Optimal Fluid and Vasopressor Interventions in Septic ICU Patients Through Reinforcement Learning Model

Maximiliano Mollura[1], Cristian Drudi[1], Li-wei Lehman[*2], Riccardo Barbieri[1]

[1] Politecnico di Milano, Milano, Italy
[2] Massachusetts Institute of Technology, Cambridge, Massachussetts

Abstract

Introduction: *Fluids and vasopressors represent the cornerstone for hemodynamic instability management in the intensive care unit (ICU). However, optimal personalized treatments strategies are still missing.* ***Goal:*** *To evaluate the ability of a reduced set of cardiovascular features in determining optimal actions with a reinforcement learning approach.* ***Methods:*** *Data were extracted from the MIMIC-III database Patients' trajectories were modeled as a Markov decision process with a target reward based on 90-day mortality. Performances with a reduced set of cardiovascular features (CARDIO), including heart rate, systolic and diastolic blood pressure, shock index, and oxygen saturation were compared with a random policy model (RANDOM) and a model with a full set of 48 clinical variables including physiologic, laboratory measurement, and ventilation parameters (FULL).* ***Results:*** *The CARDIO model achieved the highest results with a 95% lower bound (LB) of estimated policy value equal to 96.17 compared with the 86.00 obtained from the FULL model and 82.62 from the RANDOM policy model.* ***Conclusions:*** *Results show that cardiovascular features and ongoing treatments have the potential to determine the optimal dosage of fluids and vasopressors for septic patients when using reinforcement learning tools for the development of medical decision support systems.*

1. Introduction

Sepsis is an important global health problem and it is the among the most common causes of in-hospital deaths with an approximate cost of more than $24 billion annually in the United States, an incidence of about 48.9 million of cases in 2017, and with an average mortality of 19.7% [1]. Sepsis is defined as a dysregulated host response to infection and its final stage, the septic shock, is considered one of the major problems in intensive care units (ICU) with a reported mortality of about 40% [2, 3].

Timely administration of crystalloids and vasopressors are the most important interventions to deal with sepsis-induced hemodynamic instability like patients' hypothension and tissue hypoperfusion [4]. However, clinical guidelines provide only indications of the best general strategy while criteria for determining the optimal personalized strategies are still largely debated [5, 6].

In this context, the application of reinforcement learning (RL), the branch of machine learning which aims at developing models able to determine the optimal decision given a specific objective function, gained increasing interest especially in the ICU context [7]. RL applications were already developed in literature for: tracking the optimal glycemic level [8], optimizing weaning time from mechanical ventilation [9], continuous optimization of morphine dosage [10], and for optimizing the execution of laboratory tests [11].

Recently Komorowski et al. published an inspirational study, addressing the lack of personalized strategies for fluid and vasopressor dosage [12], where the authors proposed a RL model to determine the optimal treatment strategy for septic patients admitted to the ICU. The proposed policy, referred to as 'AI policy', determined the optimal dosages of fluids and vasopressors in order to minimize 90-day patients' mortality.

The goal of this study is to assess the ability of a minimal set of easily measurable cardiovascular variables with information about ongoing treatments in being used as main drivers for estimating optimal treatment strategies with a reinforcement learning approach.

2. Methods

2.1. Cohort Selection, Data Extraction and Preprocessing

Data used for this study were extracted from the MIMIC-III [13], an open-access database publicly available on PhysioNet [14] and containing data from 61,532 admissions in the intensive care units of the Beth Is-

[*] L.L. is partially supported by the NIH grant 5R01 EB017205

rael Deaconess Medical Center between 2001 and 2012. We first started by reproducing Komorowski's results and therefore followed their inclusion criteria for selecting the data. Briefly, we considered adult patients with sepsis, which was defined according to the third definition of sepsis [15] as the increase in sequential organ failure assessment (SOFA) score ≥ 2 and the contemporaneous prescription of antibiotics and sampling of bodily fluids for microbiological culture thus reflecting the required criteria which consists in the presence of organ failure and suspicion of infection.

We excluded subjects less than 18 years old, without reported mortality or intravenous fluid intake, and patients with stopped vasopressor treatments that would have died in the next 24 hours because for this patients the reason to stop treatments was mainly due to a so high illness severity that any high dosage treatment was considered futile.

The resulting 20,496 ICU admissions with also a sepsis onset estimate were included in our study. We collected 48 variables with 4-hour time steps from 24 hours prior the onset up to 48 hours after the estimated onset time. Considered variables included demographics (e.g. age, gender, weight, Elixhauser comorbidity index) physiologic variables (e.g. heart rate and systolic blood pressure), laboratory measures (e.g. pH, lactate and white blood cells count), ventilation parameters (e.g mechanical ventilation and FiO_2) and medications (e.g. fluids and vasopressors)

2.2. Reinforcement Learning Model Description

Reinforcement learning, a sub-field of artificial intelligence (AI), consists in a computational agent that learns a set of rules for taking decisions, referred to as 'policy', that in turn would maximize a specific reward function. The agent learns optimal treatment by trial-and-error procedures performed on a subset of observations drawn from an environment that given specific information about its current state generates a return depending the action performed by the agent. Therefore, differently from supervised and unsupervised learning approaches, which try to learn or find out the rules that link a specific set of observations obtained by an expert with the outcomes of interest, RL learns the optimal actions that maximize a reward signal obtained by interacting with the environment [16].

A Markov decision process (MDP) was used to model the patient environment and trajectories and policy iteration was used to solve it and to estimate actions maximizing the expected 90-day patients' survival. Each MDP process is characterized by:

- A finite set of states (S).
- A finite set of actions (A) for a given state $s \in S$.
- The probability of moving from a state s at time t to a new state s' at $t+1$ given the action $a \in A$, $(T(s', s, a))$
- The reward obtained when moving to s', $(R(s'))$
- The discount factor (γ), which goes from 0 to 1 and determines the importance of future rewards.

S was set to 750 distinct states determined by k-means++ clustering on the set of available measures plus two absorbing terminal states, i.e. death and survival. A included a discrete set of 25 possible actions as the combination of 5 distinct fluids and vasopressors dosages. R for each trajectory was set to +100 in case of a positive terminal state (survival) and -100 in case of negative state (death). Finally, (γ) was set to 0.99, indicating that a late death will have similar importance of an early death.

The state-action value function, Q^π, represents the expected sum of discounted rewards of a specific action taken in a specific state and following policy π for each MDP as follows.

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

where α is the learning rate and r the immediate reward. Therefore, Q^π quantifies the result of the chosen treatment strategy on patients' mortality.

The AI's policy corresponding to maximum state-action value $\pi(s)^*$ was estimated using in-place policy iteration, so maximizing the long-term sum of rewards representing expected patients' survival. The estimation criteria can be formulated as follows:

$$\pi(s)^* \leftarrow \arg\max_a Q^{\pi^*}(s,a) \forall s \tag{2}$$

The state-value estimated for state s following the policy π thereafter, $V^\pi(s)$, can computed according to the Bellman equation follows:

$$V^\pi(s) = \sum_a \pi(s,a) \sum_{s'} T(s',s,a)[R(s') + \gamma V^\pi(s')] \tag{3}$$

Finally, in order to evaluate the estimated optimal AI policy on the existent observations which follow clinicians' policy we used the off-policy evaluation with weighted importance sampling (WIS) as proposed by Komorowski et al., considering the clinicians' policy as the behavior policy π_C, and the AI policy as 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 $w_t = \sum_{i=1}^{N} \rho_{1:t}(i)/N$, with N as 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 in order to derive the overall WIS estimator as:

$$WIS = \frac{1}{N} \sum_{i=1}^{N} V_{WIS}^{(i)} \tag{4}$$

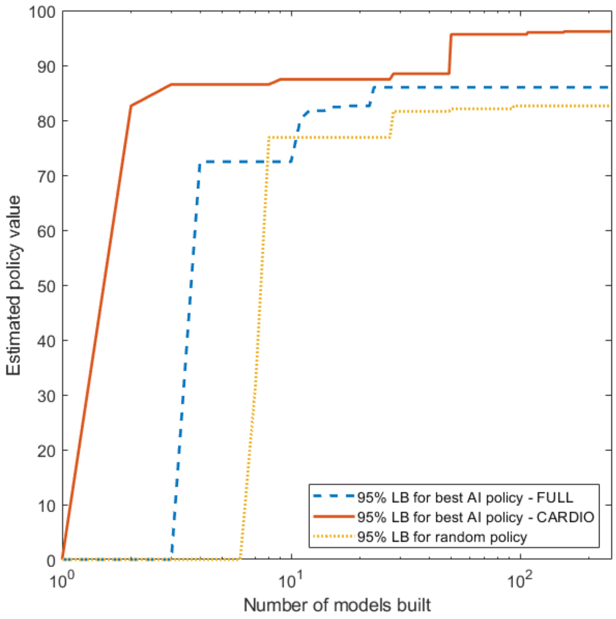

Figure 1. Estimated 95% lower bound of the 250 AI policies by using the whole feature set (orange dotted line), the subsets of cardiovascular variables and the random policy.

2.3. Reduction of the Feature Space

In this study we compare the performance obtained by reducing the feature space on features because including such large set of features might be computationally expensive, we explored the performances of the model when selecting a subspace of the features according to the following variants:

- *CARDIO* Model: reduction of the feature space by keeping only 6 cardiovascular variables: heart rate, systolic blood pressure, dyastolic blood pressure, shock index, SpO_2 and mechanical ventilation.
- *RAND* Model: Policy that takes action randomly.
- *FULL* Model: Model reproduced by Komorowski et al. [12] using 48 features.

We then computed the WIS estimator obtained on 250 different models obtained by training the clustering algorithm with a randomly selected 80% of the data (train set) and testing with the remaining 20% of data.

The 95% lower bounds (LB) of 2000 bootstrapped resamplings of the WIS estimators between the different AI models were compared in Figure 1.
The distributions of the mean bootstrapped WIS estimators obtained averaging the 2000 resamplings for each of the 250 trials are compared in Figure 2.

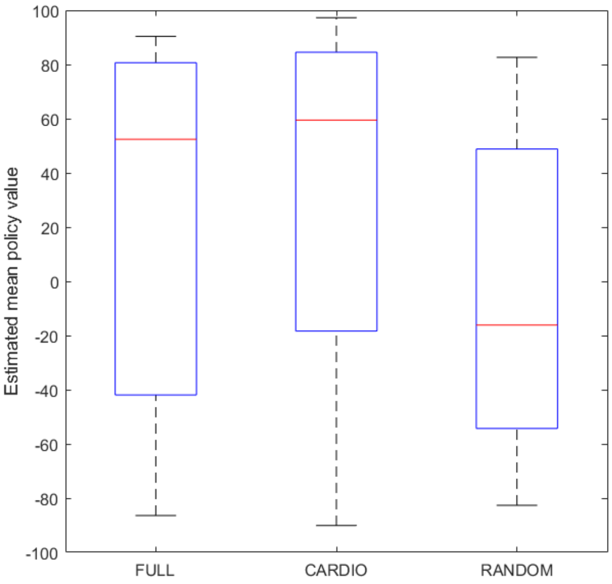

Figure 2. Boxplots of obtained mean policy values for the evaluated models over the 250 trials.

3. Results

The evolution of the best 95% LB of the models over the 250 trained models can be seen in Figure 1.
The *CARDIO* model performance is the solid orange line. After 3 models built the *CARDIO* model achieves the highest policy value 95% LB with 86.5286. The 95% LB of the model further increases and settles at the 95% LB value of 96.1722 at model number 156.
The *FULL* model performance is represented by the dashed blue line. The *FULL* model performance obtain its first positive 95% LB at trial 4 with 72.4545 and keeps increasing while training additional models, stabilizing at the value of 86.0051 at model number 23.
The *RANDOM* policy performance is represented by the yellow dotted line and it is used as baseline model. At model number 8 the random policy achieves its first positive 95% LB with the value of 76.875 and it also rapidly increases while training additional models, stopping at the value of 82.6169 at model number 94.
95% LB metrics obtained on the random policy show that all RL policies obtained higher results. Between the two AI policies the *CARDIO* model is the best performing according to the 95% LB.
In figure 2 the boxplots of mean policy values obtained over the 250 trials are shown for all model families.
The median of the *CARDIO* model is the best one with a value of 59.4721, the *FULL* model follows with a median value of 52.422. The random policy median has the lowest value with a median of -16.122.
The *CARDIO* model also shows the highest first and third

quartiles equal to -18.3422 and 84.5178, the *FULL* model resulted in good first and third quartiles (-41.9404 and 80.6584) whereas the random policy showed the lowest values of -54.2741 and 48.8.

4. Discussions

The goal of this study is to compare the performance of a RL model, in estimating optimal treatments for septic patients admitted in ICU, by comparing the ability of a reduced set of physiologic features mainly related to the cardiovascular system with the results obtained by a random policy model and by Komorowski et al. who firstly showed the potential of this approach in the ICU. Despite we thoroughly followed the published methods, we observed a difference in the number of extracted ICU admissions with respect to the original study. However, general population statistics agree with those presented by Komorowski and expected for a population of septic patients, and the obtained results with the full set of features were similar to those obtained by the authors. The higher results obtained with the reduced set of cardiovascular features show that the reduction of the feature space on just a small subset of physiologic variables improves model performance. Moreover, the proposed features can be easily and continuously recorded at the patients' bedside, thus giving the possibility to continuously optimize the treatment strategy.

5. Conclusions

In conclusion, our study shows that reinforcement learning models can benefit form a reduction of the feature space, and that physiologic variables related to the cardiovascular system, and easily measurable at the patients' bedside, contain key information for determining the optimal treatment strategy in terms of fluid and vasopressors administrations, thus giving the possibility to continuously monitor and optimize patients' treatments in the ICU.

References

[1] K. E. Rudd et al., “Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study,” The Lancet, vol. 395, no. 10219, pp. 200–211, Jan. 2020, doi: 10.1016/s0140-6736(19)32989-7

[2] Singer M. et al., “The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3),” JAMA, vol. 315, no. 8, p. 801, Feb. 2016, doi: 10.1001/jama.2016.0287.

[3] Driessen R.G.H. et al., “The influence of a change in septic shock definitions on intensive care epidemiology and outcome: comparison of sepsis-2 and sepsis-3 definitions,” Infectious Diseases, vol. 50, no. 3, pp. 207–213, Sep. 2017, doi: 10.1080/23744235.2017.1383630.

[4] Levy M.M. et al., ”The Surviving Sepsis Campaign Bundle: 2018 update.”, Intensive Care Med, 44, pp. 925–928, 2018, doi: 10.1007/s00134-018-5085-0.

[5] Malbrain M.L.N.G. et al., ”Fluid overload, de-resuscitation, and outcomes in critically ill or injured patients: a systematic review with suggestions for clinical practice.”, Anaesthesiol Intensive Ther. 2014 Nov-Dec; 46(5), pp. 361-80, doi: 10.5603/AIT.2014.0060. PMID: 25432556.

[6] Glassford N.J., et al., “Physiological changes after fluid bolus therapy in sepsis: a systematic review of contemporary data”, Critical care 18(6), pp. 696, Dec. 2014

[7] Liu S. et al., ”Reinforcement Learning for Clinical Decision Support in Critical Care: Comprehensive Review”. J Med Internet Res., 20;22(7):e18477, Jul 2020, doi: 10.2196/18477%. PMID: 32706670; PMCID: PMC7400046.

[8] Weng W. et al., ”Representation and reinforcement learning for personalized glycemic control in septic patients”, arXiv Preprint, arXiv:1712.00654, 2017

[9] Prasad N. et al., ”A reinforcement learning approach to weaning of mechanical ventilation in intensive care units”, arXiv Preprint, arXiv:1704.06300

[10] Lopez-Martinez D. et al., ”Deep reinforcement learning for optimal critical care pain management with morphine using dueling double-deep Q networks”. Conf Proc IEEE Eng Med Biol Soc 2019

[11] Cheng L. et al., ”An optimal policy for patient laboratory tests in intensive care units”. Pac Symp Biocomput, 24, pp. 320-331, 2019

[12] Komorowski M. et al. ”The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care”. Nat Med 24, pp. 1716–1720, 2018, doi: 10.1038/s41591-018-0213-5

[13] Johnson A.E.W. et al., “MIMIC-III, a freely accessible critical care database,” Sci Data, vol. 3, no. 1, May 2016, doi: 10.1038/sdata.2016.35. https://doi.org/10.1038/sdata.2016.35

[14] Goldberger A.L. et al., “PhysioBank, PhysioToolkit, and PhysioNet,” Circulation, vol. 101, no. 23, Jun. 2000, doi: 10.1161/01.cir.101.23.e215.

[15] 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. https://doi.org/10.1001/jama.2016.0287

[16] Sutton R. S. et al., ”Reinforcement learning: An introduction”, 2nd edn, MIT press, 2018.

Address for correspondence:

Maximiliano Mollura
Via Camillo Golgi, 39 Milano, Italy
maximiliano.mollura@polimi.it