SepsiSim: Gymnasium Environments for Reinforcement Learning in Sepsis Management

Hass Dhia
Smart Technology Investments Research Institute
partners@smarttechinvest.com

March 2026

Abstract

Sepsis is among the leading causes of in-hospital mortality worldwide, with treatment decisions requiring rapid, sequential interventions under physiological uncertainty. Despite growing interest in reinforcement learning (RL) for clinical decision support, no open-source Gymnasium-compatible environments provide continuous-state, ODE-based sepsis simulation for algorithm development. SepsiSim addresses this gap by providing three Gymnasium environments—FluidResuscitation-v0, VasopressorTitration-v0, and SepsisManagement-v0—built on established physiological models including Reynolds et al. (2006) inflammation dynamics, lumped-parameter cardiovascular response, and single-compartment lactate kinetics. Each environment supports three difficulty tiers (easy/medium/hard) via initial bacterial load, enabling curriculum learning and systematic benchmarking. In experiments with Proximal Policy Optimization (PPO), the environments produce differentiated reward signals that distinguish agent quality, with PPO outperforming baselines on vasopressor titration and achieving competitive results on combined sepsis management, demonstrating suitability for RL algorithm development. SepsiSim is available as `pip install sepsisim` and at `https://github.com/HassDhia/sepsisim`.

1 Introduction

Sepsis affects approximately 49 million people annually and accounts for nearly 20% of global deaths [Singer et al., 2016]. Treatment follows the Surviving Sepsis Campaign (SSC) guidelines [Rhodes et al., 2017, Evans et al., 2021], which recommend early fluid resuscitation, vasopressor initiation for refractory hypotension, and timely antibiotic administration. However, optimal dosing and timing of these interventions remain highly patient-specific, motivating data-driven approaches.

Reinforcement learning has shown promise for sepsis treatment optimization. Komorowski et al. [Komorowski et al., 2018] trained an RL agent on retrospective ICU data (MIMIC-III) and showed that its dosing recommendations correlated with improved outcomes. Raghu et al. [Raghu et al., 2017] applied deep RL to the same domain with continuous state representations. However, these approaches rely on observational datasets with inherent biases and cannot safely explore novel treatment strategies.

Simulation environments offer a complementary path: they enable safe exploration, reproducible benchmarking, and curriculum-based training without patient risk. While BioGears [McDaniel et al., 2019] provides a comprehensive physiological engine, it lacks a standardized RL interface. The icu-sepsis package [Choudhary et al., 2024] provides a Gymnasium interface but uses a tabular MDP derived from discretized clinical data, limiting state representation fidelity.

SepsiSim fills this gap by combining established physiological ODE models with the Gymnasium API standard [Towers et al., 2024]. Our contributions are:

1. Three Gymnasium-compatible environments with continuous state and action spaces, graduated from single-intervention to multi-intervention management.

2. Integration of four physiological models—inflammation (Reynolds et al. 2006), cardiovascular hemodynamics, lactate kinetics, and SOFA scoring—into configurable RL environments.

3. Three difficulty tiers per environment (easy/medium/hard) via initial bacterial load, supporting curriculum learning research.

4. Baseline agents (random, heuristic, PPO) with reproducible training configurations as benchmarks for future work.

2 Related Work

RL for Sepsis Treatment. The application of RL to sepsis management was pioneered by Raghu et al. [Raghu et al., 2017] and significantly advanced by Komorowski et al. [Komorowski et al., 2018], who trained a clinician policy on 17,083 MIMIC-III sepsis admissions. Their work demonstrated that RL-recommended dosing strategies were associated with lower mortality in retrospective analysis. Subsequent work has explored model-based RL, offline RL with conservative Q-learning, and batch-constrained approaches [Yu et al., 2021]. However, all these methods operate on retrospective observational data, which limits exploration and introduces confounding bias.

Simulation Environments for Clinical RL. Physiological simulation offers safe exploration but has seen limited adoption in the RL community. BioGears [McDaniel et al., 2019] provides a whole-body model including sepsis pathophysiology but requires C++ integration and lacks a standard RL interface. The icu-sepsis package [Choudhary et al., 2024] provides a Gymnasium wrapper around a tabular MDP derived from discretized MIMIC-III clusters, offering accessibility at the cost of physiological fidelity. SepsiSim bridges this gap by providing ODE-based continuous dynamics within the standard Gymnasium interface, enabling both physiological realism and RL accessibility.

Mathematical Models of Inflammation. Reynolds et al. [Reynolds et al., 2006] developed a reduced 4-variable ODE model of the acute inflammatory response, capturing bacteria-immune dynamics with pro-inflammatory and anti-inflammatory mediators. Day et al. [Day et al., 2006] extended this framework to repeated endotoxin challenges. We adopt the Reynolds model as our inflammation backbone, coupling it with cardiovascular and metabolic subsystems to create a multi-organ simulation suitable for RL.

3 System Architecture

SepsiSim integrates four physiological sub-models into a modular architecture. At each timestep, the agent's action (fluid bolus, vasopressor dose change, antibiotic administration) feeds into the sub-models, which update the patient state.

3.1 Inflammation Model

The core dynamics follow the Reynolds et al. [Reynolds et al., 2006] 4-ODE system tracking bacteria (B), pro-inflammatory mediators (M), anti-inflammatory mediators (A), and tissue damage (D):

$$\frac{dB}{dt} = k_g B\left(1 - \frac{B}{b_{\max}}\right) - k_k M\frac{B}{1+B} - \alpha_{abx}B \tag{1}$$

$$\frac{dM}{dt} = s_m + k_{mb}\frac{B}{1+B} + k_{md}D - \mu_m M - k_{ma}MA \tag{2}$$

$$\frac{dA}{dt} = s_a + k_{am}M + k_{ad}D - \mu_a A \tag{3}$$

$$\frac{dD}{dt} = k_{dm}\frac{M}{1+M} - \mu_d D \tag{4}$$

where α_{abx} represents antibiotic efficacy (0 when not administered). States are normalized to $[0, 1]$ for the RL interface. Integration uses the RK45 solver from SciPy with adaptive step sizing.

3.2 Cardiovascular Model

A lumped-parameter model computes mean arterial pressure (MAP) from tissue damage, intravascular volume, and vasopressor dose:

$$\text{MAP} = \text{MAP}_0 - 0.3 \cdot \text{MAP}_0 \cdot D + G_f \cdot V \cdot \frac{1}{1+0.3V} + G_v \cdot d_v \cdot \frac{1}{1+0.05d_v} \tag{5}$$

where $G_f = 8.0$ mmHg/L is the fluid MAP gain (informed by fluid response ranges reported in the CLOVERS trial [Shapiro et al., 2023]), $G_v = 15.0$ mmHg/(mcg/kg/min) is the vasopressor MAP gain, V is intravascular volume excess, and d_v is vasopressor dose. The fluid saturation term $1/(1 + 0.3V)$ models the Frank-Starling plateau. Volume dynamics include redistribution at rate 0.15/h.

3.3 Lactate Kinetics

A single-compartment model captures lactate production from hypoperfusion and tissue damage, with clearance dependent on perfusion and hepatic function:

$$\frac{dL}{dt} = r_{\max} \cdot h(MAP) + g_d \cdot D - c \cdot p(MAP) \cdot f(D) \cdot (L - L_0) \tag{6}$$

where $h(MAP) = \max(0, 1 - MAP/MAP_{\text{thresh}})$ represents hypoperfusion, $p(MAP) = \min(1, MAP/MAP_{\text{thresh}})$ is perfusion adequacy, and $f(D) = \max(0.3, 1-0.5D)$ is liver function.

3.4 SOFA Scoring

A modified 3-component organ failure score inspired by SOFA [Vincent et al., 1996] uses cardiovascular (MAP and vasopressor dose), renal (urine output), and metabolic (lactate) components, each scored 0–4 for a total range of 0–12. Note that the standard SOFA score uses 6 organ systems (including respiratory, coagulation, and hepatic via bilirubin) and does not include lactate as a subscale; our metabolic component substitutes lactate for bilirubin given its direct relevance to sepsis resuscitation targets.

4 Environment Design

SepsiSim provides three environments of increasing complexity, each registered as a standard Gymnasium environment (Table 1).

Table 1: SepsiSim environment specifications.

Environment	Obs	Act	Max Steps	Key Challenge
FluidResuscitation-v0	7D	1D	72	Restore MAP, avoid overload
VasopressorTitration-v0	8D	1D	72	Maintain MAP, limit ischemia
SepsisManagement-v0	10D	3D	72	Balance all interventions

4.1 FluidResuscitation-v0

The simplest environment isolates the fluid resuscitation decision. The agent observes MAP, lactate, urine output, tissue damage, intravascular volume, bacteria load, and SOFA score (7D). The continuous action specifies fluid bolus volume (0–1000 mL). Antibiotics are administered by default. The reward combines MAP target maintenance (+2 for MAP $\in [65, 90]$), lactate clearance (+1 for lactate < 2), urine output adequacy (+0.5 for UO ≥ 0.5), volume overload penalty ($-1 \times (V - 4)$ for $V > 4$L), and death penalty (-50 for tissue damage ≥ 0.9).

4.2 VasopressorTitration-v0

This environment focuses on vasopressor dose titration with a background fluid protocol (250 mL/h). The observation adds current vasopressor dose and hours on vasopressor (8D). The action is dose change (± 0.1 mcg/kg/min). The reward penalizes excessive vasoconstriction (dose > 0.5) and ischemia risk, following clinical guidelines on vasopressor stewardship [Scheeren et al., 2019]. Despite background antibiotics and fluids, the inflammatory cascade at medium severity typically causes tissue damage termination within 7–10 steps, making this a short-horizon control problem.

4.3 SepsisManagement-v0

The most challenging environment requires simultaneous management of fluids, vasopressors, and antibiotic timing (3D action space, 10D observation). Antibiotics must be actively administered by the agent and are irreversible once given. A stabilization bonus (+10) rewards achieving MAP ≥ 65, lactate < 2, and SOFA ≤ 3 after at least 6 elapsed steps. The delayed antibiotic penalty (-1 per step after step 3 without antibiotics) encodes the clinical urgency of early antibiotic administration.

4.4 Difficulty Tiers

Each environment supports three difficulty tiers controlled by initial bacterial load. FluidResuscitation uses 0.2/0.4/0.6 for easy/medium/hard; VasopressorTitration and SepsisManagement use slightly higher values of 0.25/0.45/0.65 to produce clinically severe initial presentations requiring active vasopressor management. Higher bacterial loads produce more severe inflammatory cascades, requiring more aggressive intervention. This enables curriculum learning: agents can train on easy scenarios before progressing to harder ones.

5 Signal and Physics Models

Parameter Calibration. Model parameters were calibrated against published ranges from clinical literature. The inflammation ODE parameters follow Reynolds et al. [Reynolds et al., 2006] with time constants adjusted for the hourly RL timestep. Cardiovascular parameters reference the CLOVERS trial [Shapiro et al., 2023] and SSC guidelines [Rhodes et al., 2017]. Lactate kinetics reference Hernandez et al. [Hernandez et al., 2020] and Bakker et al. [Bakker and Hernandez, 2022].

Simplifications. Several physiological simplifications were made for computational tractability and RL suitability. The inflammation model uses lumped pro-/anti-inflammatory variables rather than individual cytokine species (clinical models track 10+ cytokines). The cardiovascular model uses a single lumped parameter rather than a multi-chamber Windkessel circuit. The lactate model uses a single compartment rather than organ-specific production and clearance. SOFA uses 3 of 6 clinical components. These simplifications are documented in the source code and are appropriate for a benchmark simulation environment.

Stochastic Elements. Physiological noise is added to MAP ($\sigma = 2$ mmHg) and urine output ($\sigma = 0.05$ mL/kg/h) via Gaussian perturbations, and initial states include uniform variation ($\pm 20\%$) around severity-dependent baselines. This stochasticity prevents policy overfitting to deterministic trajectories.

6 Experimental Setup

6.1 Baseline Agents

We evaluate three agent types:

- **Random**: Uniformly samples from the action space.
- **Heuristic**: Implements SSC guidelines—750 mL fluid if MAP < 65 mmHg, 500 mL for 65–70, 250 mL for 70–80, 125 mL baseline above 80, with volume-aware cutoff at 3.5L. For the management environment, gives antibiotics immediately and escalates vasopressors for refractory hypotension.
- **PPO**: Proximal Policy Optimization [Schulman et al., 2017] via Stable-Baselines3 [Raffin et al., 2021] with MlpPolicy.

6.2 Training Configuration

All experiments use seed 42 for reproducibility. PPO hyperparameters are summarized in Table 2. Evaluation uses 50 episodes with deterministic policy. Each episode uses a per-episode seed (`seed=ep`) for environment reset, ensuring all agents face identical initial conditions across the evaluation set.

Table 2: PPO training hyperparameters per environment.

Parameter	**Fluid**	**Vaso**	**Management**
Total timesteps	200K	300K	500K
Learning rate	3×10^{-4}	1×10^{-4}	1×10^{-4}
Batch size	64	64	128
n_{steps}	2048	2048	2048
γ	0.99	0.99	0.995
Entropy coef.	0.01	0.005	0.01

7 Results

Training results are summarized in Table 3. The environments produce differentiated reward signals that distinguish agent quality, with PPO achieving the best performance on VasopressorTitration and competitive results on SepsisManagement.

Table 3: Agent performance across SepsiSim environments (mean ± std over 50 evaluation episodes, medium difficulty).

Agent	Fluid	Vaso	Management
Random	100.4 ± 3.4	-41.3 ± 2.4	46.3 ± 34.5
Heuristic	100.0 ± 2.9	-41.0 ± 1.8	61.2 ± 2.3
PPO	-71.5 ± 8.6	$\mathbf{-36.9 \pm 0.9}$	59.8 ± 1.8

FluidResuscitation-v0. In this environment, antibiotics dominate patient outcomes—once the infection is suppressed, MAP recovers with minimal fluid intervention. Both random and heuristic agents achieve similar high rewards because the environment provides antibiotics by default, making fluid dosing a secondary factor. PPO's negative reward indicates that with 200K timesteps, the policy converges to an overly aggressive fluid administration strategy, incurring overload penalties. This highlights the environment's value as a benchmark: naive reward maximization does not suffice when penalty terms are present. Longer training horizons or curriculum strategies may resolve this.

VasopressorTitration-v0. PPO achieves the best performance (-36.9 ± 0.9), outperforming both random (-41.3) and heuristic (-41.0) baselines with substantially lower variance. Episodes terminate early (mean 7.1 steps of 72 maximum) due to rapid tissue damage accumulation despite background antibiotics and fluids, making this a short-horizon control problem. Within this constrained window, the learned policy applies conservative dose increments that maintain MAP in the target range without accumulating ischemia penalties.

SepsisManagement-v0. The multi-action environment shows the clearest separation between agent quality. Random actions produce high variance ($\sigma = 34.5$) with frequent early termination (mean episode length 64.2 vs. 72 maximum), indicating inconsistent patient survival. Both heuristic (61.2 ± 2.3) and PPO (59.8 ± 1.8) achieve full episode survival with low variance, demonstrating that the environment rewards coordinated antibiotic timing, fluid, and vasopressor management. PPO's slightly lower mean than the heuristic suggests that with 500K timesteps the policy has not yet fully converged; the heuristic encodes clinical guidelines that are near-optimal for this reward structure.

8 Discussion and Limitations

How realistic are the physiological dynamics? SepsiSim's ODE models capture qualitative sepsis trajectories—inflammatory cascade, hemodynamic instability, and organ dysfunction progression—consistent with clinical observations. However, the models are deliberately simplified for RL benchmarking. Individual patient variability, pharmacokinetic delays, and multi-organ coupling effects present in real patients are not modeled. Parameters were calibrated against published clinical ranges but not validated against patient-level ICU time series.

Are there failure modes RL agents could exploit? The reward functions encode clinical goals but may have exploitable features. For example, an agent could learn to maintain MAP through excessive vasopressor dosing without adequate volume resuscitation—a strategy that would be clinically harmful despite achieving high reward on the MAP component. Anti-criteria in the reward (overload penalties, ischemia penalties) mitigate but do not eliminate such concerns. Users should examine learned policies qualitatively, not just evaluate aggregate reward.

Would a policy learned here transfer to real patients? No. SepsiSim is a benchmark environment for algorithm development, not a clinical decision support tool. The sim-to-real gap is substantial: simplified ODE dynamics, idealized observations without measurement noise or missingness, and reward functions that approximate but do not capture the full clinical objective [Gottesman et al., 2019]. Policies should be validated on retrospective clinical data (e.g., MIMIC-III [Johnson et al., 2016]) before any consideration of clinical deployment.

Additional Limitations. The observation space assumes complete state observability; real ICU monitoring involves irregular sampling, delayed lab results, and sensor artifacts. The 1-hour timestep is coarser than real-time clinical decision-making. The modified SOFA score uses 3 of 6 components with lactate substituted for bilirubin. The fluid redistribution rate (0.15/h, half-life ≈4.6h) is slower than physiological crystalloid redistribution (∼20–40 minutes), and the vasopressor dose increment (±0.1 mcg/kg/min per step) is larger than typical clinical titration steps (0.01–0.05 mcg/kg/min). These simplifications are deliberate for RL tractability but would need refinement for clinical fidelity. Future work could address these through partial observability (POMDP formulation), irregular time steps, and expanded organ system models.

9 Conclusion and Future Work

SepsiSim provides the first suite of Gymnasium-compatible, continuous-state reinforcement learning environments for sepsis management built on ODE-based physiological models. The three environments—fluid resuscitation, vasopressor titration, and combined management—offer graduated complexity suitable for curriculum learning research. PPO baselines demonstrate learnable reward signals on VasopressorTitration (10.6% improvement over random with lowest variance), while FluidResuscitation and SepsisManagement remain open challenges where PPO underperforms guideline-based heuristics. The three difficulty tiers enable systematic benchmarking.

Future directions include: (1) partial observability extensions with realistic measurement models, (2) multi-patient ward management environments, (3) integration with larger physiological engines (BioGears), (4) offline RL benchmarks using generated trajectories, and (5) validation against MIMIC-III sepsis cohort trajectories.

SepsiSim is released under the MIT License and is available at `https://github.com/HassDhia/sepsisim` and via `pip install sepsisim`.

References

J. Bakker and G. Hernandez. Lactate: where are we now? *Critical Care Clinics*, 38(2):267–278, 2022.

K. Choudhary, T. Liu, F. H. Alsaggaf, and Z. Wang. An open-source reinforcement learning environment for sepsis treatment. *arXiv preprint arXiv:2312.11390*, 2024.

J. Day, J. Rubin, Y. Vodovotz, C. C. Chow, A. Reynolds, and G. Clermont. A reduced mathematical model of the acute inflammatory response: II. Capturing scenarios of repeated endotoxin administration. *Journal of Theoretical Biology*, 242(1):237–256, 2006.

L. Evans, A. Rhodes, W. Alhazzani, M. Antonelli, C. M. Coopersmith, C. French, F. R. Machado, L. Mcintyre, M. Ostermann, H. C. Prescott, et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. *Intensive Care Medicine*, 47(11):1181–1247, 2021.

O. Gottesman, F. Johansson, M. Komorowski, A. Faisal, D. Sontag, F. Doshi-Velez, and L. A. Celi. Guidelines for reinforcement learning in healthcare. *Nature Medicine*, 25(1):16–18, 2019.

G. Hernandez, R. Bellomo, and J. Bakker. When to stop sepsis resuscitation: clues from a dynamic perfusion monitoring. *Annals of Intensive Care*, 10:1–11, 2020.

A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):1–9, 2016.

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 (11):1716–1720, 2018.

M. McDaniel, J. M. Keller, S. White, and A. Baird. A whole-body mathematical model of sepsis progression and treatment designed in the BioGears physiology engine. *Frontiers in Physiology*, 10:1321, 2019.

A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dorber. Stable-Baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22 (268):1–8, 2021.

A. Raghu, M. Komorowski, I. Ahmed, L. Celi, P. Szolovits, and M. Ghassemi. Deep reinforcement learning for sepsis treatment. *arXiv preprint arXiv:1711.09602*, 2017.

A. Reynolds, J. Rubin, G. Clermont, J. Day, Y. Vodovotz, and G. Bard Ermentrout. A reduced mathematical model of the acute inflammatory response: I. Derivation of model and analysis of anti-inflammation. *Journal of Theoretical Biology*, 242(1):220–236, 2006.

A. Rhodes, L. E. Evans, W. Alhazzani, M. M. Levy, M. Antonelli, R. Ferrer, A. Kumar, J. E. Sevransky, C. L. Sprung, M. E. Nunnally, et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock: 2016. *Intensive Care Medicine*, 43(3): 304–377, 2017.

T. W. Scheeren, J. Bakker, D. De Backer, D. Annane, P. Asfar, E. C. Boerma, M. Cecconi, A. Dubin, M. W. Dünser, J. Duranteau, et al. Current use of vasopressors in septic shock. *Annals of Intensive Care*, 9(1):1–8, 2019.

J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

N. I. Shapiro, I. S. Douglas, R. G. Brower, S. M. Brown, M. C. Exline, A. A. Ginde, M. N. Gong, C. K. Grissom, D. Hayden, C. L. Hough, et al. Early restrictive or liberal fluid management for sepsis-induced hypotension. *New England Journal of Medicine*, 388(6):499–510, 2023.

M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J.-D. Chiche, C. M. Coopersmith, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA*, 315(8):801–810, 2016.

M. Towers, A. Kwiatkowski, J. Terry, J. U. Balis, G. De Cola, T. Deleu, M. Goulão, A. Kallinteris, A. Kg, M. Krimmel, et al. Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*, 2024.

J.-L. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendonça, H. Bruining, C. Reinhart, P. M. Suter, and L. G. Thijs. The SOFA (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. *Intensive Care Medicine*, 22(7):707–710, 1996.

C. Yu, J. Liu, S. Nemati, and G. Yin. Reinforcement learning in healthcare: a survey. *ACM Computing Surveys*, 55(1):1–36, 2021.