative to DQN, but it still remains farther from the behavior support than GCQL. GCQL has the smallest OOD distance among the offline methods, suggesting that its guarded conservative objective better restricts the learned policy to supported state-action regions. This helps explain why GCQL is more stable in the personalization evaluation: by reducing OOD state-action visitation, GCQL avoids the unreliable value extrapolation that can occur when offline policies move outside the behavior-data support. Consequently, GCQL does not surpass the online TRPOLag behavior policy, but it narrows the offline performance gap relative to DQN and CQL and recovers a clearer individual-over-population advantage.

E GUIDANCE FOR RESEARCHERS

MedGym is designed first and foremost as a benchmark for dynamic treatment recommendation under continuous-time patient evolution, irregular intervention timing, and patient heterogeneity. Accordingly, methods that claim relevance to medical treatment RL should be evaluated on *MedGym* before drawing conclusions about their suitability for longitudinal healthcare settings. For fair comparison, researchers should begin with the default benchmark configuration and report results under the standard evaluation protocol, including treatment effectiveness, safety, and personalization-related metrics. The discount factor is configurable: setting $\gamma = 1$ recovers the finite-horizon continuous-time objective under the TACOS reformulation, whereas $\gamma < 1$ can be used to study discounted treatment objectives.

For offline RL, policies should be trained on datasets generated from the *MedGym* environment or on the provided benchmark datasets, and then evaluated through online rollouts in the same environment. For online RL, methods should be trained directly in *MedGym* until convergence and then assessed under the standard test settings. When applicable, researchers should compare both fixed-interval and adaptive-time variants, and should evaluate both population-level and individualized policies. In particular, methods that outperform the standard-of-care baseline or the population-level policy while maintaining strong safety performance provide evidence that they can exploit patient-specific and continuous-time structure rather than only fit average-case treatment patterns.

Although *MedGym* is configurable, we caution against interpreting it as a disease-specific simulator for direct sim-to-real claims. The configurable parameters in *MedGym* primarily define a family of medically grounded benchmark environments rather than an exact digital twin of a particular disease. Therefore, simply matching one dataset or one clinical cohort does not guarantee that conclusions drawn in *MedGym* will transfer directly to real deployment. Researchers interested in a specific disease should treat *MedGym* as a structured stress-test environment and, when necessary, complement it with additional domain-specific validation.

A more appropriate use of *MedGym* in applied medical RL is to tune the benchmark configuration so that it reflects the qualitative properties of the target treatment scenario. Examples include the degree of patient heterogeneity, the irregularity of measurement and intervention times, the typical treatment horizon, the sparsity of observations, and the severity of safety constraints. Such choices allow researchers to ask practically meaningful questions, such as whether a method remains effective when interaction opportunities are limited, whether adaptive timing is beneficial, or whether individualized policies outperform population-level ones under realistic patient variation.

Finally, we recommend that results in *MedGym* always be interpreted from multiple perspectives rather than from reward alone. At a minimum, researchers should report treatment effectiveness, trajectory-level safety, and the relative performance of individualized versus population-level policies. Whenever possible, evaluations should also include multiple patient models or environment instantiations in order to assess robustness and generalizability. In this way, *MedGym* can serve as a standardized yet flexible benchmark for identifying which RL methods are most promising for future dynamic treatment applications, while avoiding overly strong conclusions about direct clinical deployment.

F BROADER IMPACT

MedGym is developed to support more realistic and informative evaluation of reinforcement learning methods for dynamic medical treatment. A potential positive impact of this work is that it provides a benchmark in which researchers can study clinically important properties that are often simplified or

ignored in existing settings, including continuous-time disease progression, irregular intervention timing, patient heterogeneity, and trajectory-level safety. By enabling more faithful pre-deployment evaluation, *MedGym* may help the community identify which classes of methods are better suited to safety-critical treatment recommendation and may reduce over-optimistic conclusions drawn from overly simplified benchmarks. More broadly, the benchmark may encourage the development of learning algorithms that better reflect the temporal and personalized nature of real clinical decision-making.

The benchmark may also have practical value for comparing offline and online policy learning methods under a common simulation framework. In healthcare, direct online experimentation is often costly, ethically constrained, or infeasible. A configurable benchmark such as *MedGym* can therefore provide a useful intermediate step for stress-testing algorithms before any consideration of real-world translation. In addition, the emphasis on individualized dynamics may motivate future work on patient-specific modeling and treatment design, which could ultimately contribute to more personalized and data-efficient decision support tools.

At the same time, this work has several potential negative societal implications if misused or over-interpreted. First, *MedGym* is still a learned simulator rather than a validated clinical deployment system. If benchmark results were mistakenly interpreted as direct evidence of clinical effectiveness, this could encourage premature or unsafe real-world use of RL-based treatment policies. Second, because the benchmark is derived from historical clinical data, any biases, omissions, or cohort-specific patterns present in the source data may be inherited by the learned environment. As a result, methods that perform well in *MedGym* may still behave unevenly across patient populations that are underrepresented or insufficiently modeled in the data. Third, like other benchmark environments, *MedGym* may incentivize optimization toward benchmark-specific metrics at the expense of broader clinical validity, interpretability, or robustness.

There is also a risk that personalized modeling could be misunderstood as fully solving the personalization problem in medicine. In the current version of *MedGym*, patient-specificity is operationalized through individualized models rather than through a clinically validated latent representation of patient factors. Accordingly, good benchmark performance should not be interpreted as evidence that a method has discovered causal patient structure or that it will generalize safely across hospitals, cohorts, or diseases. In addition, although the benchmark includes safety-oriented evaluation, it does not eliminate the broader ethical challenges associated with deploying automated decision systems in healthcare, including accountability, transparency, fairness, and clinician oversight.

To mitigate these risks, *MedGym* should be used as a research benchmark rather than as a stand-alone basis for clinical decision-making. Results obtained in this environment should be interpreted together with external validation, domain expertise, and careful analysis of failure modes, subgroup behavior, and safety trade-offs. We hope that, when used appropriately, *MedGym* will have a net positive societal impact by improving the rigor with which medical RL methods are evaluated before any future translational use.