# Guiding the management of sepsis with deep reinforcement learning

Stephen Pfohl
Stanford University
spfohl@stanford.edu

Ben Marafino
Stanford University
marafino@stanford.edu

## Abstract

*Sepsis is an acute and often life-threatening condition necessitating intensive care, and is associated with high morbidity and mortality in the U.S. – where over 1.5 million cases are reported annually with a mortality rate approaching 20% – and elsewhere. Despite its toll, effective treatments for sepsis do not yet exist, nor are current treatment protocols individualized for patients on the basis of their physiological state and other clinical parameters. There thus exists an unmet need for clinical decision support tools that can guide physicians and other providers in their management of sepsis at all stages, and to personalize current treatments. As a first step towards such systems, we show that a deep Q-learning approach, based on a Double Dueling Q-Network (DDQN) architecture, can be applied to a large sample of observational data taken from patients with sepsis. We also show that this approach is able to meaningfully estimate the Q-values of the empirical policy as executed by physicians over the clinical courses of the patients in our dataset. In addition, the Q-value trajectories, despite not being well-calibrated, appear to discriminate well between those patients who survived and those who died. Finally, we also show that our network is able to estimate an optimal policy which differs from that of physicians', which may suggest alternative protocols for sepsis treatment.*

## 1. Introduction

Sepsis is an acute medical condition which principally manifests as a disproportionate immune response to infection. It is characterized by widespread inflammation and coagulopathy, both of which lead to global tissue hypoperfusion and eventually, shock and end-stage organ failure [1]. Mortality rates among patients with sepsis are high: approximately 20 to 30% of patients with sepsis die, and this rate approaches 70% among patients experiencing septic shock [2]. As of 2014 in the United States, sepsis accounts for over 1.5 million cases and 250,000 deaths annually [3], and may be linked to as many as half of all in-hospital deaths nationwide [4].

Despite the grave toll of sepsis, no specific protocols for its treatment have yet been validated in large, randomized multi-center studies, nor are any drugs marketed specifically for treating sepsis. Current approaches are primarily supportive and focus on early cardiorespiratory resuscitation followed by antibiotic therapy, with the aim of maintaining hemodynamic stability and thus organ perfusion [5]. Such resuscitation can be achieved with the administration of intravenous fluid, with or without vaspressor medications which act to increase blood pressure and oxygen delivery to tissues, and is often carried out in a goal-directed manner.

For example, critical care physicians will commonly order fluid resuscitation for patients in the early stages of sepsis, with the goal of meeting predefined hemodynamic targets. However, while the increasing prevalence of such *goal-directed therapy* appears to be associated with declining sepsis mortality over the past decade [6], it remains an "one-size-fits-all" strategy. With the increased proliferation of electronic health records (EHRs), the current bounty of clinical data could could be leveraged to further personalize treatment protocols for patients with sepsis, potentially informing existing protocols and decision support systems, thus improving outcomes.

In this work, we set out to determine if current practice patterns of sepsis treatment could be learnt from observational data and thus encapsulated in a policy, and also to determine whether these policies were qualitatively different among survivors compared to non-survivors. To do so, we aimed to primarily build off of the prior work of [7] by independently implementing and validating their published model as a foundation for our future work in this domain.

## 2. Context and related work

Data-driven approaches to tackling the problem of sepsis have attracted considerable interest in recent years. Notably, EHR data have been used to develop early-warning systems to predict sepsis onset among inpatients with high sensitivity and specificity [8]. However, the problem of sequential decision-making and decision support in sepsis, and in critical care medicine more generally, do not appear to be as well studied. Raghu *et al.* [7] applied deep reinforcement learning to learn and evaluate optimal treatment policies for sepsis using the MIMIC-III database. Nemati *et al.* [9] also applied deep RL to learn heparin dosing policies from suboptimal examples in observational data, while Prasad *et al.* [10] used fitted $Q$-iteration (FQI) – a form of $Q$-learning – to learn policies for weaning ICU patients off of mechanical ventilation; both these approaches were developed using the MIMIC database.

More broadly, reinforcement learning has also been applied to sequential decision problems in healthcare beyond the inpatient setting, outside of which the data are not as temporally dense. Ernst *et al.* [11] used fitted $Q$-iteration to learn optimal treatment strategies for patients with HIV; such strategies are complex to learn, due to antiviral resistance, leading to patients being cycled on and off of therapy over yearly timescales. Escandell-Montero *et al.* [12] again used FQI to optimize erythropoietin-stimulating agent dosing among hemodialysis patients with anemia, with the goal of stabilizing patients' hemoglobin levels and minimizing side effects. Zhao *et al.* [13] used $Q$-learning to select from among strategies for treating non-small cell lung cancer from clinical trial data.

## 3. Methods

### 3.1. MDP formulation

The sequential decision problem of optimal sepsis treatment can be modeled as a *Markov decision process* (MDP), which is defined by the 5-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \gamma, R)$, where

1. $\mathcal{S}$ is a finite state space which enumerates all possible patient states $s_t$ for each time $t$;

2. $\mathcal{A}$ is the action space, which similarly enumerates all possible actions $a_t$ that can be taken by the agent, which we take to represent the actions of medical providers;

3. $\mathcal{T}$ is the set of transition functions $P(s_{t+1} \mid s_t, a_t)$ which give the transition probabilities over states, conditional on the current state and action at time $t$;

4. $\gamma$ represents the discount factor, which controls the relative weight of short- and long-term rewards.

5. $R = R(s_t, a_t)$ is the reward function, which yields feedback following a transition $s_t \to s_{t+1}$, given that action $a_t$ was performed.

Generally, given a MDP, the goal is to then learn a policy $\pi$ which specifies an action $a_t \in \mathcal{A}$ to be taken for every state $s_t \in \mathcal{S}$, and which maximizes, in expectation, the total accumulated reward

$$R^{\pi}(s_t) = \lim_{T\to\infty} \mathbb{E}_{s_{t+1}|s_t,\pi(s_t)} \left[ \sum_{t=1}^{T} \gamma^t r(s_t, a_t) \right]$$

from that state $s_t$, which correspond to the state-action values via $\max_a Q(s,a)$. For this problem, the state space is the set of all possible patient state vectors comprising laboratory, vital sign, demographic, and severity score variables. However, by leveraging deep neural networks to embed the states in a distributed representation, we implicitly consider the state space as continuous and do not directly model the transition probabilities. The action space is defined over IV fluid intakes and vasopressor dosing rates (i.e. mcg/kg/hr), and discretized into quartiles of nonzero intakes and dosing rates, with a bin for those that were zero, thus creating a $5 \times 5$ grid of possible actions, as in [7]. The details of the data preprocessing steps are presented in §3.4 and 3.5. Finally, in-hospital mortality was used as the reward signal: a reward of $r_t = R$ was assigned on a terminal timestep if a patient survived to discharge and $r_t = -R$ otherwise. We set $R = 15$.

### 3.2. Deep Q-Networks

A deep learning approach to $Q$-learning [14] was used to learn treatment policies. As in [7], we used the Double-$Q$ learning algorithm [15] in conjunction with the Dueling network architecture [16], as these approaches have been shown to the mitigate the effect of overestimation of the Q-function that the base $Q$-learning algorithm is prone to.

The following definitions of the Dueling network architecture and the Double-$Q$ learning algorithm match those of [15, 16]. Let $Q(s, a; \theta)$ be a neural network (the *main network*) parameterized by $\theta$ that is an estimator of the value of states $s$ and actions $a$. In addition, let there be an additional model called the *target network* $Q(s', a'; \theta^-)$ parameterized by $\theta^-$ where $s'$ is the state that is transitioned to during exploration.

In the Double-Q learning algorithm, a mixture of both models is used to compute

$$y_i^{DDQN} = r + \gamma Q(s', \operatorname*{argmax}_{a'} Q(s', a'; \theta); \theta^-)$$

With the Dueling architecture, the state-action value function $Q$ is decoupled into separate value $V(s; \theta, \beta)$ and advantage $A(s, a; \theta, \alpha)$ streams, as follows:

$$\begin{aligned} Q(s, a; \theta, \alpha, \beta) = V(s; \theta, \beta) + \\ A(s, a; \theta, \alpha) - \frac{1}{|\mathcal{A}|} \sum_{a'} A(s, a'; \theta, \alpha), \end{aligned}$$

where the advantage function $A(\cdot)$ represents the relative contribution of an action to the overall values for a given state. The parameters $\alpha$ and $\beta$ are the parameters of neural network layers that are specific to the advantage and value streams, respectively. In this formulation, both the main and the target network are estimated with dueling architectures. During training, the parameters $\theta^{-}$ of the target network are held fixed and updated to match that of the main network every $T$ iterations.

We used the Huber loss [17] between $Q(s, a; \theta)$ and $y^{DDQN}$, to train the model. The Huber loss may be considered as an alternative to the mean squared error that is more robust to extreme values and outliers.

### 3.3. Dataset

The data were drawn from the Medical Information Mart for Intensive Care-III (MIMIC-III) database, [18] which contains data collected from the ICU stays of 38,597 unique adult patients from between 2001 and 2012 at the Beth Israel Deaconess Medical Center (Boston, MA). MIMIC-III exhibits several characteristics that make it particularly attractive for this task. In particular, it allows for reasoning over temporally-dense streams of clinical time series of laboratory measurements, vital signs, bedside monitor waveforms, and fluid and medication infusions alongside contextually relevant demographic and diagnostic features in coded fields and free-text clinical notes. As MIMIC-III is publicly available and frequently used in clinical applications of machine learning [19], the results of this project may be reproduced by others.

The outcome used was in-hospital mortality, which we used to define the reward. We preferred in-hospital mortality to ICU mortality, since critically ill patients who are expected to die often are transferred out of ICUs prior to death, as well as to 90-day mortality, since using such an outcome could potentially result in an even sparser reward signal, making the estimation of state-action values more difficult.

In line with recent clinical guidelines, we used the Sepsis-3 consensus definition of sepsis [20] to identify patients who have developed sepsis. The Sepsis-3 definition is based on the Sequential Organ Failure Assessment scoring system and its simplified analogue, qSOFA ("quick SOFA"). The qSOFA can be used as a quick initial screen for sepsis that does not require laboratory testing, and is based on the presence of the following diagnostic indicators: systolic blood pressure (SBP) $\leq$ 100 mm Hg; altered mental status; and respiratory rate greater than 22/minute. According to the Sepsis-3 definition, a qSOFA score of 2 or greater is suspicious for sepsis, and a SOFA score of 2 or greater definitively establishes the presence of sepsis. To form our cohort, we the first ICU stays for all adult patients having SOFA scores of 2 or greater at any point during their hospitalization. The MIMIC-III database also contains a variable that establishes the suspected infection time for patients with sepsis, and we used this variable to establish the onset of sepsis for the patients in our cohort.

Once selected, the set of all 17,122 patients with sepsis in the MIMIC-III database was split into three subsets: train, validation, and test sets in a 80:10:10 ratio. This yielded sets of size 13,700, 1702, and 1720, respectively.

The relevant SQL queries to define and extract the cohort of interest and define sepsis related endpoints were derived from the public MIMIC-CODE repository at `https://github.com/MIT-LCP/mimic-code` [21] and [22].

### 3.4. State space: specification and preprocessing

A list of variables comprising patient state are given in Table 1. These variables comprise laboratory tests, vital signs, and patient demographics. Also included are the Logistic Organ Dysfunction System (LODS) log-odds estimates of mortality [23], as well as the raw Sequential Organ Failure Assessment (SOFA) [24] score and the Elixhauser comorbidity index scores [25]; the former assesses a patient's level of organ dysfunction in terms of a subset of these state variables, while the latter measures a patient's chronic comorbidity burden. Examples of chronic comorbidities included in the Elixhauser index are diabetes, cancer, chronic obstructive pulmonary disease, and chronic heart failure, among other conditions.

The data elements corresponding to these variables were extracted from the MIMIC-III database. Since laboratory test and vital sign results could derive from multiple sources, these data were normalized so that all sources mapped to the same variable. For example, blood glucose measurements can be performed by a variety of different methods and/or machines, resulting in multiple sources for the state variable of glucose level in the data. However, these different tests have common reference ranges and so can be considered to be interchangeable, thus motivating the normalization.

Patient trajectories consist of the 24 hours of data preceding the suspected onset of sepsis, and the 48 hours following it, discretized into 4 hour windows, and for each window, we calculated the mean value of each variable based on the corresponding datapoints falling within that window. If no data were present within a window, we simply carried forward the last observed mean value for that variable. All variables were then centered and scaled to have mean 0 and variance 1, and missing values not imputed previously were imputed using the median value for that variable across non-missing observations.

### 3.5. Action space: specification and preprocessing

The action space is discrete and consists of a 5 × 5 grid made up of quartiles based on the nonzero IV fluid intakes and vasopressor dosing rates observed in the data. These

| | |
|---|---|
| Vital signs | Systolic blood pressure (mmHg) |
| | Diastolic blood pressure (mmHg) |
| | Mean blood pressure (mmHg) |
| | Heart rate (bpm) |
| | Respirations ($\text{min}^{-1}$) |
| | Oxygen saturation/$\text{SpO}_2$ (%) |
| | Body temperature (degrees Celsius) |
| | Height (m) |
| | Weight (kg) |
| Laboratory values | Albumin (g/dL) |
| | Anion gap (mEq/L) |
| | Band cell count ($\times 10^9$/L) |
| | Bicarbonate (mEq/L) |
| | Bilirubin (mg/dL) |
| | Blood urea nitrogen (mg/dL) |
| | Chloride (mEq/L) |
| | Creatinine (mg/dL) |
| | Glucose (mg/dL) |
| | Hematocrit (%) |
| | International normalized ratio |
| | Prothrombin time (sec.) |
| | Partial thromboplastin time (sec.) |
| | Platelets ($\times 10^9$/L) |
| | Potassium (mEq/L) |
| | Sodium (mEq/L) |
| | White blood cell count ($\times 10^9$/L) |
| Patient demographics | Age (years) |
| | Race (White, Hispanic, Black, Other) |
| Severity scores | Logistic Organ Dysfunction System |
| | Sequential Organ Failure Assessment |
| | Elixhauser Index |

Table 1. List of patient state variables and their units, where applicable.

intakes and doses were taken over the same 4-hour windows used to define the state variables. Over each window, we took the sum of all IV fluid intake by the patient, as well as the maximum vasopressor dosing rate that was set, and these were used to form the quartiles. Together with these quartiles, there also is a "no-op" action corresponding to no fluid or vasopressors given within the window, for a total of five possible actions for both types of intervention.

### 3.6. Evaluation

The deep $Q$-network model was trained by minimizing the temporal difference (TD) error produced by the Double $Q$-learning algorithm with a Huber loss, as previously described. While the network is trained to minimize this loss, it is difficult to directly use the loss to evaluate convergence and generalization error since the outputs of the target network that the main network tries to match are non-stationary over the course of the training procedure. To mitigate these issues, we assess the quality of the model holistically. For one, we monitor the mean estimated $Q$-values for the optimal policy (i.e. $\mathbb{E}_{s' \sim validation}[\max_{a'} Q(s', a')]$) over the set of states in the training set and the held-out validation set and perform model selection on the basis of the largest optimal mean Q-value on the validation set for a set of models over a grid of hyperparameters (see §3.7). Second, we use the held-out test set to qualitatively evaluate the $Q$-values assigned by the trained model to both the physician and predicted optimal policies in the test set, and investigate the association of those predictions with the eventual in-hospital mortality. Third, we compare the number of times that the model recommends each action with the number of times each action was taken on the test set to get a qualitative sense of the behavior of the learned policy in comparison with the standard of care.

### 3.7. Experimental details

We performed grid search in order to determine the optimal choices for the hyperparameters of the number of layers and the number of hidden units per layer for our networks; the grid used consisted of $[1, 2, 3, 5, 10]$ hidden layers and $[128, 256, 512, 1024]$ hidden units per layer, respectively. The parameters $\alpha$ and $\beta$ of the advantage and value streams in the dueling architecture were constrained to be linear transformations. All networks were trained using the Adam optimizer [26] with a minibatch size of 512 and a learning rate of $10^{-4}$ where the minibatches were formed by sampling uniformly from the set of state transitions in the training set. A training epoch was defined as a full update being applied over the full set of state transitions in the training set. The weights of the target network were updated to match the weights of the main network every $T = 10$ training epochs. The network weights were initialized with Glorot initialization [27].

Training was performed on a NVIDIA Tesla K80 GPU. All code necessary to produce the models and training procedures was developed in Python 3.6 and with the PyTorch library. The relevant code may be found at `https://github.com/spfohl/cs238_sepsis_rl`

## 4. Results

We identified 17,122 patients with sepsis in the MIMIC-III database for training, and for these patients, we were able to create 194,010 state transitions comprising the four-hour windows taken over the ICU stays for these patients. As previously described, these states were partitioned across three sets which made up the training, validation, and test sets. The network was trained to minimize the loss on the training set while monitoring the current estimate of the mean $Q$-value for the current optimal policy for all states in the validation set for model selection. The results of the hyperparameter tuning procedure are presented in Figure 1, and on the basis of these results, we used a network with 5 hidden layers and a dimension of 1024 in each hidden layer.

With this network, we were able to estimate the $Q$-values for the physician policy as well as of the optimal policy as estimated by the network for the states in the test set, and these results are presented in Figure 2 with the patients stratified by outcome. These $Q$-values are based on taking the mean of the $Q$-value for the optimal action (i.e., the action with the highest estimated $Q$-value) across all the states in the test set for each timepoint in the 72-hour window for each patient.

From this figure, we observe that the optimal policy predicted by the model achieves a larger predicted $Q$-value relative to the observed physician policy for all states in the test set regardless of whether the patient died in the hospital or not. As one would expect, the model generally assigns higher value to states that correspond to patients who survive their stay relative to the estimate for those that die. However, it is apparent that the model is out of calibration and is perhaps over-estimating the value of all state-action pairs. To see that this is the case, consider that a well-calibrated model should assign a value close to $-R = -15$ for the physician actions taken on patients who eventually died, while our trained model assigns a mean $Q$ greater than 11 for the physician policy on patients that later die. The increased variance in the estimates and the decline in the estimated $Q$-values among patients who survive later in the trajectory indicates that the model may have difficulties decoupling the value of the states from the actions taken. As a result, it is likely that increasing the representational capacity of the sub-networks $\alpha$ and $\beta$ private to the advantage and value streams in the dueling architecture would be advantageous.

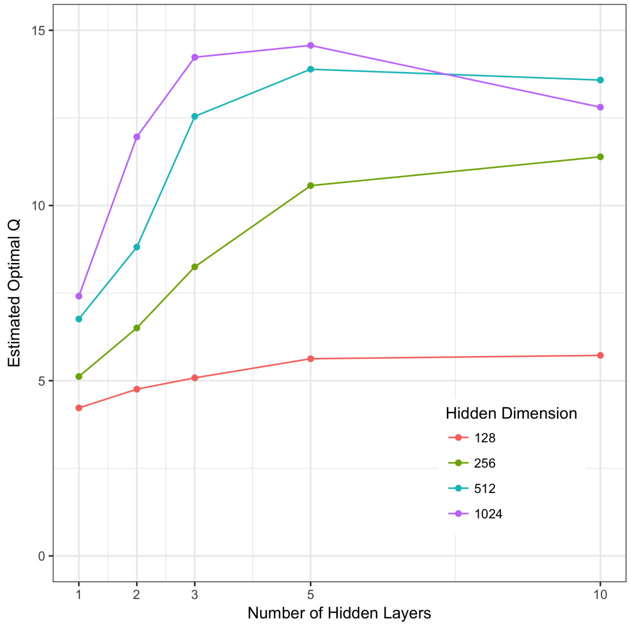

Figure 1. Results of the hyperparameter tuning procedure. Mean $Q$-values of the optimal policy predicted by the network for the validation set when varying the number of hidden layers and the number of hidden units per layer

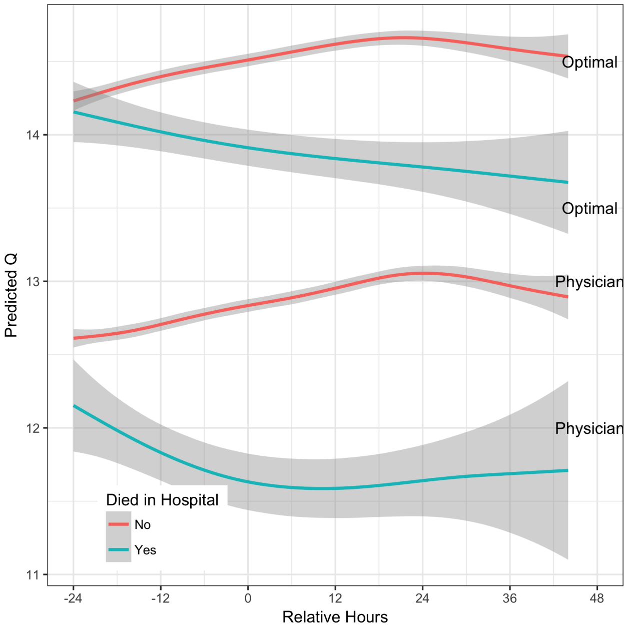

Figure 2. The $Q$-values for the physician and optimal policies as estimated by the DDQN over the 72-hour trajectories for the patients in the test set stratified by outcome. Note in particular how the physician policies provide an upper bound for the $Q$-values of the optimal policy in each subgroup. The gray bands represent the 95% confidence intervals.

As in [7], our results in Figure 3 demonstrate that, for the set of states in the test set, that the learned optimal policy recommends much heavier usage of vasopressors and less usage of IV fluids in comparison to the observed physician policy that tends to more heavily rely on IV fluids. Significantly, it is apparent that in practice (for the test set), that physicians were never observed to be using vasopressors without IV fluids and the model makes similar recommendations for cases in which no IV fluids are given, but it is not clear if this similarity is a generalizable recommendation or just an artifact of rarity of that region of the action-space in the data.

## 5. Discussion

Our approach of deep $Q$-learning, based on a dueling deep $Q$-network (DDQN) architecture, proved capable of learning the quality of empirical policies from observational data corresponding to the state of sepsis treatment as currently carried out by physicians and was able to make rec-

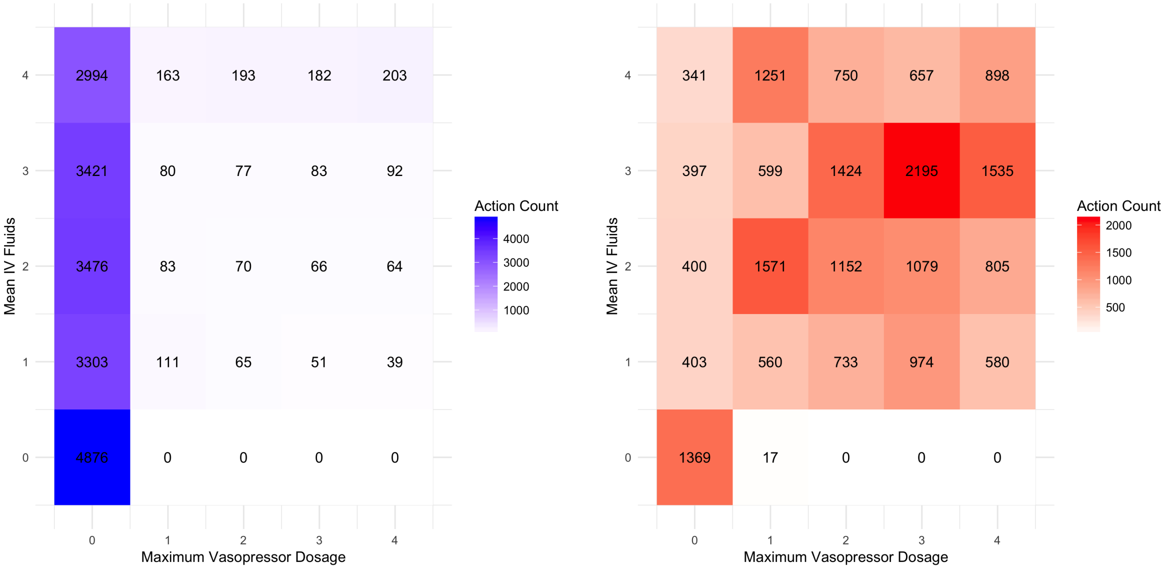

Figure 3. Distributions of actions for both the physician policy (left) and the estimated optimal policy (right).

ommendations as to how to improve the treatment of sepsis and ultimately improve mortality. In particular, our results suggest that relying less on IV fluids and using more vasopressors would improve outcomes. Further investigation is required to determine to what extent these claims hold up in practice.

A chief limitation of our approach was that each patient in the dataset was taken to be septic throughout the entirety of their clinical trajectory, including over the timesteps preceding their diagnosis, even though their vital signs and other state variables may not have been indicative of sepsis at those times. This knowledge is a luxury almost never afforded to physicians in real-world clinical settings, as sepsis often presents with significant lead time, during which its diagnosis may remain ambiguous, and easily mistaken for other, less acute conditions. With this knowledge, our agent was able to execute the learned treatment policies immediately for each patient in the dataset, thus affording it a significant advantage over physicians. In improving upon this work, future approaches should aim to learn policies from data that incorporate *both* patients with and without sepsis.

In such a data context, one approach may be to *couple* the agent with a sepsis early-warning model, such as TREWScore [8] in order to determine the precise timestep at which treatment ought to be initiated. Alternatively, these could be learned directly by the agent, perhaps by modifying the reward signal to be more sensitive to sepsis-related changes in patient state in the short term that could form dynamic thresholds for initiation of treatment. We postulate that learning policies from a dataset with a diverse case-mix as opposed to one consisting entirely of patients with sepsis could learn such thresholds implicitly. Concerns with this approach include possible overuse as well as inappropriate use of therapies, including vasopressor drugs, the use of which is associated with severe side effects. In addition, avenues for further improvement include modifying the task so that the agent can learn to act in continuous time – with the impetus based on the finding that each hour of delay in the initiation of antibiotic therapy is associated with markedly increased mortality in patients with sepsis [28].

The evaluation of learned policies is intrinsically hard for off-policy models and hard in general for the retrospectively collected observational data inherent to the electronic health record due to biases associated with its collection. In particular, we have no reliable way to assess the value of a new proposed policy with either a simulator or in the clinic due to safety and ethical concerns. Ideally, we would leverage estimators based on the Double Robust Off-Policy Value Estimation (DROVE) [29, 30] method to reduce the bias associated with estimating the ability of the learned policy to reduce sepsis-related in-hospital mortality, but we delegate that task to future work.

## 6. Conclusion

Even as treatment protocols continue to improve, the human toll of sepsis remains unacceptably high. The tools of reinforcement learning can be used to develop methods that aid clinical decision support for critical care medicine physicians treating sepsis, and to assist researchers in optimizing and evaluating candidate treatment strategies. We propose to build on, and improve upon, prior attempts to tackle this problem. Even though our agents are incapable of exploration in this setting, learning robust policies from observational data could motivate the development of decision support systems and ultimately, perhaps better and more personalized treatments for sepsis. We hope that this work, taken with that of others, will ultimately lead to tools to allow physicians and other providers to make better decisions in order to reduce the burden of sepsis.

## 7. Contributions of the group members

**Both members of the team significantly contributed towards the conceptualization of the project at all stages.** Specific individual contributions include:

1. Stephen Pfohl: Conceived and designed the project, developed the relevant Python and SQL code, and interpreted the results of analyses

2. Ben Marafino: Conceived and designed the project, interpreted the results of analyses, and wrote the majority of the text in this final report

## References

[1] Judith Jacobi. Pathophysiology of sepsis. *American Journal of Health-System Pharmacy*, 59(SUPPL. 1):1435–1444, 2002.

[2] J A Russel. The current management of septic shock. *Minerva medica*, 99(5):431–58, 2008.

[3] Centers for Disease Control and Prevention. Data Reports: Sepsis.

[4] Vincent Liu, Gabriel J. Escobar, John D. Greene, Jay Soule, Alan Whippy, Derek C. Angus, and Theodore J. Iwashyna. Hospital Deaths in Patients With Sepsis From 2 Independent Cohorts. *Jama*, 312(1):90, 2014.

[5] DC Angus and Tom Van Der Poll. Severe sepsis and septic shock. *New England Journal of Medicine*, 369:840–51, 2008.

[6] H. Bryant Nguyen, Anja Kathrin Jaehne, Namita Jayaprakash, Matthew W. Semler, Sara Hegab, Angel Coz Yataco, Geneva Tatem, Dhafer Salem, Steven Moore, Kamran Boka, Jasreen Kaur Gill, Jayna Gardner-Gray, Jacqueline Pflaum, Juan Pablo Domecq, Gina Hurst, Justin B. Belsky, Raymond Fowkes, Ronald B. Elkin, Steven Q. Simpson, Jay L. Falk, Daniel J. Singer, and Emanuel P. Rivers. Early goal-directed therapy in severe sepsis and septic shock: insights and comparisons to ProCESS, ProMISe, and ARISE. *Critical Care*, 20(1):160, 12 2016.

[7] Aniruddh Raghu, Matthieu Komorowski, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Continuous State-Space Models for Optimal Sepsis Treatment: a Deep Reinforcement Learning Approach. *Mlhc*, 68, 2017.

[8] Katharine E Henry, David N Hager, Peter J Pronovost, and Suchi Saria. A targeted real-time early warning score (TREWScore) for septic shock. *Science Translational Medicine*, 7(299):122–299, 8 2015.

[9] Shamim Nemati, Mohammad M. Ghassemi, and Gari D. Clifford. Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, volume 2016-Octob, pages 2978–2981. IEEE, 8 2016.

[10] Niranjani Prasad, Li-Fang Cheng, Corey Chivers, Michael Draugelis, and Barbara E Engelhardt. A Reinforcement Learning Approach to Weaning of Mechanical Ventilation in Intensive Care Units. 2017.

[11] Damien Ernst, Guy-Bart Stan, Jorge Goncalves, and Louis Wehenkel. Clinical data based optimal STI strategies for HIV: a reinforcement learning approach. *Proceedings of the 45th IEEE Conference on Decision and Control*, pages 667–672, 2006.

[12] Pablo Escandell-Montero, Milena Chermisi, JosÃl' M. Martínez-Martínez, Juan Gómez-Sanchis, Carlo Barbieri, Emilio Soria-Olivas, Flavio Mari, Joan Vila-Francés, Andrea Stopper, Emanuele Gatti, and JosÃl' D. Martín-Guerrero. Optimization of anemia treatment in hemodialysis patients via reinforcement learning. *Artificial Intelligence in Medicine*, 62(1):47–60, 9 2014.

[13] Yufan Zhao, Donglin Zeng, Mark A Socinski, and Michael R Kosorok. Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics*, 67(4):1422–1433, 12 2011.

[14] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518, 2015.

[15] Hado Van Hasselt, Arthur Guez, David Silver, and Google Deepmind. Deep Reinforcement Learning with Double Q-learning.

[16] Ziyu Wang, Nando de Freitas, and Marc Lanctot. Dueling Network Architectures for Deep Reinforcement Learning. *arXiv preprint arXiv:1511.06581*, 2015.

[17] Peter J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 3 1964.

[18] Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035, 2016.

[19] Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, and Aram Galstyan. Multitask Learning and Benchmarking with Clinical Time Series Data. *SIGKDD 2017*, (17):1–16, 2017.

[20] Mervyn Singer, Clifford S. Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R. Bernard, Jean-Daniel Chiche, Craig M. Coopersmith, Richard S. Hotchkiss, Mitchell M. Levy, John C. Marshall, Greg S. Martin, Steven M. Opal, Gordon D. Rubenfeld, Tom van der Poll, Jean-Louis Vincent, and Derek C. Angus. The Third Inter-

national Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *Jama*, 315(8):801, 2016.

[21] Alistair EW Johnson, David J Stone, Leo A Celi, and Tom J Pollard. The MIMIC Code Repository: enabling reproducibility in critical care research. *Journal of the American Medical Informatics Association*, 9 2017.

[22] Alistair Johnson and Tom Pollard. alistairewj/sepsis3-mimic: Sepsis-3 study v0.1.0. 8 2017.

[23] J R Le Gall, J Klar, S Lemeshow, F Saulnier, C Alberti, A Artigas, and D Teres. The Logistic Organ Dysfunction system. A new way to assess organ dysfunction in the intensive care unit. ICU Scoring Group. *JAMA*, 276(10):802–10, 9 1996.

[24] J. L. Vincent, R. Moreno, J. Takala, S. Willatts, A. De Mendonça, H. Bruining, C. K. Reinhart, P. M. Suter, and L. G. Thijs. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. *Intensive Care Medicine*, 22(7):707–710, 7 1996.

[25] A Elixhauser, C Steiner, D R Harris, and R M Coffey. Comorbidity measures for use with administrative data. *Medical care*, 36(1):8–27, 1 1998.

[26] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[27] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 2010. PMLR.

[28] Anand Kumar, Daniel Roberts, Kenneth E. Wood, Bruce Light, Joseph E. Parrillo, Satendra Sharma, Robert Suppes, Daniel Feinstein, Sergio Zanotti, Leo Taiberg, David Gurka, Aseem Kumar, and Mary Cheang. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock*. *Critical Care Medicine*, 34(6):1589–1596, 6 2006.

[29] Nan Jiang and Lihong Li. Doubly Robust Off-policy Value Evaluation for Reinforcement Learning. 11 2015.

[30] Philip S. Thomas and Emma Brunskill. Data-Efficient Off-Policy Policy Evaluation for Reinforcement Learning. pages 2139–2148, 6 2016.