Is Deep Reinforcement Learning Ready for Practical Applications in Healthcare? A Sensitivity Analysis of Duel-DDQN for Sepsis Treatment

MingYu Lu, MD, MBI[1], Zachary Shahn, PhD[2], Daby Sow, PhD[2]
Finale Doshi-Velez, PhD[3], Li-wei H. Lehman PhD[1]
[1]MIT, Cambridge, MA; [2]IBM Research, NYC, NY; [3] Harvard University, Cambridge, MA

Abstract

The potential of Reinforcement Learning (RL) has been demonstrated through successful applications to games such as Go and Atari. However, while it is straightforward to evaluate the performance of an RL algorithm in a game setting by simply using it to play the game, evaluation is a major challenge in clinical settings where it could be unsafe to follow RL policies in practice. Thus, understanding sensitivity of RL policies to the host of decisions made during implementation is an important step toward building the type of trust in RL required for eventual clinical uptake. In this work, we perform a sensitivity analysis on a state-of-the-art RL algorithm (Dueling Double Deep Q-Networks) applied to hemodynamic stabilization treatment strategies for septic patients in the ICU. We consider sensitivity of learned policies to input features, time discretization, reward function, and random seeds. We find that varying these settings can significantly impact learned policies, which suggests a need for caution when interpreting RL agent output.

Introduction

Artificial intelligence is changing the landscape of healthcare and biomedical research. Reinforcement Learning (RL) and Deep RL (DRL) in particular provide ways to *directly* help clinicians make better decisions via explicit treatment recommendations. Recent applications of DRL to clinical decision support include estimating strategies for sepsis management[1–5], mechanical ventilation control[6], and HIV therapy selection[7]. However, the quality of these DRL-proposed strategies is hard to determine: the treatment strategies are typically learned from retrospective data without access to the unobservable counterfactual reflecting what would have happened had clinicians followed the DRL strategy. Unlike DRL strategies for Atari[8] or other games[9], which can be evaluated by simply using them to play the game, testing a DRL healthcare strategy via a randomized trial can be prohibitively expensive and unethical.

Thus, it is critical that we find ways to assess DRL-derived strategies prior to experimentation. One particular axis of assessment is robustness. DRL algorithms involve many choices, and if the output treatment policy is highly sensitive to some choice, that may imply either (a) getting that choice right is truly important or (b) we should be cautious about assigning credence to the output policy because a seemingly small and possibly unimportant change can have a large impact on results. In contrast, if the output policy is robust to analysis decisions then any errors are likely due to traditional sources of bias in observational studies (such as unobserved confounding), which can be more readily considered by subject matter experts assessing the credibility and actionability of DRL results.

In this paper, we explore the sensitivity of a particular DRL algorithm (duelling double Deep Q-networks, or Duel-DDQN[10,11], a state-of-the-art of Deep Q-Learning which has led to many success in scaling RL to complex sequential decision-making problems[9]) to data preparation and modeling decisions in the context of hemodynamic management in septic patients. Septic patients require repeated fluid and/or vasopressor administration to maintain blood pressure, but appropriate fluid and vasopressor treatment strategies remain controversial[12,13]. Past DRL applications by Komorowski,[1,2] Raghu,[3,4] and Peng *et al*[5] make different implementation decisions while seeking to identify optimal fluid and vasopressor administration strategies in this setting (see additional discussion in Gottesman *et al*[14]). However, these works do not perform systematic sensitivity analyses around their choices.

Starting with a baseline model similar to the works above, we perform sensitivity analyses along multiple axes, including: (1) inclusion of treatment history in the definition of a patient's state; (2) time bin durations; (3) definition of the reward; (4) embedding network architecture; and (5) simply setting different random initialization seeds. In all cases, we find that the Duel-DDQN is sensitive to algorithmic choices. In some cases, we have clear guidance: for example, making sensible decisions about a patient now requires knowing about their prior treatments. In other cases, we find high sensitivity with no clear physiological explanation; this suggests an area for caution and concern.

The paper is organized as follows. We first quickly review background and related work in DRL and sepsis and introduce some notation and terminology used throughout the paper. We then describe the components of a 'baseline' DDQN implementation in detail. For select components that we examine in a sensitivity analysis, we discuss why different specifications could be reasonable and how we go about evaluating sensitivity to alternative specifications. Then we present results of our sensitivity analysis and conclude with a discussion highlighting several limitations and pitfalls to avoid when applying DRL in clinical settings.

Background and Related Work

Markov Decision Process(MDP) & Q-Learning

A Markov decision process (MDP) is used to model the patient environment and trajectories, which consists of[15]

- A set of states S, plus a distribution of starting states $p(s_0)$.
- A set of actions A.
- Transition dynamics $T(s_{t+1}|s_t, a_t)$ that map a state action pair at time t onto a distribution of states at time $t+1$.
- An immediate/instantaneous reward function $r_t = R(s_t, a_t, s_{t+1})$.
- A discount factor $\gamma \in [0, 1]$, where lower values place more emphasis on immediate rewards.

Every roll-out of a policy accumulates rewards from the environment, resulting in the return $R = \sum_{t=0}^{T-1} \gamma^t r_{t+1}$. The goal of RL is to find an optimal policy π^*, which achieves the maximum expected return from all the states. $\pi^* = \arg\max_\pi E[R|\pi]$. To find π^*, one of the reinforcement learning algorithm is Q-learning. The basic idea to evaluate the policy is to use temporal difference (TD) learning over the policy iteration to minimize the TD error[15,16]. $Q^\pi(s, a) \leftarrow Q^\pi + \alpha(r + \gamma + Q^\pi(s', a') - Q^\pi(s, a))$

More formally, Q learning aims to approximate the optimal action-value function given the observed state s and the action a at time t[15,16]. The future reward r_t is discounted at every time step t by a constant factor. $Q^*(s, a) = E[r_t + \gamma * r_{t+1} + \gamma^2 * r_{t+2} + ...|s_t, a_t = a, \pi]$

Dueling Double Deep Q-Learning (Dueling DDQN) with Prioritized Experience Relay (PER)

To approximate the optimal action-value function, we can use a deep Q-network: $Q(s, a; \theta)$ with parameter θ. To estimate this network, we optimize the following sequence of loss functions at iteration i: $L_i(\theta_i) = E_{s,a,r,s'}[(y_i^{DQN} - Q(s, a; \theta_i))^2]$;$y_i^{DQN} = r + \gamma max Q(s', a'; \theta')$, updating parameters by gradient descent such that $\nabla\theta_i L_i(\theta_i) = E_{s,a,r,s'}[y_i^{DQN} - Q(s, a; \theta_i)\nabla\theta_i Q(s, a; \theta_i)]$.

Dueling Double Deep Q-learning[11] is a particular state-of-the-art deep Q-learning algorithm consisting of separate 'dueling' architectures that decouple the $value$ and $advantage$ streams in deep Q-networks[11] to determine the value of the next state[10]. Prioritized experience replay[10,17], i.e. sampling mini-batches of experience that have high expected impact on learning, further improves efficiency.

Learning Sepsis Management with DRL

Sepsis is a life-threatening organ dysfunction disease caused by dysregulated host response to infections[12]. How to maintain septic patients' hemodynamic stability via administration of intravenous fluid (IV fluid) and vasopressors is a key research and clinical challenge.[12,13] A number of DRL studies have been carried out to address this issue in the past few years. Raghu *et al*[3,4] applied a Dueling Q Network and a sparse autoencoder to tackle with continuous state space. Peng *et al*[5] further presented a mixture-of-experts framework, combining a kernel and a Dueling DDQN with PER to personalize sepsis treatment. One thing worth noting is that these studies are conducted based on different reward settings. Raghu *et al*[3,4] used hospital and 90-day mortality as a sparse reward issued at the end of patients' trajectories, while other studies used short-term rewards such as SOFA score in combination with lactate levels[3] or changes in probability of mortality[5].

Data Description & Cohort Data for our cohort were obtained from the Medical Information Mart for Intensive Care (MIMIC-III v1.4)[18] database. The data set contained all MIMIC-III patients meeting Sepsis-3 criteria[19] from years 2008-2012.* It comprised 7,956 patients with 649,661 clinical event entries. In this analysis, we extracted and collected static features (e.g. demographic), past treatment history and a summary of hourly observation (mean,

*ICU admissions between 2008-2012 were recorded using the MetaVision system with higher resolution treatment information.

maximum, and minimum within an hour) of all laboratory values within patients' first 72-hour ICU stay. The detailed features are shown in Table 1. All measured values were standardized, and we carried forward covariate values from the most recent measurement. The data set was split using 80% for training and validation and 20% for testing.

Table 1: Details of attributes including vital signs, laboratory values, and treatment history

Demographic	Age, Weight, Height, Ethnicity
Vital Sign\Laboratory	GCS, Heart rate, Temperature, Respiratory Rate, Diastolic Blood Pressure, Systolic Blood Pressure, Mean Arterial Blood Pressure, Potassium, Sodium, Chloride, Magnesium, Calcium, Anion gap, Hemoglobin, Hematorcit, WBC, Platelets, Bands, PTT, PT, INR, Arterial pH, SpO_2, FiO_2, PaO_2, $TotalCO_2$, pCO_2, Arterial Base excess, Bicarbonate, Arterial Lactate, SOFA score, Glucose, Creatinine, BUN, Total Bilirubin, Indirect Bilirubin, Direct Bilirubin, AST, ALT, Total Protein, Troponin, CRP, Elixhauser Score, Albumin
Treatment	Vasopressor & IV Fluid

Methods: Baseline Implementation

We first describe our baseline implementation of a Dueling DDQN with PER to learn an optimal resuscitation strategy. This implementation combines elements from several published RL applications to sepsis treatment[4,5], with slight modifications. In describing the baseline, we also illustrate the many components involved in specification of a Dueling DDQN analysis. In our sensitivity analysis, we will systematically vary these components from our baseline model one at a time.

Time Discretization We divided patient data into one-hour bins. To avoid inappropriately adjusting for covariate values that were measured after treatment actions were taken[20], we performed a time-rebinning procedure. If a treatment action occurred within an existing time bin, covariate measurements made after the treatment action in that bin were moved to the following bin and the time of the treatment action became the new endpoint of the time bin. Figure 1 illustrates this process with the blue bar covering 1.75 hr to 2.75 hr defining the time period for which covariate measurements contribute to the bin ending at 2.75 hours. (Time rebinning is rare in RL literature but necessary to avoid adjusting for post-treatment variables.)

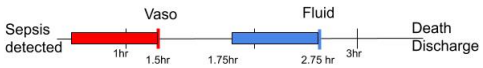


Figure 1: Time discretization

Compressing Patient Histories We follow Peng *et al.*[5] in encoding patient states recurrently using an LSTM autoencoder representing the cumulative history for each patient. LSTMs can summarize sequential data through an encoder structure into a fixed-length vector and then reconstruct into its original sequential form through the decoder structure[21]. The summarized information can be used to represent time series features[23]. LSTM-RNN models can prevent a vanishing or exploding gradient and are commonly used to capture long-term sequence structures[24].

Action definition and Treatment Discretization Following Raghu *et al.*[4] and Peng *et al.*[5], we focus on intravenous fluids and vasopressors as the actions of the MDP. We computed the hourly rate of treatment as the action and sum the rate when there are overlapped treatment events of the same type. The hourly rate of each treatment is divided into 5 bins defined by quartiles under current physician practice. Accordingly, a 5 by 5 action space is defined for the medical intervention[4]. An action of (0,0) means no treatment is given; whereas, an index of (4, 4) represents top quartile dosages of both fluids and vasopressors.

Reward formulation We follow Peng *et al.*[5] in defining the reward at time t as the change in negative log-odds of 30 day mortality between t and $t+1$ according to a predictive model for 30 day mortality. The probability of mortality was estimated with a 2-layer neural network with 50 and 30 hidden units with L1 regularization given the recurrent embedding of the compressed history at the corresponding time. Let $f(o)$ be the probability of mortality given observations through the current time point o and $f(o')$ be the probability of mortality given observations through the next time step. Then we define the reward

$$r(o, a, o') = -log\frac{f(o')}{1-f(o')} + log\frac{f(o)}{1-f(o)}. \quad (1)$$

Dueling DDQN Architecture Following Raghu *et al*[4], our final Duelling Double-Deep Q Network (Dueling DDQN)[10,11] with PER architecture has two hidden layers of size 128, using batch normalization after each, Leaky-ReLU activation functions, a split into equally sized advantage and value streams, and a projection onto the action-space by combining

these two streams. The Duel-DDQN architecture divides the value function V into the value of the patients underlying physiological condition, called the $Value$ stream, and the value of the treatment given, called the $Advantage$ stream.

Methods: Sensitivity analysis

In this section, we describe the ways in which we altered the baseline Dueling DDQN implementation described in the previous section in our sensitivity analysis. For each component that we varied, we specify the alternatives we considered, explain why the choice could be important, and also explain why each alternative might be considered reasonable. We emphasize that the aim of this sensitivity analysis is not to determine which choices are best, as there is no ground truth available to make such a determination. Rather, it is to understand the robustness of the treatment policy with respect to a priori reasonable-seeming alternatives. In particular, we explore the effects on learned policies of: including treatment history in the state definition; varying the time bin size; varying the reward specification; specifying different recurrent embedding models; and setting different random seeds.

Including Past Treatment History

One decision is whether to include the history of past treatments in the representation of patient state. Several prior works[3,5,6] did not do so, but given the Markov assumption on which DQNs rest, this amounts to assuming that past treatments cannot impact future outcomes through pathways that do not run through measured covariates included in the state. This assumption will usually be false. For example, vasopressors have potentially serious long term cardiovascular side effects[13], but the added risk after administering more vasopressors would not be captured in short term changes in measured patient covariates. Thus, in our sensitivity analysis we compare learned policies from DQN implementations with state summaries that include and exclude treatment history. In this analysis, we consider the cumulative dosage of all previous treatment until $t-1$ as a proxy for treatment history at t.

Duration of Time Bins

When applying a discrete time RL algorithm to data with actions taken and measurements recorded in continuous time, an implementation decision that inevitably arises is how to bin time into discrete chunks in which to define patient states S_t. With infinite data, shorter time bins would be superior for two reasons. First, the state at each time step reflects the patient's condition closer to when the treatment action at that time step was decided. This improves confounding adjustment, since states are more reflective of information that actually influenced treatment decisions. Second, more time steps allow for more flexible and dynamic learned strategies that are more responsive to changes in patient state. However, with finite data, the capacity to learn more flexible strategies with shorter time bins can be detrimental, leading to instability of the estimated optimal policy.

In the case of sepsis, past work has used 4 hour time bins[4]. Treatment decisions in this clinical context are made on a finer time scale than 4 hours, which is why we defined 1 hour time bins in our baseline model. But stability is also a major concern in this dataset, so either choice is defensible. Hence, we compared the learned policies of Dueling DDQNs fit to 1 hour and 4 hour time binned data sets in our sensitivity analysis.

Horizon of Rewards

A key decision is specifying the reward function. Ideally, the reward function would summarize the entire long term utility of the stay, as this is what we really seek to optimize. However, for reasons of practicality, researchers often choose short term rewards measured at each time-step. When rewards are short term, there is more 'signal' in that it is easier to estimate associations between rewards and actions. It would be nice if learned policies were broadly similar whether we choose to optimize our true reward of interest or a shorter term proxy. To investigate the impact of using long vs short term rewards, we also compared reward functions that were weighted mixtures of long and medium term information about outcomes for a range of weights.

We define a utility function reward U as follows. Let M be the worst possible SOFA score. Let Y be observed SOFA at the end of the stay. Let $S=1$ if the patient survived more than 1 year after admission, 0 otherwise. Let H be hours survived after admission. Let C be a constant that controls relative weight assigned to SOFA score at the end of stay and survival. We define $r'(C)$ as

$$\textbf{if } H \geq 24*365 \textbf{ then } U = log(1+\frac{M-Y}{C}), \textbf{ else } U = log(\frac{H}{24*365}+1) \tag{2}$$

For large values of C, survival time is all that matters. For low values of C, patient state at the end of the stay matters

a lot for patients who survive more than 1 year. For all values of C, rewards are medium to long term (since they are based on patient state at the end of the stay or later as opposed to the following time bin), but differing C levels reflect different subjective prioritization of patient health outcomes.We compare learned DQN policies using short term reward (1) with long term rewards (2) for varying values of C.

Choice of Embedding Model

Another question is how to summarize patient history. In DQNs, it is important for the information contained in the state S_t at each time t to satisfy several conditions. First, S_t should satisfy the "sequential exchangeability" assumption[22], which would be satisfied if S_t contains all relevant information about variables influencing treatment decisions at time t and associated with future rewards, i.e. S_t should contain sufficient information to adjust for confounding. If sequential exchangeability fails, then estimates of the impact of actions on future rewards will be biased, and therefore the estimate of the optimal treatment strategy will be biased as well.

To learn an optimal treatment strategy, it is also important that S_t contain relevant information about variables that are *effect modifiers*. An effect modifier is a variable with the property that the conditional average effect of an action on future rewards varies with the variable's value. Good treatment rules assign treatment based on the value of effect modifiers. (Effect modifiers may or may not be confounders, which are necessary to include in the model to avoid bias but may not be good inputs to treatment decision rules.)

Finally, DQNs make a very strong Markov assumption on states[8,16]. S_t must be defined to be sufficiently rich that this Markov assumption can approximately hold. Thus, to allow for realistic long term temporal dependencies, states at each time should be rich summaries of patient history. Without a priori knowledge of exactly which aspects of patient history to retain (to adjust for confounding, model effect modification, and satisfy the Markov assumption), a reasonable strategy is to define patient states as embeddings generated by a RNN[23,25].

However, RNN embeddings are not optimized to retain the types of information specifically required to be contained in DQN states. Different choices of black box embedding method may generate states that satisfy the DQN requirements to varying degrees and produce different learned policies, with no principled way to choose between them.

In our sensitivity analysis, we compare two common RNN embedding models–long short-term memory (LSTM)[21] and gated recurrent unit (GRU)[26]. The architectures have been shown to perform comparably across a range of tasks[24]. Each consists of two hidden layers of 128 hidden units and is trained with mini-batches of 128 and the Adam optimizer for 200 epochs until convergence. We have no reason a priori to believe that either option would produce more suitable embeddings than the other, and the point of comparing them is to determine whether the decision is important.

Random restarts

Finally, solving a DDQN is a non-convex optimization, and thus random restarts are frequently used to find a good local optimum. As Henderson *et al*[27] reports, a wide range of results can be obtained from the same deep RL algorithm depending on the random initialization of the network's weights initialization. To observe the impact of random weight initialization in our dataset, we fit our baseline model repeatedly using different seeds. While one would generally simply take the best of the random restarts as the solution, high variation across random restarts might mean that reproducing a result will be more challenging as the problem has many diverse local optima.

Methods: Evaluation Metrics & Experimental Settings

Metrics Previous works have used off-policy estimators, such as Weighted Doubly Robust (WDR[28]) to estimate the quality of a proposed policy. However, these estimators can have high variance as well as bias. Instead, we compare policies based on the distribution of actions they recommended. If these distributions are very different from each other or from clinicians (whom we know act reasonably) then that may be cause for skepticism.

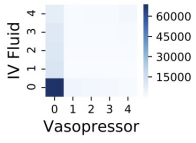


Figure 2: Action distribution

Specifically, for each time point in the patient history in the test set, we compute the Dueling DDQN's recommended action. We use heat maps like Figure 2 to display action distributions, where the y axis represents the IV fluid dosage (quartile), the x axis represents the vasopressor dosage (quartile), and the density of the color represents the frequency with which the treatment action is applied. These action distributions are aggregates in that we are summing over all

time points over all patients.Still, differences in this metric can provide insights into the ways in which policies differ.

Parameters and Optimization We train the Dueling DDQN for 100,000 steps (except the reward horizon experiment, where we perform early stopping at 15,000 steps to prevent over-fitting) with batch size of 30. We conducted 5 restarts for every experimental setting (except the random restart experiment, where we look at variation across individual restarts). Following Peng *et al*[5], of the policies resulting from the 5 restarts we choose the one with highest value as estimated by a weighted doubly robust off policy evaluation method[28]. For models trained with long term rewards, $r'(C)$, where the WDR estimator is unfeasible, we selected a policy from the 5 restarts based on the Q-value.

Table 2: Summary of the variance across different experimental settings

	Timing	Encoder	Treatment history	Reward
Baseline	1 hr	LSTM (full history)	Yes	Short term: (1) Immediate change in prognosis
Alternatives	1hr, 4hr	LSTM\GRU (full history)	Yes, No	Short term: (1) Immediate change in prognosis; Long term: (2) Combinations of SOFA at end of stay and survival time
Raghu *et al*[4]	4 hr	Sparse Autoencoder (current obs. only)	No	Long term: In-hospital mortality
Peng *et al*[5]	4 hr	LSTM (full history)	No	Short term*: (1) Immediate change in prognosis

* In Peng *et al*'s reward, prognosis was estimated conditional on current observations only. In the baseline implementation's reward, prognosis was estimated conditional on full patient history.

Results

We altered aspects of the baseline Dueling DDQN implementation described in the previous section and compared the resulting learned policies according to their action distributions. In the following, we abbreviate Dueling DDQN trained with embedding to DQN-embedding. For, example the Dueling DDQN trained with LSTM and 1 hourly binned data is called DQN-LSTM-1hr.

Treatment History: Excluding treatment history leads to aggressive treatment policies

Here, we compare treatment strategies output by Dueling DDQNs that do and do not include treatment history in patient state representations (Figure 3). The DQN-LSTM-1hr trained without treatment history recommends nonzero vasopressor doses at all time points. It frequently recommends high doses of each treatment compared to physicians and an agent trained with treatment history included in the state definition. Excluding past treatment information increases the average recommended dosage of vasopressor and fluid by 1.6 - 1.8 times and 1.7 - 3.1 times, respectively. We hypothesize and explanation for this behavior in the Discussion section.

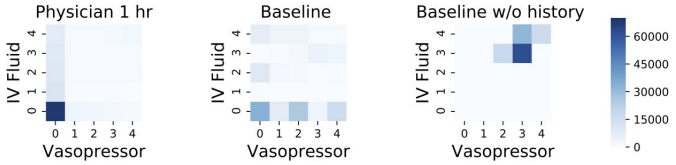


Figure 3: The action distribution of Dueling DDQNs trained with/without past treatment history information. Note that the agent trained without treatment history aggressively prescribes high (3rd quartile, 4th quartile) dosage of vasopressor and IV fluid. Low dosage treatment are rarely administered to patients.

Time bin durations: Longer time bins result in more aggressive policies.

Figure 4 illustrates that while different ways of segmenting time do not affect the clinician action distributions (by definition), they have a large effect on the Dueling DDQN action distributions. The DQN-LSTM's 4 hour bins increased the frequency of nonzero vasopressor doses by 40% and decreased the overall usage of IV fluid only, a less aggressive action, by 34%.

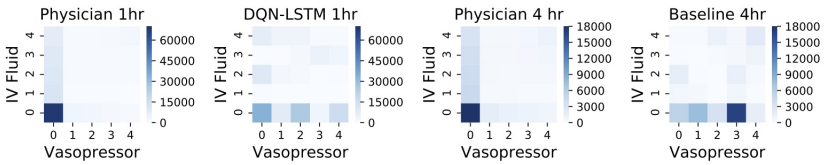


Figure 4: Comparison of action distribution across DQN-LSTM 4-hour and DQN-LSTM 1-hour time. (Note the same color density does not represent the same count in 4 hour bins and 1 hour bins). The 4 hour bins lead to much more frequent recommendations of high vasopressor doses by the DQN-LSTM, while the physician's policy remains conservative.

Rewards: Long-term objectives lead to more aggressive and less stable policies

In figure 5, we see that longer term objectives resulted in more aggressive policies—specifically in more frequent high fluid doses than our short term baseline reward for all values of C (for C = 100, it is difficult to see visually in the heat map, but the agent administered the maximum fluid dosage 40% of the time). Policies also vary considerably across level of emphasis on medium versus long term outcomes determined by values of C. We noticed higher variation across random restarts in the long term reward settings than the short term baseline settings (see Figure 8). This could indicate that optimization is more challenging and unstable for long term rewards.

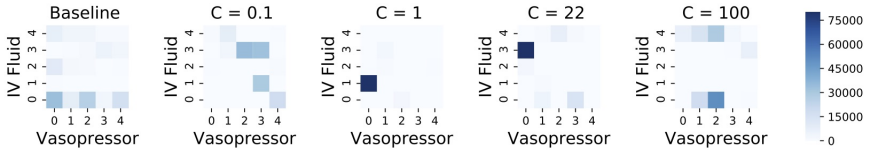


Figure 5: Comparison of Duel-DDQN trained with short term & long/intermediate reward across varying Cs.

Embedding model: High sensitivity to architecture

Results comparing LSTM and GRU embeddings can be found in Figure 6. We can observe that both our baseline (LSTM) implementation and the GRU implementation recommended nonzero doses of vasopressors significantly more frequently than physicians. However, the GRU implementation was more aggressive, recommending nonzero fluid doses significantly more often than both physicians and the DQN-LSTM.

The choice of embedding architecture also interacts with other analysis settings, whose effects differ depending on embedding architecture. We illustrate interactions with treatment history and time segmentation.

Exclusion of Treatment History Excluding prior treatment history has an even more extreme effect when embeddings use the DQN-GRU architecture, with maximum dosage of both treatments being delivered most of the time.

Different time segmentation 4 hour time bins led to more frequent high vasopressor doses in the baseline LSTM implementation, but more frequent high fluid doses in the GRU implementation.

There is no way to apply clinical, physiological, or statistical knowledge to reason about which embedding architecture is more appropriate as they are conceptually quite similar. Thus, the variation stemming from the choice of embedding is a source of concern.

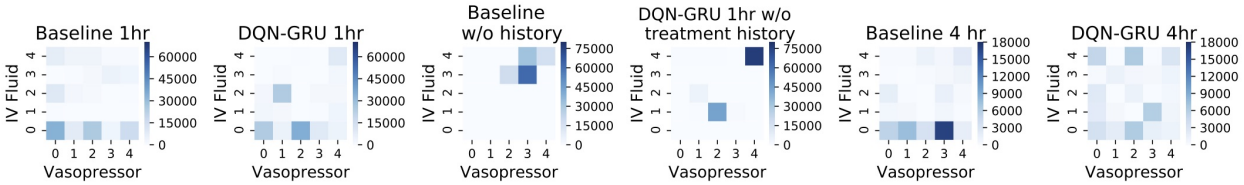
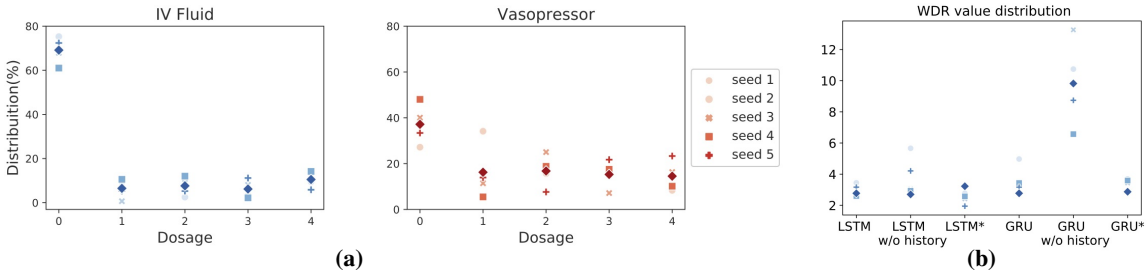


Figure 6: ***GRU implementation***: Comparing to the baseline (LSTM), the most observed treatment by DQN-GRU 1hr was to deliver a medium dosage vasopressors without IV fluid; ***Exclusion of Treatment history***: DQN-GRU administers maximum dosage of both vasopressor and IV fluid most of the time; ***Time bin duration*** : 4-hour bins increased the overall usage of IV fluid by 20% and more than doubled increased maximum dosage of IV fluid.

Random Restarts: DRL policies have many local optima

Our final sensitivity analysis looked at variation across restarts of the algorithm, which assesses sensitivity to where the algorithm was initialized. While the broad qualitative differences between the Dueling DDQN and physician policies remained constant across seeds in our baseline model, there was still much variation in the resulting action distributions, especially for vasopressors (Figure 7a). In Figure 7b, we see that despite these differences, the estimated values of these policies are similar; this demonstrates that the variation is not because the optimization sometimes landed in a poor optimum, but because there are many optima with similar estimated quality that lead to qualitatively different policies. This is another cause for concern, as it suggests that the agent has no way of telling if any of these very different policies are better than the others.



(a) (b)

Figure 7: (a) Treatment distribution across random restarts in baseline. While variance of IV fluid is small, distribution of vasopressor varies across different seeds. (b) Comparision of values of WDR estimator across random seeds in each settings (* represents 4 hour time bins duration). In the setting of exclusion of treatment history, agents are highly sensitive to the seeds.

We also found that policies with medium and long-term objective policies were much more sensitive to random seed. Figure 8b depicts the distribution over the action space of the variance across 5 random restarts of the frequency with which that action was recommended. The variances were much greater for the implementations using long term rewards. The presence of many local optima increases variance and makes it more challenging to differentiate policies in implementations with long term rewards.

(a) (b)

Figure 8: (a) Q values distribution across different seeds (b) Comparison between long term and short term (baseline) objectives. Note that long term reward implementation demonstrates higher variance compared to baseline. X-axis labels coefficient of variance or relative standard deviation, $c_v = \frac{\sigma}{\mu}$ where σ is the standard deviation and μ is the mean of action distribution across random restarts of each of the 25 discretized actions (as in Figure2.)

Subgroup Analysis: Grouping by SOFA score finds DQN agents are underaggressive in high risk patients and over-aggressive in low risk patients

We further perform an analysis in subgroups defined by severity of sepsis as indicated by Sequential Organ Failure Assessment (SOFA score[29]). The SOFA score is a commonly used tool to stratify and compare patients in clinical practice, with higher scores indicating worse condition. When the assessment is greater than 15, mortality is increased up to 80%.[29] In Figure 9, the Dueling DDQN agents are significantly more aggressive than physicians in treating lower risk patients. As was also observed in Raghu *et al*[3], in high risk patients the reverse is true. Physicians commonly give maximum doses of both vasopressors and fluids, while the DQN agents rarely do. This suggests that the Dueling DDQN models may not be correctly accounting for patient severity or adjusting for confounding by indication.

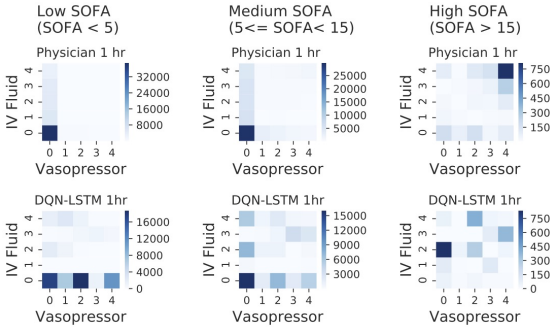


Figure 9: Subgroup analysis: For patients with SOFA < 5, both baseline and GRU implementation are 5-7 times more likely than physicians to give patients vasopressors.

Conclusion and Discussion

State-of-the-art deep reinforcement learning approaches are largely untested in rich and dynamic healthcare settings. We presented a sensitivity analysis exploring how a Dueling DDQN agent would react to alternative, including: 1) approaches to adjusting for treatment history; 2) time discretization bin durations; 3) recurrent neural network state representations; 4) reward specifications; and 5) random seeds. We have shown that choices between equally a priori justifiable implementation settings can have large clinically significant impacts on learned DQN policies. Given this lack of robustness, results from individual implementations should be received skeptically.

The one area where our results do seem to point toward some clear guidance concerns the inclusion of treatment history in the state. Exclusion of treatment history from the state is only warranted under the implausible Markov assumption that past treatments only influence future outcomes through measured intermediate variables. If this assumption fails and there are cumulative dangers from too much treatment (e.g. pulmonary edema from fluid overload and cardiovascular side effects from vasopressors in our application), then past treatment will affect both the response (treatment is less likely to be beneficial given excessive past treatment) and the current treatment decision (treatment is less likely to be administered given extensive past treatment, and outcomes are likely to be worse given extensive past treatment). Thus, omitting treatment history would make excessive treatment appear more beneficial than it actually is. Indeed, the behavior we observed in our Dueling DDQN agents was consistent with the behavior that would be predicted by theory. Agents trained without treatment history included in their states recommended vasopressors or fluids at every timestep, an obviously harmful strategy. Yet all three prior DQN implementations in sepsis omitted treatment history from state definitions.

While we cannot provide definitive statistical or physiological explanations for most of the DQN outputs observed in our sensitivity analysis, here we discuss possible sources of DQN instabilities. One theme appeared to be unreasonable policies stemming from extrapolation beyond treatment decisions observed under current practice. For example, in our SOFA score subgroup analysis we saw the DQN agents recommending clearly harmful actions rarely seen in the data, i.e. failing to give high doses to the highest risk patients and frequently giving high doses to low risk patients. Also, the fact that different initializations found solutions with similar estimated Q-values but qualitatively different action distributions (also observed in Arjumand *et al*[30]) suggests that the problem is not sufficiently constrained. We need better ways to incorporate knowledge of what features are important and what actions are reasonable to constrain learned policies[31].

There is a long road from the current state of DRL healthcare applications to clinically actionable insights or treatment recommendations. Currently, AI researchers apply DRL algorithms to clinical problems and claim that the policies they learn would greatly improve health outcomes compared to current practice[3,5]. In this work, we demonstrate that had these researchers made slightly different (but a priori reasonable) decisions, they would have obtained very different policies that also appeared superior to current practice. Beginning to map this sensitivity is a small but important step along the road to clinically actionable DRL policies. We hope our observations will lead to future work on characterizing and (more importantly) obtaining the type of robustness required to justify empirical testing of a DRL policy via a clinical trial.

We close with some speculative suggestions for that future work. First, a wide range of analysis settings can be compared in extensive and physiologically faithful simulation experiments where the ground truth value of resulting learned policies would be available. This could shed light on certain operating characteristics and best practices. For example, alternative approaches to preventing models from extrapolating too far from current practice as suggested above could be evaluated in this framework.

Further, when policies are sensitive to algorithmic choices, one could search for areas of broad agreement across policies recommended under a range of settings. These areas of agreement would be policy fragments, i.e. recommendations only applying to specific contexts. These strategy fragments could then be rigorously assessed by subject matter experts for plausibility and their effects could be estimated by epidemiologists using more stable techniques for treatment effect estimation. Seeking approaches to make DRL robust and human-verifiable will help us properly leverage information in health records to improve care.

References

1. M. Komorowski, A. Gordon, L. Celi, and A. Faisal, A markov decision process to suggest optimal treatment of severe infections in intensive care, in Neural Information Processing Systems Workshop on Machine Learning for Health, 2016.
2. Komorowski, M., Celi, L.A., Badawi, O. et al. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med 24, 17161720 (2018).
3. Aniruddh Raghu, Matthieu Komorowski, Leo Anthony Celi, Peter Szolovits, Marzyeh Ghassemi et al. Deep Reinforcement Learning for Sepsis Treatment. Machine Learning For Health at the conference on Neural Information Processing Systems, 2017;arXiv:1711.09602
4. Raghu Aniruddh, Komorowski, Matthieu, Celi Leo Anthony, Szolovits Peter, Ghassemi Marzyeh. Continuous State-Space Models for Optimal Sepsis Treatment - a Deep Reinforcement Learning Approach 2017;arXiv:1705.08422
5. Peng X, Ding Y, Wihl D, et al. Improving Sepsis Treatment Strategies by Combining Deep and Kernel-Based Reinforcement Learning. AMIA Annu Symp Proc. 2018;2018:887896.
6. Niranjani Prasad, Li-Fang Cheng, Corey Chivers, Michael Draugelis, and Barbara E Engel-hardt. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. 2017;arXiv:1704.06300
7. Parbhoo S, Bogojeska J, Zazzi M, Roth V, Doshi-Velez F. Combining Kernel and Model Based Learning for HIV Therapy Selection. AMIA Jt Summits Transl Sci Proc. 2017;2017:239248.
8. Volodymyr Mnih, Koray Kavukcuoglu, David Silver et al; DeepMind Technologies. Playing Atari with Deep Reinforcement Learning; NIPS Deep Learning Workshop 2013; arXiv:1312.5602
9. Mnih, V., Kavukcuoglu, K., Silver, D. et al. Human-level control through deep reinforcement learning. Nature 518, 529533 (2015).
10. Hado V. Hasselt. Double Q-learning. Advances in Neural Information Processing Systems 23; (2010);Pages 26132621
11. Ziyu Wang, Tom Schaul, Matteo Hessel, Hado van Hasselt, Marc Lanctot, Nando de Freitas. Dueling Network Architectures for Deep Reinforcement Learning; arXiv:1511.06581
12. Rhodes, A., Evans, L.E., Alhazzani, W. et al. Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock: 2016. Intensive Care Med 43, 304377 (2017).
13. Overgaard Christopher B., Davk Vladimr. Inotropes and Vasopressors. 2008; Circulation. 2008;118:10471056
14. Omer Gottesman1, Fredrik Johansson, Joshua Meier1, Jack Dent1, et al. Evaluating Reinforcement Learning Algorithms in Observational Health Settings 2018arXiv:1805.12298
15. K. Arulkumaran, M. P. Deisenroth, M. Brundage and A. A. Bharath. Deep Reinforcement Learning: A Brief Survey. IEEE Signal Processing Magazine, vol. 34, no. 6, pp. 26-38, Nov. 2017.
16. Chris Watkins and Peter Dayan. Technical Note Q-Learning. Machine Learning 1992; 8, 279-292.
17. Tom Schaul, John Quan, Ioannis Antonoglou, David Silver. Prioritized experience replay. arXiv preprint arXiv:1511.05952, 2015
18. Johnson, A., Pollard, T., Shen, L. et al. MIMIC-III, a freely accessible critical care database. Sci Data 3, 160035 (2016).
19. M. Singer, C. S. Deutschmann, C. W. Seymour et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016;315(8):801-810. doi:10.1001/jama.2016.0287
20. Hong, G. and Raudenbush, S. W. Causal Inference for Time-Varying Instructional Treatments.Journal of Educational and Behavioral Statistics, 33(3), pp. 333362. doi: 10.3102/1076998607307355. 2008
21. Hochreiter, S. and J. Schmidhuber, Long short-term memory. Neural computation, 1997. 9(8): p. 1735-1780.
22. James M Robins. Robust estimation in sequentially ignorable missing data and causal inference models. Biometrics 61, 962972 December 2005; DOI: 10.1111/j.1541-0420.2005.00377.x
23. Che, Z., Purushotham, S., Cho, K. et al. Recurrent Neural Networks for Multivariate Time Series with Missing Values. Sci Rep 8, 6085 (2018).
24. Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, Yoshua Bengio. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling; NIPS Deep Learning and Representation Learning Workshop 2014; arXiv:1412.3555
25. Bram Baker. Conference on Neural Information Processing Systems 2002 Dept.of Psychology,LeidenUniversity/ IDSIA 2002
26. Kyunghyun Cho, Bart van Merrinboer, Dzmitry Bahdanau, Yoshua Bengio; On the Properties of Neural Machine Translation: Encoder:Decoder Approaches; Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation 2014;DOI:10.3115/v1/W14-4012
27. Henderson Peter, Islam Riashat, Bachman Philip, Pineau Joelle ,Precup Doina, Meger David. Deep Reinforcement Learning that Matters.AAAI Conference On Artificial Intelligence (AAAI) 2018; arXiv preprint arXiv:1709.06560
28. N. Jiang, L. Li. Doubly Robust Off-policy Evaluation for Reinforcement Learning; arXiv:1511.03722
29. Flavio Lopes Ferreira, MD; Daliana Peres Bota, MD; Serial Evaluation of the SOFA Score to Predict Outcome in Critically Ill Patients. Annette Bross, MD; et al. 2001 The Journal of the American Medical Association
30. Masood, Muhammad and Doshi-Velez, Finale. (2019). Diversity-Inducing Policy Gradient: Using Maximum Mean Discrepancy to Find a Set of Diverse Policies. 5923-5929. 10.24963/ijcai.2019/821.
31. Fujimoto Scott, Meger David, and Precup Doina. (2018). Off-Policy Deep Reinforcement Learning without Exploration. arXiv:1812.02900.

1 Appendix

The 95 % confidential interval(CI) is calculated based on 200 bootstrapped test datasets.

Table 3: Action distribution

	Physician	DQN-LSTM 1 hr	DQN-GRU 1 hr
	No.%	No.% (95% CI)	No.% (95% CI)
Vasopressor			
No action	84%	40% (39 to 41)	23% (21 to 24)
1st	4%	11% (11 to 12)	24% (23 to 25)
2nd	4%	25% (24 to 26)	28% (26 to 29)
3rd	3%	7% (7 to 8)	9% (8 to 10)
4th	4%	16% (15 to 17)	17% (16 to 17)
IV fluid			
No action	60%	68% (66 to 69)	61% (59 to 62)
1st	10%	0.7% (0.5 to 0.9)	6% (5 to 7)
2nd	10%	10% (10 to 12)	23% (22 to 24)
3rd	10%	8% (7 to 9)	3% (2 to 3)
4th	10%	13% (12 to 14)	8% (8 to 8)

Table 4: Relative Risk of Treatment of the baseline, DQN-LSTM, and DQN-GRU

	IV fluid		Vaospressor	
	Relative Risk	95% CI	Relative Risk	95% CI
No action	0.802	0.785 - 0.822	0.836	0.775 - 0.891
1st	0.645	0.558 - 0.747	0.708	0.679 - 0.745
2nd	9.246	7.724 - 10.96	1.765	1.661 - 1.858
3th	0.825	0.685 - 0.997	0.614	0.555 - 0.676
4th	0.814	0.743 - 0.934	1.980	1.781 - 2.146

Table 5: Action distribution of Duel-DDQNs without history information. Without history information, for DQN-LSTM agent, the learned policy highly concentrates on the combination of vasopressor of 3rd quartile dosage and IV fluid of 3rd quartile dosage. DQN-GRU administers maximum dosage of both vasopressor and IV fluid most of the time.

	Physician	DQN-LSTM -1hr	DQN-GRU-1hr
	No.%	No.% (95% CI)	No.% (95% CI)
Vasopressor			
No action	84%	0%	1% (1 to 1)
1st	4%	$1.5 * 10^{-3}$% (0 to $4.6 * 10^{-3}$)	5% (4 to 5)
2nd	4%	14% (13 to 16)	34% (33 to 35)
3rd	3%	73% (71 to 74)	1% (1 to 1)
4th	4%	13% (12 to 14)	60% (59 to 61)
IV fluid			
No action	60%	0.03% (0.002 to 0.07)	0%
1st	10%	0.01% (0. to 0.02)	35% (34 to 37)
2nd	10%	0.002% (0 to 0.004)	4% (4 to 4)
3rd	10%	62% (60 to 64)	1% (1 to 1)
4th	10%	38% (37 to 40)	60% (59 to 62)

Table 6: Action distribution of 4-hour time bins

	Physician	DQN-LSTM 4 hour	DQN-GRU 4 hour
	No.%	No.% (95% CI)	No.% (95% CI)
Vasopressor			
No action	74%	29% (28 to 30)	17% (16 to 18)
1st	8%	8% (7 to 9)	18% (17 to 19)
2nd	6%	31% (30 to 33)	11% (10 to 11)
3rd	6%	18% (17 to 19)	43% (41 to 45)
4th	7%	14% (13 to 15)	12% (11 to 13)
IV fluid			
No action	51%	32% (30 to 33)	74% (73 to 77)
1st	12%	20% (19 to 21)	4% (3 to 4)
2nd	12%	10% (9 to 11)	8% (7 to 9)
3rd	12%	7% (6 to 8)	4% (3 to 4)
4th	13%	32% (30 to 33)	10% (8 to 11)

Table 7: Vasopressor distribution by DQN agents

	DQN-LSTM				DQN-GRU			
	w history		w/o history		w history		w/o history	
	No. (%)	95% CI	No. (%)	95% CI	No. (%)	95% CI	No. (%)	95% CI
Vasopressor								
No action	40.17	38.70 - 41.27	0.	0	22.66	21.44 - 23.69	0.661	0.573 - 0.764
1st	11.41	10.79 - 12.06	0.0015	0.0 - 0.0046	24.20	23.16 - 25.38	4.555	4.325 - 4.778
2nd	25.09	23.88 - 26.23	14.15	13.04 - 15.55	27.50	26.29 - 28.68	34.03	32.56 - 35.35
3rd	7.041	6.548 - 7.704	72.75	71.11 - 74.49	8.999	8.309 - 9.847	0.867	0.760 - 0.978
4th	16.29	15.44 - 16.96	13.04	12.27 - 13.96	16.50	15.68 - 17.39	59.84	58.51 - 61.45
IV Fluid								
No action	68.00	66.20 - 69.47	0.031	0.0015 - 0.074	60.50	59.04 - 61.92	0.	0.
1st	0.692	0.465 - 0.910	0.0054	0.00 - 0.016	6.098	5.391 - 7.030	35.15	33.54 - 36.54
2nd	10.73	9.729 - 11.70	0.0015	0.00 - 0.004	22.89	21.65 - 24.11	3.886	3.709 - 4.066
3rd	8.126	7.257 - 8.910	62.17	60.36 - 63.59	2.746	2.327 - 3.277	0.606	0.517 - 0.696
4th	12.54	11.62 - 13.51	37.76	36.65 - 39.56	7.722	7.533 - 7.946	60.38	58.91 - 62.00

Table 8: Difference action distribution among random seed of the DQN agent

	Vasopressor			IV fluid		
	Average(%)	SD(%)	Min - Max	Average(%)	SD(%)	Max - Min
No action	37.1	7.8	27-48	69.1	5.4	61-75
1st	16.2	10.7	5-34	6.5	3.9	1-11
2nd	16.7	6.3	8-25	7.6	3.9	3-12
3rd	15.2	5.3	7-22	6.2	3.6	2-11
4th	14.5	5.9	8-23	10.5	3.2	6-14

Table 9: Result for summary of WDR estimator of policies performed by DQN agents across different settings of the sensitivity analysis. For the following sections, we abbreviate the name of Duel-DDQN trained with different embedding to DQN-embedding. For example the Duel-DDQN trained with LSTM of 1 hourly binned data is called DQN-LSTM 1 hour.

Model Type	Avg	Std	Max
DQN-LSTM 1hour	2.9	0.32	3.8
DQN-GRU 1hour	3.6	0.75	4.9
DQN-LSTM 1hour w/o history	3.7	1.1	6.2
DQN-GRU 1hour w/o history	9.8	2.5	13.
DQN-LSTM 4hour	2.6	0.42	2.8
DQN-GRU 4hour	3.3	0.38	3.6