Analysis of Reward Formulation Based on Mean Arterial Pressure in Reinforcement Learning for Critically Ill Septic Patient

Cristian Drudi[1], Maximiliano Mollura[1], Riccardo Barbieri[1]

[1] Politecnico di Milano, Milano, Italy

Abstract

Aims: *Optimal reward formulation in Reinforcement Learning (RL) is still uncertain. The aim of this study is to show that formulating a reward in RL for sepsis treatment using Mean Arterial Pressure (MAP) is a viable solution and can improve patient outcomes.* ***Methods:*** *The data were extracted from the MIMIC-III database. Patient data from 20,496 intensive care unit (ICU) stays were modeled with two different Markov Decision Processes that differed in reward formulation. The Mortality Model had a reward function linked only to 90-day mortality, and the Target MAP Model had an additional reward component that penalized the RL agent if the patient's MAP fell below 65 mmHg.* ***Results:*** *The Target MAP Model achieved the best results with a 95% lower bound (LB) of estimated policy value equal to 88.64 compared to 86.01 obtained from the Mortality Model despite having a more penalizing reward. The Target MAP Model in hypotensive patients uses less intravenous fluids and resorts more often to aggressive dosages of vasopressors.* ***Conclusions:*** *The results show that tying the reward to MAP is a viable approach, and the less sparse reward that comes with tying the reward to high temporal resolution cardiovascular features allows to evaluate single actions rather than the whole sequences of actions leading to the final outcome, allowing the RL agent to learn a better policy.*

1. Introduction

Sepsis is a complex condition that can develop when the body's immune response to an infection becomes dysregulated and begins to damage its own tissues and organs. It is a major cause of morbidity and mortality worldwide, affecting millions of people each year [1]. Despite advances in medical technology and treatment strategies, sepsis remains a significant healthcare challenge and there is an urgent need for more effective approaches to its management [2].

One promising approach to improving sepsis management is the use of Reinforcement Learning (RL), an Artificial Intelligence (AI) technique that involves training an agent to make decisions based on feedback from its environment. In the context of sepsis management, RL can be used to train an agent to make treatment decisions that optimize a specific goal, such as reducing mortality and improving patient outcomes.

In the literature, RL agents in sepsis treatment have been trained using a reward function that is based on mortality outcomes [3, 4]. This formulation rewards the agent for reducing the likelihood of the patient dying, the rationale behind this choice is that in this way the RL agent will minimize mortality [5]. However, recent research has provided strong recommendations on some aspects of sepsis management, such as a target mean arterial pressure (MAP) of 65 mmHg [6], so including these vital parameters in the reward formulation may lead to more effective treatment strategies learned by the RL agent.

MAP is a key physiologic parameter in sepsis management as it reflects the adequacy of tissue perfusion and the ability of the cardiovascular system to maintain blood flow to vital organs. In sepsis, hypotension and hypoperfusion can lead to organ dysfunction and failure, and MAP is a critical parameter to monitor and manage in the clinical setting. By providing feedback to the RL agent on MAP levels, it will be incentivized to maintain a target stable blood pressure, possibly leading to better patient outcomes.

In this study, we will explore the use of MAP reward formulation in sepsis treatment using reinforcement learning. We will compare the learned policies of an agent that only minimizes mortality and an agent that also considers MAP levels. We will also discuss the potential benefits and limitations of the two different reward formulations in sepsis management. In this study, we define the reward formulation of the Markov Decision Process (MDP) as "optimal" if the agent generated minimizes mortality rates among treated patients.

By analyzing the treatment decisions made by the two different RL agents, we can gain insight into the factors that contribute to better patient outcomes and identify areas where existing treatment strategies can be improved.

2. Methods

2.1. Data description

This study uses data from the Multi-parameter Intelligent Monitoring in Intensive Care (MIMIC III) database [7]. The database includes information from 53,423 hospital admissions of patients aged 16 years or older, collected between 2001 and 2012.

The cohort consists of individuals who meet the sepsis-3 criteria: if a patient had a microbiological specimen collected prior to antibiotic administration or within 24 hours of a previous antibiotic administration. In the case of microbiological specimens, sepsis is defined only if the antibiotic is administered within 72 hours of the specimen. The time of sepsis onset is determined as the time of the earliest event according to [8].

Patients were excluded if they were under 18 years of age at the time of enrollment, had no documented mortality or intravenous fluid (IV) administration, or had treatment discontinued due to nonrecovery. Treatment discontinuation was defined as those who died within 24 hours of the end of data collection and did not receive a vasopressor (VP) during the last 24 hours of data collection, but had received at least one previous administration.

The total number of ICU stays that meet the requirements is 20,496.

The data used in this study cover a period from 24 hours before the estimated onset of sepsis to 48 hours after the estimated onset of sepsis. The data are organized in 4-hour time intervals as a time series. When multiple observations were available for a variable in a given time interval, they were appropriately summarized by averaging or summing, depending on the variable type. The dataset includes 48 clinical variables including the two treatments of interest: IVs and VPs.

To address the problem of sparsity in clinical time series data, a zero-order interpolation approach is used. Remaining missing data are then imputed by interpolation. The data are divided into 80% for training and 20% for testing.

2.2. Markov Decision Process

The data are used to construct two different Markov Decision Processes (MDPs). At each time step, patients are assigned to a state using a k-means++ algorithm, resulting in a total of 750 clusters. Additionally, two absorbing states are included in the analysis, one for patient survival and the other for patient death.

The available treatments for IVs and VPs are binned to define the actions in the MDP. A total of five bins are created for each treatment type. One bin is reserved for zero dose, while the remaining four bins are defined by the 25%, median, and 75% percentiles. This process results in a total of 25 possible interventions.

Transitions that occur less than five times are excluded to force the RL agent to choose only the actions commonly used by clinicians, resulting in a safer policy.

In this work, we have formulated two different MDPs, which differ in the reward formulation:

- *Mortality Model*: In this MDP, the reward is always zero, except for the terminal states. If the model transitions to the "survived" state, the agent is rewarded +100, and if the model transitions to the "dead" state, the agent is penalized with a reward of -100.
- *Target MAP Model*: In this MDP, the model is rewarded zero if it visits a state with a MAP equal to or above the target of 65 mmHg, otherwise the agent is penalized with a negative reward $R = MAP_{current} - 65$. The agent also receives the reward linked to the final outcome of the patient, formulated in the same way as the other model; patient survival is rewarded with +100, patient death is penalized with a reward of -100.

2.3. Optimal Policy and Policy Evaluation

To determine the optimal policy in both the *Mortality Model* and the *Target MAP Model*, we used a policy iteration algorithm.

The policy iteration algorithm first evaluates a random policy, and then improves the policy by updating the optimal action for each state. This process must be repeated until we converge to the optimal policy [9].

Using temporal difference (TD) learning, we were able to estimate the policy value of the clinicians:

$$Q^{\pi}(s,a) \leftarrow Q^{\pi}(s,a) + \alpha(r + \gamma Q^{\pi}(s',a') - Q^{\pi}(s,a)) \tag{1}$$

r is the immediate reward, *s'* is the future state, *a'* is the future action, $Q^{\pi}(s,a)$ is the state-action value function, α is the learning rate and γ is the discount factor.

To evaluate the optimal AI policy on existing observations generated with the clinicians' policy, we used off-policy evaluation with weighted importance sampling (WIS). The clinicians' policy was considered as the behavior policy π_C, and the AI policy as the evaluation policy π_{AI}. The cumulative importance ratio up to step t was defined as $\rho_{1:t} := \prod_{t'=1}^{t} \pi_{AI}(a_{t'}|s_{t'})/\pi_C(a_{t'}|s_{t'})$, and its average at horizon t as $w_t = \sum_{i=1}^{N} \rho_{1:t}(i)/N$, where N is the number of trajectories. The trajectory-wise WIS estimator, $V_{WIS} = \frac{\rho_{1:T}}{w_T} \sum_{t=1}^{T} \gamma^{t-1} r_t$, is then averaged for all trajectories to derive the overall WIS estimator as $WIS = \frac{1}{N} \sum_{i=1}^{N} V_{WIS}^{(i)}$.

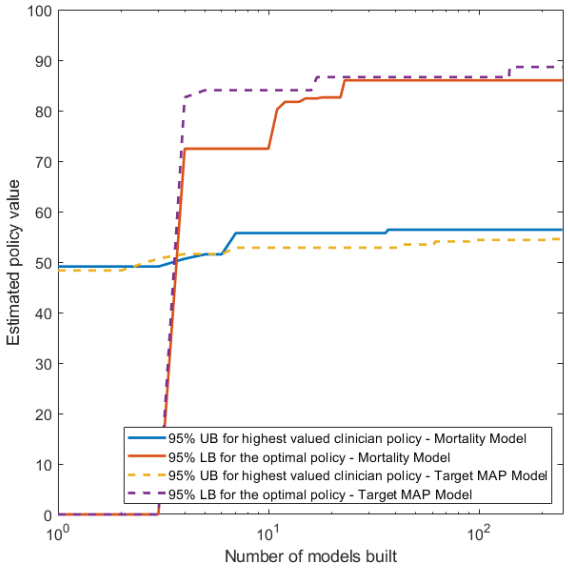

Figure 1. The figure shows the evolution of the 95% LB of the optimal RL policy and the 95% UB of the clinician's policy for both built models.

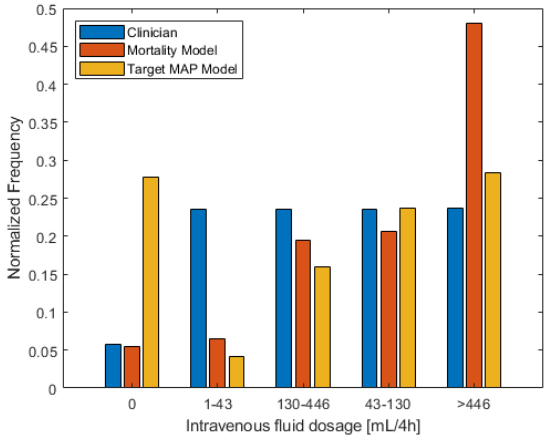

Figure 2. Normalized frequencies of the intravenous fluid dosages selected by the RL agent to treat patients with MAP $<$ 65 mmHg for both the models built and the clinicians' choices.

3. Results

To evaluate the effectiveness of the optimal policies of the *Mortality Model* and the *Target MAP Model*, we plotted the evolution of the 95% Confidence Lower Bound (LB) of the policy value during training, which represents the safety and effectiveness of the estimated optimal policy, and compared it to the 95% Confidence Upper Bound (UB), which is an optimistic estimate of the clinicians' policy performance.

Figure 1 shows the performance of the optimal policy of the *Mortality Model*. The 95% confidence value of the optimal policy, shown with a solid orange line, converges to a value of 86.01, while the clinician's policy, shown with a solid blue line, obtains a value of 56.40. The performance of the optimal policy of the *MAP model* is also shown in figure 1 with a dashed purple line and reaches a value of 88.64, while the clinician's policy in the *Target MAP model* is shown with a dashed yellow line and reaches a value of 54.55.

To assess the difference between the two strategies in managing patients with a MAP equal to or less than 65 mmHg, we visualized the normalized frequencies of actions performed on patients with a MAP equal to or less than 65 mmHg in the test set.

Figure 2 shows the normalized frequencies of IVs for the *Mortality Model*, *Target MAP Model* and clinicians' policy.

The *Target MAP Model*, uses far fewer IVs, avoiding them 27.8% of the time, while the *Mortality Model*, avoids IVs only 5.5% of the time, similar to the clinicians. The latter model also uses the maximum dosage of IVs 48% of the time.

The *Target MAP Model* has an interesting behavior, using moderate doses of VPs (0.08-0.20 μg/kg/min) more sparingly, only 13.7% of the time, while the *Mortality Model* uses them 39.2% of the time.

The *Target MAP Model* tends to have a more aggressive treatment strategy, using high doses ($>$0.45 μg/kg/min) 13.7% of the time and moderately high doses (0.20-0.45 μg/kg/min) 23%, while the *Mortality Model* uses them only 8.6% and 12.6%, respectively. Clinicians tend to avoid VPs, not using them 67.3% of the time.

4. Discussion

The goal of this study is to compare the effect of different reward formulations on the optimal policy learned by the AI agent. Some authors in the literature state that putting all the reward on mortality is desirable because in this way the RL agent should only minimize mortality, the rationale behind this is that formulating a reward linked to any other variable actually puts undesirable constraints on the agent that reduce its freedom to find the optimal treatment strategy [5].

However, a reward assigned to mortality results in a sparse reward, and the agent receives feedback on its actions only at the end of the trajectory. This makes it very difficult to assign credit to specific actions, since with the sparse reward we are evaluating the sequence of actions that lead to the final outcome (mortality), rather than the single action taken in a particular state.

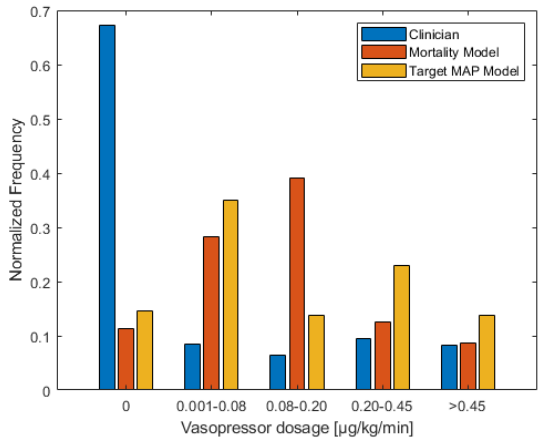

Figure 3. Normalized frequencies of the vasopressors dosages selected by the RL agent to treat patients with MAP < 65 mmHg for both the models built and the clinicians' choices.

To address this issue, we propose to link the reward of the MDP to mean arterial pressure (MAP), a physiological parameter considered important in the management of sepsis by clinical guidelines [6] and previous work in the literature has also shown the primary importance of cardiovascular signals (including MAP) in the application of RL to sepsis treatment strategies [10, 11].

The *Target MAP Model*, despite having a more punitive reward formulation, achieves slightly better performance than the *Mortality Model*, showing that intermediate rewards have the power to provide timely feedback to the RL agent before the final step of the trajectory.

To gain further insight into the choice of agents when treating patients with a MAP < 65 mmHg, we plotted the normalized frequencies of actions for both the *Target MAP Model* and the *Mortality Model*. The former model uses intravenous fluids more sparingly, prioritizing the administration of high doses of VPs (> 0.20 μg/kg/min). Since the *Target MAP Model* is the best performing model, this suggests that using aggressive doses of VPs to bring MAP to a target level of 65 mmHg may be desirable.

Frequent use of high dosages of IVs of the *Mortality Model* may be undesirable because hypotensive patients do not fully respond to IVs later in the ICU stay, worsening their condition and causing edema [2], which partially explains the worse performance than the *Target MAP Model*.

5. Conclusion

In conclusion, this study shows that formulating the reward function using prior medical knowledge is not necessarily worse than linking the entire reward to the outcome of interest, in this case mortality. Formulating rewards using cardiovascular signals also has the advantage of providing immediate feedback to the agent and, thanks to the high temporal resolution of cardiovascular variables, allows the evaluation of individual actions rather than a collection of actions.

References

[1] Dugani S, Veillard J, Kissoon N. Reducing the global burden of sepsis. Canadian Medical Association Journal Jan 2017;189(1):E2–E3.

[2] Gotts JE, Matthay MA. Sepsis: pathophysiology and clinical management. BMJ May 2016;i1585.

[3] Raghu A, Komorowski M, Celi LA, Szolovits P, Ghassemi M. Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach. In Proceedings of the 2nd Machine Learning for Healthcare Conference, volume 68. 2017; 147–163.

[4] Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nature Medicine Oct 2018;24(11):1716–1720.

[5] Liu S, See KC, Ngiam KY, Celi LA, Sun X, Feng M. Reinforcement learning for clinical decision support in critical care: Comprehensive review. Journal of Medical Internet Research Jul 2020;22(7):e18477.

[6] Evans Lea. Surviving sepsis campaign: International guidelines for management of sepsis and septic shock 2021, Oct 2021.

[7] Johnson AE, Pollard TJ, Shen L, Lehman LwH, Feng M, Ghassemi M, Moody B, Szolovits P, Anthony Celi L, Mark RG. Mimic-iii, a freely accessible critical care database, May 2016.

[8] Seymour CW, et al. Assessment of Clinical Criteria for Sepsis: For the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA 02 2016; 315(8):762–774. ISSN 0098-7484.

[9] Sutton RS, Barto AG. Reinforcement learning: An introduction. MIT press, 2018.

[10] Mollura M, Drudi C, Lehman LW, Barbieri R. A reinforcement learning application for optimal fluid and vasopressor interventions in septic icu patients. In 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). 2022; 321–324.

[11] Mollura M, Drudi C, Lehman LW, Barbieri R. Optimal fluid and vasopressor interventions in septic icu patients through reinforcement learning model. In 2022 Computing in Cardiology (CinC), volume 498. 2022; 1–4.

Address for correspondence:

Cristian Drudi
Via Camillo Golgi, 39, 20133 Milano MI
cristian.drudi@polimi.it