Safe Reinforcement Learning for Sepsis Treatment

by

Liling Lu

BS in Biology Engineering, Nanchang University, China, 2011

MS in Microbiology, Zhejiang University, China, 2014

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

2022

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Liling Lu

It was defended on

April 12, 2022

and approved by

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

Chung-Chou H. Chang, PhD, Professor,
Departments of Medicine and Biostatistics, University of Pittsburgh

Victor Talisa, PHD, Research Assistant Professor,
Department of Critical Care Medicine, School of Medicine, University of Pittsburgh

Safe Reinforcement Learning for Sepsis Treatment

Liling Lu, MS

University of Pittsburgh, 2022

Sepsis, defined as an overactive immune system response to infection followed by acute life-threatening organ failure, kills eight million people annually. Mortality of acute sepsis is up to 50%, and significantly higher in low-income countries. The correction of the absolute hypovolemia with intravenous fluids and vasopressors is the most difficult aspect of sepsis treatment. There were promising Reinforcement Learning (RL) approaches to learn the optimal administration of vasopressor and intravenous fluids to treat septic patients. However, the existing RL approaches did not take some safety constraints into consideration. Firstly, they only captured end-point outcome and ignored patients' intermediate outcomes, which are also very important to patients. Secondly, they did not consider the dose change of vasopressor within a short amount of time. This is not in accordance with clinical safety protocol, which states that the dose change of vasopressor should be gradual, while a dramatic major change of vasopressor dose is unsafe to patients. In this project, we extended an existing model-free Q-learning algorithm by addressing its two safety concerns. We learned a more robust and safer AI agent which takes intermediate outcomes into consideration by incorporating SOFA score and lactate level as intermediate health status. Additionally, we developed another safer and more competitive AI agent to address the sudden major change in vasopressor dose use by adding vasopressor penalty. The two learned AI agents are more adherent to current clinical practices and knowledge.

Public Health Significance: This work has demonstrated that we can train a safer machine learning AI agent by incorporating knowledge-based constraints and thus giving safer treatment

strategy in sepsis treatment. It is an important progress towards integrating safety into machine learning applications in health sciences.

Table of Contents

List of Tables

List of Figures

Preface

Firstly, I would like to thank my advisor, Dr. Lu Tang, for his patient guidance, timely advice, meticulous scrutiny, and great support during my job search. Dr. Lu Tang has been a tremendous mentor for me. I would also like to express my gratitude to Drs. Chung-Chou H. Chang and Victor Talisa for sitting on my committee, giving brilliant suggestions, and making my defense an enjoyable experience. I also want to thank Dr. Christopher W. Seymour for letting the UPMC data set available to me and thank Jason Neal Kennedy for the assistance with the methodology.

I also want to thank Dr. Hyun Jung Park for offering me the research assistant opportunity in his team and leading me working on multiple exciting projects. And I want to thank all the biostatistics faculties for providing me support during my graduate study.

It is also my privilege to thank my family for the constant support and encouragement.

1.0 Introduction

As a life-threatening medical emergency, sepsis is one of the leading causes of death in hospitalized patients in the ICU (intensive care unit). According to the CDC, at least 1.7 million adults in the USA develop sepsis each year; one in three patients who die in a hospital has sepsis. Sepsis accounts for about eight million deaths each year worldwide; the mortality of acute sepsis is up to 50% or higher in low-income countries (Dugani S, et al., 2017). The 2016 Sepsis-3 conference defined sepsis as "life-threatening organ dysfunction caused by a deregulated host response to infection", while septic shock is defined as "a subset of sepsis in which underlying circulatory and cellular/metabolic abnormalities are profound enough to significantly increase mortality" (Singer M, et al., 2016). It is very challenging to come up with a successful management system for sepsis for three reasons: early detection, severity prognostication and providing optimal targeted therapy (Gotts & Matthay, 2016). Of the three, providing optimal targeted therapy with the correction of absolute hypovolemia through intravenous fluids and vasopressors is given the top research priority (Byrne & Haren, 2017; Marik & Bellomo, 2016). Over the years, many protocols have been developed in terms of the use of intravenous fluids and vasopressors in sepsis treatment, such as the surviving sepsis campaign (SSC) guidelines, highly aggressive Early-Goal directed therapy (EGDT) protocols, etc. But consensus on how to prescribe the right amount and balance of intravenous fluids and vasopressors is yet to be achieved, which leads to large variation in clinicians' practice. Randomized Controlled Trials (RCTs), as a gold standard in clinical discovery settings (Alsowas, & Alahdab, 2016), cannot efficiently and exhaustively explore the exponential combinations of patients, sepsis severeness and treatments. To date, there is still no access to individualized sepsis treatment. Large-scale Electronic Health Records (EHRs) data

embraced with Machine Learning (ML) and biostatistics set light to novel approaches to address the challenges in achieving the precision sepsis treatment (M.Ghassemi et al., 2015).

In sepsis treatment, the goal of clinicians is to maximize patients' probability of good outcomes by making reasonable therapeutic prescriptions at different stage of sepsis (Bennett & Hauser, 2013). The problem here boils down to searching for the optimal sequential decisions. Reinforcement Learning (RL) is a powerful ML algorithm that is broadly used to identify an optimal policy in complex sequential decision-making tasks. Similar to the clinician's goal, in RL, a virtual agent learns an optimal policy that maximizes an expected cumulative reward through trial and error (Sutton & Barto, 2018). There are two reasons that make RL perfect for application in medical decision-making tasks. First, according to Sutton & Barto, RL is well suited to overcome problems of complexity and heterogeneity of patients' responses to therapeutic decisions and latent to interventions, which is attributed to RL's intrinsic design of sparse reward signals. Second, with the beneficiary of large-scale clinical databases where clinical practice variation exists, RL can learn optimal treatment from even suboptimal training practices, which is the professional clinicians' policies provided in the clinical database (Celi et al., 2014). There were already some promising RL applications in medical decisions tasks. In sepsis treatment settings, Komorowski et al. (2018) has created a computational ML model to provide optimal treatment of vasopressor and intravenous fluid to patients with acute sepsis based on MIMIC-III data (Elixhauser et al., 1998). Following that, Kennedy (2021) reproduced Komorowski's model on a retrospective EHR data with information of a large cohort of UPMC patients. He demonstrated that the policy recommended by ML model is significantly superior to clinician policy.

In this project, we would like to address two safety concerns that have existed in an earlier developed AI agent. Firstly, the agent only captures end-point outcome (90-day mortality).

Ignoring intermediate outcomes when making treatment recommendations might reduce patients' quality of life or may even worsen patients' prognosis (Yende et al., 2016). A more robust AI policy should optimize both patients' short-term and long-term outcomes, which is exactly what professional clinicians do in practice. Secondly, when recommending vasopressor dose to patients, the model only infers the optimal action based on the current state, thus it might cause a sudden dramatic vasopressor dose change within a short time interval. Dramatic dosage change (either increase or decrease) can be extremely dangerous. For some patients, it might cause acute hypotension following rapid dose decrease. Additionally, a rapid dose increase might lead to hypertension or cardiac arrhythmias (Fadale et al., 2014; Allen et al., 2014).

We hypothesized that the first safety constraint could be addressed by adding intermediate rewards. These intermediate rewards would incorporate sepsis patient's instant health status. We modified Raghu et al.'s (2017) intermediate reward framework to our usage. We penalized an increase of Sequential Organ Failure Assessment (SOFA) score and high SOFA scores, and vice versa. Same to arterial lactate level, we penalized increase arterial lactate level and rewarded a decrease arterial lactate level. SOFA score is a derived measurement that uses accessible parameters to identify result of sepsis in terms of key organs' failure or dysfunction. An increasing SOFA score during the first 48 hours in the ICU, regardless of the initial score, predicts a mortality rate of at least 50% (Minne L et al., 2008; Doerr F et al., 2011). Lactate is also the biomarker of organ dysfunction. Increases in arterial lactate levels are always followed by a progression of organ dysfunction. It is highly associated with mortality (Rello et al., 2017).

For the second safety constraint, we proposed adding penalty to sudden changes of the major vasopressor dose within a single interval step. This reward framework is modified from Raghu et al.'s reward framework as well. We also try to address the second safety constraint by

adding vasopressor changes to the feature space and redefining the state space using K-means clustering. We hypothesized that we could create a safer, more robust, and more applicable AI policy by modifying a model developed by Kennedy (2021).

2.0 Method

2.1 Data extraction and patient cohort definition

For this project, we extracted data from the Cerner Electronic Health Record System (Cerner, Kansas City, MO). This database contains all medical records of patients from 14 community and academic hospitals within the UPMC health care system from 2013 to 2017. Important informations like demographics, labs, clinical features, and medical treatments are included in the database. Based on our research objective, we extracted patients who meet all of the following criteria: 1. adult (over the age of 18); 2. met sepsis-3* criteria within 6 hours of hospital admission; 3. ICU encounters; 4. medication information; 5. stayed in the hospital for more than 8 hours;

The diagram below shows the full data filteration flow:

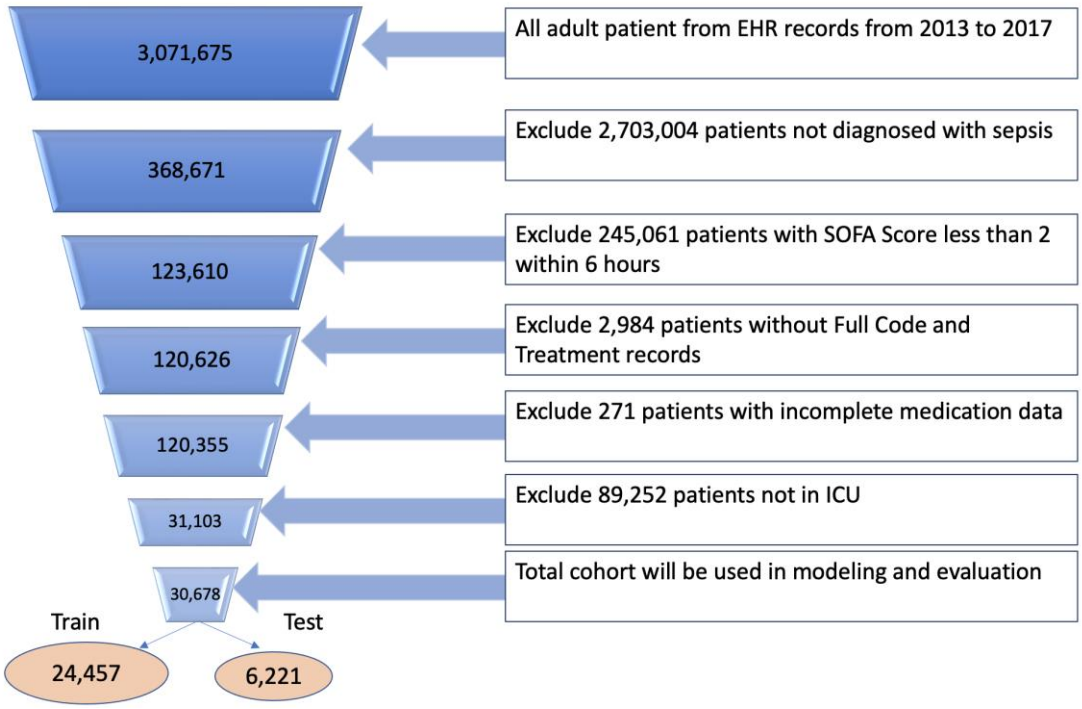

Figure 1 Date extraction flow

2.2 Data pre-processing

Define intervals for each trajectory

With the cohort we defined, we subset the data to include all information within a 54-hour window, which starts from 6 hours before the onset of sepsis to 48 hours after the onset of sepsis. We further divided the 54-hour broad window into 13 intervals, with each interval spanning 4 hours, with the exception that we defined the first 6 hours before suspected sepsis onset as interval 1. The reason why we chose 4 hours as an interval resolution is that it can balance the ability represent sepsis clinical changes and deal with limited changes within samples. For example, if we choose a shorter interval, say 10 minutes, then most of the consecutive data samples will have very similar values, which is undesired for machine learning modeling. On the other hand, if we choose

a wider interval, say 24 hours, then it would be hard to capture an accurate patient health status using the K-means model. Using 4 hours as an interval resolution is a reasonable choice in our content.

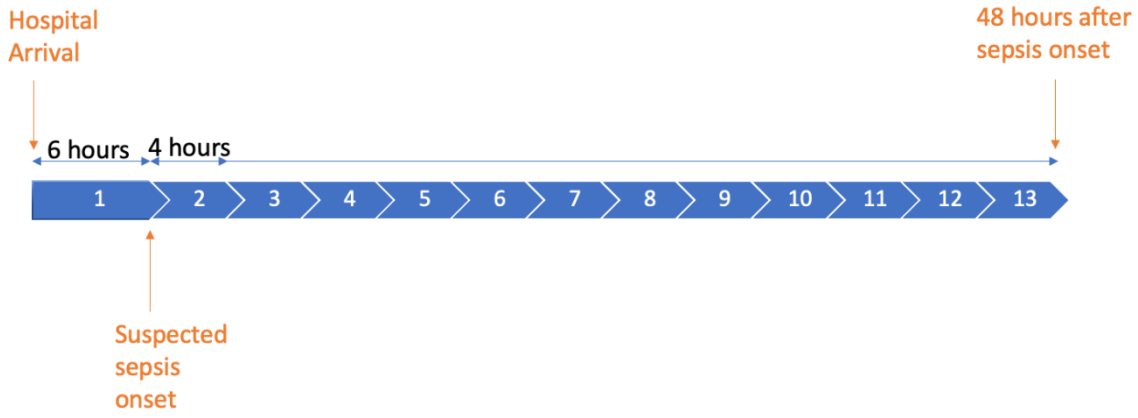

Figure 2 Resolution of time interval

Feature selection

We selected features to match the feature set used by Kennedy, adding an extra feature "vaso_change", which is the absolute difference between the vasopressor of the current step and the previous step. We included patient demographics ("age", "gender", "weight", "Elixhauser score" (Elixauser et al., 1998)), vital signs ("SOFA Score", "SIRS", "shock index", "heart rate", "temperature", "respiratory rate", "diastolic blood pressure"), laboratory measurements ("white blood cells count", "platelets", "glucose", "BUN", "$PaCO_2$", "INR", "FiO_2", "Hemoglobin", "Bilirubin", "pH", "Lactate", "albumin", "bicarbonate", "creatinine", "base_excess", "PaO_2/FiO_2", "MAP", "alanine aminotransferase", "aspartate aminotransferase" , "potassium", "sodium", "Chloride", "Glasgow Coma Scale score") and ventilation parameters("Mech Vent in Window"). These features were associated with sepsis onset, severeness, treatment according to

Angus et al. (2001) and Angus et al. (2013). We also chose features according to their availability in our dataset. The summary statistics of the 39 features are shown in appendix table 1.

Data aggregation

As mentioned previously, we divided our data into 13 intervals of non-overlapping 4 hours as a block. For each feature, we chose the worst (e.g., SOFA score, highest value is the worst) value within the 4-hour block to represent that interval. For vasopressors, we converted them to norepinephrine-equivalents and the maximum dosage per interval was recorded (Brown et al., 2012). For intravenous fluid, we calculated the total dosage by taking the difference between administration start and end, and we computed the mean hourly dosages by averaging over the administration window.

Handling missing data

There are three types of missingness in our data set. First is missing not at random (MNAR): not missing at random, the probability of being missing is different for many unknown reasons. In our dataset, the reason might be human error. To address this kind of missingness, we used a time-limited, parameter-specific sample-and-hold approach (L.Breiman, 2001), which is often used in addressing missing values in longitudinal data. By using this method, the missing value is replaced with the patient's previously observed value, meaning that the last observation is carried forward. We carried forward laboratory values for up to 24 hours and vital signs for up to 4 hours (Kennedy, 2021).

The other two are missing completely at random (MCAR) and missing at random (MAR). One of the most appropriate methods used to deal with such missingness is the Random Forest Algorithm (we used the R package missRanger ()). As it can adapt to the data structure and take the high variance and bias into consideration. Furthermore, Random Forest imputation also has

advantages in handling mixed types of missing data, addressing interactions and nonlinearity, scaling to high dimensions while avoiding overfitting, yielding measures of variable importance useful for variable selection, etc.

After the first step of filling NA with carry forward, we used Random Forest imputation to deal with the rest of the missingness. Random Forest imputation mainly based on proximity from Random Forest, by performing classification (categorical feature) and regression (continuous feature) tasks. For categorical features, the missing value was filled with the category that has the largest mean proximity. For continuous features, we filled the missing value with the weighted average of the non-missing values. For both categorical and continuous features, we iterated 50 times (total 50 separate trees for each feature); within each iteration, we randomly sampled 10% of the whole data to create large variance among trees so that our model can predict the target feature more accurately

Summary of feature statistics before and after imputation

We then summarized the statistics of each feature before and after imputation, as shown in appendix table 1. We showed the mean and standard deviation for features that were evenly distributed, and the median and inter-quantile range (IQR) for features that were skewed. Features after imputation are represented by the median and IQR. Score-based features such as "SOFA score", Systemic Inflammatory Response Syndrome ("SIRS"), Mean Arterial Pressure ("MAP"), Base Excess, and Shock Index were recalculated after imputation.

Data preparation for K-means algorithm

K-means input requires data to have symmetric distribution of features and all the features are on the same scale. We checked the distribution for all the features. For skewed features, we first de-skewed them using a log or inverse-log transformation, and then standardized them to the

same scale with a mean 0 and a standard deviation 1. For normal distributed features, we simply standardized them to the same scale with mean 0 and a standard deviation deviation of 1.

2.3 Basics of reinforcement learning

Together with supervised and unsupervised learning, reinforcement learning (RL) as a self-learning technology forms the machine learning system. In contrast to supervised and unsupervised learning, RL is a decision-making science that is based on goal-directed learning.

In RL, the agent can learn the optimal sequential action by maximizing cumulative rewards: by repeatedly interacting with the dynamic environment and observing the reward, the agent learns how to make better actions to optimize the feedback over time. In the context of sepsis treatment, the doctor is the agent who engages to act (prescribe medications) to interact with the environment (the patient at current state). Following the doctor's intervention, the patient enters a new state. If the patient's health status improves (SOFA score decreases, for example), this action and state pair will be rewarded; conversely, if the patient's health status deteriorates, this action and state pair will be penalized.

Terminology

1. Agent: the entity that we train to make appropriate policy decisions (for example, in sepsis treatment, the reinforcement learning algorithm is the agent)
2. Environment: the conditions or surroundings with which the agent may interact with (e.g., in sepsis treatment, the environment represents each patient)
3. State: the agent’s current status (e.g., patient’s physiology and demographic characteristics fall into)

4. Action: the action taken by the agent at the current time step (e.g., taken by agent, defined as the amount of intravenous fluids and dosages of vasopressors over specific time window)
5. Reward: each action yields a reward, which can either be positive or negative, depending on whether the current state deteriorates or improves (e.g., in our content, rewards are defined by the subsequent change in a patient's health status or risk of mortality)

The traditional interaction framework depicted in Figure 3 is used to define the RL process. The agent performs an action on the environment in the current state, and the environment responds by transitioning into a new state, resulting in an immediate reward (positive or negative). The same procedure will be repeated, and the decision that resulted in the transition will be reinforced over time, until the best decision-making technique is discovered.

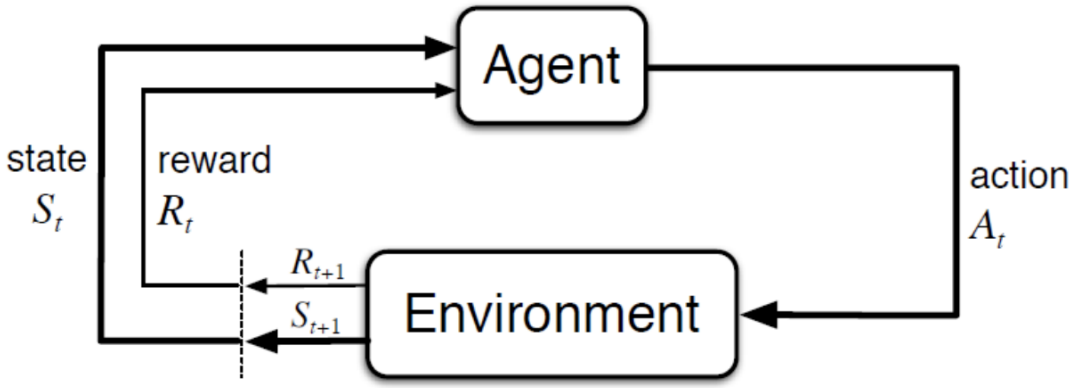

Figure 3 Reinforcement learning framework; Sutton & Barto, 2018

How are RL algorithms deployed?

MDP (Markov decision process) is the most popular mathematical framework with a stochastic control process. RL algorithms can be deployed using a variety of Markov model frameworks, the most basic of which is a discrete MDP (Sutton & Barto, 2018). It is described as a five elements tuple (S, A, P, R, γ), where S represents the finite state space, A represents the finite action space, P represents the probability that the state will change to s' (next state) at time $t+1$ after taking action a at time t, which can be mathematically expressed as $P_{ss'}^{a} = P[S_{t+1} = s' | S_t = s, A_t = a]$; R represents the expected immediate reward received due to action a given state s, which can be represented mathematically as $R_s^a = E[R_{t+1} | S_t = s, A_t = a]$; γ denotes the discount factor which affects how much of the future rewards is allocated to current state or state action value. Policy π is the distribution of the probability of state transition: $\pi(a|s) = P[A_t = a | S_t = s]$, π(a|s) means the probability to take action a at state s.

In MDP, policy only depends on current state; it has nothing to do with previous environment or future status. For an entire episode where pai includes all possible actions, the transition probability function can be described as $P_{s,s'}^{\pi} = \sum_{a \epsilon A}^{\pi} (a|s) P_{ss'}^{a}$, which means the probability of state s transferring to s' equals to the sum of the probability of taking action a multiplied by the probability of state s transferring to s' follow policy π. The reward function denoted as $R_s^{\pi} = \sum_{a \epsilon A}^{\pi} (a|s) R_s^a$, which means the immediate rewards for current state s if follow policy π is equivalent to the sum of all possible reward of specific action multiplied by the probability of that action being taken.

The value function includes state value function: $v_{\pi}(S) = E_{\pi}[G_t | S_t = s]$ and action value function: $q_{\pi}(s, a) = E_{\pi}[G_t | S_t = s, A_t = a]$. The former refers to the long term expected return value over a state. whereas the latter means the long term expected return value over a state action

pair, which is what we will optimize. In the state value function, v is the state value function for policy π at states, G_t is the discounted cumulative value starting from state s. In the action value function, q is the expected cumulative return after following policy π taking action a and starting from state s. They can be further described in a standard way: $v_\pi(s) = E_\pi[R_{t+1} + \gamma v_\pi(S_{t+1})|S_t = s]$ and $q_\pi(s,a) = E_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1})|S_t = s, A_t = a]$, which is the famous Bellman equation. It essentially composes of two components: the immediate reward and the discounted value of the future state.

The optimal solutions to the equations are as bellow (Bellman Optimality Equation)

$$\begin{aligned} v_*(S) &= max_a q_{\pi_*}(s,a) \\ &= max_a E_{\pi_*}(R_{t+1} + \gamma G_{t+1}|S_t = s, A_t = a) \\ &= max_a E(R_{t+1} + \gamma v_*(S_{t+1})|S_t = s, A_t = a) \\ &= max_a R_s^a + \gamma \sum_{s' \epsilon s} P_{ss'}^a \, v_*(s') \end{aligned}$$

$$\begin{aligned} q_*(s,a) &= E\big(R_{t+1} + \gamma max_{a'} q_*(S_{t+1}, a'|S_t, A_t = a)\big) \\ &= R_s^a + \gamma \sum_{s' \epsilon s} P_{ss'}^a \, max_{a'} q_*(s', a') \end{aligned}$$

The Bellman equation specifies a recursive expected value for the cumulative reward following specified policy π. When we use the optimal policy (maximizes expected total reward), as the specified policy π, the equation is unsolvable due to the non-linearity of the maximum function over each state and action pair. Additionally, we may not have the reward function and probability function. While there are multiple iterative methods that can be used to obtain the optimal solutions for the MDP equation. For example, value iteration, policy iteration, state-action-reward-state-action (SARSA) algorithm, Q-learning etc.

Q-learning

The process of determining the optimal policy might be model-based or model-free. A model-based algorithm uses the transition function and the reward function to estimate the optimal policy. Whereas a model-free algorithm estimates the optimal policy without prior knowledge of the environment. In this project, we have no idea of the transition function and the reward function. We will use a model-free Q-learning algorithm to solve the problem. Q-learning (Watkins, 1989) is defined as

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma max_a Q(S_{t+1}, a) - Q(S_t, A_t)]$$

Q-learning as an off-policy/model-free reinforcement learning algorithm, aims to find the optimal action in a given state while the agent has no idea about the preferred states or the rewarding principle. According to the definition, Q (the expected learned action value function) approximates optimal q (q_*) directly. The optimal action value function is independent of the policy being followed. This significantly simplifies the process to solve the bellman equation and converge to the optimal solution. The Q-learning iteration process is shown below.

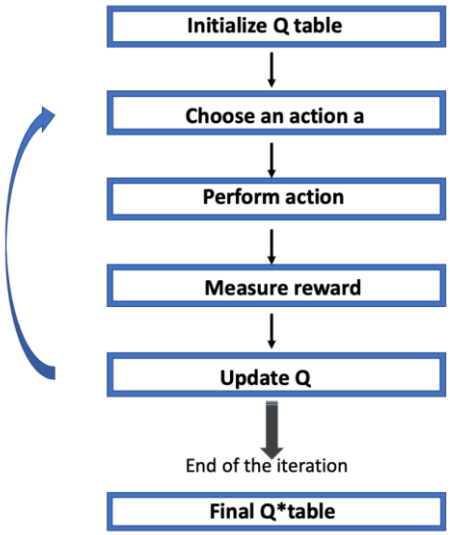

Figure 4 Q-learning framework

Parameters in the Q-learning algorithm

Learning Rate, α: this term can be used to define how much we want to take from the new values and add to the old values. As shown in the algorithm, we are adding the difference between the old and new q values to the old q value, which essentially helps us update the q table. We used α = 0.1 in our algorithm to balance how much to learn from the newly learned and how much to keep from the value we already learned.

Discount Factor, γ: γ is used to balance immediate rewards and future rewards. Typically, people set γ in a range of 0.8 to 1. The higher the value, the more weight is given to the future reward. In our contest, we set γ=0.99 to encourage long-term survival.

Exploration Parameter, ε: ε is used to balance exploration and exploitation or set the chance of how often to explore or exploit. ε can be set at any value between 0 and 1. If ε=0, the algorithm learns the optimal policy purely by random exploration. If ε=1, the algorithm uses the q-table as a reference and selects the action based on the maximum value of available actions. We

applied an exploration rate of 0.1 for our project. This means that 90% of the time, the algorithm will follow the best treatment plan. It will also randomly try out all the 25 actions when iterating.

2.4 Define action-space

The action space is all the possible treatments that the clinician (agent) can apply to the environment (patient). As we noted in the introduction, the registration of intravenous fluids and vasopressors is one of the most challenging aspects in the treatment of sepsis. As reported (Waechter et al., 2014), intravenous fluids include boluses and background infusions of colloids, crystalloids, and blood products normalized by tonicity. The vasopressors include epinephrine, norepinephrine, dopamine, vasopressin, and phenylephrine, were converted when necessary to norepinephrine-equivalent using previously established dose correspondence (Brown et al., 2013). We defined the action space based on the intravenous fluids and vasopressors used in the EHR data.

We wanted to optimize the administration of the maximum dose of vasopressors and the total volume of intravenous fluids over a 4-hour time block. Thus, we stratified IV fluids and vasopressor doses into 5 groups separately (zero dose plus the medians of the non-zero remaining four quantiles in the data), and the action space is a permutation of the two as a 5*5 matrix (table 1). Table 1 shows the distribution of non-zero drug uses for vasopressors and intravenous fluids separately in the EHR data.

Table 1 Action space

Range		Dose of vasopressor (mcg/kg/min)				
		0(0)	1(0-0.09)	2(0.09-0.2)	3(0.2-0.5)	4(>0.5)
Dose of IV fluid (mL/4h)	0(0)	1	2	3	4	5
	1(0-250)	6	7	8	9	10
	2(250-400)	11	12	13	14	15
	3(400-750)	16	17	18	19	20
	4(>750)	21	22	23	24	25

2.5 Define state-space

In RL, state means the current health status of the environment. In the sepsis context, it represents a patient's aggregation of clinical and demographic features at the specific time point. These features include laboratory values, vital signs, and the severity of both chronic and acute illnesses. Here, we defined the state space as discrete by clustering all patients into appropriate groups. There are many clustering algorithms to choose from, such as Mixture of Gaussians, Spectral Clustering, Mean Shift, Mini-Batch K-Means etc. To echo Kennedy's work, we opt to use K-means to divide patients with similar health status into the same group.

K-means Algorithm

The K-means clustering algorithm generates a result through iterative refinement. The algorithm takes two inputs: cluster K and data set. The data set contains all the patients' selected characteristics at various time points. The objective of K-means is to minimize the sum of the squared distances between each point and its corresponding cluster centroid. For each of the observed data points $(x_1, x_{2,} x_3, \dots, x_n)$, a multi-dimensional feature vector is created. K-means clustering attempts to group n data points into K sets $S = \{S_1, S_2, \dots, S_k\}$ to minimize the within-

cluster sum of squares: $\arg\min_{S} \sum_{i=1}^{k} \sum_{\{x \in S_i\}} ||x - \mu_i||^2$, where μ_i is the mean of points in S_i. As shown in figure 5, the algorithm starts by randomly place initial K centroids into the data set space. The second step is to assign each object to the group that is closest to it: $\arg\min_{c_i \in C} dist(c_i, x)^2$, where dist (.) is the standard L2 Euclidean distance, c_i denotes the set of data points assigned to i^{th} cluster centroid. Third, once all data points have been assigned, recalculate K centroids ($c_i = \frac{1}{|S_i|} \sum_{x_i \in S_i} x_i$, taking the mean of all data points in each set S_i. Iterate between step two and step three until the centroids remain stationary. This algorithm will be sure to converge to some centroids.

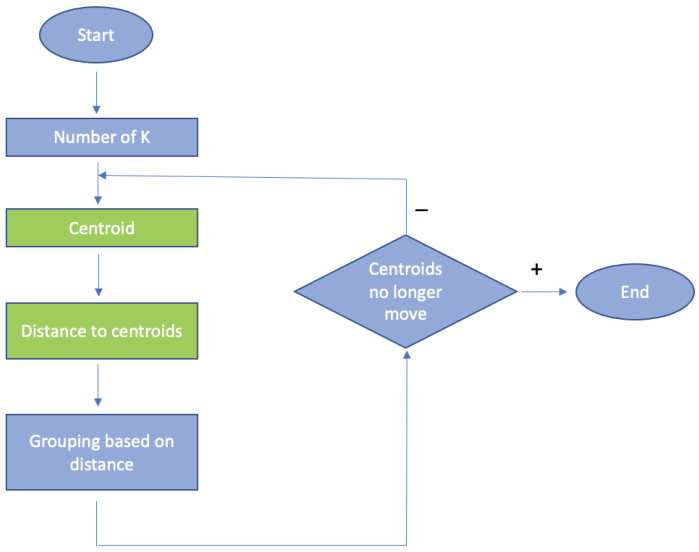

Figure 5 K-means framework

How to choose the optimal K?

To determine the optimal number of clusters for the entire data set, we run the K-means model with a range of different K values and compare the results. In general, numerous metrics can be used to evaluate the results. We chose to use Akaike (AIC), Bayesian information criteria

(BIC) and within-cluster-sum of squares (WSS) as the evaluation metrics. Additionally, we calculated these metrics for different values of K. Then the plots of AIC/BIC/WSS versus K were plotted. We chose the K at the point where AIC/BIC/WSS stop declining rapidly, which is the well-known Elbow method.

2.5.1 Old state-space

State definition

As described in the section on data processing (section 2.2), we used 38 features to ascertain patient states. Since K-means employs the Euclidean distance as its measure, feature scaling is critical. Thus, for features with a normal distribution, we standardized them; for features with a log-normal distribution, we log-transformed them first and then standardized them. We centered binary features so that their mean was zero.

We run the K-means algorithm with K from a range of 50 to 2,000 with intervals of 50 and then plotting a clustering score (AIC, BIC, WSS) as a function of the number of clusters (figure 6). Elbow point is illustrated as the vertical dash line. We selected 750 as the optimal clustering number.

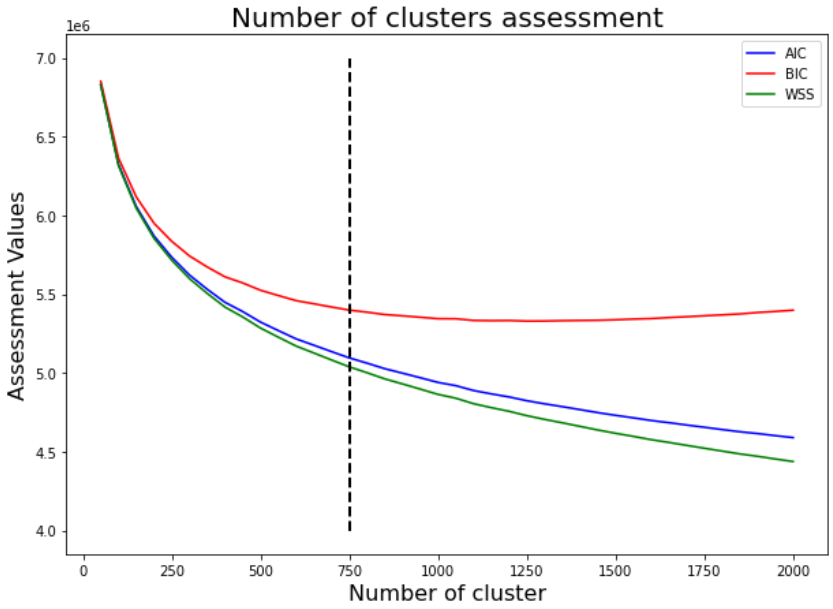

Figure 6 K-means assessment for old state space

State fit evaluation

Figure 7 shows there is heterogeneity among the features by state, indicating that the clustering was not driven by a single one or very few of the features.

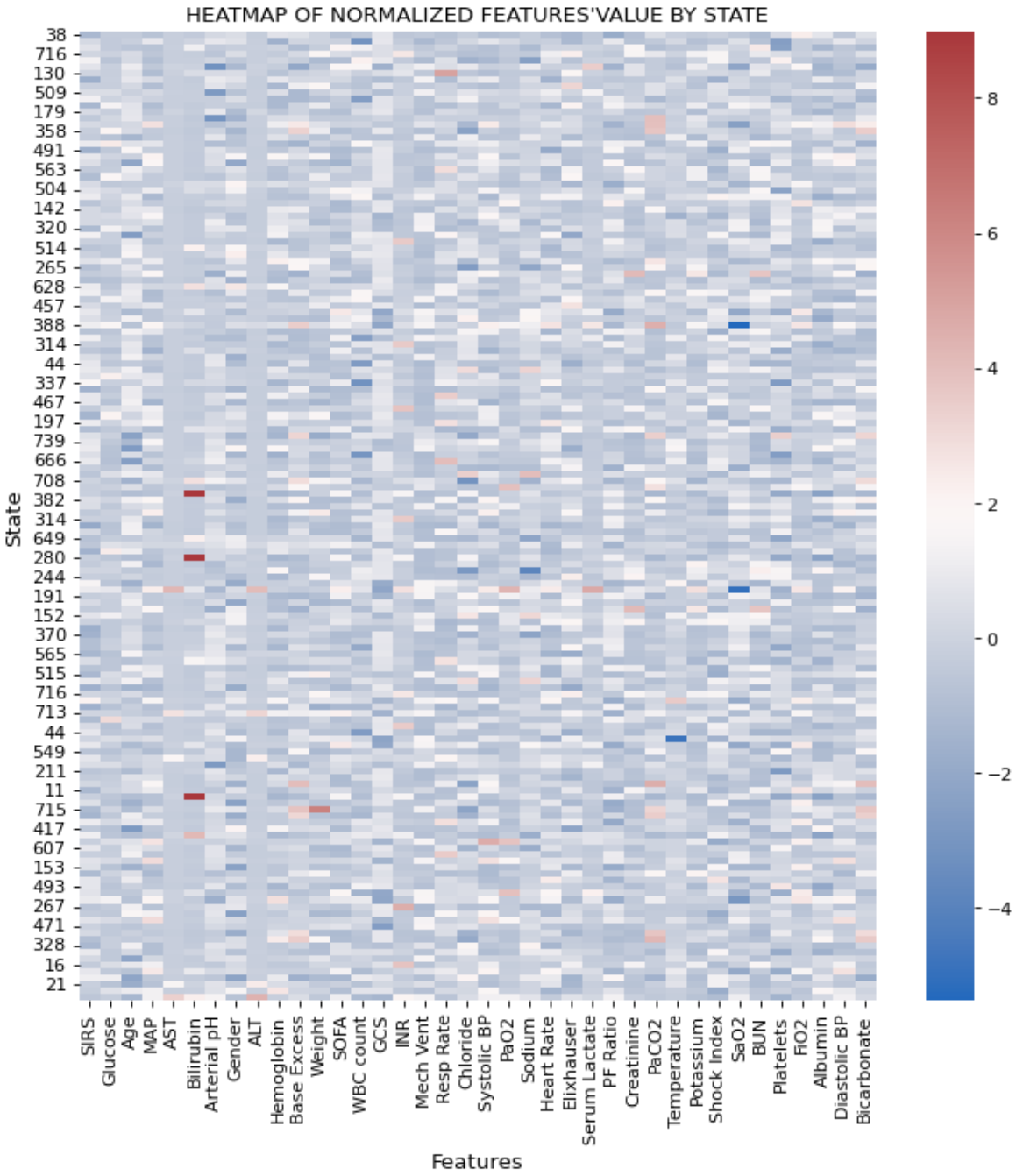

Figure 7 K-means fit evaluation for old state space

2.5.2 Modified state-space

State definition

To address the safety concern of sudden vasopressor dose change, we modified the state-space to incorporate the safety constraint. We created an extra feature called "vaso-change" by computing the change in vasopressor dose between two neighboring intervals. We verified that the variable "vaso-change" is normally distributed. Then we standardized it before feeding it into the K-means algorithm. Together with all the 38 features used to define the old state space, now we had 39 features to represent the patient state. We hypothesized that by doing so, we can enable new state space to capture the difference in vasopressor dose between the current and previous time points, resulting in smoother vasopressor dosage recommendations.

Similar to the old state-space, we run the K-means algorithm with K ranging from 50 to 2,000 with intervals of 50 and then plot the clustering score (AIC, BIC, WSS) as a function of the number of clusters (Figure 8). Elbow point is illustrated as the vertical dash line. We selected a total 750 states as well so that the results of different models under different state pace can be comparable and we can attribute the difference between the two models to extra feature "vaso-change" instead of a different number of states.

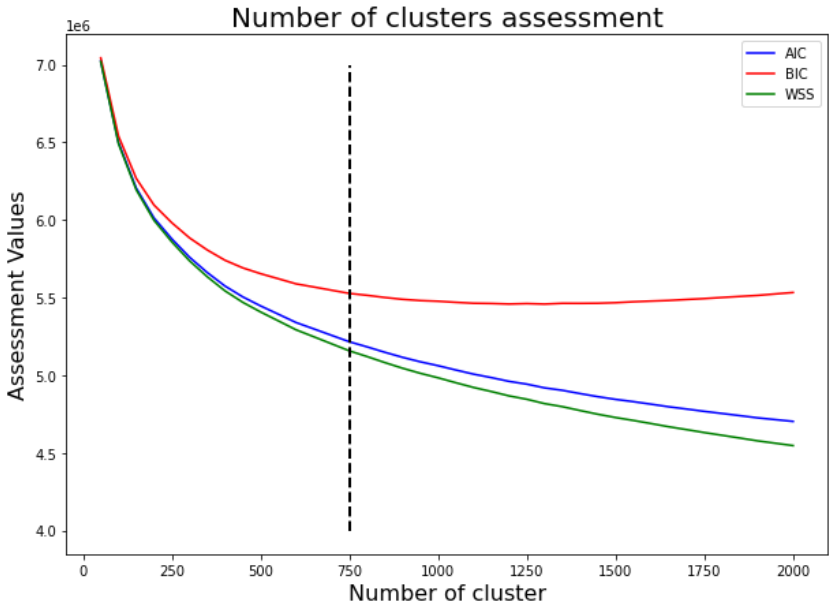

Figure 8 K-means assessment for new state space

State fit evaluation

Figure 9 shows that there is heterogeneity among the features by state, indicating that the clustering was not driven by a single one or very few of the features.

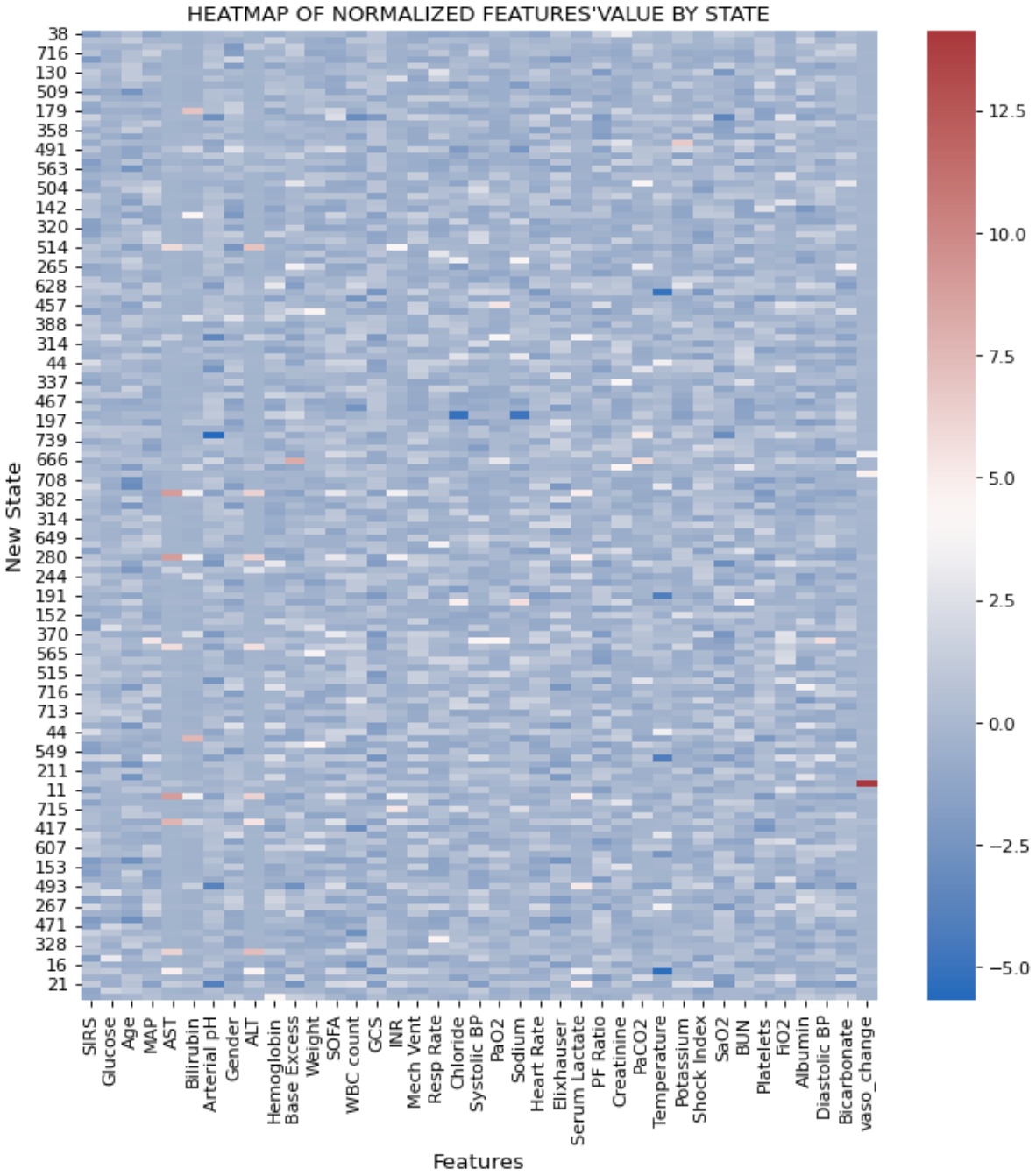

Figure 9 K-means fit evaluation for new state space

2.6 Define reward framework

2.6.1 Unconstrained reward framework

We only focused on the long-term outcome (either survival or death) of patients, which means that only the final time block was rewarded or penalized, while all other intermediate steps were assigned 0s. In the terminal time block, if a patient survived through 90-days, a reward of (+15) was assigned. Otherwise, a minus reward (-15) was applied.

Now that we have defined state, action and reward, an example of the trajectory of a random patient is showing below.

Table 2 Patent's trajectory under unconstrained AI policy

Interval	State	Action	Reward
9	437	21	0
10	170	1	0
11	576	1	0
12	498	11	0
13	315	1	0
14	113	1	0
15	727	1	0
16	315	1	0
17	498	11	15

2.6.2 Intermediate reward framework

However, in clinical settings, doctors consider not only long-term survival but also hospital stability when they prescribe drugs. To mimic doctors' actions, we incorporated intermediate rewards. This incentive should be reflected in the best indicators of patient's immediate health status. In this instance, we referred to Raghu et al.'s (2018) intermediate reward framework. We used the patient's total sofa score to represent the patient's health condition. The other indicator we used was lactate level, which are a measure of cell-hypoxia and are typically higher in patients with sepsis, as sepsis causes low blood pressure and further depletes tissue oxygen. Below is the reward function we used to compute rewards for intermediate timesteps:

$$r(s_t, s_{t+1}) = C_0 \mathbb{1}(s_{t+1}^{\text{SOFA}} = s_t^{\text{SOFA}} \;\&\; s_{t+1}^{\text{SOFA}} > 0) + C_1(s_{t+1}^{\text{SOFA}} - s_t^{\text{SOFA}}) + C_2 \tanh(s_{t+1}^{\text{Lactate}} - s_t^{\text{Lactate}})$$

In this function, high SOFA scores and increases in SOFA score would be penalized, and vice versa, as C_0 and C_1 are negative. Similar for lactate, we rewarded a decrease in lactate and penalized an increase in lactate. For terminal rewards, the procedure was the same to unconstrained reward system: we assigned +15 reward if the patient survived through 90-days and -15 if not. We experimented with three distinct sets of parameters:

C_{01} = -1/40, C_{11} = -5/40, C_{21} = -2 (Reward framework for AI_intm1).

C_{02} = -1/30, C_{12} = -5/30, C_{22} = -2 (Reward framework for AI_intm2).

C_{03} = -1/20, C_{13} = -5/20, C_{23} = -2 (Reward framework for AI_intm3).

We explored the optimal weight of penalty for SOFA score. AI_intm1 had the smallest penalty, AI_intm3 had the highest, while AI_intm2 has a penalty that is in the middle.

Below is an example of the same patient's trajectory as in previous section under this reward system.

Table 3 Patent's trajectory under AI policy with intermediate reward

Interval	State	Action	Reward
9	437	21	-1.97
10	170	2	-1.08
11	576	1	0.25
12	498	11	0.13
13	315	21	-0.03
14	113	1	-0.25
15	727	1	-0.03
16	315	19	0.38
17	498	11	15

2.6.3 Vasopressor penalty reward framework

Previous attempts to use RL to learn optimal policies in sepsis treatment are very promising. However, the learned AI policy only took the patients' current states into consideration when recommending a dosage of vasopressor. While in clinical practice, the administration of vasopressor dose should be gradually decreased or increased. According to Bassi et al. (2013), Norepinephrine doses over 0.5, 1.0 mcg/kg/min are usually considered to be excessive and rarely excessive respectively. In our work, we defined high threshold as the median dose of the fourth quantile of vasopressor, which is 1.05 mcg/kg/min. Figure 10 shows original unconstrained AI policy had much more sudden dramatic change of vasopressor dose use then clinician policy.

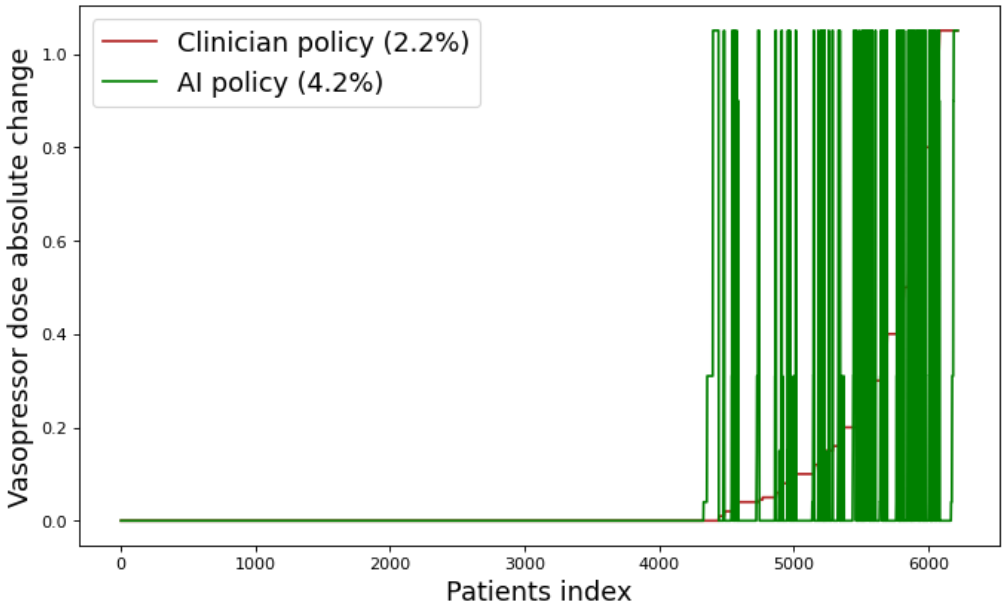

Figure 10 Sudden vasopressor change

To address this constraint, we proposed to penalize big vasopressor dose change use within a time step interval and reward those with vasopressor dose change very slightly. The reward function is as below:

$$r(S_t, S_{t+1}) = C_0 - C_1 tanh(\left| S_{t+1}^{Vasopressor} - S_t^{Vasopressor} \right|)$$

We experimented with three sets of parameters:

C_{01} = 0.25, C_{11} = 2 (Reward framework for AI_vaso1).

C_{02} = 0.5, C_{12} = 5 (Reward framework for AI_vaso2).

C_{03} = 0.5, C_{13} = 10 (Reward framework for AI_vaso3).

We explored the optimal weight for vasopressor change. AI_vaso1 had the lowest penalty, AI_vaso3 had the highest penalty. Rewards at terminal time points were same as unconstrained reward system. Rewarded +15 if patient survived to 90-days and penalized 15 if patient did not survive through 90-days.

Because the maximum change in vasopressor dose use was significantly greater than the average change among all time blocks. To ensure that the absolute reward was limited under the maximum absolute reward (15), we used function tanh to cap the maximum absolute reward to $|C_0 - C_1|$.

Below is an example of the same patient's trajectory as in the previous section under this reward system.

Table 4 Patent's trajectory under AI policy with vasopressor penalty

Interval	State	Action	Reward
9	437	21	0.25
10	170	20	0.01
11	576	1	0.25
12	498	11	0.23
13	315	1	0.23
14	113	1	0.25
15	727	1	0.13
16	315	1	0.18
17	498	11	15

2.7 Model evaluation and comparison

It's not realistic to evaluate the AI policy on real patients because of risk, legal and ethical issues. In this section, we compared trajectory-wise weighted importance sampling policy values (TWIS) and importance sampling (IS) estimated mortalities under all policies. We evaluated whether adding a penalty to sudden vasopressor dose change and redefining state space can help to reduce the proportion of patients who have sudden major changes in use of vasopressor. We finally ranked the feature importance of each model using a random forest algorithm to evaluate whether the model is interpretable in terms of clinical knowledge. All evaluations are applied on the 20% testing data set.

2.7.1 Importance sampling-based policy value evaluation

As we want to evaluate the learned AI policy (target policy) given the data generated from clinical policy (behavior policy), This is referred to as an off-policy evaluation. Importance sampling (IS) is a technique for estimating expected values under one distribution using samples from a different distribution. It is the basic evaluation method for almost all off-policy learning methods. The key principle of IS is to correct the discrepancy between the behavior (π_b) and the evaluation (π_e) policies when learning from off-policy returns (Jiang & Li, 2015). Weighted importance sampling is a variant of importance sampling.

In IS, we first computed the probability for all the state-action trajectories $\{A_t, S_{t+1}, A_{t+1}, \ldots., S_t\}$ under each policy π, shown as a probability function below. For any AI policy, we set the exploratory rate to epsilon (€). So, the probability of exploiting state-optimal

action was 1-€+€/25, while the probability of exploring all the other non-optimal actions was set to be €/25.

$$Pr(A_t, S_{t+1}, A_{t+1}, \dots, S_T | S_t, A_{t:T-1} \sim \pi)$$
$$= \pi(A_t|S_t)p(S_{t+1}|S_t, A_t)\pi(A_{t+1}|S_{t+1}) \dots p(S_T|S_{T-1}, A_{T-1})$$
$$= \prod_{k=t}^{T-1} \pi(A_k|S_k)p(S_{k+1}|S_k, A_k)$$

The second step is to calculate the importance sampling ratio, which is simply dividing the probability of the trajectory under the target policy (π) by the behavior policy (b). Here we used $T-1$ because we dropped the last record of each patient as there was no next state for the last record.

$$\rho_{t:T-1} \doteq \frac{\prod_{k=t}^{T-1} \pi(A_k|S_k)p(S_{k+1}|S_k, A_k)}{\prod_{k=t}^{T-1} b\,(A_k|S_k)p(S_{k+1}|S_k, A_k)} = \prod_{k=t}^{T-1} \frac{\pi(A_k|S_k)}{b(A_k|S_k)}$$

The third step is to average returns from all the observed episodes following policy b:

$$V_{IS} = \frac{1}{n} \sum_{i=1}^{n} \rho_i \left(\sum_{k=t}^{T-1} \gamma^k R_k^i \right)$$

The standard IS estimator is unbiased but has a high variance, and there are numerous variants of IS that can tradeoff between bias and variance. For more information, check (Jiang & Li, 2015). The variance of ordinary importance sampling is unbounded because the variance of the ratios may be unbounded, whereas the weighted estimator's maximum weight on any single return is one.

In practice, the weighted estimator is strongly favored because it has a significantly lower variance. The equation of weighted importance sampling is described below, instead of just simply average through all trajectories like IS, TWIS uses weighted average through all encounters.

$$V_{TWIS} = \frac{1}{\sum_{i=1}^{n} \rho_i} \sum_{i=1}^{n} \rho_i \left(\sum_{k=t}^{T-1} \gamma^k R_k^i \right)$$

We evaluated TWIS policy values for each policy in this project. For AI policies, where ρ_i is the ratio of π(AI) and b(clinician), we substituted 0.0001 to 0 if b(clinician) equals zero. For the clinician policy, ρ_i was set to one. We estimated the patient's average mortality on test data using the IS estimator, as the mortality rate was not trajectory-wise, and we observed that the variance of mortality was very small. We used 1 for 90-day survival and 0 for 90-day death to substitute the original reward values.

2.7.2 Maximum vasopressor dose change evaluation

We evaluated the proportion of maximum vasopressor dose change for all the policies in the test data set, which has 6,221 patients. We first calculated the maximum absolute vasopressor dose change for each patient. We then evaluated the proportion of patients who have a max value greater than 1.05 mcg/kg/min (the median of the fourth quartile of vasopressors).

2.7.3 Clinical interpretability evaluation

To further assess the feature importance and interpretability of each policy, we fitted a random forest classification model for each policy to get the rank of relative feature importance when recommending vasopressor regardless of dosage. For the random forest model, the independent variables are all the 38 features we used to define the state space, and the dependent feature is whether vasopressor is used (0 for no, represents action 1,6,11,16,21; 1 for yes,

represents all the rest of the actions in the action space). The feature importance was computed based on feature permutation, which overcomes one of the impurity-based method's limitations: a bias toward high-cardinality features.

3.0 Results

3.1 Optimal action identification for test data

We applied the same parameters (as described in section 2.3) to each Q-Learning algorithm when training (24,882 encounters) and learned the optimal policies under different reward frameworks. We then identified the recommended action for each record in the test data (6,221 encounters) and reported the distributions of recommended actions on the test data by intravenous fluids and vasopressors separately.

Distribution of intravenous fluids use of different policies

Table 5 below shows the proportions of different intravenous fluids dosages recommended by different policies under old state space and new state space. If we compare the AI policies with clinician policy, we can find that the AI polices with intermediate rewards recommended more intravenous fluids than clinicians, and the proportions of 0 intravenous fluids block were lower than clinician policy; Whereas the AI policies with vasopressor penalty recommended fewer intravenous fluids than clinicians, and the proportions of 0 intravenous fluids block were higher than clinician policy. Figure 11 illustrates the results more clearly. This trend is similar among the old state space and the new state space.

Table 5 Marginal distribution of intravenous fluids

	Intravenous Fluids (mL/4h)								
		Proportions of Actions							
	Range	Clinician	AI	AI_intm1	AI_intm2	AI_intm3	AI_vaso1	AI_vaso2	AI_vaso3
Old state space	0.00	0.47	0.47	0.46	0.45	0.45	0.50	0.49	0.48
	1-250	0.11	0.08	0.08	0.08	0.08	0.06	0.07	0.07
	251-400	0.16	0.20	0.21	0.21	0.21	0.19	0.19	0.19
	401-700	0.10	0.10	0.10	0.11	0.10	0.09	0.09	0.09
	>701	0.17	0.16	0.16	0.16	0.16	0.16	0.16	0.16
New state space	0.00	0.47	0.48	0.47	0.46	0.47	0.52	0.51	0.48
	1-250	0.11	0.10	0.10	0.10	0.10	0.08	0.08	0.09
	251-400	0.16	0.16	0.16	0.17	0.16	0.15	0.15	0.16
	401-700	0.10	0.10	0.10	0.10	0.10	0.10	0.09	0.10
	>701	0.17	0.16	0.17	0.17	0.17	0.16	0.16	0.16

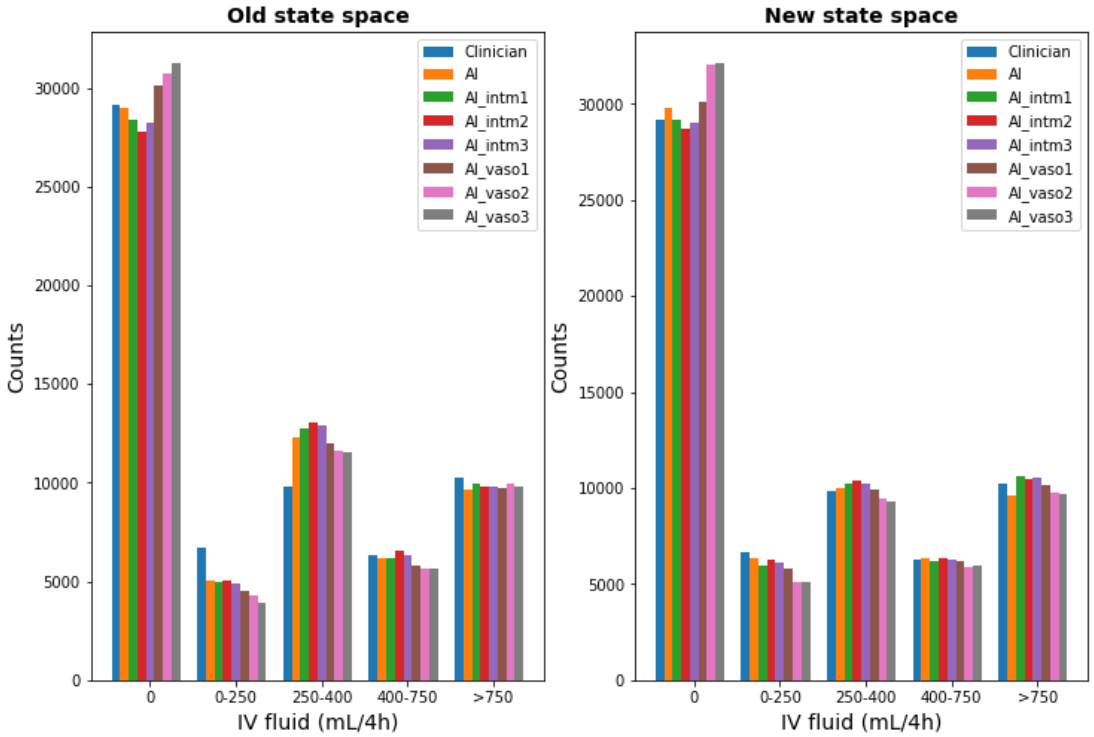

Figure 11 Marginal distribution of intravenous fluids for different policies

Distribution of vasopressor of different policies

Table 6 and Figure 12 below illustrate the distribution of proportions of different vasopressor dose recommendations across all the policies. We can conclude that all of the AI policies recommended less vasopressor than clinicians, no matter if they were under the old state space or the new state space. AI policies with intermediate rewards had similar proportions of vasopressor recommendation to unconstrained AI policy, while AI policies with vasopressor penalties recommended fewer vasopressors than the other policies.

Table 6 Marginal distribution of vasopressors

	Vasopressor (mcg/kg/min)								
		Proportions of Actions							
	Range	Clinician	AI	AI_intm1	AI_intm2	AI_intm3	AI_vaso1	AI_vaso2	AI_vaso3
Old state space	0	0.855	0.891	0.892	0.888	0.89	0.904	0.902	0.895
	.001-.09	0.038	0.023	0.024	0.025	0.026	0.025	0.026	0.025
	0.1-0.2	0.029	0.024	0.026	0.024	0.027	0.023	0.025	0.023
	0.21-0.5	0.043	0.028	0.028	0.029	0.028	0.024	0.023	0.09
	>0.5	0.034	0.034	0.03	0.022	0.03	0.022	0.022	0.15
New state space	0	0.855	0.867	0.866	0.86	0.867	0.882	0.889	0.871
	.001-.09	0.038	0.102	0.022	0.026	0.022	0.026	0.028	0.028
	0.1-0.2	0.029	0.16	0.029	0.03	0.029	0.026	0.026	0.028
	0.21-0.5	0.043	0.101	0.034	0.038	0.034	0.03	0.026	0.034
	>0.5	0.034	0.156	0.049	0.046	0.047	0.036	0.031	0.039

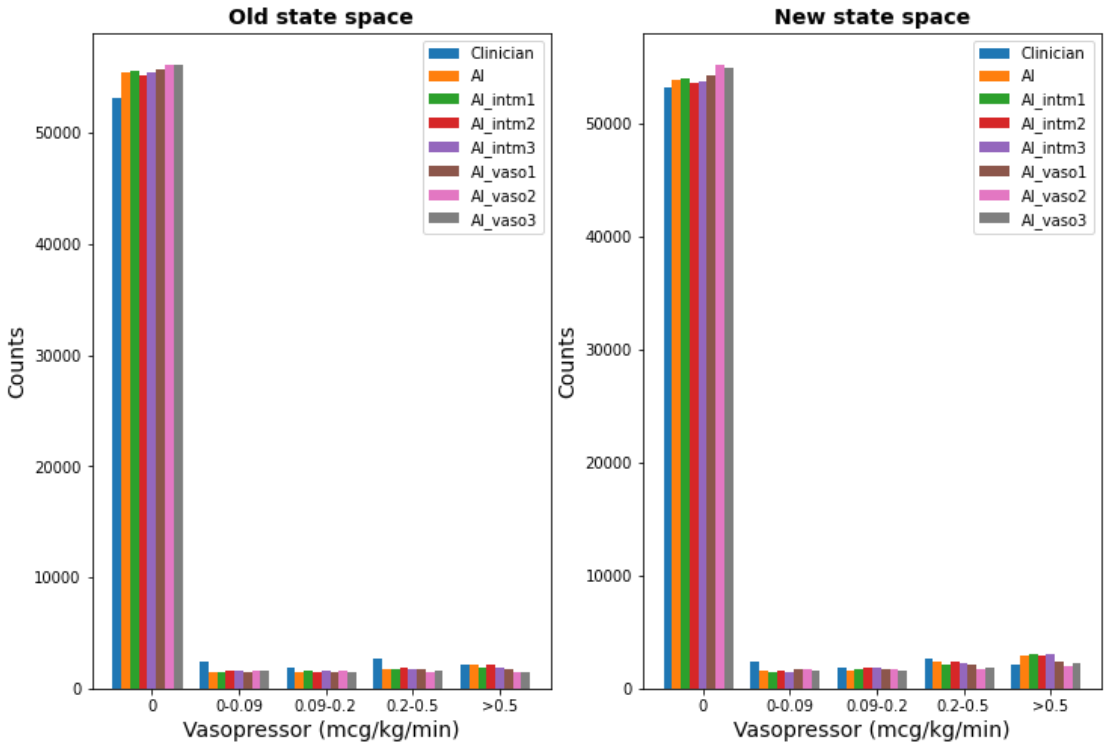

Figure 12 Marginal distribution of vasopressors for different policies

Discussion about distribution of intravenous fluids and vasopressor

As we concluded in the previous section, AI policies with intermediate rewards recommended more intravenous fluids and fewer vasopressors than clinician policy. This is in accordance with EGDT protocols that septic patients generally require large doses of intravenous fluids at initial resuscitation to reverse refractory tissue hypoperfusion and organ failure (Rivers et al., 2001). While this is different from Komorowski et al.'s that AI policy recommended more vasopressors. This might be because the patients in our cohort are different from the patients Komorowski et al. used (MIMIC-III), and we considered intermediate outcomes which makes all model more comprehensive.

The modified AI policies with vasopressor penalties recommend fewer vasopressors and fewer intravenous fluids than clinician policy. If we penalize dramatic vasopressor dose change,

the AI policy tends to prescribe no vasopressor to avoid such dramatic changes. So, AI policies prescribed no vasopressor more of the time. After being constrained by the vasopressor penalty, the modified AI agent became more conservative.

Vasopressor recommendation VS mortality

We further compared the vasopressor recommendation among 750 states under the old state space and the new state space, as shown in Figure 13. The x-axis of the heatmaps below shows the 90-day death mortality at the very left column, and the other 8 columns on the right are proportions of vasopressor recommendations under different policies (the color indicates the proportion of vasopressor recommendation; vasopressor is encoded with 0 and 1, regardless of the scale of vasopressor). The y axis represents the state ranked by mortality, from low to high. We can find that under both state spaces, AI policies recommended vasopressors to patients with high death risk more often and to patients with low death risk less of the time compared to the clinician policy. We also noticed that AI policies in the new state space recommended less vasopressor than in old state space when the patient had a high risk of death.

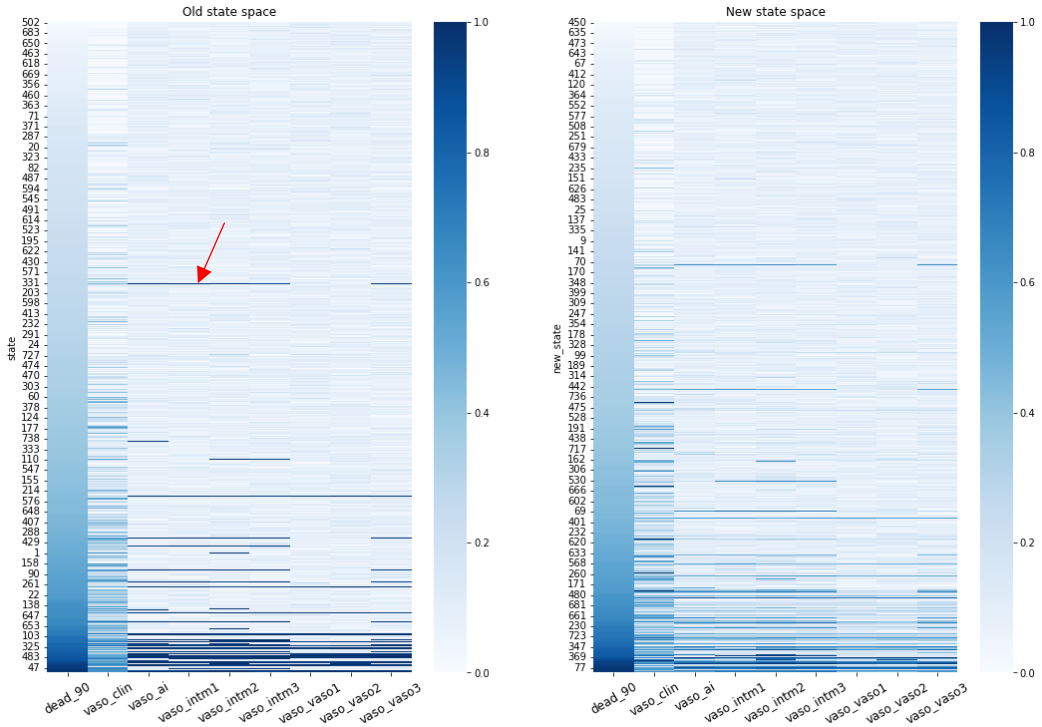

Figure 13 Vasopressors use ranked by mortality rate

The state depicted in Figure 13 has a low mortality rate but is treated aggressively with vasopressors. We conducted additional examinations on these patients. They were assigned to state 403 in the old state space. The mean SOFA score for this group is 8.7, which is significantly higher than the mean SOFA score for the entire data set, which is 5.4. The mean SOFA score of patients who survived and those who did not survive at 90 days was then compared separately in this group. We received a score of 9 for survivors and 8.5 for non-survivors. Thus, patients in this group should be treated as if they have severe sepsis (a high SOFA score is associated with an increased risk of death) and given additional vasopressors. Many of these patients eventually survived for a variety of reasons. Thus, even though mortality is low in this group, the use of vasopressors is high.

3.2 Major vasopressors change evaluation

We evaluated the proportions of major vasopressor dose changes following the unconstrained AI policy, the modified AI policies, and the clinician policy on the test data set, which has 6,221 patients. Major vasopressor dose change is defined as 1.05 micrograms/kg/min, which is the median vasopressor dose of the highest quantile. According to Bassi et al. (2013), this sudden change is not recommended in clinician treatment. The dose change of 1.05 micrograms/kg/min is considered very rare as it might result in acute hypotension or cardiac arrhythmias. Under the old state space, the clinician policy had 136 (2.19%) patients who had major vasopressor changes. The AI policy had 277 (4.45%) patients who had major vasopressor changes, which doubled the number of the clinician policy. The modified AI policies with intermediate rewards slightly decreased the number compared to the unconstrained AI policy but were still higher than the clinician policy. The modified AI policies with vasopressor penalties performed the best in terms of reducing the number of patients with major vasopressor changes; the more penalties applied to vasopressor changes, the more effective they were in reducing the number of patients with major vasopressor changes. Under the new state space, the conclusion was the same as under the old state space. When comparing the new state space to the old state space, the proportions of major vasopressor change have increased a lot for all the AI policies. We conclude that adding vasopressor change to state definition space is not helpful in decreasing dramatic vasopressor change. In further estimation, we will only include policies in the old state space.

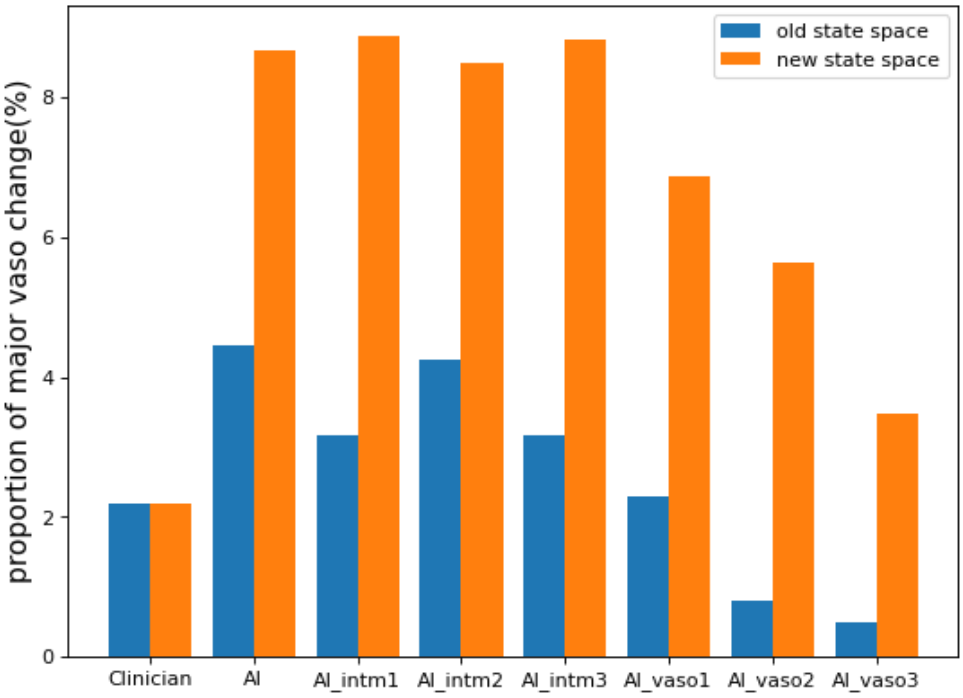

Figure 14 Proportion of major vasopressors change comparison

Table 7 Proportion of sudden vasopressors change

Policies	Clinician	AI	AI_intm1	AI_intm2	AI_intm3	AI_vaso1	AI_vaso2	AI_vaso3
Old state	2.19%	4.45%	3.17%	4.24%	3.17%	2.80%	0.80%	0.50%
New state	2.19%	8.68%	8.87%	8.50%	8.82%	6.88%	5.64%	3.47%

3.3 **Importance sampling-based model evaluation**

3.3.1 **Trajectory-wise importance sampling (TWIS) policy value comparison**

We evaluated the performance of all of the learned AI policies under a variety of reward frameworks using model-free off-policy evaluation. Figure 15 depicts the final trajectory-wise WIS policy values for various policies, with 95 percent confidence intervals calculated using 100 independent bootstrapping. The policy value associated with the clinician policy was calculated in the test data set and served as the benchmark for all the other learned AI policies. As we observed, adding an intermediate reward to penalize high a SOFA score and Lactate can slightly increase the TWIS policy value compared to the unconstrained AI policy. However, an excessive penalty would result in unstable results, such as policy AI_intm3, which had a TWIS policy value of 8.54 and a very large standard deviation of 7.98. Additionally, we discovered that the AI policies with vasopressor penalties, such as AI_vaso1, AI_vaso2, and AI_vaso3, achieved comparable TWIS policy values to the unconstrained AI policy.

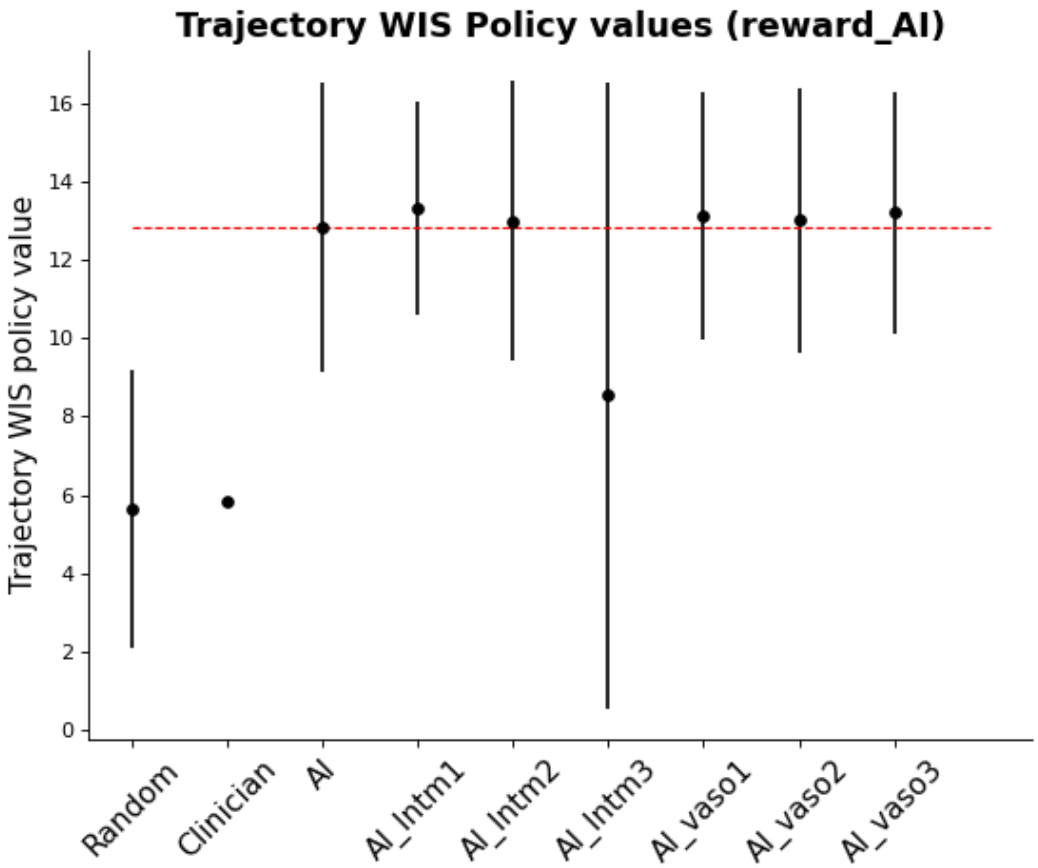

Figure 15 Trajectory Weighted Importance Sampling policy value comparison

Figure 15 shows the TWIS policy values for all policies when evaluated using the unconstrained reward framework with only terminal rewards. We also evaluated them using intermediate reward frameworks and vasopressor penalty frameworks (as in supplementary Figure 1). The TWIS policy values were very similar to the results evaluated using terminal reward frameworks, with the exception that the AI_intm policies obtained slightly higher TWIS policy values using the intermediate reward frameworks and the AI_vaso policies got slightly higher TWIS policy values using the vasopressor penalty frameworks than the other policies. This indicates that optimizing the terminal outcome does not conflict with optimizing both the terminal and intermediate outcomes simultaneously.

3.3.2 Importance sampling-based estimated mortality comparison

We would like to further compare the estimated mortalities under different policies in the test data set. We calculated the estimated mortality using the importance sampling method, and the 95% confidence interval of each mortality was achieved using 100 independent bootstrapping. The estimated mortality for each reward framework is shown in Figure 16. The estimated mortality of the clinician policy in the test data is set to be the benchmark for all the other AI policies. It shows that the AI_intm3 got the lowest mortality rate of 0.277 with a standard deviation of 0.0009, while the AI_intm2 got competitive mortality as the unconstrained AI policy. All the other modified policies had a higher death rate than the unconstrained AI policy, but they were still significantly lower than the clinician policy.

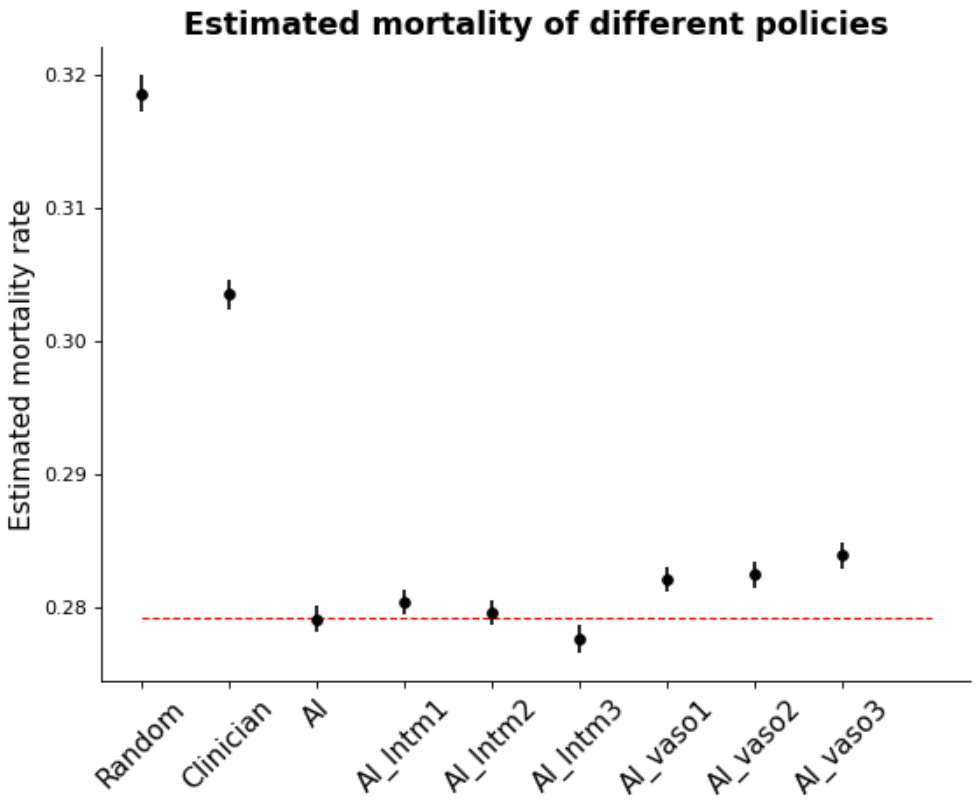

Figure 16 Estimated mortality rate comparison

3.4 **Model selection**

Among the three intermediate rewards framework experiments, AI_intm3 had the lowest mortality, but we did not choose it due to its instability of TWIS policy value. Rather than that, we chose the AI_intm2 policy because it achieved a higher TWIS policy value and competitive mortality than the unconstrained AI policy by simultaneously considering intermediate and long-term outcomes. In terms of reducing major vasopressor change, we chose AI_vaso1 after considering all evaluation criteria (TWIS policy value, estimated mortality, and proportion of vasopressor change), because AI_vaso1 had a lower proportion of dramatic vasopressor dose

change, a competitive TWIS policy value, and a competitive estimated mortality compared to an unconstrained AI policy.

3.5 **Feature importance interpretability evaluation**

We've already tested that our modified AI policies have better quantitative performance in the previous sections. We think it is important as well to demonstrate that they are clinically interpretable or reasonable, especially in such a high-risk setting. By fitting classification random forest models, we identified the relative feature importance. The independent variables in the random forest models were the 38 features used for state definition, and the dependent variable was the recommendation of vasopressor (yes or no, scale of dose is discarded). To avoid bias, we used permutation feature importance. We estimated the out-of-bag score of each feature in the test data set by permuting the value of each feature (L. Breiman et al., 2018).

The results of the modified AI policies were then compared to those of the clinician policy. As illustrated in Figure 17, SOFA plays a significant role in three policies, which is to be expected given that SOFA reflects sepsis-related organ failure. Lactate is also on the top of the list because it is associated with the requirement for vasopressor administration. Gender is given the least weight in all policies because it is expected to have no effect on the recommendation of vasopressor. As illustrated in the figure, clinicians focused primarily on the SOFA score and a small amount on platelets and creatinine, while giving little weight to the majority of other features. This is acceptable because physicians typically make decisions based on their experience after recognizing specific biomarkers. Whereas, when the AI algorithm recommends vasopressors, it

takes into account all available features. As we observed, the modified AI policy's importance weight is more evenly distributed across all features. This demonstrated that the recommendations generated by the modified AI policies are clinically interpretable and are primarily based on clinically relevant features.

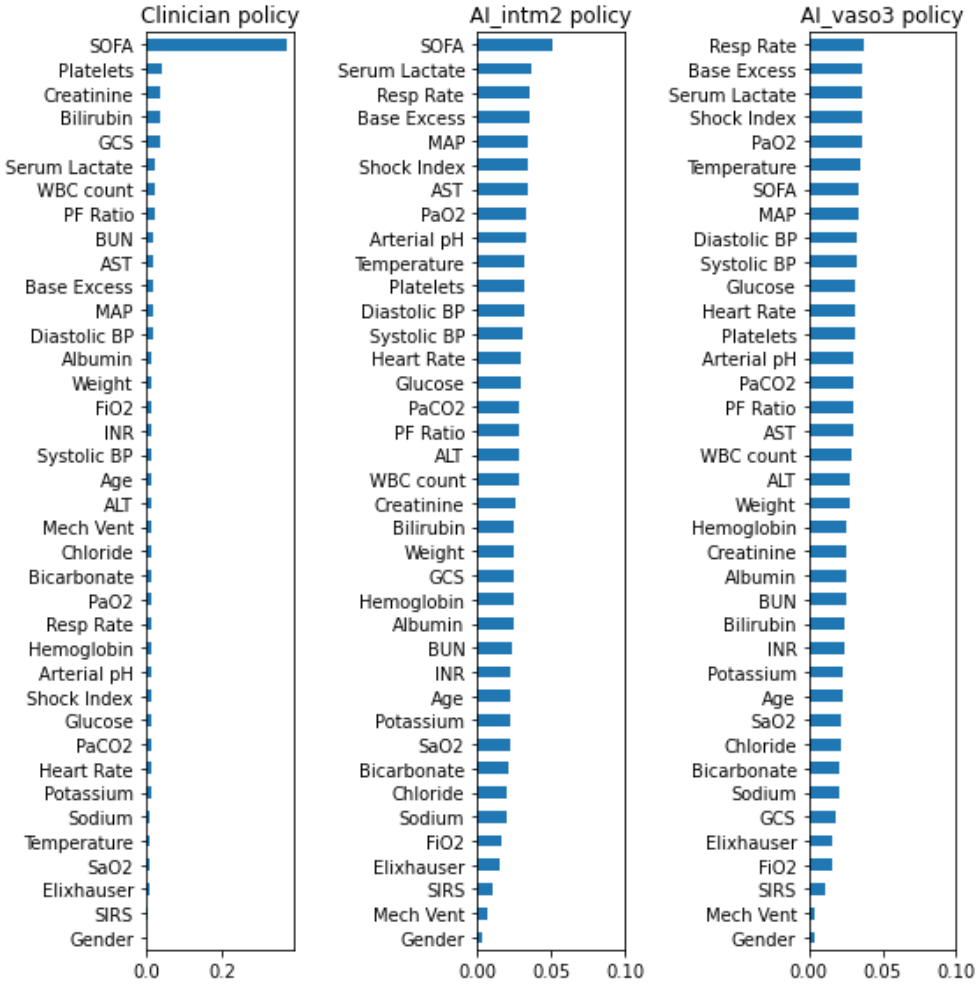

Figure 17 Permutation-based feature importance comparison

4.0 **Discussion and future work**

In this project, we demonstrated that by modifying the reward system, we can enhance the reinforcement learning process. Kennedy has previously demonstrated that AI policy learned via a Q-learning algorithm outperforms clinician policy. We extended this work by addressing two previously identified constraints: 1, failure to incorporate intermediate outcomes for patients; 2, abrupt changes in vasopressor dose use. Rather than focusing exclusively on end-point rewards, we propose two modified AI policies capable of independently addressing the two constraints. The modified AI_intm2 policy achieved competitive TWIS policy value when optimizing end-point outcomes and higher TWIS policy value when optimizing both intermediate and end-point outcomes than the unconstrained AI policy. Additionally, the AI_intm2 policy achieved a comparable estimated mortality to the unconstrained AI policy. Thus, we think AI_intm2 is more robust and applicable in real-world clinical settings. In comparison to the unconstrained AI policy, our second proposed policy, AI_vaso1, achieved a competitive TWIS policy value and 1.66 percent fewer patients who have such significant vasopressor dose changes. We observed that AI_intm2 policy and AI_vaso1 policy tended to avoid administering vasopressors to patients without acute sepsis (Figure 13), which was consistent with Pruinelli et al (2016).'s finds and clinically interpretable according to prior research that low-dose vasopressors are effective and safe in the treatment of sepsis (Mutlu & Factor, 2004). Whereas in patients with severe sepsis (high mortality), AI_intm2 and AI_vaso1 tended to recommend more vasopressor than the clinician policy (Figure 13). This is also in accordance with clinical professionals' understanding that acute septic patients are more likely to be administrator with vasopressor (Alaniz, C et al., 2013). We further confirmed that the decisions suggested by AI_intm2 policy and AI_vaso1 were clinically

interpretable and primarily relied on critical biomarkers associated with sepsis diagnosis, such as SOFA score, White Blood Cells, and Shock Index, as shown in Figure 17.

The strength of the proposed modified AI policies is that we addressed the two safety constraints that existed in prior work. In the AI_intm2 policy, we added intermediate rewards for outcomes such as patients' SOFA scores and arterial lactate levels that are related to intravenous fluid and vasopressor administration. We penalized high SOFA scores and increased SOFA scores, vice versa. In the same way, we penalized increased arterial lactate levels and rewarded decreased arterial lactate levels. The SOFA score and arterial lactate levels are both significant indicators that are associated with poor outcomes in sepsis treatment (Christopher W., et al, 2016; Sauer C.M., et al, 2021). With the regulation of the intermediate reward system, the modified policy AI_intm2 would more closely mimic clinicians' treatment policy by considering both instant health status and long-term outcome, which helps the AI_intm2 policy converge to a robust and safe AI policy. In AI_vaso1, we monitored the vasopressor changes over the trajectories of each patient. We penalized dramatic changes in vasopressor dose use within one time step and rewarded smooth vasopressor dose changes. In this way, we successfully decreased the proportion of patients who have such major changes in our test data set. 174 (2.8%) out of 6221 patients have this dramatic change under the AI_vaso1 policy, compared to 277 (4.5%) out of 6221 patients under the unconstrained AI policy. This approach ensures the AI_vaso1 prescribes vasopressor doses in a gradually increasing and decreasing manner. According to Ellender TJ et al, a sudden large change in vasopressor dose might be very harmful to sepsis patients. Thus, the AI_vaso1 policy is safer than the unconstrained AI policy in terms of clinical safety practice. Overall, we have modified the reward framework to address previous work's constraints and have learned two safer and more robust policies by incorporating clinical practice and knowledge.

Our endeavor is constrained by several constraints. To begin, we did not thoroughly investigate the reward function parameters (for both intermediate rewards and vasopressor penalty rewards systems). We only tested three different parameter groups for each reward function and chose the optimal final policies. We could improve our results if we conducted a more explicit examination of the parameters. The same holds true for the parameters used in the Q-learning algorithm and for the k value used in K-means clustering. Second, the dataset was filtered to include only patients admitted to the intensive care unit (ICU) to introduce more heterogeneity into an all-sepsis population. As a result, our learned AI policy is not generalizable to a different patient cohort. Third, we used only 38 numerical features to represent patients' health status and considered only two drug administrations, which is oversimplified in comparison to reality's complexity. Additionally, in a clinical setting, clinicians would have access to much more qualitative data, such as nurse reports. And, because many additional medications are prescribed concurrently, it is critical to understand how these medications interact. Fourth, we reduced the temporal resolution of our data to four hours. This enabled the AI agent to access certain laboratory values, which are not immediately available to doctors in a real clinical setting. As a result, our findings may be biased. Fifth, we have addressed the two safety constraints separately thus far; it would be more reasonable if we could find a reward function that addressed both constraints concurrently. Sixth, the discrete state space defined by K-means clustering may have underestimated the health status of patients. Furthermore, because Q-learning is a tabular method, it trains data in a highly correlated sequential order, which makes the Q-learning algorithm unstable. The two disadvantages of Q-learning can be overcome using the Deep Q-Network algorithm (Van et al., 2016), which employs an experience replay buffer and freezes the target

network to reduce instability. Additionally, DQN's input is a continuous matrix of patient characteristics.

In the future, we'd like to experiment with two approaches. To begin, we will attempt to develop a reward function that takes both intermediate outcomes and vasopressor dose changes into account simultaneously. Then, we will attempt to thoroughly explore all parameters, such as k in K-means clustering; the C parameters in the reward function; α, γ and ε in Q-learning. Once we have this final optimal policy, we will attempt to test its generalizability on an external data set, such as the MIMIC III data set, without further learning or tuning. Attempting to implement DQN on our dataset is our second future direction. Raghu et al. (2017) have already developed a DQN model and demonstrated excellent work in sepsis treatment drug recommendation. DQN outperforms Q-learning by incorporating two critical improvements to the Q-learning algorithm. DQN employs an experience replay buffer to mitigate the instability caused by training on highly correlated sequential data. Additionally, DQN freezes the target network to mitigate the instability caused by locating a moving target. Rather than inputting discrete predefined patient states, patient data (a multidimensional matrix) will be fed into a convolutional neural network, with the output neurons representing the number of actions that an AI agent can perform. Continuous states may alleviate the problem of heterogeneity within discrete states.

Appendix: Python code

This work is implemented in Python 3.7, the code can be found in my GitHub repository at https://github.com/lilinglu/Reinforcement-Learning-For-Sepsis-Treatment.

Appendix Table 1 Description of included features

Table 8 Appendix Table

Feature	Original	Missingness	Post-imputation	Tansformation
Demographics				
Age (mean(sd))	64(16)	0.0%	64(16)	Standardization
Elixhauser (mean(sd))	5.3(2.3)	0.2%	5.3(2.3)	Standardization
Weight (mean(sd))	85(29)	2.6%	85(29)	Ln + Standardization
Gender/male (n(%))	156,664(51%)	0.0%	156,664(51%)	Standardization
Vital Signs				
Distolic BP (median(IQR))	69(60-80)	0.5%	69(60-80)	Ln + Standardization
Heart Rate (mean(sd))	95(21)	0.4%	95(21)	Ln + Standardization
GCS (mean(sd))	12.1(3.5)	16.0%	12.4(3.3)	Standardization
MAP (median(IQR))	89(79-101)	0.5%	89(79-101)	Ln + Standardization
Respiratory Rate (mean(sd))	21(6)	0.4%	21(6)	Standardization
Temperature (mean(sd))	36.8(0.9)	0.6%	36.8(0.9)	Standardization
SOFA (mean(sd))	3.5(2.9)	0.0%	5.4(3.3)	Standardization
SIRS (mean(sd))	1.6(1.0)	0.4%	1.8(1.1)	Standardization
Systolic BP (median(IQR))	128(113-146)	0.4%	128(113-146)	Ln + Standardization
Shock Index (mean(sd))	0.8(0.2)	0.4%	0.8(0.2)	Ln + Standardization
Laboratory Measurements				
Albumin mean(sd))	2.6(0.6)	37.0%	2.7(0.6)	Standardization
ALT (median(IQR))	31(17-77)	38.0%	24(17-41)	Ln + Standardization
AST (median(IQR)	42(22-115)	38.0%	30(21-54)	Ln + Standardization
Base Excess (median(IQR))	-2.1(7.5)	50.0%	-1.0(6.0)	Standardization
Bilirubin (median(IQR))	0.8(0.5-1.6)	38.0%	0.6(0.5-1.0)	Ln + Standardization
Bicarbonate (mean(sd))	23(6)	4.3%	24(6)	Standardization
BUN (median(IQR))	28(17-47)	4.3%	26(16-42)	Ln + Standardization
Chloride (mean(sd))	106(8)	3.8%	105(7)	Standardization
Creatinine (median(IQR))	1.4(0.9-2.5)	4.3%	1.3(0.8-2.1)	Ln + Standardization
FiO_2 (median(IQR))	50(40-70)	40.0%	40(40-50)	Standardization
Glucose (median(IQR))	148(114-201)	3.7%	135(109-176)	Ln + Standardization
Hemoglobin (mean(sd))	10(2)	3.6%	11(2)	Ln + Standardization
INR (median(IQR))	1.5(1.2-2.1)	39.0%	1.3(1.2-1.6)	Ln + Standardization
Potassium (mean(sd))	4(1)	3.5%	4(1)	Standardization
Lactate (median(IQR))	2.1(1.3-3.7)	34.0%	1.4(1.1-2.1)	Ln + Standardization
Sodium (mean(sd))	139(7)	3.9%	139(5)	Standardization
SaO_2 (median(IQR))	95(93-98)	0.4%	95(93-98)	Inverse Ln + Standardization
$PaCO_2$ (mean(sd))	44(16)	50.0%	42(11)	Ln + Standardization
PaO_2 (mean(sd))	130(79)	51.0%	103(46)	Ln + Standardization
PF Ration (meidan(IQR))	223(143-332)	40.0%	250(163-375)	Ln + Standardization
Arterial pH (mean(sd))	7.3(0.1)	50.0%	7.4(0.1)	Standardization
Platelets (median(IQR))	173(114-241)	5.0%	7.4(0.1)	Ln + Standardization
WBC Count (median(IQR))	12(8-17)	5.3%	12(8-16)	Ln + Standardization
Ventilation Parameter				
Mesh Vent in Window (n(%))	122,464(40%)	0.0%	122,464(40%)	Standardization
Generated feature				
Vaso_change (mean(sd))	0.004(0.16)	0.0%	0.004(0.16)	Standardization

Abbreviations: GCS, Glasgow Coma Scale score; MAP, mean arterial pressure; SOFA, sequential organ failure assessment; SIRS, systemic inflammatory response syndrome; BP, blood pressure; ALT, alanine aminotransferase; AST, aspartate aminotransferase; BUN, blood urea nitrogen; FiO_2, fraction of inspired oxygen; $PaCO_2$, partial pressure of arterial carbon dioxide; PaO_2, partial pressure of arterial oxygen; PF Ration, ratio of $PaCO_2$ and FiO_2; WBC, white blood cell.

Appendix Figure 1 TWIS policy value evaluation

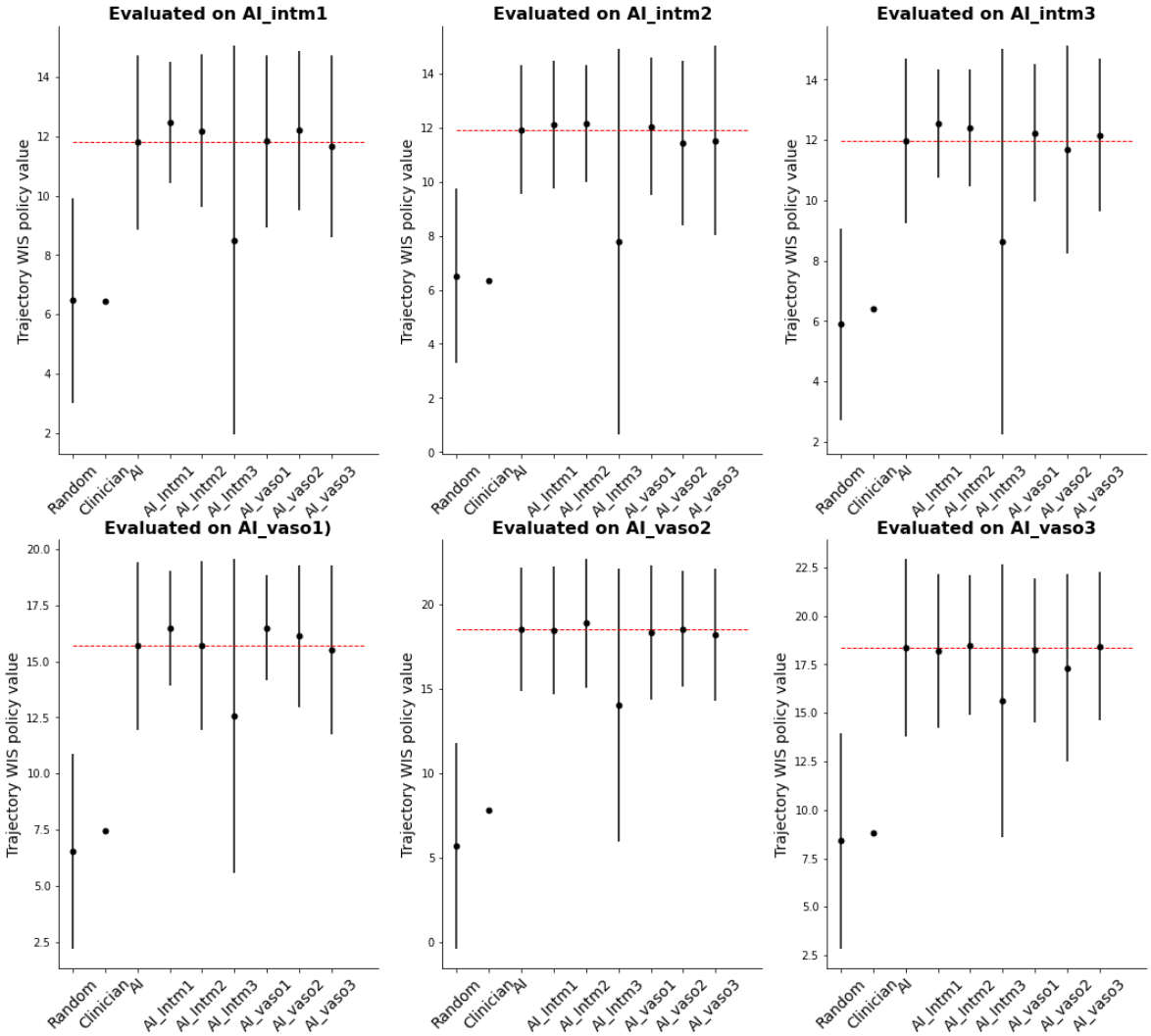

Figure 18 Appendix Figure

Bibliography

Alaniz, Cesar PharmD; Pollard, Sacha PharmD, BCPS Vasopressor Dosing in Septic Shock, Critical Care Medicine: December 2013 - Volume 41 - Issue 12 - p e483-e484 doi: 10.1097/CCM.0b013e3182916fe7

Allen, Jeanne Maree, and Suzie Elizabeth Wright. "Integrating theory and practice in the pre-service teacher education practicum." Teachers and teaching 20.2 (2014): 136-151.

Angus, Derek C., and Tom Van der Poll. "Severe sepsis and septic shock." N Engl J Med 369 (2013): 840-851.

Angus, Derek C., et al. "Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care." Critical care medicine 29.7 (2001): 1303-1310.

Bassi E, Park M, Azevedo LC. Therapeutic strategies for high-dose vasopressor-dependent shock. Crit Care Res Pract. 2013;2013:654708. doi: 10.1155/2013/654708. Epub 2013 Sep 15. PMID: 24151551; PMCID: PMC3787628.

Bennett, Casey C., and Kris Hauser. "Artificial intelligence framework for simulating clinical decision-making: A Markov decision process approach." Artificial intelligence in medicine 57.1 (2013): 9-19.

Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.

Brown, Samuel M., et al. "Relationships among initial hospital triage, disease progression and mortality in community-acquired pneumonia." Respirology 17.8 (2012): 1207-1213.

Byrne, Liam, and Frank Van Haren. "Fluid resuscitation in human sepsis: time to rewrite history?" Annals of intensive care 7.1 (2017): 1-8.

Celi, Leo Anthony, et al. "From pharmacovigilance to clinical care optimization." Big Data 2.3 (2014): 134-141.

Doerr F, Badreldin AM, Heldwein MB, et al. A comparative study of four intensive care outcome prediction models in cardiac surgery patients. J Cardiothorac Surg. 2011;6:21.

Dugani, Sagar et al. “Reducing the global burden of sepsis.” CMAJ: Canadian Medical Association journal = journal de l'Association medicale canadienne vol. 189,1 (2017): E2-E3. doi:10.1503/cmaj.160798

Ellender TJ, Skinner JC. The use of vasopressors and inotropes in the emergency medical treatment of shock. Emerg Med Clin North Am. 2008 Aug;26(3):759-86, ix. doi: 10.1016/j.emc.2008.04.001. PMID: 18655944.

Fadale, Kristin Lavigne, et al. "Improving nurses' vasopressor titration skills and self-efficacy via simulation-based learning." Clinical Simulation in Nursing 10.6 (2014): e291-e299.

Ghassemi, Marzyeh, Leo Anthony Celi, and David J. Stone. "State of the art review: the data revolution in critical care." Critical Care 19.1 (2015): 1-9.

Gotts, Jeffrey E., and Michael A. Matthay. "Sepsis: pathophysiology and clinical management." Bmj 353 (2016).

Jiang, Nan, and Lihong Li. "Doubly robust off-policy value evaluation for reinforcement learning." International Conference on Machine Learning. PMLR, 2016.

Kennedy, Neal Jason. Towards a Learning Health System: Using Reinforcement Learning to Optimize Treatment Decisions in Sepsis Patients, Master thesis, University of Pittsburgh, 2021

Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. Nature Med. 2018;24:1716-1720.

Marik, P., and Rinaldo Bellomo. "A rational approach to fluid therapy in sepsis." BJA: British Journal of Anaesthesia 116.3 (2016): 339-349.

Minne L, Abu-Hanna A, deJonge E. Evaluation of SOFA-based models for predicting mortality in the ICU: A systematic review. Crit Care. 2008;12(6):R161.

Mutlu, Gökhan M., and Phillip Factor. "Role of vasopressin in the management of septic shock." Intensive care medicine 30.7 (2004): 1276-1291.

Pruinelli, Lisiane, et al. "A data mining approach to determine sepsis guideline impact on inpatient mortality and complications." AMIA Summits on Translational Science Proceedings 2016 (2016): 194.

Raghu, Aniruddh, et al. "Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach." Machine Learning for Healthcare Conference. PMLR, 2017.

Rello, J., Valenzuela-Sánchez, F., Ruiz-Rodriguez, M. et al. Sepsis: A Review of Advances in Management. Adv Ther 34, 2393–2411 (2017). https://doi.org/10.1007/s12325-017-0622-8

Sauer, C.M., Gómez, J., Botella, M.R. et al. Understanding critically ill sepsis patients with normal serum lactate levels: results from U.S. and European ICU cohorts. Sci Rep 11, 20076 (2021). https://doi.org/10.1038/s41598-021-99581-6

Seymour, Christopher W., et al. "Assessment of clinical criteria for sepsis: for the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)." Jama 315.8 (2016): 762-774.

Singer M, Deutschman CS, Seymour CW, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). JAMA. 2016;315(8):801.

Sutton, Richard S., and Andrew G. Barto. Reinforcement learning: An introduction. MIT press, 2018.

Van Hasselt, Hado, Arthur Guez, and David Silver. "Deep reinforcement learning with double q-learning." Proceedings of the AAAI conference on artificial intelligence. Vol. 30. No. 1. 2016.

Watkins, Christopher John Cornish Hellaby. "Learning from delayed rewards." (1989).

Yende, Sachin, et al. "Long-term quality of life among survivors of severe sepsis: analyses of two international trials." Critical care medicine 44.8 (2016): 1461.