Improving Treatment Decisions for Sepsis Patients by Reinforcement Learning

by

Ruishen Lyu

BS in Preventive Medicine, China Medical University, China, 2018

Submitted to the Graduate Faculty of the

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2020

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Ruishen Lyu

It was defended on

April 17, 2020

and approved by

Thesis Advisor: Lu Tang, PhD, Assistant Professor, Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

Thesis Co-Advisor: Chung-Chou H. Chang, PhD, Professor, Departments of Medicine and Biostatistics
School of Medicine and Graduate School of Public Health, University of Pittsburgh

Committee Members: Florian Mayr, MD, MPH, Assistant Professor, Department of Critical Care Medicine
School of Medicine, University of Pittsburgh

Lu Tang, PhD

Chung-Chou H. Chang, PhD

Improving Treatment Decisions for Sepsis Patients by Reinforcement Learning

Ruishen Lyu, MS

University of Pittsburgh, 2020

Abstract

Sepsis is defined as a dysregulated immune response to infection leading to acute life-threatening organ dysfunction. Patients with sepsis have 25.8% intensive care unit (ICU) mortality, which was significantly higher than in the general ICU population. Making optimal medication decisions becomes an emergent and important task. The purpose of this study is to develop a data-driven decision-making tool that can dynamically suggest optimal treatments for each individual ICU patient with sepsis, and help clinicians make better treatment decisions to improve patients' long-term survival outcomes.

Model-free Q-learning was applied to data extracted from the eICU Research Institute (eRI) database. We selected 3,800 patients admitted to ICUs with septic shock and summarized their first 7 days of lab results and vitals into 18,014 daily records. To identify best treatment decisions of vasopressor use, we first clustered patients' demographics and daily medical conditions into 100 distinct states. We then mapped ICU survival to time-dependent rewards and estimated the Q-values for each action taken at each state using temporal-difference learning. Finally, we obtained the optimal policy that maximizes the action-value function by policy iteration. An off-policy evaluation method was implemented to evaluate the performance of several treatment policies.

The result showed that the Q-learning policy has significantly higher long-term average reward than the clinician policy or the random policy, meaning that patients who received treatments matching those suggested by the Q-learning policy had better survival outlook than those who did not.

In conclusion, we showcased that the prospect of long-term survival may be improved through using modern reinforcement learning methods that optimizes the rewards against the dynamics of the environment.

Public Health Significance: We developed a data-driven automatic reinforcement learning tool and applied it to an electronic health database of sepsis patients. The result showed that medical decisions that matched our Q-leaning policy led to better survival outlook; this suggests that machine learning can be used to help clinicians decide the most effective treatments and reduce the burden on medical and economical resources.

Table of Contents

List of Tables

List of Figures

Preface

I want to thank my advisors, Dr.Tang and Dr.Chang's help to me, not just for my master thesis but also for my personal development. I couldn't finish the thesis without their patient guidance and constructive advice. I cannot express enough my appreciation to them.

I also want to thank the department of Biostatistics. I feel I am so grateful that I can have the opportunity to learn here and begin my journey in the field of Biostatistics. Thanks to all the professors that taught me before and I worked with. I am truly blessed that I can learn from them. Also, I want to thank my classmates and my friends who gave me a lot of support in the last two years.

1.0 Introduction

Sepsis is defined as a dysregulated immune response to infection leading to life-threatening acute organ dysfunction (Seymour et al., 2019). It is also a common global health issue associated with a high mortality rate and a low quality of life after survival. Based on a study from the Intensive Care Over Nations (Vincent et al., 2014), 29.5% of patients had sepsis on admission to the intensive care unit (ICU). The study also noted that one out of four patients with sepsis died in the ICU and one out of three could not survive until hospital discharge. With increased sepsis severity, rising costs and worsening outcomes show a great burden in economic and epidemic perspectives (Paoli, Reynolds, Sinha, Gitlin, & Crouser, 2018).

Most of the patients (72%) with sepsis before admission had comorbidities such as respiratory tract, urinary tract or gastrointestinal infections (Novosad et al., 2016). There are different severities of sepsis ranging from systemic inflammatory response syndrome (SIRS), sepsis, severe sepsis, to septic shock. The 28-day mortality increases from approximately 10% with SIRS to 40%-60% with septic shock (Brun-Buisson, 2000). There is a high mortality rate during the first 48 hours after admission and a high incidence of multiple organ failures at diagnosis, which suggests there are delays in the initial diagnosis of sepsis and starting antibiotic therapy (Blanco et al., 2008).

Making optimal medication decisions for patients with different medical conditions and different characteristics is a challenge to date. The severity of diseases and heterogeneity of medical conditions of patients in the ICU requires more advanced and precise medical decisions (Ng et al., 2018). Also, it is important for clinicians to keep in mind the long-term goals of curing sepsis, rather than providing treatments that only offer immediate relief of the patients' symptoms.

The progression of the disease can develop rapidly for patients while in the ICU, so it is important for clinicians to make an optimal decision in a short period of time. Therefore, quick and personalized treatments with the best long-term outcomes are in high demand.

Reinforcement learning is a class of machine learning methods that can be used to help sequential decision making toward predefined long-term terminal goals (Burkov, 2019). Machine learning methods usually involve training an agent to learn and react to an environment through trial and error. Since an action taken by the agent could potentially have different effects on individuals over time, we designed the agent to be adaptive and immediately responsive to change during such interactions. By adjusting rewards for different actions in different conditions, the learning agent can learn and develop rules for achieving the long-term goal.

Q-learning is a reinforcement learning method that can solve sequential decision-making problems by estimating the Q-values, which represent the overall quality, or more rigorously defined as the expected cumulative rewards, of different actions made in each of the different states. There are model-based Q-learning methods, such as Markov Decision Process (MDP), and non-model-based Q-learning methods, such as Temporal-Difference (TD) learning and deep Q-learning. The research conducted by Komorowski et al., (2018) adopted MDP to establish an artificial intelligence agent to suggest decisions to sepsis patients. In order to solve the sequential decision-making problem, the MDP decomposes the learning procedure into estimating the state transition, and the reward value is dependent only on the current state and applied action.

MDP is best applicable to deterministic environments where transition and reward are known in advance. However, in most real world settings, state transition and reward value are probabilistic and cannot be estimated precisely based on data collected from non-controlled settings, which is especially the case in ICUs (Pardo, Tavakoli, Levdik, & Kormushev, 2017). The

model-free Q-learning method can be used to train an agent without needing to estimate the reward and state transition functions. Examples of model-free methods include the Monte Carlo method and the TD method. Rather than being limited in the observed state-action pairs, they can explore all possibilities of state-action pairs and estimate the long-term reward that will be received by choosing an action in certain state. Model-free Q-learning can develop a 'critic' by measuring the 'quality' of different state-action pairs and identify the optimal action-selection policy that maximizes the long-term reward (Watkins, 1989).

By using reinforcement methods, clinicians can address issues resulted from the heterogeneity of patients and the delayed indications of the efficacy of treatments. Reinforcement learning can be designed to seek higher long-term returns over the immediate rewards so that the resulting policy will maximize the final outcome of patients. Reinforcement learning methods have been applied to solving a number of medical problems and provided the optimal suggestions for clinicians and patients (Komorowski et al., 2018; Saria, 2018; Srinivasa Rao & Diamond, 2020).

The purpose of this study is to develop a data-driven decision-making tool that can dynamically suggest optimal treatments for each ICU patient with sepsis, and help clinicians make better treatment decisions to improve patients' long-term survival outcomes. Section 2 (Method) introduces the dataset and the model-free Q-learning method applied in the study. Section 3 (Result) presents the results from the data analysis, and Section 4 (Discussion) concludes the study based on the results.

2.0 Method

2.1 Data

Our data was extracted from the publicly available eICU Research Institute Database (eRI) (Pollard et al., 2018), which is a multi-center data set comprised of de-identified electronic health data with over 200,000 admissions to ICUs across the United States from 2014-2015. The database includes vital sign measurements, care plan documentation, severity of illness measures, diagnosis information, and treatment information. We selected 3,800 patients with septic shock based on the ICD-9 diagnosis code (785.52) within the first 24 hours of admission. For each patient, we summarize their daily information over the first 7 days since hospital admission, which resulted in a total of 18,014 daily records. The in-hospital survival status was also extracted. The final analytic data set includes 25 variables of patients' demographics, Elixhauser premorbid status, vital signs, and laboratory values. The treatments received in each daily record of the individual patients were also included in the dataset. The observations with age less than 18 were excluded. The observations with no documentation in the discharge outcome or incomplete daily records were also excluded. After applying exclusion criteria, we used 17,825 records of 3,726 patients in the later model building.

2.2 Data Preparation

We examined the above-mentioned 25 variables including the distribution characteristics, missing percentage, data range, and outliers. The temperature unit was unified to Fahrenheit. We used multivariate imputation by chained equations (MICE) with 5 multiple imputations and 5,000 maximum iterations to impute the missing values in the dataset (Azur, Stuart, Frangakis, & Leaf, 2011). For the purpose of clustering, all variables were converted to normality if necessary. Variables related to erythrocyte sedimentation rate (ESR), C-reactive protein (CRP), and troponin concentrations were excluded because of high percentage of missingness or unadjusted abnormality.

Treatment of sepsis patients with vasopressors is our major interest in the study. There are different types of vasopressors. Based on the administration of vasopressors, we categorized treatments into first-line vasopressors (norepinephrine and epinephrine), second-line vasopressors (phenylephrine and vasopressin), and others (Stratton, Berlin, & Arbo, 2017). Due to unbalanced distributions of different treatments and the clinical significance, we identified four possible treatment actions related to the use of vasopressors, including: 1) first-line only (NE): norepinephrine or epinephrine, 2) second-line only (PVO): phenylephrine, vasopressin, dopamine or others, 3) first-line and second-line (NEPVO) norepinephrine or epinephrine with phenylephrine, vasopressin or others, and 4) no treatment used (None).

Unsupervised clustering was used to divide patients' clinical conditions into a finite set of more homogeneous groups. Based on the clustering results, we assigned each patient at each day to a cluster so that the patient's data point has the shortest distance to the center of that cluster as compared with the distance to the center of any other clusters. The K-means clustering method was used with 1,000 random starts, maximum iteration of 5,000 times, and the predefined number

of clusters $K = 100$. We further ranked the 100 clusters based on their in-hospital mortality rates (Table 1), which is the proportion of the number of patients who died over the total number of patients in each cluster. The lower ranks represent higher mortality rates and higher ranks represent lower rates.

Table 1 Mortality rates for states

State	Mortality Rate	Sample size	Mortality ranking	State	Mortality Rate	Sample Size	Mortality Ranking
15	77.8%	63	1	98	18.7%	75	51
85	62.5%	80	2	31	18.1%	144	52
45	50.9%	112	3	6	17.7%	164	53
39	50.0%	42	4	62	17.6%	102	54
21	43.7%	135	5	68	16.8%	173	55
4	41.3%	104	6	67	16.1%	155	56
80	41.0%	78	7	16	16.1%	211	57
86	40.6%	96	8	75	16.1%	205	58
19	36.9%	122	9	10	15.6%	192	59
28	35.5%	172	10	27	15.5%	161	60
35	34.4%	131	11	63	15.3%	131	61
14	33.3%	150	12	88	14.8%	81	62
24	33.3%	78	13	87	14.8%	169	63
72	33.3%	30	14	43	14.7%	170	64
74	33.3%	141	15	66	14.4%	153	65
60	33.1%	133	16	52	13.9%	202	66
100	32.9%	82	17	44	13.7%	139	67
57	32.7%	150	18	2	12.9%	178	68
5	31.4%	102	19	11	12.8%	109	69
92	30.9%	175	20	23	12.7%	142	70
20	30.0%	150	21	91	12.7%	150	71
97	29.5%	105	22	81	12.3%	171	72
71	29.2%	120	23	7	12.3%	163	73
56	28.5%	151	24	83	12.1%	190	74
40	28.2%	85	25	41	11.7%	188	75
55	27.9%	61	26	82	11.6%	215	76
95	27.6%	127	27	69	11.6%	121	77
33	26.7%	172	28	18	11.3%	194	78
49	26.4%	159	29	12	10.8%	120	79
22	26.3%	156	30	64	10.4%	249	80
58	25.6%	117	31	65	10.2%	118	81
76	24.2%	157	32	8	10.1%	237	82

Table 1 Continued

50	24.1%	162	33	99	9.9%	162	83
42	23.4%	124	34	13	8.9%	224	84
34	23.4%	154	35	84	8.8%	160	85
37	23.3%	146	36	73	8.2%	183	86
51	23.1%	143	37	25	7.5%	201	87
53	22.7%	141	38	96	7.4%	149	88
48	22.5%	173	39	36	7.3%	123	89
61	22.1%	145	40	29	7.3%	193	90
32	22.0%	182	41	54	6.7%	165	91
38	22.0%	164	42	46	6.6%	136	92
30	21.7%	161	43	78	6.5%	169	93
93	21.6%	148	44	70	4.4%	180	94
9	21.0%	214	45	79	4.4%	181	95
17	20.7%	111	46	94	3.6%	137	96
77	20.0%	130	47	90	3.4%	148	97
47	19.3%	114	48	3	3.0%	164	98
1	19.0%	163	49	26	3.0%	101	99
89	19.0%	195	50	59	0.7%	146	100

2.3 Basics of Reinforcement Learning

Reinforcement learning has been used to solve the sequential decision-making problem in order to maximize the long-term goal. In our study, we use the reinforcement technique to determine the optimum daily vasopressor use for each patient that will maximize the patient's in-hospital survival. There are three key entities in the setting: state, reward, and action, that enable the agent to interact with the environment and develop a policy. We specify these three entities and other key elements below.

Time: The study period was divided into equally spaced discrete intervals since hospital admission. Each patient has up to 7 daily records.

State: We categorized patients' medical conditions into 100 unique states which were determined based on the K-mean clustering method by using the 25 demographics, Elixhauser

premorbid status, vital signs, and laboratory values for all 17,825 records of 3,726 patients. From time t to time $t+1$, a patient may transition from one state to another. Notation S_t and S_{t+1} represent a patient's current state (at time t) and the next state (at time $t+1$). Table 2 shows the reinforcement learning input dataset for patient $i, i = 1, \ldots, 3726$. Each patient has at most 6 transitions because the value of the 'next state' for the last time point is unknown. In Table 2, the example patient who survived the first 7 days has 6 records in the end.

Reward: At each time t, we assigned reward R_t to be 0 for all patients except for the last record. For the last record, we assigned a reward of 100 to those patients who survived in hospital and -100 to those patients who died in hospital (see Table 2). The rationale for giving zero rewards before the final transition is to rule out the possibility of an agent learning some treatment policy that only provides short-term conditional relief to patients but fails to gain long-term improvement in survival.

Action: The action at each time t for each patient, A_t refers to the vasopressor the patient received at that day. Based on the vasopressor administration guidelines (Stratton et al., 2017), we categorized different vasopressors into the following 4 combinations of treatments: first-line vasopressors only (NE: norepinephrine or epinephrine), second-line vasopressors only (PVO: phenylephrine, vasopressin, or others), combination of the first-line and the second-line vasopressors (NEPVO), and no vasopressor (None).

Table 2 Example transition of a patient who survived the first 7 days

Patient ID	Time	State	Next State	Reward	Action
00001	1	60	45	0	NE
00001	2	45	70	0	NE
00001	3	70	16	0	PVO
00001	4	16	25	0	PVO
00001	5	25	16	0	NE
00001	6	16	9	100	None

Generally speaking, state, reward, and action can be defined as either continuous or discrete. We applied reinforcement learning method with discretized state and time defined by homogeneous patient conditions (aka clusters) and daily time windows, but we will also discuss the possibility of extending to the continuous patient states.

2.4 Q-learning

Agent and Environment: The agent is one who takes actions based on the rewards in different states. The environment is what the learning agent interacts with and learns from. Different states, rewards, and actions shape the environment that limit the learning agent (Sutton & Barto, 2018). As shown in Figure 1, the interactions between the learning agent and the environment is continuous as if in a sequential situation. The learning agent selects actions and the environment responds with corresponding states and rewards. The intrinsic characteristics of reinforcement learning drives the agent toward yielding the most optimal policy that maximizes the cumulative rewards it receives in the long run.

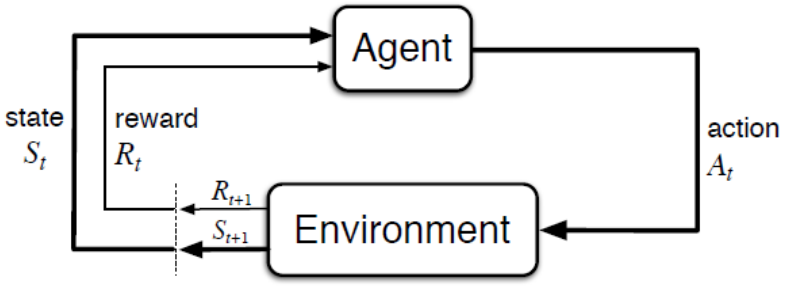

Figure 1 Interaction between Agent and Environment

(Sutton & Barto, 2018)

Policy: The policy is the set of rules that the learning agent develops to select an action in a state. The optimal policy is the rule that is recommended to follow for achieving the most optimal long-term reward. In this study, we will compare three different policies: clinician policy, random policy, and the optimal policy where clinician policy is based on clinicians' selection of vasopressors for each patient at each day; random policy is based on the random treatment selection from NE, PVO, NEPVO, and None with equal chance; and optimal policy was the best treatment selection based on the reinforcement Q-learning procedure.

Policy Value: Given a policy, the policy value represents the expected future reward collected at each state, or for each state-action pair. In our study, the value for each state is a vector of dimension 100; and the value for state-action pair is a matrix of 100 by 4. This value function also suggests possible improvement to the current policy, leading to an update of the current policy. In Q-learning, the mapping from states, or state-action pairs, to values are commonly known as the Q-value function, hence the name Q-learning for the algorithm to estimate the Q-value function. In this study, the policy value is the value of the policy that was estimated to be the survival after receiving vasopressors from the Q-learning policy.

Markov property assumes that all aspects of the past agent-environment interaction that make a difference for the future are included in the current states (Sutton & Barto, 2018). In other words, how a patient may respond to a treatment under certain conditions (*aka* state) is probabilistically fixed. By breaking the continuing interactions into different episodes with $\{s, a, r, s'\}$, the sequential problem was decomposed into a memoryless series. The next state s' and the reward value r depend only on the current state s and the applied action a. In the dataset, we defined each patient's information at each time period as $\{s, a, r, s'\}$, and the collection of information across all time periods of a patient as an episode. An episode summarizes the trajectory

of a patient's hospital stay. It describes the actions in states and corresponding reward from day 1 up to day 7 leading to the survival outcome of every patient. Q-learning method was used to solve the sequential decision-making problems by estimating the Q-values, the expected cumulative rewards, associated with different actions made in different states.

There are model-based Q-learning methods, such as Markov Decision Process (MDP), and non-model-based Q-learning methods, such as temporal-difference learning (TD-learning). Both MDP and TD learning methods assume that state, reward, and action can only take a discrete and a finite set of values. Time horizon can be finite or infinite depending on real applications. MDP is a specific formalism to describe an environment's dynamics and a decision problem to solve within. It is the benchmark reinforcement learning method applicable in many classical settings, such as solving a maze. MDP is a model-based learning method that requires knowing the dynamics between states, actions and corresponding rewards. However, when such environment dynamics are not well defined, as in the case of treatment-response dynamics, MDP requires fully observable state-action pairs $\{s, a\}$ to estimate the model for the environment. This is not suitable in real clinical scenarios because there are circumstances when certain types of treatments are strictly prohibited for certain types of patients, thus no observed data on the rewards. Therefore, such a complicated setting with different combinations of treatments and states is not enough information for the MDP. Even with a very large dataset, the observable state-action pairs $\{s, a\}$ may still be insufficient to describe all possible combinations of the many treatment choices and heterogeneous patient states. On the contrary, TD-learning was used to bypass the need of a model. It directly approximates the Q-values using weighted average of past experiences (Sutton & Barto, 2018):

$$q_k(s, a) = (1 - \alpha_k) q_{k-1}(s, a) + \alpha_k (r + \gamma q_{k-1}(s', a'))$$

TD-learning approximates the Q-values by using the past value of itself. The Q-values of the actions in different states that the learning agent made are computed. The higher Q-values mean the patients will receive higher rewards if they take certain actions in different states. By exploring all possibilities of state-action pairs, TD-learning can estimate the long-term rewards that patients will receive by choosing an action in a certain state. The resulting state-action value function $Q^{\pi}(s,a)$ is a mapping of the expected sum of rewards starting from a certain state s, taking the action a, and thereafter following policy π (Bennett & Hauser, 2013; Sutton & Barto, 2018). The estimation was conducted based on the transition data consisting of four elements, including 'State', 'Action', 'Reward', and 'Next State', $\{s,a,r,s'\}$, for different daily records of 3,726 patients, 14,099 records in total.

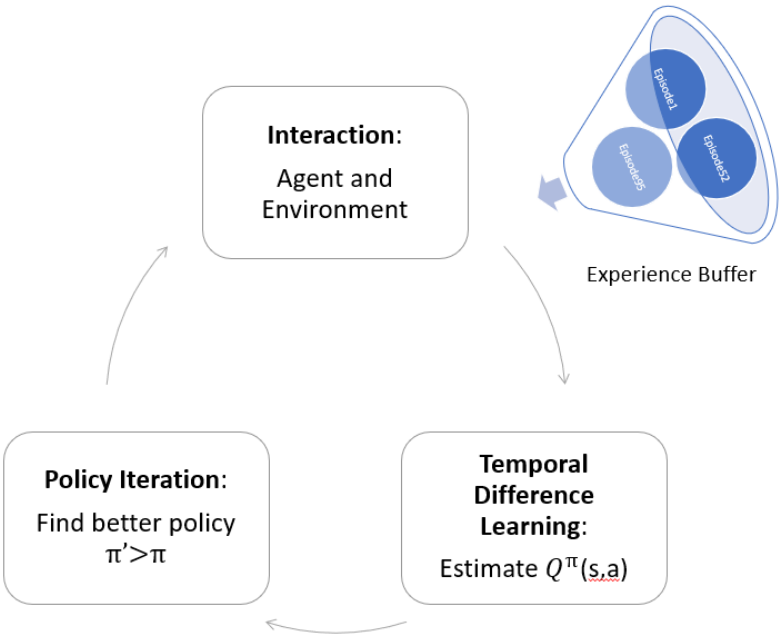

Figure 2 Flowchart of policy building

By observing the actual policy followed by clinicians, the learning agent learns from the state-action pair episodes $\{s,a\}$ from the clinician policy to develop the iteration and generate the most optimal policy π^{*}. The policy iteration involved initialization, policy evaluation and policy

improvement. The policy iteration starts with a random policy and compares different policies based on the action value function $Q^{\pi}(s,a)$, and finds another new policy based on the previous action value function. In the end, the policies converge to the most optimal policy (Sutton & Barto, 2018).

$$\pi^{*}(s) \leftarrow arg\max_{a} Q^{\pi^{*}}(s,a), \qquad \forall s$$

The policy iteration algorithm requires pre-specification of three controlling parameters, learning rate α, discount factor γ, and the probability of random actions ε. The learning rate can control how quickly the learning agent adapts to the random changes imposed by the environment. The discount factor $\gamma \in [0,1)$ quantifies how much importance is given to future rewards. In the study, we set gamma to be 0.99, which is high enough to encourage the learning agent to consider future rewards rather than immediate rewards. The ε-greedy method was used for actions exploration in the study with ε as 0.1. The ε means that the learning agent would exploit the treatments to obtain maximum immediate reward with a 90% probability, and there is a 10% probability that the learning agent would explore other treatments in order to maximize the long-term reward. The ε will decay during the iteration, because after exploring the other possible actions, it will exploit the known actions more in the same state (Sutton & Barto, 2018).

We used the R package ***ReinforcementLearning*** with the experience replay algorithm, which creates a buffer that stores different 'experience' when the learning agent interacts with the environment. The experiences are different episodes with $\{s,a,r,s'\}$, from different policies. The experience buffer allows the learning agent to randomly select a batch of data from the buffer, which not only can reduce time of policy iteration and increase the data efficiency, but also improve the training performance with lower variance based on more diverse batch data (Mnih et

al., 2015). The iterations were set as 5,000 to make the Q-value converge to the maximum value for each action in the study.

2.5 Model Evaluation

Off-policy evaluation was conducted to measure the performance of the Q-learning policy based on the patient trajectories observed under the clinician policy. The value of a policy is defined in terms of the likelihood of survival under the proposed policy. Since difference policies could lead to different distributions of observable patient populations, we can view the data generated from two different policies as two population samples. In order to quantify the rewards a patient receives hypothetically if they follow the proposed policy, the rewards were reweighted by importance sampling weights. This is a common technique used in causal inference for generalizing results from one population to another. The hypothetical sample through reweighing mimics the patient population arising from the proposed Q-learning policy (Jiang & Li, 2016; Thomas, Theocharous, & Ghavamzadeh, 2015). In other words, the observed rewards from the clinician policy will be reweighted to approximate the rewards from the proposed policy of interest. Using this technique, we evaluate both the Q-learning policy and the random treatment policy.

To calculate the importance sampling weights for a proposed policy, first, the per-step importance ratio was calculated: $\rho_t^{(i)} = \pi_1(a_t^{(i)}|s_t^{(i)})/\pi_0(a_t^{(i)}|s_t^{(i)})$, where $\pi_1(a_t\,|s_t)$ denotes the probability of choosing treatment a_t at state s_t based on the proposed policy; and the $\pi_0(a_t|\,s_t)$ denotes the probability of choosing treatment a_t at state s_t based on the existing clinician policy. Although policy probabilities are insensitive to time, they are trajectory specific. To evaluate the

Q-learning policy, π_1 will be the derived policy from the Q-values of our learning algorithm; to evaluate the random policy, π_1 will be a policy with the same probability, 0.25, for choosing every treatment in every state. For a given patient i, a cumulative per-step importance ratio up until time t was calculated by multiplying the ratios of daily records, $\rho_{1:t}^{(i)} = \prod_{t'=1}^{t} \rho_{t'}^{(i)}$. A normalizing quantity w_t for cumulative ratio at time horizon t was estimated by averaging $\rho_{1:t}$ from subjects with trajectories of at least length t. By definition, $w_t = \sum_{i=1}^{|D_t|} \rho_{1:t}^{(i)} / |D_t|$, where $|D_t|$ is the number of subjects with trajectories of at least length t. In our setting, $w_t, t \in \{1, \dots, 6\}$, can only take 6 unique values. The weighted reward of each trajectory based on weights of importance sampling (WIS) for patient i was estimated by

$$V_{WIS}^{(i)} = \frac{\rho_{1:H^{(i)}}^{(i)}}{W_{H^{(i)}}} \left(\sum_{t=1}^{H^{(i)}} \gamma^{t-1} r_t^{(i)}\right),$$

where $H^{(i)}$ is the number of records the patient had. Patients with different histories of visits have different weights of importance sampling. As previously defined, γ is the discount factor with value 0.99, r_t is the reward the patient received at each time point. The average reward of all patients was estimated by

$$WIS = \frac{1}{n} \sum_{i=1}^{n} V_{WIS}^{(i)} .$$

We then evaluated the performance of the clinician, random, and reinforcement learning policies by comparing their average trajectory-wise rewards.

By using bootstrapping with the off-policy evaluation performed 50 times, the true distribution of the policy value was estimated (Hanna, Stone, & Niekum, 2017). The values of the Q-learning policy, the clinician policy, and the random policy were constructed and compared by using the one-way analysis of variance at the significance level of 5%. The post-hoc test with

Bonferroni correction was performed to compare the difference between each pair of the policy values.

3.0 Results

Below we show results on four domains. The first is a descriptive summary that shows the original variables we included in the study from the dataset and the variables after standardization and imputations. Next, we present the means of variables in different survival ranks to demonstrate the state identification. Third, we present the differences between the clinician policy and the Q-learning policy. Last, we present a quantitative evaluation and comparison of Q-learning policy, clinician policy, and random policy. Our model-free Q-learning approach produces the most optimal policy that allows patients to receive the maximum long-term rewards.

3.1 Descriptive Statistics

Table 3 shows the summary statistics of the 25 variables of patients; demographics, Elixhauser premorbid status, vital signs, and laboratory values from the original dataset. The frequency and percentage of missing data for each of the variables are presented at the last column of each table.

In preparation for subsequent clustering analysis, we first excluded erythrocyte sedimentation rate (ESR), C-reactive protein (CRP), and troponin concentrations because of high percentage of missingness. We then standardized the rest of the 22 variables by subtracting its mean and then dividing by its standard deviation (SD). The missing data of the standardized variables were then imputed using the MICE method with 5 replications. The descriptions of the final imputed data are summarized in Table 5.

An overall description of vasopressor use was summarized in Table 3. The no treatment (None) option was used the most (51.3%), while the first-line vasopressor only (NE) was the least prescribed treatment option by clinicians. The description of treatment frequency over 7 days is shown in Table 4. The frequency of first-line vasopressor (NE) decreases from day 1 to day 7, while the frequency of no vasopressor treatment (None) increases from day 1 to day 7. Among the 3,726 patients included in this study, the in-hospital mortality rate was 22.8%.

Table 3 Descriptive statistics for the baseline variables

	Missing	(%)	Mean or Median	(SD) or (IQR)		Min	Max
Normally distributed continuous variables, mean (SD)							
Age	860	4.80%	64.9	15.2		18	89
Bicarbonate	2314	13.00%	24.1	5		3	47
Chloride	757	4.20%	107.2	6.9		74	155
RR	16807	94.30%	20.5	7.2		0	50
Sodium	722	4.10%	140.4	5.6		116	179
Temperature	15742	88.30%	37.1	0.6		32.2	41.2
Heart rate	4221	23.70%	112.4	22.3		43	300
Systolic BP	13233	74.20%	156.6	36.4		-6	351
Nonnormally distributed continuous variables, median (IQR)							
Albumin	8509	47.70%	2.3	2	2.8	0.6	6
Bands	15512	87.00%	8	3	17	0	92.5
ALT	9665	54.20%	33	18	84	3	10940
AST	9613	53.90%	41	22	100	3	31664
BUN	772	4.30%	28	16	45	1	261
Creatinine	725	4.10%	1.2	0.8	2.2	0.1	15.7
CRP*	17576	98.60%	17.6	8.4	479	0.2	4723
ESR*	17584	98.60%	49	28	72	1	139
Glucose	803	4.50%	128	102	167	6	1791
Hgb	1236	6.90%	9.6	8.6	10.9	4.4	19.5
Lactate	13218	74.20%	2.1	1.3	4	0.2	36.7

Table 3 Continued

O2 sat	12476	70.00%	97	95	99	12	101
PaCO2	11826	66.30%	39.3	33.3	47	11.6	161.6
Platelets	1402	7.90%	163	100	239	2	1102
INR	12215	68.50%	1.5	1.3	2.2	0.9	23.6
Troponin*	17578	98.60%	0.1	0.1	0.6	0	26.9
WBC	1368	7.70%	12	8.2	17.6	0	402.8
Categorical variables, n(%)							
			n	**(%)**			
Female			8447	0.474			
Treatment							
NE			1113	0.062			
PVO			3185	0.179			
NEPVO			4377	0.246			
None			9150	0.513			

Abbreviations: ALT: alanine aminotransferase, AST: aspartate aminotransferase, BUN: blood urea nitrogen, Hgb: hemoglobin, O2 sat: oxygen saturation, PaCO2: partial pressure of carbon dioxide, INR: prothrombin time international normalized ratio, RR: respiratory rate, WBC: white blood cell count, BP: blood pressure.
*Variables CRP, ESR, and Troponin were excluded.

Table 4 Treatment frequency over 7 days

	N (Probability)						
	Day 1	**Day 2**	**Day 3**	**Day 4**	**Day 5**	**Day 6**	**Day 7**
NE	530 (14.3%)	296 (8.9%)	134 (4.7%)	68 (2.7%)	45 (2.1%)	20 (1.1%)	20 (1.3%)
PVO	551 (14.9%)	567 (17.1%)	564 (19.6%)	463 (18.7%)	403 (18.9%)	353 (19.5%)	284 (18.8%)
NEPVO	1472 (39.8%)	1118 (33.7%)	693 (24.1%)	432 (17.4%)	296 (13.9%)	211 (11.7%)	155 (10.3%)
None	1143 (30.9%)	1339 (40.3%)	1488 (51.7%)	1515 (61.1%)	1393 (65.2%)	1224 (67.7%)	1048 (69.5%)
Total	3696	3320	2879	2478	2137	1808	1507

Table 5 Descriptive summary of variables after standardization and imputation

	Mean	(SD)	Median	(IQR)		Min.	Max.
Age	0.01	1.00	0.14	-0.59	0.79	-3.09	1.58
Albumin	0.00	1.00	0.00	-0.54	0.76	-5.16	3.68
ALT	-0.05	0.96	-0.26	-0.74	0.39	-2.04	4.12
AST	-0.05	0.96	-0.25	-0.73	0.36	-2.19	4.77
Bands	-0.07	1.00	-0.03	-0.85	0.68	-2.30	2.47
Bicarbonate	-0.01	1.00	-0.02	-0.61	0.58	-4.18	4.54
BUN	0.00	1.00	0.10	-0.63	0.72	-4.28	3.03
Chloride	0.00	1.00	-0.03	-0.61	0.69	-4.79	6.90
Creatinine	0.00	1.00	-0.14	-0.73	0.70	-3.51	3.33
Glucose	0.10	1.10	-0.51	-0.52	0.38	-0.53	4.61
Hgb	0.13	1.04	0.18	-0.74	0.81	-1.63	2.74
Lactate	0.00	1.00	-0.11	-0.71	0.59	-8.20	6.86
O2 sat	0.00	1.00	-0.07	-0.72	0.67	-4.65	4.08
PaCO2	-0.25	0.91	-0.33	-0.93	0.24	-2.99	3.31
Platelets	0.05	1.01	0.01	-0.84	0.59	-2.66	4.79
INR	0.03	0.99	0.00	-0.59	0.57	-4.37	4.93
RR	0.00	1.00	0.00	-0.68	0.66	-2.79	5.01
Sodium	-0.07	0.96	-0.34	-0.83	0.41	-1.46	5.75
Temperature	-0.08	1.02	-0.34	-0.76	0.49	-2.83	4.08
WBC	0.00	1.00	-0.06	-0.60	0.64	-4.31	6.84
Heart rate	-0.01	0.97	-0.12	-0.12	-0.12	-8.09	6.89
Systolic BP	-0.09	0.53	-0.29	-0.32	-0.18	-0.34	12.14

Abbreviations: ALT: alanine aminotransferase, AST: aspartate aminotransferase, BUN: blood urea nitrogen, Hgb: hemoglobin, O2 sat: oxygen saturation, PaCO2: partial pressure of carbon dioxide, INR: prothrombin time international normalized ratio, RR: respiratory rate, WBC: white blood cell count, BP: blood pressure.

3.2 State Identification via K-mean Clustering

Scatter plots with loess lines and the heatmap of means of medical vitals for states with different survival ranks are shown in Figure 3 and Figure 4, respectively. Based on the scatter plots and the heatmap, variables such as platelets and bicarbonate have higher values with higher ranks, while variables such as creatinine and BUN have lower values with higher ranks. The results show that the clustering has good performance that can clearly separate patients into different states based on their demographic information and medical conditions.

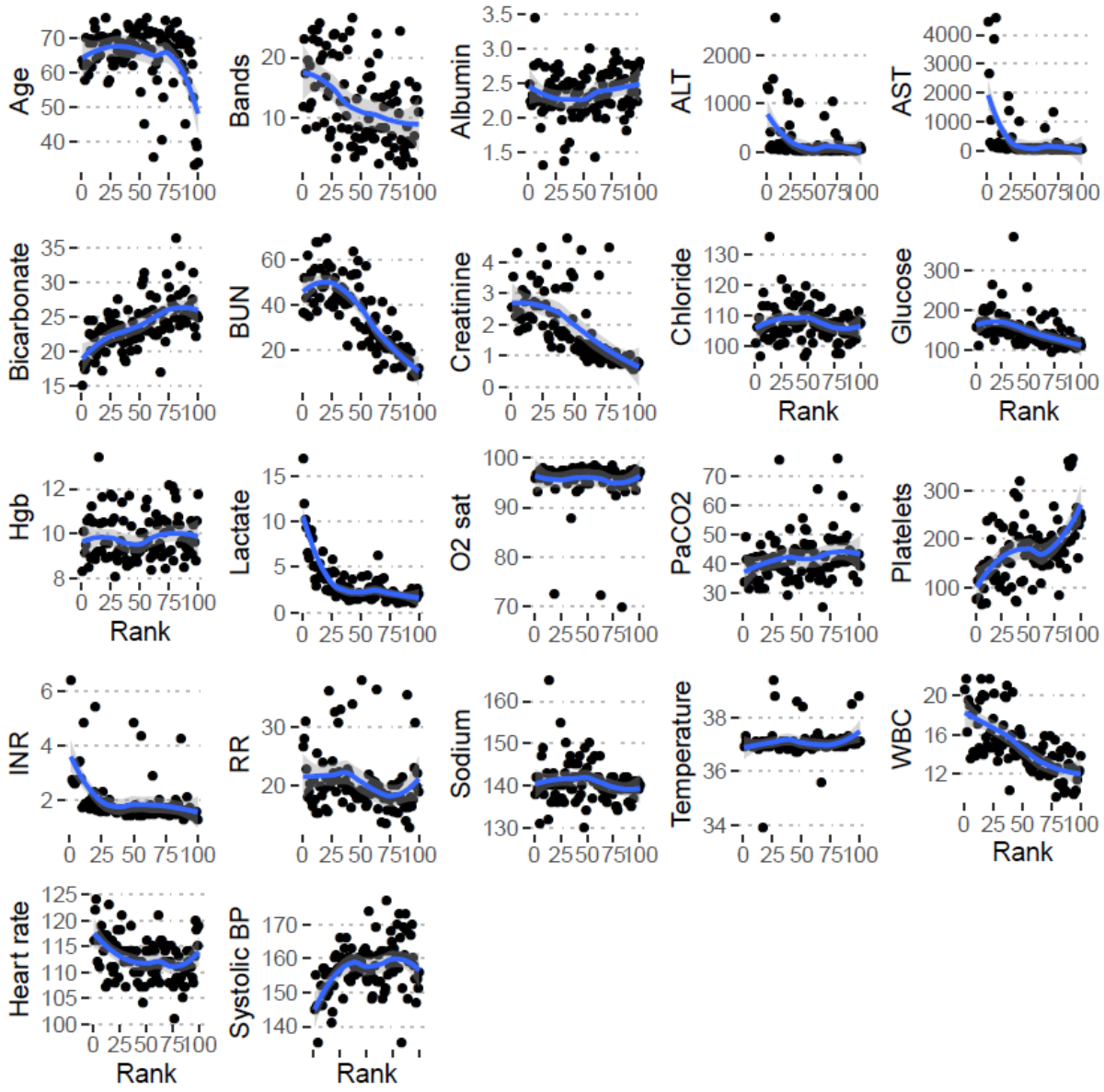

Figure 3 Means of medical conditions in different ranked states

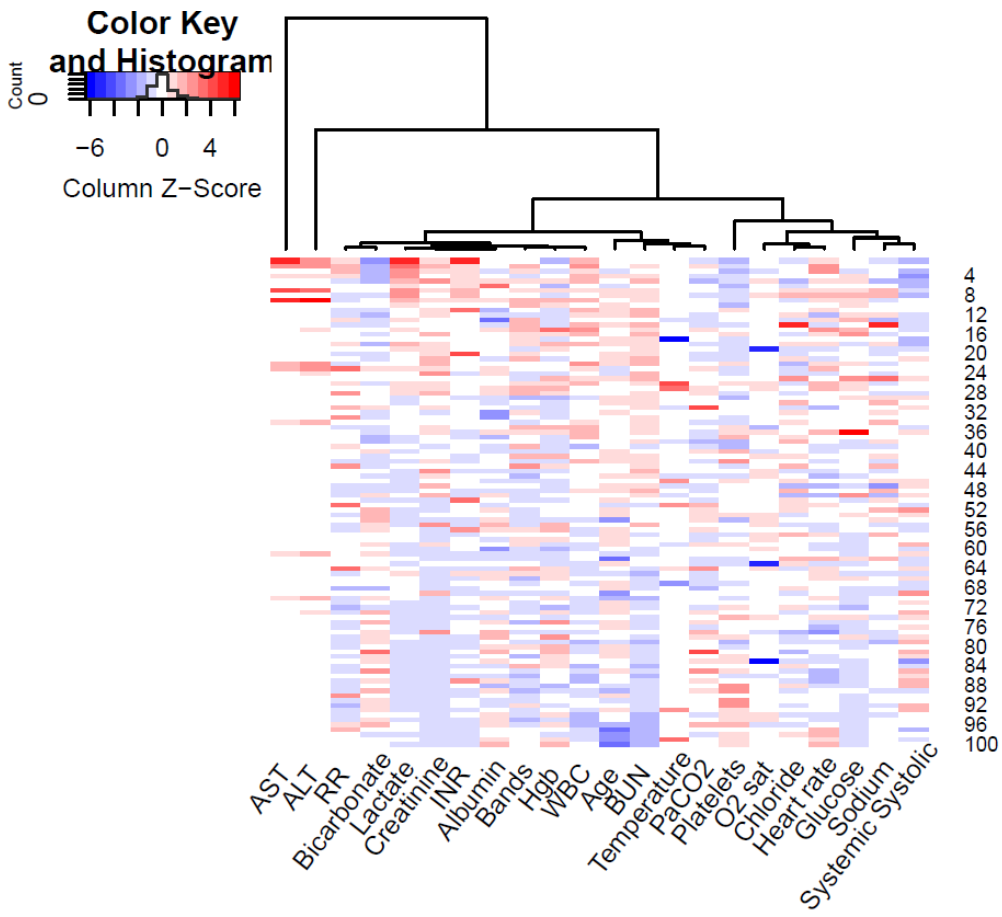

Figure 4 Heatmap of medical vitals means

3.3 Optimal Policy Identification via Q-learning

The table with Q-values, Q-learning policy and clinician policy is shown in Table 6. The Q-values for choosing different treatments in different states were estimated. Higher Q-values represent higher overall expected reward, survival, after patients receive certain treatments in different states. The final Q-learning policy was conducted by choosing the treatment with the highest Q-value in each state. The probability of treatments from Q-learning policy in different

states was converted by the Q-values, while the probability of treatment from clinician policy was also calculated in the table.

Table 6 Q-values, Q-learning policy, and clinician policy for different states[1]

Rank	NEPVO (Q)	PVO (Q)	NE (Q)	None (Q)	NEPVO (C)	PVO (C)	NE (C)	None (C)	Q-learning Policy	Clinician Policy
1	2963.2	2949.0	3012.0	3004.3	0.64	0.08	0.03	0.25	NE	NEPVO
2	2978.9	3010.1	3096.0	2979.7	0.52	0.15	0.05	0.28	NE	NEPVO
3	3009.6	2997.0	3045.5	3010.7	0.51	0.12	0.07	0.30	NE	NEPVO
4	2988.1	3031.2	3000.9	3009.2	0.50	0.11	0.05	0.34	PVO	NEPVO
5	2992.2	2993.5	3006.8	2998.1	0.53	0.10	0.10	0.27	NE	NEPVO
6	3000.3	3018.1	3007.6	3020.6	0.51	0.20	0.04	0.25	None	NEPVO
7	2978.5	3030.3	3073.2	3020.1	0.45	0.14	0.02	0.39	NE	NEPVO
8	3006.0	3017.4	3017.4	2981.0	0.54	0.21	0.04	0.21	PVO	NEPVO
9	3006.4	3014.9	3042.6	3022.6	0.42	0.2	0.05	0.32	NE	NEPVO
10	3011.5	3013.7	3026.1	3025.7	0.46	0.15	0.04	0.35	NE	NEPVO
11	3005.2	3022.3	3019.6	3030.8	0.36	0.15	0.07	0.42	None	None
12	3004.3	3005.3	3001.4	3004.9	0.39	0.18	0.07	0.36	PVO	NEPVO
13	3002.9	3025.1	2988.8	3012.1	0.20	0.14	0.10	0.56	PVO	None
14	3012.7	3047.3	3017.3	3023.6	0.37	0.22	0.13	0.28	PVO	NEPVO
15	3023.7	3014.1	2977.7	3035.2	0.52	0.16	0.05	0.27	None	NEPVO
16	2999.7	3030.8	3004.4	3015.8	0.46	0.13	0.09	0.32	PVO	NEPVO
17	3010.1	3009.0	3035.9	3042.3	0.36	0.18	0.07	0.39	None	None
18	3003.0	3029.2	3004.4	3011.3	0.48	0.12	0.10	0.29	PVO	NEPVO
19	2997.9	3012.6	3054.0	3030.3	0.34	0.24	0.03	0.39	NE	None
20	2986.5	3025.7	3018.3	3029.7	0.34	0.16	0.10	0.40	None	None
21	2992.7	3039.9	3015.0	2993.2	0.38	0.15	0.08	0.40	PVO	None
22	3020.1	3023.1	3009.5	3015.5	0.34	0.16	0.10	0.40	PVO	None
23	3014.4	3007.5	3024.1	3013.4	0.40	0.20	0.03	0.38	NE	NEPVO
24	3017.5	3004.1	3006.2	3015.7	0.27	0.21	0.03	0.48	NEPVO	None
25	3013.2	3017.4	3029.7	3013.3	0.18	0.31	0.06	0.45	NE	None
26	3013.0	3028.2	3004.2	3017.7	0.35	0.23	0.05	0.37	PVO	None
27	3006.2	3034.0	3034.8	3022.8	0.42	0.18	0.04	0.36	NE	NEPVO
28	2998.2	3042.2	3024.8	3043.3	0.41	0.14	0.07	0.38	None	NEPVO
29	3034.4	3033.6	3017.7	3025.6	0.31	0.16	0.08	0.44	NEPVO	None
30	3007.9	3007.6	3042.0	3007.0	0.22	0.24	0.05	0.48	NE	None
31	3016.8	3047.4	2972.8	3034.3	0.28	0.12	0.07	0.54	PVO	None
32	2998.3	3051.6	3033.1	3053.3	0.27	0.16	0.04	0.52	None	None

1 Q: Q-values; C: Probability derived from clinician policy.

Table 6 Continued

33	2996.0	3022.8	3016.2	3026.4	0.28	0.12	0.11	0.49	None	None
34	3008.9	3014.8	3007.9	3034.5	0.34	0.21	0.07	0.38	None	None
35	3014.9	3019.8	3029.9	3030.9	0.40	0.14	0.09	0.37	None	NEPVO
36	3019.6	3014.0	3021.8	3036.0	0.42	0.22	0.06	0.30	None	NEPVO
37	3011.4	3022.3	3022.6	3010.0	0.30	0.13	0.08	0.50	NE	None
38	3010.6	3022.2	3010.4	3016.5	0.24	0.23	0.11	0.42	PVO	None
39	3021.8	3036.5	3018.1	3033.5	0.20	0.17	0.09	0.54	PVO	None
40	3023.9	3039.8	3031.2	3048.6	0.20	0.26	0.07	0.47	None	None
41	3019.2	3019.5	3010.5	3021.2	0.28	0.22	0.06	0.44	None	None
42	3009.9	3008.2	3016.3	3033.6	0.33	0.17	0.09	0.41	None	None
43	3032.6	3027.5	3016.2	3036.5	0.31	0.28	0.07	0.33	None	None
44	3011.9	3027.4	3011.9	3031.9	0.29	0.19	0.09	0.43	None	None
45	3006.8	3009.7	3015.7	3016.6	0.30	0.22	0.06	0.41	None	None
46	3029.0	3033.1	3028.5	3064.0	0.25	0.23	0.04	0.48	None	None
47	3000.4	3033.1	3016.9	3018.5	0.24	0.18	0.12	0.46	PVO	None
48	3008.8	3031.1	3012.1	3035.1	0.27	0.21	0.08	0.43	None	None
49	3012.8	3011.9	3033.5	3047.0	0.27	0.15	0.08	0.50	None	None
50	3034.3	3025.1	3042.3	3035.1	0.17	0.22	0.09	0.51	NE	None
51	2985.5	3015.7	3038.7	3074.4	0.24	0.17	0.08	0.51	None	None
52	3018.7	3038.2	3026.8	3032.2	0.20	0.32	0.02	0.46	PVO	None
53	3034.9	3023.6	3051.5	3040.9	0.21	0.26	0.02	0.50	NE	None
54	3033.4	3042.4	2929.4	3048.7	0.18	0.34	0.01	0.46	None	None
55	3015.2	3012.0	3020.6	3015.9	0.25	0.25	0.08	0.42	NE	None
56	3027.6	3052.6	3056.1	3052.2	0.20	0.21	0.04	0.55	NE	None
57	3030.3	3034.0	3037.6	3038.1	0.21	0.21	0.05	0.52	None	None
58	3030.4	3031.7	3044.8	3064.4	0.21	0.25	0.07	0.47	None	None
59	3021.7	3038.1	3012.0	3055.0	0.24	0.20	0.07	0.50	None	None
60	3025.3	3035.1	3031.9	3035.8	0.20	0.15	0.09	0.55	None	None
61	3021.1	3040.6	3030.6	3033.7	0.20	0.19	0.02	0.58	PVO	None
62	3018.0	3049.7	2997.9	3041.5	0.42	0.19	0.02	0.37	PVO	NEPVO
63	3012.2	3005.3	3004.0	3022.1	0.22	0.17	0.06	0.55	None	None
64	3021.2	3029.1	3035.8	3059.9	0.19	0.15	0.04	0.62	None	None
65	3040.3	3040.5	3025.8	3068.3	0.17	0.13	0.04	0.67	None	None
66	3018.8	3040.7	3048.7	3038.9	0.22	0.19	0.08	0.51	NE	None
67	3013.6	3035.3	3056.5	3055.1	0.19	0.16	0.04	0.61	NE	None
68	2998.4	3045.6	3016.3	3036.8	0.20	0.14	0.10	0.56	PVO	None
69	3021.3	3062.5	3009.2	3050.3	0.24	0.18	0.04	0.55	PVO	None
70	3016.6	3060.1	3027.8	3067.0	0.21	0.16	0.02	0.61	None	None
71	3033.0	3037.1	3050.9	3065.5	0.18	0.16	0.08	0.58	None	None
72	3020.6	3016.3	3035.9	3038.2	0.15	0.19	0.08	0.59	None	None
73	3027.7	3053.7	3023.1	3060.2	0.14	0.25	0.02	0.60	None	None
74	3032.3	3054.8	3044.4	3040.3	0.25	0.17	0.08	0.50	PVO	None
75	3030.1	3065.6	3047.2	3062.5	0.11	0.24	0.04	0.61	PVO	None

Table 6 Continued

76	3021.6	3041.6	3031.4	3047.0	0.20	0.17	0.09	0.54	None	None
77	3020.3	3024.3	3031.3	3052.0	0.23	0.11	0.13	0.53	None	None
78	3014.3	3040.4	3026.1	3056.5	0.19	0.19	0.04	0.58	None	None
79	3047.1	3046.8	3040.8	3052.9	0.17	0.11	0.04	0.69	None	None
80	3038.5	3034.2	3058.3	3053.9	0.16	0.15	0.06	0.64	NE	None
81	3041.6	3056.8	3041.9	3086.6	0.14	0.18	0.03	0.65	None	None
82	3031.0	3034.7	3047.4	3038.8	0.12	0.17	0.05	0.65	NE	None
83	3017.4	3048.2	3040.1	3063.5	0.09	0.12	0.03	0.76	None	None
84	3024.3	3056.4	3062.9	3070.8	0.12	0.13	0.07	0.68	None	None
85	3038.5	3046.4	3059.0	3052.2	0.09	0.24	0.04	0.63	NE	None
86	3033.7	3038.3	3049.1	3064.9	0.13	0.07	0.07	0.74	None	None
87	3052.7	3050.0	3056.8	3060.0	0.09	0.16	0.05	0.70	None	None
88	3032.7	3034.9	3057.1	3063.9	0.17	0.25	0.07	0.51	None	None
89	3035.1	3056.0	3085.9	3090.6	0.11	0.2	0.01	0.68	None	None
90	3023.5	3045.5	3040.1	3030.7	0.12	0.17	0.11	0.60	PVO	None
91	3042.0	3050.8	3060.5	3056.8	0.10	0.20	0.05	0.66	NE	None
92	3059.1	3049.6	3061.8	3051.3	0.11	0.09	0.09	0.70	NE	None
93	3046.1	3062.4	3031.2	3057.0	0.17	0.17	0.05	0.61	PVO	None
94	3058.6	3066.3	3046.4	3046.9	0.06	0.13	0.03	0.78	PVO	None
95	3052.2	3070.0	3042.7	3087.9	0.13	0.11	0.04	0.72	None	None
96	3056.4	3052.7	3034.6	3095.2	0.07	0.10	0.02	0.81	None	None
97	3042.2	3051.2	3055.4	3072.6	0.12	0.14	0.05	0.68	None	None
98	3053.3	3063.2	3072.3	3052.5	0.19	0.16	0.07	0.59	NE	None
99	3060.5	3083.8	3092.4	3088.3	0.13	0.11	0.03	0.73	NE	None
100	3070.5	3078.4	3090.3	3088.5	0.14	0.18	0.04	0.64	NE	None

Boxplots of Q-learning policy at different states by clinicians' first choices, the treatment with the highest probability selected by clinicians, are shown in Figure 5. In the clinicians' first choice, there were only two treatments, NEPVO and None. When NEPVO is the clinicians' first choice, there is significant difference between four treatments in Q-learning policy from the result of ANOVA (*P*-value = 0.000891). The post-hoc test result in Table 7 shows that NEPVO is significantly lower than the other three treatments, which means that NEPVO is the last choice suggested by the Q-learning policy. When no vasopressor (None) is the clinicians' first choice, there are significant difference between four treatments in Q-learning policy from the result of ANOVA (*P*-value = 2e-16). The post-hoc test result shows that None is significantly higher than

the other three treatment, which means that the Q-learning policy suggests the same treatment, None, as clinicians.

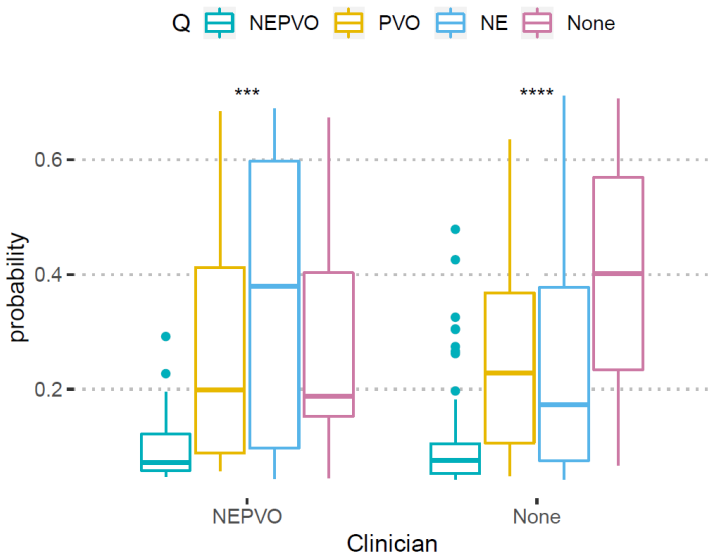

Figure 5 Multiple clinician treatment groups boxplot for Q-learning treatment

Table 7 The ANOVA Post-Hoc test results of different treatments suggested by the Q-learning policy

	ANOVA post-hoc[1] p-value (Clinician's first choice: NEPVO)			**ANOVA post-hoc p-value (Clinician's first choice: None)**		
	NE	NEPVO	None	NE	NEPVO	None
NEPVO	0.0010*	-	-	5.7e-06*	-	-
None	1	0.0324*	-	5.0e-08*	< 2e-16*	-
PVO	1	0.0085*	1	1	1.4e-07*	2.2e-06*
1The post-hoc test was performed with Bonferroni correction.						

The heatmaps of treatments in different ranks based on clinician policy and Q-learning policy are shown in Figure 6 and Figure 7, respectively. From the figures below, we can see that the vasopressor treatments prescribed to patients with different conditions differs between the clinician policy and Q-learning policy. Most of the time, for patients with severe conditions (i.e., lower ranks), clinicians gave vasopressor treatments; however, the Q-learning policy tends to still prefer no treatments (None) to these severely ill patients. The heatmaps also show that clinicians

tend to prescribe NEPVO to very severe patients, while the Q-learning policy only prescribed NEPVO to a few patients. When comparing NE and PVO, clinicians did not prescribe these treatments often, while the Q-learning policy prescribed NE and PVO more frequently to patients with severe conditions.

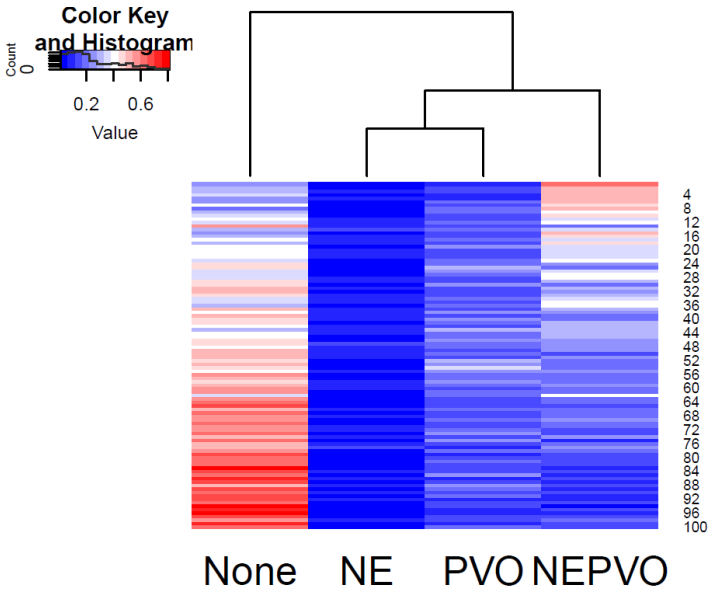

Figure 6 Heatmap of treatments for ranks with clinician policy

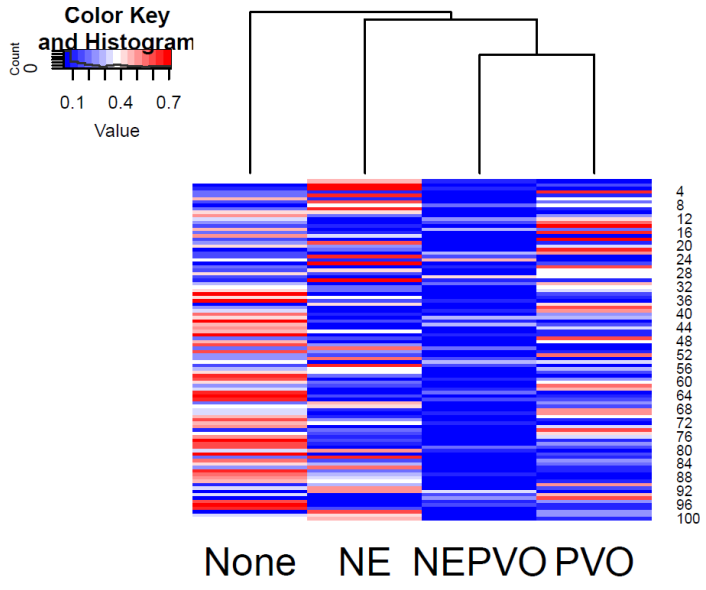

Figure 7 Heatmap of treatments for ranks with Q-learning policy

3.4 Policy Evaluation

Results of off-policy evaluation of Q-learning policy, clinician policy and random policy with 18 bootstrapped datasets of the original, resampling with replacement, are summarized. The mean policy value of the Q-learning policy is 79.0 (SD = 12.2), while the mean policy value of the clinician policy and random policy are 60.1 (SD = 1.3) and 64.6 (SD = 5.9), respectively. The result of the one-way ANOVA shows that the p-value is less than 2e-16. We conclude that there is significant difference between the Q-learning policy, the clinician policy, and the random policy with 0.05 significance level. The post-hoc test shows that the value of the Q-learning policy is significantly higher than the clinician policy with p-value less than 2e-16, and significantly higher than the random policy with p-value 1.2-15, which means that patients can receive higher cumulative reward with Q-learning policy than with the clinician policy or the random policy. However, the p-value of the pairwise t-test between the clinician policy and the random policy is 0.015, which is greater than 0.016 with the Bonferroni correction. So, we draw the conclusion that the policy value of random policy is significantly higher than clinician policy.

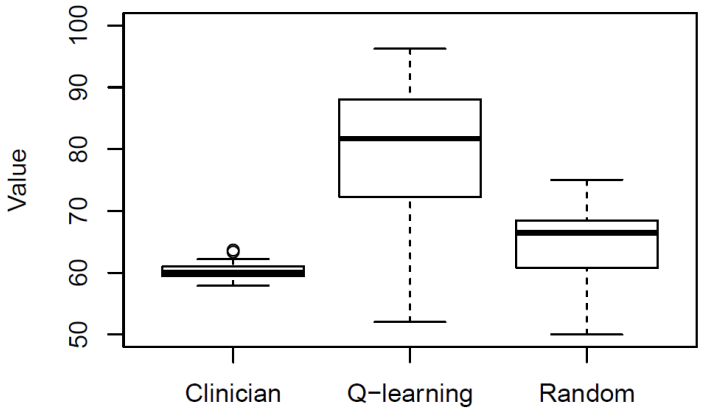

Figure 8 Boxplot of policy values

4.0 Discussion

Taking advantage of big electronic health record data and the development of modern machine learning technology, the clinicians can make fast, individualized, and interpretable treatment decisions through the use of reinforcement learning. The goal of this study was to apply reinforcement learning methods to develop an optimal vasopressor policy for sepsis patients based on observational data, and to better understand different rationales of treatment choices by clinicians and data-driven agents. Specifically, we developed a Q-learning method which aims to search for the optimal decisions targeting long-term survival, rather than focusing on acute resuscitation efforts. Our results showed that the policy based on the Q-learning has significantly higher policy value compared with the one obtained from the clinician policy or the random policy.

We found that most clinicians prescribed vasopressors to patients in worse conditions, which correlates with our understanding that sicker patients are more likely to be treated with vasopressors. Prior research showed that low-dose (<0.04 U/min) vasopressors are safe and effective for the treatment of vasodilatory shock (Mutlu & Factor, 2004). Specifically, the clinicians tend to give combinations of first-line and second-line vasopressor treatments (NEPVO) to patients with worse conditions, which was consistent with the guidelines and prior research about first-line with other vasopressors being the vasopressor therapy in septic shock (Oba & Lone, 2014).

In contrast, the Q-learning policy tends to recommend different kinds of treatments than clinicians to patients with the worst conditions. The result indicates that the Q-learning policy tends to not provide vasopressor treatments, which was also observed in Pruinelli's study where there was only 16% compliance to the guidelines of using vasopressors in patients with low mean

arterial pressure (MAP) (Pruinelli et al., 2016). There are many complications associated with administering vasopressors, such as tachycardia, atrial fibrillation, myocardial infarction, limb ischemia, and necrosis. Because of these serious complications, Q-learning policy may be more hesitant to start vasopressor therapy (Hollenberg, 2011), and be more exploratory and chooses other treatments than the ones typically chosen by clinicians. The Q-learning policy recommends using NE and PVO more frequently, which is consistent with the current research that recommends vasopressors should not be delayed until after fluid resuscitation and should be started to achieve a target MAP of $\geq$ 65 mmHg (Scheeren et al., 2019). Multiple vasopressors, NEPVO, was not suggested by the Q-learning policy as frequently for patients with good conditions. This could be due to the fact that these patients are in a generally good condition so that prescribing NEPVO may be seen as too extreme by the Q-learning policy because of the high risk for side-effects (Hollenberg, 2011).

Our study found that the policy value is significantly lower for the clinicians' choices than random policy, which is counterintuitive. Caution must be taken in the interpretations because that the two policies are largely different in many respects. We offer two factors that might contribute to this result: 1) potential confounding factors that were not included in our study are available for clinicians in their treatment decision making; and 2) clinician policy not only targeted in-hospital survival but also several intermediate outcomes, thus it may not necessarily be fair to compare on a single endpoint. The policy value of Q-learning policy is significantly higher than clinician policy showing that, retrospectively, if the patients received the treatments suggested from the Q-learning policy, their survival probability will increase compared to the existing clinician policy. The evaluation demonstrates that the vasopressor administration based on the Q-learning method values the long-term survival of the patients more than the clinician policy.

There are some limitations in this study. First, vasopressors are the treatments for hypotensive episodes in septic patients and provide relief for a short period of time, rather than treating the underlying causes of sepsis (Jeter, Josef, Shashikumar, & Nemati, 2019). Therefore, more treatments including antibiotic and fluids should be considered. Second, we used daily records of each patient as our analysis time step, which is not dense enough. In real life, medical decision making for septic shock patients in the ICU requires immediate diagnosis so that instant decisions can be made, which does not correspond to the long trajectory discretion. Moreover, patient's vital status such as blood pressure, heart rate, and oxygen saturation can change dramatically within one day. With a finer time frame, estimation results will more closely resemble the real-time effect of the vasopressors. Third, the missing data in the datasets due to exclusion of patients with poor-quality data may compromise the representation of the datasets of the overall patient population. A more granular data is required to avoid a larger percentage of missing data at each time point. Fourth, the complexity of the disease and the variety of medical conditions for patients might require a larger study sample so more clusters can be formed. Fifth, the Q-value may be overestimated because the agent tends to choose the overestimated actions, which can be resolved by using double DQN (Deep Q-Network) or dueling DQN (Van Hasselt, Guez, & Silver, 2016). Finally, the prioritized experience replay, the reweighting sampling improvement for the DQN algorithm training, can be applied to reduce the TD error of the batch data selected from the experience buffer (Schaul, Quan, Antonoglou, & Silver, 2015).

In the future, our goal is to expand our method by incorporating short-term or immediate rewards based on the important intermediate outcomes. Ignoring these outcomes during the treatment decision process might worsen patient's prognosis or reduce his/her quality of life (Yende et al., 2016). In addition, we may also consider more complicated methods that can balance

the long-term and short-term rewards with a quality of life study for patients after hospitalization to determine the influence of the choice of treatments on their post-discharge life. To improve the treatment decision making and form a more well-rounded treatment for each patient, we may also consider adding more treatments other than vasopressors (e.g., antibiotics and fluid use) in the future. We may also improve the model by replacing the discrete patient states with continuous patient states through the use of deep Q-learning (Haarnoja, Tang, Abbeel, & Levine, 2017) so that the monitoring can be on a patient's unique conditions. Our future research will focus on those models that can map the latent structure of the balance of rewards, the complexity of treatments, and continuous trajectories.

The use of artificial intelligence (AI) for guidance and outcome improvements has attracted great attention in the medical community. The purpose of using AI in clinical decision making is to provide support and suggestions, not to replace clinicians. Quality of health care also depends on some other unquantifiable factors, such as emotional intelligence (Norgeot, Glicksberg, & Butte, 2019), for patient outcomes can be affected not only by clinical decisions making but also the patient-doctor relationship. As for AI, we can enhance our modeling and apply machine reinforcement learning to diagnoses and treatments of other diseases in different clinical settings.

Appendix: R code

The data analysis R code is presented in the GitHub repository: https://github.com/alryson52/Improving-treatment-decisions-for-sepsis-patients-by-reinforcement-learning.

Bibliography

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research, 20*(1), 40-49. doi:10.1002/mpr.329

Bennett, C. C., & Hauser, K. (2013). Artificial intelligence framework for simulating clinical decision-making: A Markov decision process approach. *Artificial Intelligence in Medicine, 57*(1), 9-19. doi:https://doi.org/10.1016/j.artmed.2012.12.003

Blanco, J., Muriel-Bombin, A., Sagredo, V., Taboada, F., Gandia, F., Tamayo, L., . . . Grupo de Estudios y Analisis en Cuidados, I. (2008). Incidence, organ dysfunction and mortality in severe sepsis: a Spanish multicentre study. *Crit Care, 12*(6), R158. doi:10.1186/cc7157

Brun-Buisson, C. (2000). The epidemiology of the systemic inflammatory response. *Intensive Care Med, 26 Suppl 1*, S64-74. doi:10.1007/s001340051121

Burkov, A. (2019). *The hundred-page machine learning book*: Andriy Burkov Quebec City, Can.

Haarnoja, T., Tang, H., Abbeel, P., & Levine, S. (2017). *Reinforcement learning with deep energy-based policies.* Paper presented at the Proceedings of the 34th International Conference on Machine Learning-Volume 70.

Hanna, J. P., Stone, P., & Niekum, S. (2017). *Bootstrapping with Models: Confidence Intervals for Off-Policy Evaluation*. Paper presented at the Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems, São Paulo, Brazil.

Hollenberg, S. M. J. C. C. N. C. (2011). Inotrope and vasopressor therapy of septic shock. *23*(1), 127-148.

Jeter, R., Josef, C., Shashikumar, S., & Nemati, S. (2019). *Does the "Artificial Intelligence Clinician" learn optimal treatment strategies for sepsis in intensive care?*

Jiang, N., & Li, L. (2016). *Doubly robust off-policy value evaluation for reinforcement learning*. Paper presented at the Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48, New York, NY, USA.

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*(11), 1716-1720. doi:10.1038/s41591-018-0213-5

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., . . . Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature, 518*(7540), 529-533. doi:10.1038/nature14236

Mutlu, G. M., & Factor, P. (2004). Role of vasopressin in the management of septic shock. *Intensive Care Medicine, 30*(7), 1276-1291. doi:10.1007/s00134-004-2283-8

Ng, S., Strunk, T., Jiang, P., Muk, T., Sangild, P. T., & Currie, A. (2018). Precision Medicine for Neonatal Sepsis. *Front Mol Biosci, 5*, 70. doi:10.3389/fmolb.2018.00070

Norgeot, B., Glicksberg, B. S., & Butte, A. J. (2019). A call for deep-learning healthcare. *Nature Medicine, 25*(1), 14-15. doi:10.1038/s41591-018-0320-3

Novosad, S. A., Sapiano, M. R., Grigg, C., Lake, J., Robyn, M., Dumyati, G., . . . Epstein, L. (2016). Vital Signs: Epidemiology of Sepsis: Prevalence of Health Care Factors and Opportunities for Prevention. *MMWR Morb Mortal Wkly Rep, 65*(33), 864-869. doi:10.15585/mmwr.mm6533e1

Oba, Y., & Lone, N. A. J. J. o. c. c. (2014). Mortality benefit of vasopressor and inotropic agents in septic shock: a Bayesian network meta-analysis of randomized controlled trials. *29*(5), 706-710.

Paoli, C. J., Reynolds, M. A., Sinha, M., Gitlin, M., & Crouser, E. (2018). Epidemiology and Costs of Sepsis in the United States-An Analysis Based on Timing of Diagnosis and Severity Level. *Crit Care Med, 46*(12), 1889-1897. doi:10.1097/CCM.0000000000003342

Pardo, F., Tavakoli, A., Levdik, V., & Kormushev, P. J. a. p. a. (2017). Time limits in reinforcement learning.

Pollard, T. J., Johnson, A. E. W., Raffa, J. D., Celi, L. A., Mark, R. G., & Badawi, O. (2018). The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific Data, 5*(1), 180178. doi:10.1038/sdata.2018.178

Pruinelli, L., Yadav, P., Hangsleben, A., Johnson, J., Dey, S., McCarty, M., . . . Simon, G. J. (2016). A Data Mining Approach to Determine Sepsis Guideline Impact on Inpatient Mortality and Complications. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science, 2016*, 194-202.

Saria, S. (2018). Individualized sepsis treatment using reinforcement learning. *Nature Medicine, 24*(11), 1641-1642. doi:10.1038/s41591-018-0253-x

Schaul, T., Quan, J., Antonoglou, I., & Silver, D. J. a. p. a. (2015). Prioritized experience replay.

Scheeren, T. W. L., Bakker, J., De Backer, D., Annane, D., Asfar, P., Boerma, E. C., . . . Teboul, J.-L. (2019). Current use of vasopressors in septic shock. *Annals of Intensive Care, 9*(1), 20. doi:10.1186/s13613-019-0498-7

Seymour, C. W., Kennedy, J. N., Wang, S., Chang, C.-C. H., Elliott, C. F., Xu, Z., . . . Angus, D. C. (2019). Derivation, Validation, and Potential Treatment Implications of Novel Clinical Phenotypes for Sepsis. *JAMA, 321*(20), 2003-2017. doi:10.1001/jama.2019.5791 %J JAMA

Srinivasa Rao, A. S. R., & Diamond, M. P. (2020). Deep Learning of Markov Model-Based Machines for Determination of Better Treatment Option Decisions for Infertile Women. *Reproductive Sciences, 27*(2), 763-770. doi:10.1007/s43032-019-00082-9

Stratton, L., Berlin, D. A., & Arbo, J. E. J. E. M. C. (2017). Vasopressors and inotropes in sepsis. *35*(1), 75-91.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*: MIT press.

Thomas, P. S., Theocharous, G., & Ghavamzadeh, M. (2015). *High-Confidence Off-Policy Evaluation*.

Van Hasselt, H., Guez, A., & Silver, D. (2016). *Deep reinforcement learning with double q-learning.* Paper presented at the Thirtieth AAAI conference on artificial intelligence.

Vincent, J. L., Marshall, J. C., Namendys-Silva, S. A., Francois, B., Martin-Loeches, I., Lipman, J., . . . Sakr, Y. (2014). Assessment of the worldwide burden of critical illness: the intensive care over nations (ICON) audit. *Lancet Respir Med, 2*(5), 380-386. doi:10.1016/s2213-2600(14)70061-x

Watkins, C. J. C. H. (1989). Learning from delayed rewards.

Yende, S., Austin, S., Rhodes, A., Finfer, S., Opal, S., Thompson, T., . . . Angus, D. C. (2016). Long-Term Quality of Life Among Survivors of Severe Sepsis: Analyses of Two International Trials. *Critical care medicine, 44*(8), 1461-1467. doi:10.1097/CCM.0000000000001658