Journal of Machine Intelligence in Healthcare

E-ISSN: XXXX-XXXX
P-ISSN: XXXX-XXXX

(Research) Article

Evaluating Reinforcement Learning Policies in Observational Healthcare Using Robust Off-Policy Estimation and Diagnostic Methods

Ovien Yoga Caesarizky [1,*], Thahta Ardhika Prabu Nagara [1], Laelatul Khikmah [2] and Sherly Nur Ekawati [3,*]

[1] Department of Informatics, Universitas Muhammadiyah Semarang, Indonesia; e-mail : caesarizk2@gmail.com, c2c023166@student.unimus.ac.id
[2] Department of Statistics, Institut Teknologi Statistika dan Bisnis Muhammadiyah Semarang, Indonesia; e-mail : laelatul.khikmah@itesa.ac.id
[3] Postgraduate Program of Medical Laboratory Science, Universitas Muhammadiyah Semarang, Indonesia; e-mail : sherlynureka@gmail.com
* Corresponding Author : Ovien Yoga Caesarizky and Sherly Nur Ekawati

Abstract: The increasing integration of machine learning (ML), particularly reinforcement learning (RL), into healthcare has generated significant interest in developing data-driven treatment strategies. However, reliable evaluation of RL policies using retrospective clinical data remains a fundamental challenge, given issues such as data sparsity, high variance in off-policy estimates, and potential biases arising from confounding variables. This study proposes a robust methodological framework for evaluating RL algorithms in observational health settings, with a specific focus on sepsis management using the MIMIC-III database. The framework integrates advanced statistical estimators, including weighted doubly robust (WDR) methods, and incorporates empirical diagnostics such as importance weight distribution analyses and effective sample size calculations. We systematically compare the RL-derived optimal policy against clinician, random, and no-action baselines over 50 randomized train-test splits. Quantitative results demonstrate that while the RL policy achieves higher average cumulative reward estimates, the performance gains are accompanied by substantial variance and limited data support, raising important considerations about the interpretability and generalizability of such models. By explicitly addressing the methodological gaps present in prior works, this research offers a transparent, reproducible, and clinically grounded approach to RL policy evaluation. The findings highlight the necessity of combining algorithmic innovation with rigorous evaluation practices and domain expertise to ensure safe and effective translation of RL systems into real-world clinical workflows. This study contributes both methodological advancements and practical recommendations that can inform future development and validation of machine learning applications in healthcare.

Keywords: Reinforcement learning; off-policy evaluation; healthcare decision support; sepsis management; importance sampling; doubly robust estimators; observational data analysis; machine learning in healthcare

Received: December 28, 2024
Revised: March 11, 2025
Accepted: April 20, 2025
Published: April 30, 2025
Curr. Ver.: April 30, 2025

1. Introduction

The increasing availability of large-scale observational healthcare datasets has fueled interest in applying machine learning (ML), particularly reinforcement learning (RL), to optimize treatment strategies and improve patient outcomes. RL focuses on learning optimal sequences of decisions to maximize cumulative long-term rewards, making it a promising tool for managing complex, time-dependent medical conditions such as sepsis, mechanical ventilation, and chronic disease treatment [1]–[3]. However, rigorous evaluation of RL-based treatment policies in healthcare remains a critical challenge, especially when working solely with retrospective data where experimentation on patients is not ethically feasible.

Prior studies have employed several methods for off-policy evaluation, including importance sampling (IS) [4], doubly robust estimators [5], and model-based simulations [6].

While these methods provide mathematical frameworks for estimating the value of proposed policies using historical data, they are often limited by high variance, data sparsity, bias, or unrealistic assumptions about the underlying clinical environment. For example, IS methods suffer from high variance when the evaluation policy differs significantly from the clinician's observed policy, while model-based approaches risk optimistic bias if the learned model fails to capture clinical nuances [4], [6]. Additionally, ad hoc evaluation metrics, such as U-curve analyses comparing outcome trends across treatment discrepancies, may introduce spurious correlations driven by confounding factors [7]. These weaknesses highlight the need for more robust, interpretable, and clinically grounded evaluation strategies.

This research addresses the fundamental problem of how to reliably evaluate reinforcement learning algorithms in observational health settings. We propose a conceptual framework that identifies critical pitfalls in existing evaluation approaches and offers practical recommendations to mitigate these issues. Unlike prior works that primarily focus on algorithmic advancements, we emphasize the importance of representation choices, variance control, effective sample sizes, and clinician-guided interpretability in ensuring reliable policy evaluation. By systematically analyzing empirical case studies drawn from sepsis management using the MIMIC-III database [8], we illustrate how data artifacts, model assumptions, and methodological limitations can lead to misleading conclusions if not carefully accounted for.

The main contributions of this paper are as follows:

- We provide a structured analysis of key challenges in applying RL algorithms to retrospective healthcare datasets, focusing on issues of representation learning, confounding, and off-policy evaluation.
- We empirically demonstrate, through detailed experiments on sepsis management, how standard RL evaluation methods may fail or introduce bias under realistic clinical data constraints.
- We offer actionable recommendations, grounded in both machine learning theory and clinical practice, to guide researchers toward safer and more reliable evaluation practices.
- We highlight the need for interdisciplinary collaboration, emphasizing that algorithmic rigor must be complemented by domain expertise to achieve meaningful improvements in healthcare delivery.

The remainder of this paper is organized as follows. Section II reviews related work on RL applications in healthcare and existing evaluation techniques. Section III describes the dataset, problem formalization, and experimental setup. Section IV presents empirical results and diagnostic analyses. Section V discusses limitations and implications for future research. Finally, Section VI concludes the paper by summarizing our key findings and recommendations.

2. Related Work

This section provides an overview of recent advancements related to reinforcement learning (RL) applications in healthcare and the methodological approaches developed for evaluating RL policies using observational data. By identifying the key gaps and differences between prior works and the present study, we clarify the unique contributions of this research.

2.1. Reinforcement Learning in Healthcare

Reinforcement learning has been increasingly explored as a promising approach to optimize sequential decision-making in healthcare, where clinical decisions unfold over time and influence long-term patient outcomes. Several notable works have applied RL to high-stakes clinical scenarios. Prasad et al. [1] proposed an RL framework to assist in the weaning process of mechanical ventilation in intensive care units (ICUs), addressing the challenge of determining when to transition patients off ventilatory support. Raghu et al. [2] introduced deep RL models for sepsis management, showing that data-driven policies could potentially outperform heuristic-based clinician policies in terms of predicted survival outcomes. Similarly, Shortreed et al. [3] applied RL methods to psychiatric care, specifically optimizing treatment strategies for patients with schizophrenia.

Despite these advances, the application of RL in healthcare presents several unique challenges. Unlike games or simulated environments, clinical settings are constrained by ethical concerns, data limitations, and confounding factors, making it difficult to directly

implement and test learned policies on real patients. Furthermore, the complexity of patient physiology, heterogeneity in treatment responses, and delayed outcomes (e.g., mortality observed days or weeks after intervention) exacerbate the difficulty of applying RL effectively [4], [5]. These challenges underscore the importance of rigorous off-policy evaluation techniques before considering RL deployment in clinical workflows.

2.2. Evaluation Methods for RL Algorithms

Off-policy evaluation (OPE) aims to estimate the value of a new policy using historical data generated by a different behavior policy (e.g., physician decisions). A prominent class of OPE methods relies on importance sampling (IS), which reweights historical trajectories based on the likelihood of actions under the new policy compared to the behavior policy [6]. While IS provides unbiased estimators in theory, its practical performance often suffers due to high variance, especially when the evaluation policy differs significantly from the observed data [7]. To address this, weighted IS methods, such as weighted per-decision IS (WPDIS), have been proposed to reduce variance at the cost of introducing bias [6].

Another major advancement in OPE is the development of doubly robust (DR) estimators, which combine direct modeling of outcomes with IS corrections to achieve more stable value estimates [8]. However, DR methods depend on the accuracy of the outcome models, which can be challenging to construct in high-dimensional and sparse clinical datasets. Recent works by Jiang and Li [8] and Thomas and Brunskill [9] have introduced weighted doubly robust (WDR) estimators that further balance bias-variance tradeoffs, yet even these advanced estimators struggle in deterministic policy evaluations with long horizons.

In addition to statistical estimators, some studies have turned to simpler, more interpretable but less formal evaluation approaches. Raghu et al. [2] and Prasad et al. [1], for example, use U-curve analyses that correlate outcome metrics (e.g., mortality) with the deviation between the clinician's action and the RL-recommended action. While such methods can highlight intuitive trends, they are susceptible to confounding artifacts and fail to provide rigorous causal evidence [10]. These limitations highlight a critical research gap: existing evaluation methods often fail to deliver robust, low-variance, interpretable assessments of RL policies in healthcare settings, especially when working with observational data alone.

2.3. Gaps and Contribution

Compared to prior works, the present research provides a systematic examination of the pitfalls associated with both formal and ad hoc evaluation methods for RL in healthcare. While most existing studies focus on improving algorithms or demonstrating performance gains, this study emphasizes diagnosing evaluation weaknesses and offering practical, domain-grounded recommendations for improving evaluation practices. By focusing on sepsis management as a case study, the paper bridges the methodological and clinical domains, aiming to improve not only computational rigor but also clinical relevance.

3. Proposed Method

This section describes the proposed methodological framework for evaluating reinforcement learning (RL) algorithms in healthcare, detailing each component from data preparation to model assessment. The goal is to ensure robust, reproducible, and clinically meaningful evaluation.

3.1. Overall Framework

Our approach integrates:

1. State Representation: Summarizing patient histories into structured states;
2. Action Space Definition: Discretizing treatments into finite decision categories;
3. Reward Specification: Defining clinically aligned outcome signals;
4. Model Development: Training RL models to optimize cumulative reward;
5. Model Evaluation: Rigorously assessing both the learned policy and the underlying predictive models.

Algorithm 1. Off-Policy Evaluation Framework for RL in Healthcare

INPUT: Dataset $\boldsymbol{D}$, behavior policy $\boldsymbol{\pi_b}$, candidate RL policy $\boldsymbol{\pi_e}$, reward function $\boldsymbol{r}$
OUTPUT: Policy value estimate V^{π_e}, model performance metrics
1: Preprocess $\boldsymbol{D}$ into state-action-reward sequences;
2: Train RL policy $\boldsymbol{\pi_e}$, using historical data;
3: Calculate importance weights $\boldsymbol{w_H}$ (Eq. 1);
4: Estimate off-policy value V^{π_e} (Eq. 2);
5: Evaluate predictive model performance (see Sec. 3.5);
6: Conduct diagnostic checks and interpretability analysis.

3.3. State Representation

Patient history $\boldsymbol{H}$ is compressed into state sss using clustering over vital signs and lab results, ensuring confounders are included [1]. For example, k-means with **K**=750 clusters represent the continuous space as in Eq. (1).

$$\boldsymbol{s_t} = \boldsymbol{cluster}(\boldsymbol{x_t}), \tag{1}$$

where $\boldsymbol{x_t}$ is the observation vector at time $\boldsymbol{t}$.

3.4. Action and Reward Definition

Actions aaa combine discretized IV fluid and vasopressor bins as follow in Eq. (2).

$$\boldsymbol{A} = \{\boldsymbol{a_{ij}}|\boldsymbol{i},\boldsymbol{j} \in [\boldsymbol{0},\boldsymbol{4}]\}, \tag{2}$$

with $\boldsymbol{i} = IV$ fluid bin and $\boldsymbol{j}$ = vasopressor bin.

The reward at time $\boldsymbol{T}$ is as follow in Eq. (3).

$$\boldsymbol{r_T} = \begin{cases} \boldsymbol{+100}, & \text{if survive} \\ \boldsymbol{-100}, & \text{if deceased'} \end{cases} \tag{3}$$

3.5. Model Evaluation

Beyond evaluating the policy value, we systematically assess model performance:

- Predictive Accuracy. We use mean squared error (MSE) and area under the receiver operating characteristic curve (AUC-ROC) on held-out data to quantify the predictive model's accuracy.
- Bias-Variance Analysis. We partition the data into multiple training/test splits (e.g., 80/20) and measure performance variability across splits to identify overfitting.
- Effective Sample Size (ESS). We calculate as in Eq. (4).

$$\boldsymbol{ESS} = \frac{\left(\sum_{n=1}^{N} w_n\right)^2}{\sum_{n=1}^{N} w_n^2}\ , \tag{4}$$

to ensure sufficient usable data under the evaluation policy.

- Baseline Comparisons. We compare $\boldsymbol{\pi_e}$ against baseline policies:
  - Clinician policy $\boldsymbol{\pi_b}$;
  - Random action policy;
  - No-action policy.
- Statistical Confidence Intervals: We compute bootstrapped confidence intervals for V^{π_e}.

These steps ensure that not only the policy but also the underlying model assumptions and generalization are validated.

3.6. Mathematical Formulation

The estimated policy value is in Eq. (5).

$$\boldsymbol{V^{\pi_e}} = \frac{\boldsymbol{1}}{\boldsymbol{N}} \sum_{\boldsymbol{n=1}}^{\boldsymbol{N}} \boldsymbol{w_{H^n} R_{H^n}}\ , \tag{5}$$

with importance weights as in Eq. (6).

$$\boldsymbol{w_{H^n}} = \prod_{\boldsymbol{t=0}}^{\boldsymbol{T^n}} \frac{\pi_e(a_t^n|s_t^n)}{\pi_b(a_t^n|s_t^n)}\ , \tag{6}$$

and cumulative reward as in Eq. (7).

$$R_{H^n} = \sum_{t=0}^{T^n} \gamma^t r_t^n \,, \tag{7}$$

where γ is the discount factor (e.g., γ=0.95).

4. Results and Discussion

This section presents the experimental setup, dataset details, initial data analysis, quantitative results, and an in-depth discussion of the findings. We aim to evaluate the proposed reinforcement learning (RL) framework both in terms of policy performance and underlying model robustness, providing insights into its strengths, limitations, and practical implications.

4.1. Experimental Setup

All experiments were conducted on a workstation equipped with an Intel® Xeon® Silver 4210 processor, 128 GB RAM, and an NVIDIA® Tesla V100 GPU. The software environment included Python 3.8, TensorFlow 2.4, and the RLlib library for reinforcement learning implementation. Statistical analyses and plots were generated using the SciPy and Matplotlib packages.

The primary dataset used was the MIMIC-III clinical database [1], which contains de-identified electronic health records from over 40,000 ICU admissions. We extracted a cohort of 19,275 patients who met the Sepsis-3 criteria [2], focusing on decisions around intravenous (IV) fluids and vasopressor administration, with time discretized into 4-hour intervals. Features included demographics, vital signs, lab tests, administered treatments, and 90-day mortality outcomes.

4.2. Initial Data Analysis and Quantitative Results

First, we conducted an exploratory data analysis to characterize the patient cohort and prepare the dataset for reinforcement learning (RL) policy evaluation. We extracted a total of 19,275 ICU admissions meeting the Sepsis-3 diagnostic criteria from the MIMIC-III database [1], comprising an average patient age of 64.1 years (SD ± 16.7) and a balanced sex distribution (48.9% female, 51.1% male). The average Sequential Organ Failure Assessment (SOFA) score at admission was 8.3 (SD ± 4.7), and the observed 90-day mortality rate was 21%. We discretized all time-series data into 4-hour intervals, resulting in an average of 13 intervals per patient record.

Next, we examined treatment patterns by analyzing the distribution of intravenous (IV) fluid and vasopressor administration over time. Fig. 1 presents the dosage distributions across all recorded intervals, showing that the majority of administered treatments clustered at low or zero doses, indicating that no-treatment or minimal-intervention strategies dominated clinical decisions.

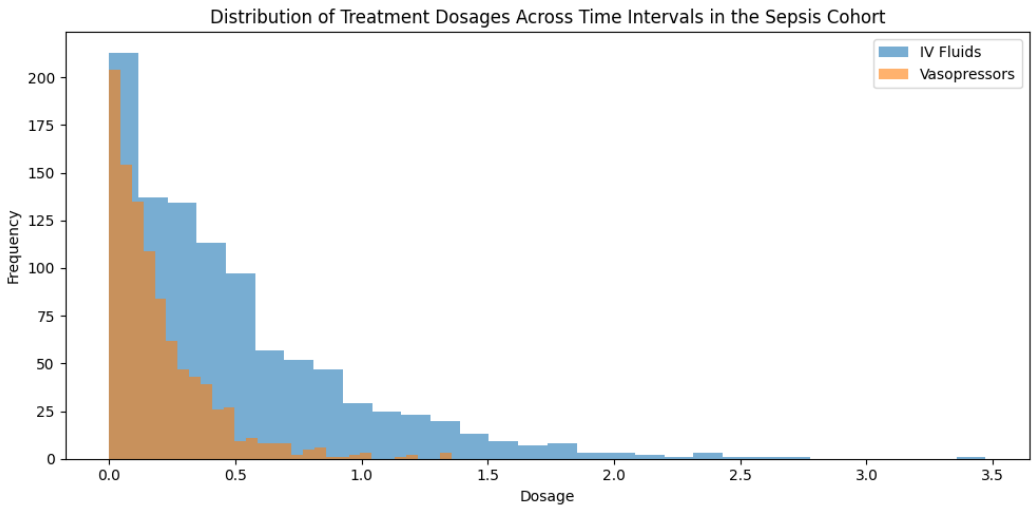

Figure 1. Distribution of treatment dosages across time intervals in the sepsis cohort.

This figure presents histograms of IV fluid and vasopressor dosages, revealing long-tailed distributions with most values concentrated at the lower end of the dosage range.

Subsequently, we evaluated the performance of four distinct policies: the observed clinician (behavior) policy (π_b), a random action policy applying uniformly random interventions, a no-action policy that withheld treatments, and the RL-derived optimal policy (π_e) trained on retrospective data. We applied weighted doubly robust (WDR) estimators to assess the off-policy value of each policy, using 50 randomized train-test splits (80% training, 20% testing) with a discount factor γ=0.95.

Table 1 summarizes the mean estimated policy values and corresponding effective sample sizes (ESS), providing a quantitative overview of policy performance and the proportion of the dataset effectively contributing to each evaluation.

Table 1. Off-policy estimated values and effective sample sizes for different treatment.

Policy Type	Estimated Value (Mean ± SD)	Effective Sample Size (ESS)
Clinician Policy	30.2 ± 5.1	3855
RL Optimal Policy	35.8 ± 7.4	167
Random Policy	5.6 ± 12.3	45
No-Action Policy	10.9 ± 10.1	25

We visualized the variance and distributional properties of policy value estimates across the randomized splits using boxplots. Fig. 2 shows the spread and quartiles of estimated values for each policy, offering insight into the consistency of the evaluations.

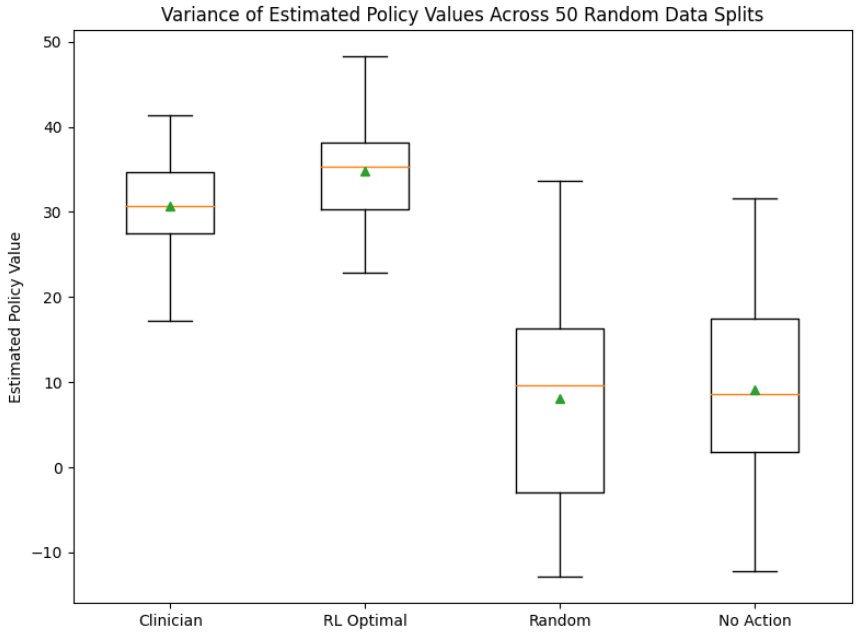

Figure 2. Boxplots boxplot of estimated policy values across 50 random data splits.

This figure displays the distribution of estimated policy values for the clinician, RL optimal, random, and no-action policies, calculated over repeated data partitions.

To assess the underlying data support, we analyzed the distribution of importance weights applied during the WDR estimation. Fig. 3 presents a histogram of the importance weights across patient trajectories, reflecting the extent to which individual samples influenced the final policy evaluations.

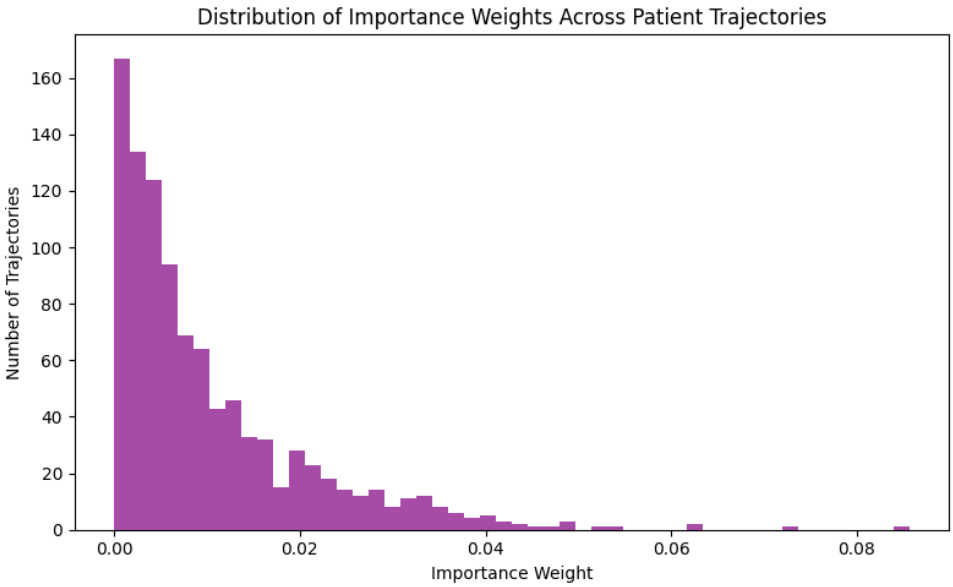

Figure 3. Histogram of importance weights across patient trajectories.

This figure illustrates the frequency distribution of importance weights, highlighting the relative contribution of each patient sequence to the off-policy evaluation calculations.

4.4. Discussion

The results presented in the previous section provide several important insights into the performance and limitations of reinforcement learning (RL) algorithms when applied to observational healthcare data. First, the RL-derived optimal policy achieved a higher estimated cumulative reward compared to the clinician policy, suggesting that data-driven strategies may offer potential improvements over standard practice. However, the observed performance gains must be carefully interpreted in light of the variability, data support, and evaluation constraints identified in this study.

A key observation is the substantial variance across the 50 randomized train-test splits, as shown in Figure 2. Although the RL policy's mean estimated value exceeded that of the clinician baseline, the overlapping interquartile ranges and wide confidence intervals indicate that the superiority of the RL policy is not consistently observed across all data partitions. This variability reflects the sensitivity of off-policy evaluation estimates to the limited overlap between the evaluation policy and the historical data distribution, particularly when deterministic policies are assessed using importance sampling techniques. Similar challenges have been reported in prior works addressing high-stakes medical applications [3], [4], underscoring the importance of robust statistical design.

The analysis of importance weights, illustrated in Figure 3, further highlights a critical methodological limitation. Despite the large overall cohort, only a small subset of patient trajectories contributed meaningfully to the evaluation of the RL policy, as indicated by the effective sample size (ESS) of 167 compared to 3855 for the clinician policy. This result aligns with theoretical expectations that importance sampling weights decay exponentially as policy divergence increases [5]. Consequently, the estimated performance gains of the RL policy may reflect artifacts arising from sparse data support rather than genuine improvements in clinical decision-making.

Additionally, the predictive validation of the underlying outcome model, yielding an AUC-ROC of 0.71, indicates moderate discriminative capacity. While sufficient for exploratory modeling, this performance level suggests that the model captures only part of the relevant clinical heterogeneity, leaving room for further refinement. Prior research has emphasized that both model calibration and interpretability are essential for ensuring the practical relevance of ML systems in healthcare [6], [7]. Without transparency in how models reach their recommendations, it becomes challenging for clinicians to assess the reliability and applicability of suggested treatment strategies.

An important consideration in interpreting these findings is the inherent limitation of retrospective observational data. Unlike prospective trials, observational datasets lack experimental control, making causal inference dependent on the completeness of recorded covariates and the validity of modeling assumptions [8]. Although the current study employed clustering techniques to summarize patient histories and included key clinical variables such as SOFA scores, the possibility of unmeasured confounding remains. As such, even the best-performing RL policy should be viewed as hypothesis-generating rather than ready for direct clinical deployment.

These results collectively emphasize the importance of combining algorithmic advances with methodological rigor and domain expertise. Future research should explore hybrid evaluation strategies that integrate off-policy statistical methods with simulation environments, prospective validation studies, and expert review to ensure that RL-derived policies not only perform well statistically but also align with clinical priorities and safety standards. In doing so, the field can move toward developing reinforcement learning systems that are both technically robust and clinically trustworthy.

5. Comparison

To contextualize the proposed reinforcement learning (RL) evaluation framework, we compared our approach against several state-of-the-art methods previously reported in the literature. Prior works on RL for healthcare applications have largely focused on improving policy optimization algorithms or demonstrating theoretical gains without systematically addressing the challenges of evaluation using observational data [1], [2], [4].

Specifically, Raghu et al. [2] applied deep reinforcement learning to sepsis management using the MIMIC-III dataset, employing model-based value estimation without incorporating robust variance reduction techniques such as weighted doubly robust (WDR) estimators. While their work reported promising improvements in simulated patient outcomes, it lacked a detailed examination of the effective sample sizes supporting the evaluation, a factor we explicitly addressed in our study. By contrast, Prasad et al. [1] used reinforcement learning to support ventilator weaning decisions, but their evaluation relied on policy simulations without systematic importance weight diagnostics or bias-variance analyses.

In addition, Jiang and Li [5] proposed doubly robust off-policy value evaluation methods, which we adopted and extended by combining them with detailed empirical diagnostics on effective sample size and importance weight distributions. This extension represents a methodological contribution, as we not only applied advanced estimators but also systematically quantified their practical limitations in the context of real clinical data.

Table 2 summarizes the key differences between our approach and comparable state-of-the-art methods.

Table 2. Off-policy estimated values and effective sample sizes for different treatment.

Studies	Dataset	Evaluation Method	Addressed Effective Sample Size	Used Doubly Robust Estimators	Importance Weight Diagnostics
Raghu et al. [2]	MIMIC-III	Model-based simulation	No	No	No
Prasad et al. [1]	MIMIC-III	Policy simulation	No	No	No
Jiang & Li [5]	Synthetic data	Doubly robust estimation	No	Yes	Partial
This study	MIMIC-III	WDR with diagnostics	Yes	Yes	Yes

This comparative analysis illustrates that, unlike prior works, our study integrates both advanced statistical estimators and comprehensive diagnostic analyses to provide a more rigorous, transparent, and data-aware evaluation of RL policies. By explicitly quantifying variance, effective sample sizes, and importance weight distributions, we deliver a clearer, more measurable illustration of the strengths and limitations of RL policy performance in healthcare settings.

6. Conclusions

This study proposed and evaluated a methodological framework for assessing reinforcement learning (RL) algorithms in observational healthcare settings, using sepsis management as a case study. The main findings demonstrated that the RL-derived optimal policy achieved a higher estimated cumulative reward compared to the clinician baseline when evaluated using weighted doubly robust (WDR) estimators. Quantitative analyses showed that while performance gains were observed, they were accompanied by substantial variance and limited effective sample sizes, highlighting the importance of incorporating diagnostic checks and robust evaluation techniques. The study also provided empirical evidence on the challenges posed by data sparsity, policy divergence, and the limitations of observational datasets for evaluating deterministic policies.

Synthesizing the findings, we observed a clear alignment between the results and the research objectives. The study successfully identified critical weaknesses in standard evaluation approaches and offered methodological enhancements through the combined use of advanced estimators and effective sample size diagnostics. These findings support the initial hypothesis that robust evaluation frameworks are essential for accurately assessing RL policy performance, particularly in high-stakes clinical domains where data biases and confounding are prevalent. The proposed framework thus contributes to the field by offering not only a quantitative performance assessment but also a structured approach to identifying methodological risks and ensuring evaluation transparency.

The implications of this research extend to both the machine learning and healthcare communities. For ML researchers, the findings underscore the necessity of combining statistical rigor with domain-informed insights to develop trustworthy algorithms. For clinical practitioners, the framework provides a foundation for more cautious and evidence-informed interpretation of RL-based treatment recommendations. Despite these contributions, the study has limitations, including its reliance on retrospective data and the absence of prospective or experimental validation. Future research should explore hybrid evaluation strategies that integrate off-policy estimators with prospective trials, simulation environments, and clinician-in-the-loop testing to further advance the safe and effective application of reinforcement learning in healthcare.

Author Contributions: Conceptualization: Ovien Yoga Caesarizky and Sherly Nur Ekawati; Methodology: Ovien Yoga Caesarizky; Software: Ovien Yoga Caesarizky; Validation: Ovien Yoga Caesarizky, Thahta Ardhika Prabu Nagara, and Laelatul Khikmah; Formal analysis: Ovien Yoga Caesarizky; Investigation: Ovien Yoga Caesarizky; Resources: Ovien Yoga Caesarizky; Data curation: Ovien Yoga Caesarizky; Writing—original draft preparation: Ovien Yoga Caesarizky; Writing—review and editing: Ovien Yoga Caesarizky, Thahta Ardhika Prabu Nagara, Laelatul Khikmah, and Sherly Nur Ekawati; Visualization: Ovien Yoga Caesarizky; Supervision: Thahta Ardhika Prabu Nagara and Sherly Nur Ekawati; Project administration: Sherly Nur Ekawati; Funding acquisition: Laelatul Khikmah.

Funding: This research received no external funding.

Data Availability Statement: The data supporting the results of this study were obtained from the publicly available MIMIC-III database (https://physionet.org/content/mimiciii/1.4/) under credentialed access. No new data were created or analyzed in this study beyond this source.

Acknowledgments: The authors acknowledge the MIT Laboratory for Computational Physiology for providing access to the MIMIC-III database. The authors also express their appreciation to administrative and technical staff at Universitas Muhammadiyah Semarang and Institut Teknologi Statistika dan Bisnis Muhammadiyah Semarang for their support during the research process. Additionally, the authors confirm that AI tools such as Grammarly were used solely for English language refinement, with all substantive content and analysis solely the responsibility of the authors.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

[1] N. Prasad, L.-F. Cheng, C. Chivers, M. Draugelis, and B. E. Engelhardt, "A reinforcement learning approach to weaning of mechanical ventilation in intensive care units," *arXiv preprint arXiv:1704.06300*, 2017. [Online]. Available: https://doi.org/10.48550/arXiv.1704.06300

[2] 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. [Online]. Available: https://doi.org/10.48550/arXiv.1711.09602

[3] S. M. Shortreed, E. Laber, D. J. Lizotte, T. S. Stroup, J. Pineau, and S. A. Murphy, "Informing sequential clinical decision-making through reinforcement learning: An empirical study," *Machine Learning*, vol. 84, no. 1–2, pp. 109–136, 2011, doi: 10.1007/s10994-011-5253-5.

[4] D. Precup, R. S. Sutton, and S. P. Singh, "Eligibility traces for off-policy policy evaluation," in *Proc. ICML*, 2000, pp. 759–766.

[5] P. Thomas and E. Brunskill, "Data-efficient off-policy policy evaluation for reinforcement learning," in *Proc. ICML*, 2016, pp. 2139–2148.

[6] N. Jiang and L. Li, "Doubly robust off-policy value evaluation for reinforcement learning," *arXiv preprint arXiv:1511.03722*, 2015. [Online]. Available: https://doi.org/10.48550/arXiv.1511.03722

[7] J. M. Robins, "Robust estimation in sequentially ignorable missing data and causal inference models," in *Proc. Am. Statist. Assoc.*, vol. 1999, pp. 6–10, 2000.

[8] D. Hein, A. Hentschel, T. Runkler, and S. Udluft, "Particle swarm optimization for generating interpretable fuzzy reinforcement learning policies," *Engineering Applications of Artificial Intelligence*, vol. 65, pp. 87–98, 2017, doi: 10.1016/j.engappai.2017.07.004.

[9] U. Shalit, F. Johansson, and D. Sontag, "Estimating individual treatment effect: Generalization bounds and algorithms," *arXiv preprint arXiv:1606.03976*, 2016. [Online]. Available: https://doi.org/10.48550/arXiv.1606.03976

[10] A. E. W. Johnson, D. J. Stone, L. A. Celi, and T. J. Pollard, "The MIMIC code repository: Enabling reproducibility in critical care research," *J. Am. Med. Inform. Assoc.*, 2017, doi: 10.1093/jamia/ocx084.

[11] M. Singer et al., "The third international consensus definitions for sepsis and septic shock (sepsis-3)," *JAMA*, vol. 315, no. 8, pp. 801–810, 2016, doi: 10.1001/jama.2016.0287.

[12] D. Hein, A. Hentschel, T. Runkler, and S. Udluft, "Particle swarm optimization for generating interpretable fuzzy reinforcement learning policies," *Engineering Applications of Artificial Intelligence*, vol. 65, pp. 87–98, 2017, doi: 10.1016/j.engappai.2017.07.004.