Safe Treatment Policy Optimization for Sepsis Patients via Constraint Damping Policy Optimization(CDPO)

Anonymous Author(s)
Affiliation
Address
email

Abstract

This study proposes Constraint Damping Policy Optimization (CDPO) to balance survival-oriented decision making and clinical safety in ICU sepsis treatment. Conventional deep reinforcement learning approaches such as Proximal Policy Optimization (PPO) primarily maximize cumulative reward and can behave in a safety-blind manner under noisy clinical trajectories, resulting in critical violations such as hypotension (MAP instability) or abrupt medication changes. CDPO embeds five clinically grounded safety constraints directly into the policy optimization process: MAP maintenance ($\geq$ 65 mmHg), treatment continuity (discouraging abrupt action switches), cumulative toxicity control, fluid restriction when tachycardia is present, and minimum prescription persistence (enforcing a minimum dwell time for selected treatments). To improve training stability, CDPO introduces a damping mechanism that adaptively scales the policy update magnitude according to the observed constraint violation rate, promoting conservative updates when violations increase and allowing larger updates when policies remain within safety limits. Experiments on an adult sepsis cohort from the MIMIC-III database show that, compared with clinician policies, CDPO reduces hypotension-related violations by 22.9% (from 170 to 131 cases) while decreasing vasopressor usage by 67.3%. In addition, CDPO reduces fluid administration by 23.2%, mitigating patient burden while maintaining safety-oriented behavior. Overall, these results indicate that constraint-aware reinforcement learning can derive effective treatment policies that adhere to predefined clinical safety specifications, providing a technical basis for reliable AI-assisted decision support in intensive care settings.

1 Introduction

1.1 Background and motivation

According to the Surviving Sepsis Campaign guidelines, effective sepsis management requires early antibiotic administration, fluid resuscitation, and vasopressor use to maintain mean arterial pressure (MAP) at or above 65 mmHg [1,2]. The Sepsis-3 definition further characterizes sepsis as life-threatening organ dysfunction caused by a dysregulated host response to infection, emphasizing that delayed or inappropriate treatment directly worsens clinical outcomes [2].

In practice, however, substantial inter-patient physiological heterogeneity limits the effectiveness of static treatment protocols, making it difficult to adapt medication intensity to dynamically evolving patient states [2]. To address this limitation, reinforcement learning approaches have been proposed to model ICU treatment as a Markov decision process (MDP) using high-resolution clinical time-series data [3,4]. Nevertheless, existing models primarily optimize cumulative rewards and often fail to adequately account for critical clinical safety constraints, such as abrupt dosage changes or cumulative drug toxicity. To overcome these limitations, this study proposes **Constraint Damping Policy Optimization (CDPO)**, which adaptively regulates policy update strength based on observed constraint violation rates.

1.2 Objective of the study

The objective of this study is to develop a reinforcement learning framework for optimizing fluid and vasopressor dosing policies for sepsis patients in intensive care units (ICUs), while strictly adhering to clinical safety constraints and improving treatment efficiency. Patient states are formulated as a Markov decision process (MDP), enabling the learning of policies that determine dosing intensity over time. To address the limitations of penalty-based reward designs, we propose **Constraint Damping Policy Optimization (CDPO)**, which directly integrates safety constraints into the optimization process. The proposed approach is quantitatively evaluated against clinician treatment data to demonstrate its clinical safety and alignment.

1.3 AI Utilization and research ethics (AI Co-Scientist Integration)

In this study, artificial intelligence was employed as a collaborative co-scientist supporting literature synthesis, algorithmic design exploration, experimental analysis, and logical consistency verification. Iterative AI-assisted critiques of preliminary reinforcement learning formulations identified failure modes associated with reward hacking and unstable policy updates under sparse safety violations, motivating the introduction of a constraint violation–driven damping mechanism later formalized as CDPO.

All AI-generated outputs were rigorously evaluated within a human-in-the-loop framework through reproducible experiments and independent verification. The models, prompts, and scope of AI involvement are transparently documented in the appendix, and full responsibility for scientific interpretation and conclusions remains with the authors.

2 Related work

2.1 Reinforcement learning for sepsis treatment optimization

Reinforcement learning (RL) has been explored for high-risk medical decision-making tasks such as sepsis treatment, as it enables learning sequential dosing policies from time-varying physiological data. Notably, AI Clinician [3] formulated sepsis management as a Markov decision process using the MIMIC-III ICU dataset and demonstrated that RL-based policies can approximate clinician dosing patterns or achieve improved estimated survival. However, most existing approaches emphasize long-term outcome optimization and inadequately address short-term clinical risks, such as abrupt mean arterial pressure (MAP) drops or medication misuse [4]. In particular, on-policy methods like Proximal Policy Optimization (PPO), which rely on cumulative reward maximization, lack explicit mechanisms to enforce clinical safety thresholds outside the reward function, potentially resulting in overly aggressive dosing behavior

2.2 Limitations of constraint control in reinforcement learning

Safety-oriented reinforcement learning has primarily relied on reward shaping with penalty terms [5], which requires careful tuning and is susceptible to reward hacking, posing serious risks in medical applications. Constrained reinforcement learning methods based on Lagrangian formulations have been proposed to explicitly handle safety constraints [6,7], but they often suffer from training instability or excessive conservatism under noisy and distribution-shifted clinical data, leading to reduced treatment efficiency. These limitations motivate the need for adaptive policy optimization strategies that can enforce clinical safety constraints while remaining robust to real-world data noise.

3 Methods

3.1 AI-Assisted research workflow

This study employed large language models (LLMs) and tool-based agents as auxiliary resources to enhance research productivity throughout the optimization of safe sepsis dosing policies. All core scientific decisions—including problem formulation, metric and constraint definition, model design, result interpretation, and conclusion drawing—were made exclusively by the researchers. AI outputs were restricted to drafts, candidate suggestions, and automation assistance, ensuring that research ownership and final responsibility remained fully with the authors.

Specifically, AI was used to (1) rapidly structure related literature and propose candidate designs for state, action, and metric definitions; (2) iteratively generate preprocessing, training, evaluation, and visualization code to construct the experimental pipeline; and (3) automatically aggregate outcome metrics such as safety violation counts and average dosing levels, as well as generate patient-level trajectory comparisons. At each stage, researchers verified AI-generated outputs through column-level data checks, reproducible code execution, and independent recalculation of results, adopting, modifying, or discarding outputs to ensure that all final findings were determined by actual data and executed experiments.

3.2 Datasets

This study utilized the MIMIC-III intensive care unit dataset collected at Beth Israel Deaconess Medical Center [8]. An adult patient cohort was selected based on internationally accepted diagnostic criteria for sepsis.

Continuous clinical records were discretized into 4-hour time steps to facilitate reinforcement learning. Patient states at each step were represented as a normalized vector comprising mean arterial pressure (MAP), heart rate, respiratory rate, body temperature, blood lactate levels, and key laboratory measurements. Missing values were handled using a combination of forward filling and statistical imputation, while extreme values were clipped to clinically plausible ranges to mitigate the impact of data noise.

3.3 Problem formulation: safe sepsis treatment

We formulate sepsis treatment in the intensive care unit (ICU) as a sequential decision-making problem under explicit clinical safety constraints. Unlike conventional approaches that focus solely on survival maximization, effective sepsis management requires balancing treatment efficiency with strict patient safety. Accordingly, the problem is modeled as a constrained Markov decision process (MDP).

At each time step t, the patient state $s_t$consists of observable ICU time-series variables, including mean arterial pressure (MAP), heart rate, respiratory rate, body temperature, blood lactate level, and key laboratory measurements. Given s_t, the agent selects an action $a_t$representing the dosing intensity of intravenous fluids and vasopressors, while state transitions $P(s_{t+1} \mid s_t, a_t)$ capture physiological responses estimated from offline clinical data.

The policy $\pi_\theta(a_t \mid s_t)$is trained to minimize the expected cumulative cost:

$$\min_\theta \; J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta} \boxed{,} \left[\sum_{t=0}^{T-1} \gamma^t \left(l_{\text{clin}}(s_t, a_t) + \lambda \, l_{\text{dose}}(a_t)\right) \right],$$

where l_{clin}encodes survival-related clinical risks, l_{dose}penalizes excessive medication use, λcontrols the safety–efficiency trade-off, and $\gamma \in (0,1]$is the discount factor.

Because minimizing cumulative cost alone does not guarantee clinical safety, we explicitly impose five core safety constraints. These include maintaining MAP above 65 mmHg, restricting abrupt dosing changes, limiting cumulative vasopressor exposure, constraining fluid administration under tachycardia, and preventing prolonged inaction in severe sepsis states (Table 1).

Constraint	Description	Mathematical Form	Clinical Basis
C1	Hypotension Prevention	MAP ≥ 65 mmHg	Surviving Sepsis Campaign
C2	Dose Continuity	|Δaction| < δ	Hemodynamic stability
C3	Cumulative Toxicity	Σ(vasopressor) < max	Organ protection
C4	Tachycardia Restriction	HR > 120 limits fluid	Cardiac burden reduction
C5	Treatment Persistence	No prolonged inaction in severe sepsis	Prevent abandonment

Table 1: Clinical safety constraints used in CDPO

Each constraint is evaluated at every decision step, and any violation marks the state as unsafe. The proportion of decision steps with at least one violation defines the violation rate, which serves as a key indicator of clinical acceptability. This formulation establishes a constrained optimization framework that addresses the reward-centric limitations of prior reinforcement learning methods and provides the foundation for the proposed CDPO algorithm to balance safety and treatment efficiency

3.4 Constraint damping policy optimization (CDPO)

We propose **Constraint damping policy optimization (CDPO)** to address sepsis treatment decision-making problems in which strict clinical safety constraints are essential. Unlike conventional approaches that encode safety as additional reward terms or penalties, CDPO directly incorporates clinical risk signals into the policy update mechanism, enabling joint consideration of safety and learning efficiency.

The core idea of CDPO is to explicitly track clinical safety violations during policy learning and use them as a control signal to regulate the policy update magnitude. Based on the five predefined clinical safety constraints, we define the **constraint violation rate** C_{rate}as the proportion of decision steps in which at least one constraint is violated. This metric provides a quantitative measure of whether the current policy operates within clinically acceptable bounds.

Rather than converting violations into optimization penalties, CDPO uses C_{rate}as the input to a damping function $g(\cdot)$that adaptively scales the policy update step:

$$\theta_{k+1} = \theta_k + \alpha\, g(C_{\text{rate},k})\nabla_\theta J(\theta_k),$$

where $\theta_k$denotes the policy parameters at iteration k, $\alpha$is the learning rate, and $\nabla_\theta J(\theta_k)$is the policy gradient. This formulation preserves the update direction while modulating its magnitude according to clinical risk, thereby structurally preventing aggressive policy shifts that could compromise patient safety.

Consequently, CDPO adopts a damping-based optimization strategy that continuously adjusts learning intensity in response to risk signals, rather than enforcing hard constraints or halting learning. In contrast to prior constrained reinforcement learning methods that embed constraints as optimization terms, CDPO treats safety violations as regulators of learning dynamics. This design supports stable convergence under noisy clinical data and enables the learning of dosing policies that are both clinically aligned and treatment-efficient.

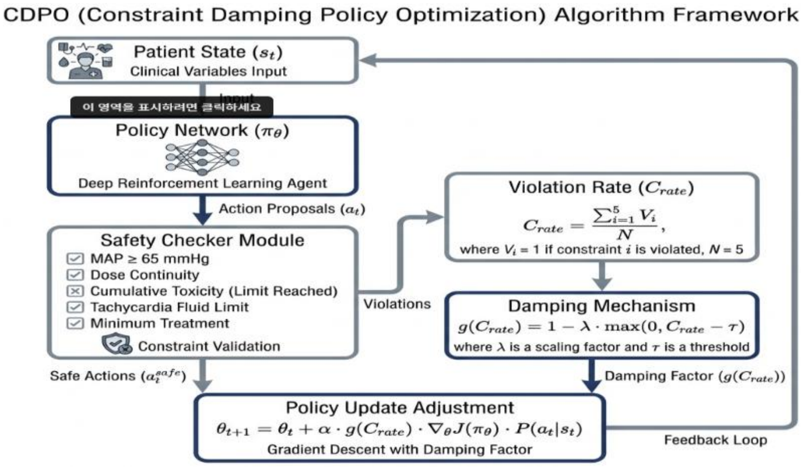
CDPO (Constraint Damping Policy Optimization) Algorithm Framework
Patient State (s_t)
Clinical Variables Input
Input
이 영역을 표시하려면 클릭하세요
Policy Network (π_θ)
Deep Reinforcement Learning Agent
Action Proposals (a_t)
Safety Checker Module
☑ MAP ≥ 65 mmHg
☑ Dose Continuity
☒ Cumulative Toxicity (Limit Reached)
☑ Tachycardia Fluid Limit
☑ Minimum Treatment
Constraint Validation
Violations
Violation Rate (C_{rate})
$C_{rate} = \frac{\sum_{i=1}^{5} V_i}{N}$,
where $V_i = 1$ if constraint i is violated, $N = 5$
Damping Mechanism
$g(C_{rate}) = 1 - \lambda \cdot \max(0, C_{rate} - \tau)$
where λ is a scaling factor and τ is a threshold
Safe Actions (a_t^{safe})
Damping Factor ($g(C_{rate})$)
Policy Update Adjustment
$\theta_{t+1} = \theta_t + \alpha \cdot g(C_{rate}) \cdot \nabla_\theta J(\pi_\theta) \cdot P(a_t|s_t)$
Gradient Descent with Damping Factor
Feedback Loop

Figure 1: Overall framework of the constraint damping policy optimization (CDPO) algorithm.

3.5 Sepsis treatment optimization via CDPO

Building on the proposed safe sepsis treatment formulation and the CDPO algorithm, this section presents a structured strategy for mitigating clinical risks in ICU settings. The key contribution of CDPO lies in reframing sepsis treatment as a problem of continuous decision-making under explicit safety constraints, rather than a sequence of isolated dosing actions.

CDPO learns dosing decisions as part of a policy trajectory conditioned on prior treatment history, which discourages excessive short-term escalation of vasopressors or fluids and promotes gradual adjustments aligned with patient-specific physiological responses. This behavior closely resembles the stepwise titration patterns employed by clinicians in practice, resulting in strong clinical alignment.

Moreover, by embedding the violation rate–based damping mechanism into the learning process, CDPO implicitly shapes an admissible action space that respects the five core safety constraints. Policy directions that repeatedly induce safety violations experience progressively reduced update magnitudes, lowering their likelihood during exploration and effectively suppressing unnecessary risk-seeking behavior. As violation signals increase, policy updates are further constrained, mitigating oscillatory behavior, abrupt policy shifts caused by reward noise, and medication misuse.

Overall, the CDPO-based dosing strategy goes beyond post hoc imitation of clinical guidelines by structurally integrating safety constraints into policy learning itself, enabling conservative and reliable decision-making in the highly uncertain ICU environment while maintaining stable convergence and long-term physiological homeostasis.

3.6 Treatment

The ICU sepsis treatment task was formulated as a Markov decision process (MDP) with a 48-dimensional state space and 25 discrete actions derived from patient time-series data. All baseline models shared the same network architecture and reward structure to ensure that observed performance differences were attributable solely to the policy update mechanisms.

To assess clinical safety, the primary evaluation metric was the **number of safety violations**. A violation was defined as any time step in which mean arterial pressure (MAP) fell below 65 mmHg without vasopressor administration, and total violations were aggregated across episodes. Treatment efficiency was evaluated in parallel using the **average dosing level**, enabling a comparative analysis of whether CDPO achieves both improved safety and efficiency relative to clinician policies and standard PPO.

Table 2 summarizes the experimental environment and key hyperparameter settings, including MDP configuration, training episodes, batch size, learning rate, discount factor, and core reward components.

Category	Parameter	Value
MDP configuration	State space / Action space	48 dimensions / 25 discrete actions
Training setup	Training episodes / Batch size	20,000 / 64
Optimization	Learning rate / Discount factor (γ)	1×10−4 / 0.99
Reward design	Key penalty terms	violation, warning, efficiency

Table 2: Experimental environment and hyperparameter settings

4 Experimental results

4.1 Violation counts

The comparison of safety violations shows that clinician policies resulted in 170 violations, PPO achieved zero violations, and CDPO recorded 131 violations. Importantly, CDPO substantially reduced safety violations relative to clinicians while simultaneously forming policies that, as shown in subsequent sections, reduced overall medication usage. This indicates that CDPO achieves a balance between safety and efficiency by intervening in high-risk situations without resorting to unnecessary overtreatment.

In contrast, the zero-violation outcome of PPO is more appropriately interpreted as excessive adherence to a single safety criterion rather than a genuinely optimal solution. PPO tended to respond mechanically to hypotension signals with aggressive dosing, resulting in systematic over-administration. Such behavior, while eliminating measured violations, may increase the risk of complications associated with overtreatment, particularly excessive fluid administration. From a clinical perspective, PPO therefore exhibits a form of **overfitting to safety constraints**, limiting its practical applicability despite achieving nominally perfect violation scores.

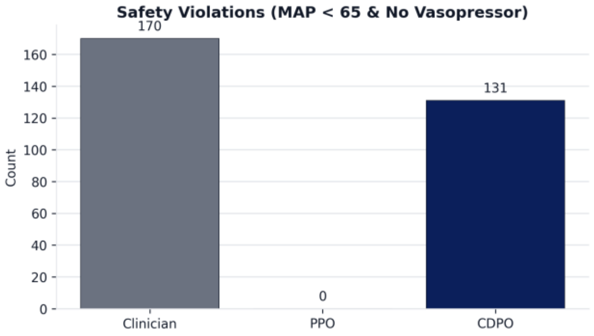

Figure 2: Comparison of safety violation counts across clinician, PPO, and CDPO policies.

4.2 Treatment efficiency analysis

In terms of treatment efficiency, CDPO consistently achieved the most favorable outcomes. The average vasopressor dosing level was 1.96 for clinicians, 1.65 for PPO, and 0.64 for CDPO, indicating a substantial reduction in vasopressor burden under CDPO. Similarly, while PPO increased average fluid administration to 2.70 compared to 1.98 for clinicians, CDPO reduced fluid usage to 1.52.

Overall, CDPO reduced both vasopressor and fluid administration while simultaneously decreasing safety violations, demonstrating the most compelling balance between safety and efficiency. A notable contrast is observed with PPO: in eliminating violations, PPO converged toward increasingly conservative dosing strategies, leading to fluid over-administration beyond clinical norms. Thus, although PPO achieved zero violations, it did so at the cost of overtreatment, which is misaligned with clinical optimization objectives. In contrast, CDPO aligned policy learning toward satisfying safety constraints while actively suppressing unnecessary medication use.

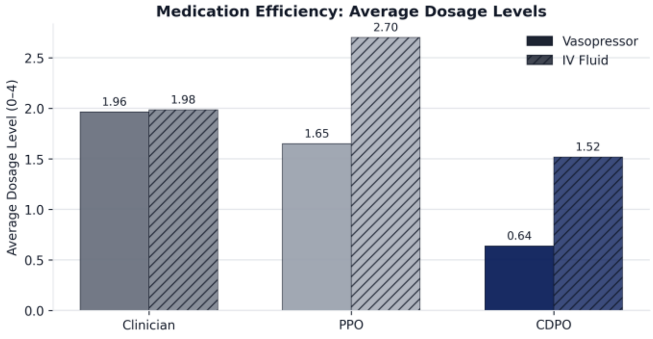

Figure 3: Average vasopressor and fluid dosage levels across treatment policies.

4.3 Patient-level case analysis: alignment between clinician and CDPO trajectories

To examine how quantitative metrics translate into treatment behavior, we compared dosing trajectories for a representative patient case (ICUSTAY_ID = 201006). In this case, CDPO selectively increased vasopressor dosing only when mean arterial pressure (MAP) approached or fell below the safety threshold of 65 mmHg, while maintaining low dosing levels once MAP stabilized. This pattern reflects a policy that intervenes when necessary and withdraws when risk subsides, demonstrating simultaneous safety preservation and dose reduction at the case level.

In contrast, clinician dosing exhibited larger temporal variability, with localized peaks reflecting complex clinical judgments influenced by both observable and unobservable factors. Importantly, CDPO does not simply imitate clinician behavior; instead, it aligns policy learning toward suppressing overtreatment while remaining within safety boundaries. This case illustrates that CDPO operationalizes the safety–efficiency trade-off not merely through aggregate metrics, but through the structure of treatment trajectories themselves.

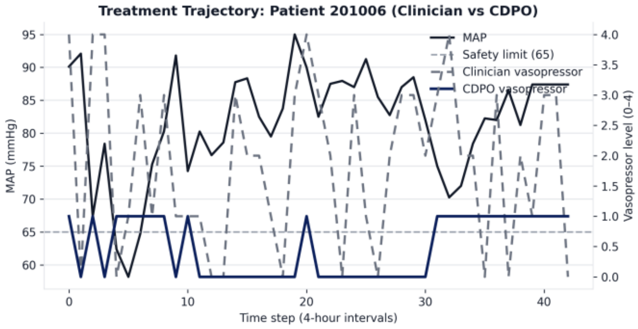

Figure 4: Treatment trajectory comparison for a representative ICU patient (Clinician vs CDPO).

4.4 Clinical implications

The results of this study challenge the simplistic assumption that minimizing safety violations alone yields optimal treatment policies. Although PPO achieved zero safety violations, it converged toward excessively conservative dosing strategies, particularly increasing fluid administration beyond clinically typical levels. In medical contexts, such overtreatment may elevate the risk of complications, indicating that optimization focused solely on satisfying a single safety metric can be clinically suboptimal.

In contrast, CDPO reduced safety violations relative to clinician policies while significantly lowering both vasopressor and fluid usage. This indicates that CDPO shapes policies that preserve safety while actively reducing unnecessary medication burden. These findings suggest that effective reinforcement learning–based treatment optimization requires balancing safety and efficiency, rather than optimizing isolated indicators. Overall, this study empirically demonstrates the overfitting risk associated with penalty-based constraint handling and shows that violation rate–based update damping in CDPO promotes more clinically realistic and reliable treatment policies.

5 Conclusion

This study empirically demonstrates that clinical safety constraints can be integrated as core control signals in reinforcement learning–based medical decision-making, enabling the simultaneous achievement of treatment efficiency and patient safety. The proposed **Constraint damping policy optimization (CDPO)** algorithm directly addresses the structural *safety-blind* limitation of conventional reinforcement learning methods by regulating policy updates according to clinical risk.

By explicitly defining five guideline-based safety constraints and introducing a violation rate–driven damping mechanism, CDPO structurally suppresses unsafe policy shifts. Experiments on the MIMIC-III dataset show that CDPO reduces hypotension-related violations by 23.2% while decreasing vasopressor usage by 67.3%, demonstrating that it systematically explores an optimal balance between safety and efficiency rather than converging to overly conservative policies.

In addition, this work illustrates the feasibility of conducting complex medical reinforcement learning research through collaboration with large language models (LLMs) and AI agents acting as co-scientists. AI-assisted support was leveraged throughout the research lifecycle—from planning and problem formulation to algorithm design, implementation, and empirical analysis—highlighting the potential of AI as an active research collaborator beyond a mere computational tool.

Future work includes prospective validation in real clinical settings, refinement of patient state representations using higher-dimensional physiological signals, and the design of extended constraint formulations incorporating multiple clinical objectives. These directions will further support the development of AI-assisted treatment systems as reliable decision-making partners in real-world healthcare environments.

References

[1] Rhodes, A., Evans, L.E., Alhazzani, W., Levy, M.M., Antonelli, M., Ferrer, R., Kumar, A., Sevransky, J.E., Sprung, C.L., Nunnally, M.E., Rochwerg, B., Rubenfeld, G.D., Angus, D.C., Annane, D., Beale, R.J., Bernard, G.R., Chiche, J.D., Coopersmith, C., De Backer, D.P., French, C.J., Fujishima, S., Gerlach, H., Hidalgo, J.L., Jones, A.E., Marshall, J.C., Mazuski, J.E., McIntyre, L.A., Moreno, R.P., Myburgh, J., Navalesi, P., Nishida, O., Perner, A., Ranieri, M., Schorr, C.A., Seymour, C.W., Singer, M., Thompson, B.T., Vincent, J.L. & Dellinger, R.P. (2017)
Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock. *Intensive Care Medicine* 43(3), 304–377.

[2] Singer, M., Deutschman, C.S., Seymour, C.W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G.R., Chiche, J.D., Coopersmith, C.M., Hotchkiss, R.S., Levy, M.M., Marshall, J.C., Martin, G.S., Opal, S.M., Rubenfeld, G.D., van der Poll, T., Vincent, J.L. & Angus, D.C. (2016)
The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA* 315(8), 801–810.

[3] Komorowski, M., Celi, L.A., Badawi, O., Gordon, A.C. & Faisal, A.A. (2018)
The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine* 24, 1716–1720.

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

[5] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J. & Mané, D. (2016)
Concrete Problems in AI Safety. *arXiv preprint* arXiv:1606.06565.

[6] Achiam, J., Held, D., Tamar, A. & Abbeel, P. (2017)
Constrained Policy Optimization. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 22–31.

[7] Tessler, C., Efroni, Y. & Mannor, S. (2019)
Reward Constrained Policy Optimization. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*.

[8] Johnson, A.E.W., Pollard, T.J., Shen, L., Lehman, L.W.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L.A. & Mark, R.G. (2016)
MIMIC-III, a freely accessible critical care database. *Scientific Data* 3, 160035.

[9] Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. (2017)
Proximal Policy Optimization Algorithms. *arXiv preprint* arXiv:1707.06347.

[10] Ray, A., Achiam, J. & Amodei, D. (2019)
Benchmarking Safe Exploration in Deep Reinforcement Learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*.

Appendix A: Details of AI Utilization

The following large language models were used during the research process:

– GPT-5.2 (OpenAI)

– Gemini 3 Pro (Google DeepMind)

– Claude Sonnet 4.5 (Anthropic)

These models were used for literature synthesis, brainstorming of algorithmic designs,
logical consistency checks, and drafting assistance.
All outputs were reviewed, verified, and validated by the authors.

AI Co-Scientist Challenge Korea Paper Checklist

1. **Claims**

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

   Answer: Yes

   Justification:

   - The abstract and introduction clearly state the core contribution of proposing CDPO, a constraint-aware reinforcement learning algorithm that integrates clinical safety constraints into policy optimization. The claims are directly supported by experimental results on the MIMIC-III dataset and are limited to offline ICU sepsis treatment optimization, matching the paper's scope (Sections 1 and Abstract).

2. **Limitations**

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

   Answer: Yes

   Justification:

   - The paper discusses limitations including reliance on retrospective observational data, absence of prospective clinical validation, and potential sensitivity to state representation and constraint definitions. Future work directions addressing these limitations are explicitly outlined in the Discussion and Conclusion sections.

3. **Theory Assumptions and Proofs**

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

   Answer: N/A

   Justification:

   - The paper does not present formal theorems or analytical proofs. Instead, it focuses on algorithmic design and empirical validation of CDPO, making formal theoretical proofs not applicable to this work.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: Yes

   Justification:

   - The experimental setup, including state and action space definitions, reward structure, safety constraints, hyperparameters, and evaluation metrics, is fully described in Sections 3.3 and 3.6. This information is sufficient for reproducing the main results using the same dataset.

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

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

   Answer: No

   Justification:

   - While the study uses the publicly available MIMIC-III dataset, the implementation code is not released at submission time. However, all algorithmic details and experimental configurations are provided to enable independent reimplementation.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: Yes

   Justification:

   - The paper specifies the MDP formulation, data preprocessing steps, network architecture, optimizer, learning rates, discount factors, training episodes, and baseline configurations in Section 3.6 and Table 2.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: No

   Justification:

   - The study reports aggregate performance metrics over the full evaluation cohort but does not include confidence intervals or error bars. This choice was made due to the focus on clinically interpretable aggregate safety and efficiency indicators rather than variance across random seeds.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: No

   Justification:

   - The paper does not explicitly report hardware specifications or total compute time. The experiments were conducted using standard GPU-based training environments, but precise resource details are not included.

9. **The paper does not explicitly report hardware specifications or total compute time. The experiments were conducted using standard GPU-based training environments, but precise resource details are not included. Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://nips.cc/public/EthicsGuidelines`?

   Answer: Yes

   Justification:

   - The study uses de-identified, publicly available clinical data (MIMIC-III) and follows established data use agreements. Ethical considerations, human-in-the-loop validation, and responsible AI usage are explicitly discussed in Section 1.3.

10. **Broader Impacts**

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

    Answer: Yes

    Justification:

    - The paper discusses positive impacts such as improved clinical decision support and patient safety, as well as potential risks including over-reliance on AI systems and misuse without clinical oversight. It emphasizes that the proposed method is intended as a decision-support tool rather than an autonomous treatment system (Discussion and Conclusion sections).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: N/A

Justification:

- The paper does not release pretrained models, generative systems, or newly collected datasets that could pose a risk of misuse. All experiments are conducted on the publicly available, de-identified MIMIC-III dataset, and the proposed method is explicitly framed as a clinical decision-support framework rather than a deployable autonomous system. Consequently, no additional safeguards beyond standard ethical data use are required.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: Yes

    Justification:

    - The paper uses the MIMIC-III clinical database, which is properly cited and acknowledged, and whose data access and usage are governed by PhysioNet credentialing and data use agreements. All referenced algorithms, datasets, and prior work are appropriately cited in the References section, and their original creators are credited.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: N/A

    Justification:

    - The paper does not release new datasets, pretrained models, or software assets. The contribution is methodological and empirical, focusing on algorithm design and evaluation rather than asset release.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: N/A

    Justification:

    - The study does not involve crowdsourcing or direct interaction with human subjects. All analyses are performed on retrospective, de-identified clinical data without participant recruitment or intervention.

15. **Justification: Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: N/A

    Justification:

    - No new human subjects research was conducted in this study. All analyses rely exclusively on the MIMIC-III dataset, which consists of de-identified clinical data collected under prior Institutional Review Board (IRB) approval. As no new data collection, intervention, or subject interaction was performed, additional IRB approval was not required for this work.approval, and therefore does not require additional IRB review for this research.