Reinforcement Learning for Sepsis Treatment: Baselines and Analysis

Aniruddh Raghu [1]

Abstract

In this work, we consider the task of learning effective medical treatment policies for sepsis, a dangerous health condition, from observational data. We examine the performance of various reinforcement learning methodologies on this problem, varying the state representation used. We develop careful baselines for the performance of these methods when using different state representations, standardising the reward function formulation and evaluation methdology. Our results illustrate that simple, tabular Q-learning and Deep Q-Learning both lead to the most effective medical treatment strategies, and that temporal encoding in the state representation aids in discovering improved policies.

1. Introduction

In the field of data-driven healthcare, there is significant interest in developing *decision-support systems* for clinicians. Such systems take in a patient's physiological information at a point in time, and provide insight to the attending clinician as to what treatment (medication types and dosages) to prescribe the patient so as to maximally improve their eventual outcome.

The reinforcement learning (RL) framework provides a natural way to approach this problem. The medical treatment process can be modelled as either a fully or partially-observed Markov Decision Process (MDP), and RL algorithms can be used to discover effective medical treatment strategies. Indeed, prior literature has utilised RL to discover treatment strategies for sepsis (Komorowski et al., 2018), mechanical ventilation usage (Prasad et al., 2017), heparin dosing (Nemati et al., 2016), and more.

Although RL has been used to tackle these problems in prior work, there exist relatively few baselines comparing how different RL methods perform on the task of discovering effective medical treatment strategies. Ideally, such baselines would control for potential sources of variability in performance (e.g., the state representation used, the reward function formulation, and the evaluation paradigm). Due to the lack of such baselines, it is challenging to identify what learning algorithms perform best, and where more development is needed.

In this work, we develop careful benchmarks of performance for different RL methods for the task of discovering sepsis treatment policies, explored in recent prior work (Komorowski et al., 2016; Raghu et al., 2017b; Komorowski et al., 2018; Peng et al., 2018). We choose to focus on this case study because of the medical significance of this problem – sepsis treatment is a very challenging clinical problem, and the condition is a leading cause of mortality (Cohen et al., 2006; Vincent et al., 2006). Furthermore, there is a lack of prior work on benchmarking different methods for this task, and the dataset we consider is publicly available (Komorowski et al., 2018). We standardise the cohort, reward formulation, and evaluation paradigm used, and compare the performance of different methods and state representations, aiming to understand what modelling choices and algorithms perform best.

Our investigation reveals that simple discretised state-space models, developed by clustering the original state-space and then using tabular Q-learning, and Deep Q-Learning methods can learn potentially effective and clinically plausible treatment policies. We demonstrate how temporally-sensitive state representations can improve the performance of such methods on this healthcare task. We also highlight the importance of combining qualitative and quantitative analysis when comparing the performance of RL methods in real-life settings.

2. Preliminaries

2.1. Reinforcement Learning (RL)

We assume the reader is familiar with the standard characterisation of the RL problem using Markov Decision Processes, represented by the tuple $\langle \mathcal{S}, \mathcal{A}, R, P, P_0, \gamma \rangle$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $R(s, a, s')$ is

[*] Equal contribution [1] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. Correspondence to: Aniruddh Raghu <araghu@mit.edu>.

Reinforcement Learning for Real Life (RL4RealLife) Workshop in the 36^{th} International Conference on Machine Learning, Long Beach, California, USA, 2019. Copyright 2019 by the author(s).

the reward function, $P(\cdot|s,a)$ is the transition probability distribution, P_0 is the initial state distribution, and $\gamma \in [0,1)$ is the discount factor. The goal is to develop an agent that learns a policy $\pi(a|s)$ that maximises the expected return, $V^\pi = \mathbb{E}_{H \sim P_H^\pi}\big[R(H)\big]$, where trajectories $\{H_i\}$ are generated by the distribution over trajectories P_H^π (a function of the policy π), and the return $R(H)$ is the sum of discounted rewards, $\sum_{t=0}^{T-1} \gamma^t r_t$, from a particular trajectory H.

2.2. Medical case study – sepsis treatment

We consider developing benchmarks for RL methods on a real-world medical dataset, explored in prior literature (Komorowski et al., 2016; Raghu et al., 2017b; Komorowski et al., 2018), dealing with the medical treatment of sepsis patients in intensive care units (ICUs). This dataset contains trajectories of patients with sepsis for their time in the ICU, and has information at every point in time (discretised into 4 hour blocks) about a patient's physiological state and the treatments they were given by clinicians. It also contains a patient's eventual outcome (survival/mortality).

We use the same basic modelling of the sepsis treatment problem as in Raghu et al. (2017b); the medical treatment process for a sepsis patient is framed as a continuous state-space MDP. A patient's state in each 4 hour block is represented as a vector of demographic features, vital signs, and lab values. The action space, $\mathcal{A}$, is of size 25 and is discretised over doses of two drugs commonly given to sepsis patients (IV fluids and vasopressors). The reward $r(s,a,s')$ is positive at intermediate timesteps when the patient's well-being improves, and negative when it deteriorates (dictated by differences in key physiological parameters from s to s'). At the terminal timestep of a patient's trajectory, a positive reward is assigned for survival, and a negative reward for mortality.

When we use RL for learning treatment strategies, we are restricted to only using observational data (as we have no simulator), so apply batch-mode RL, sampling patient trajectories from this fixed observational dataset during learning.

2.3. Related work

There is a large body of prior work considering the use of RL methods to learn medical treatment strategies.

Prasad et al. (2017) used the Fitted-Q Iteration (FQI) algorithm (Ernst et al., 2005) to learn treatment strategies for mechanical ventilation weaning from observational data; the authors compared the use of random forest models and neural networks to approximate the Q function. In contrast, we consider the sepsis treatment problem, and survey a broader set of methods, including discrete state-space models, FQI, and deep Q-learning. We also consider the impact of the state representation on the policy learned.

For the sepsis treatment problem, Komorowski et al. (2016; 2018) discretize the feature representation described in Section 2.2 using k-means clustering, model the transition distribution of the MDP in this discrete space, and then use Q-value iteration to discover an effective treatment policy.

Raghu et al. (2017b;a); Peng et al. (2018) use variants of Deep Q-learning to learn medical treatment policies for sepsis. These methods use different preprocessing techniques to generate state representations from the above physiological features; for example, Raghu et al. (2017b) uses the raw physiological features (following standardization) and Peng et al. (2018) uses a recurrent autoencoder to capture temporal information. Peng et al. (2018) also incorporates a Mixture of Experts approach, combining the Deep Q-learning policy with a kernel-based policy.

Raghu et al. (2018b) considers continuous state-space model-based RL for sepsis treatment, by first fitting an environment model to capture the transition dynamics of the MDP and then using the Proximal Policy Optimization algorithm (Schulman et al., 2017) to learn suitable treatment policies. As a feature representation, this work uses a concatenation of the raw physiological features over several timesteps to incorporate temporal information.

In this work, we also focus on the sepsis treatment problem, but aim to develop baselines for different RL methods and state representations through standardising the reward function and evaluation paradigm used.

3. Methods

We now present the state representations and RL methods considered in this work, and discuss the evaluation methodology used to develop the performance baselines.

3.1. State representation and reward function

Prior work has used several different representations for a patient's physiological state, considering either the **original state** — observed features at a point in time, as described in Section 2.2 — the **concatenated state** — obtained by concatenating the observed features from the last several timesteps — or the **autoencoded state** — preprocessed using a recurrent autoencoder to incorporate temporal information. We consider all three varieties of state preprocessing in this work, in combination with different RL algorithms. The autoencoder uses a one layer LSTM encoder and a one layer LSTM decoder, with a hidden state size of 128, as in Peng et al. (2018). To train the autoencoder, a batch of trajectories is sampled from the training dataset, and gradients are computed based on the mean squared error in predicting the current state s_t given the embeddings up until time t: $x_1, x_2, \ldots, x_t$. These embedddings are computed using the LSTM encoder, which takes the original state sequence

$s_1, s_2, \ldots, s_t$ as input.

Prior works also differ in terms of the reward function used; for consistency, we adopt the reward formulation from Raghu et al. (2017a), based on eventual outcome and variation in important physiological features, due to its clear clinical correspondence.

3.2. RL methods

We focus on commonly used methods in the literature such as Q-value iteration, tabular Q-learning, Deep Q-learning, and Fitted-Q Iteration (FQI), considering both neural networks and random forests as the Q function approximators. We do not compare to continuous state-space model-based RL, due to the difficulty in developing reasonable environment models in the continuous state-space (Raghu et al., 2018b). As further description of the methods we consider:

- Q-value iteration (QVI): We consider each of the different state spaces mentioned before, discretise each state space using k-means clustering (1000 cluster centroids), and then fit a transition matrix based on the observed data. The resulting MDP is solved using QVI to discover an 'optimal' policy.
- Tabular Q-learning: the state space is discretised as with QVI, and then standard batch-mode Q-learning is applied (Watkins & Dayan, 1992); trajectories are sampled from the training dataset and the Q-table is updated.
- FQI: We compare both random forest FQI and neural FQI, as in Prasad et al. (2017). The FQI algorithm is run for 50 iterations for random forest FQI, and 20 iterations for neural FQI. Each iteration alternates training on the current data, and then bootstrapping to generate new target values (Ernst et al., 2005). The random forest used has 50 estimators, and is implemented in scikit-learn (Pedregosa et al., 2011). The neural network is a simple 2 layer MLP, with batch normalization (Ioffe & Szegedy, 2015) and Leaky ReLU activation. We note here that an improved version of FQI incorporating Double Q-learning (van Hasselt et al., 2015) and more sophisticated sampling methods (e.g., Prioritized Experience Replay (Schaul et al., 2015)) could prove to be more effective; however, this was not tested in our benchmarking.
- Deep Q-learning: a Deep Q-Network (DQN) (Mnih et al., 2015) is used; the architecture is as in Raghu et al. (2017b) – a 2 layer fully connected network with 256 and 128 hidden units respectively. The network uses batch normalization, and leaky ReLU activations. The final model is a Dueling Double-DQN (van Hasselt et al., 2015), and uses Prioritized Experience Replay (PER) (Schaul et al., 2015) to speed up learning. During training, we sample a batch of transitions from the dataset (according to PER) and then use this to compute a gradient update.

A basic search over parameters (training time, some architecture specifications, etc.) was conducted for the different methods (using cross validation) before settling on this final set of hyperparameters.

3.3. Evaluation

For evaluation, we are restricted to using only observational data, so we utilise off-policy evaluation (OPE) methodologies (Precup et al., 2000). We consider two estimators, Per-Horizon Weighted Importance Sampling (PHWIS) and Per-Horizon Weighted Doubly Robust (PHWDR) (Jiang & Li, 2016; Doroudi et al., 2017). We follow the guidelines in (Raghu et al., 2018a) when constructing these estimators; we model the behaviour policy (the clinical policy) using approximate kNN (Indyk & Motwani, 1998) with 250 neighbours in the concatenated state space, and FQI with random forests (50 estimators) for control variate terms. As these estimators can be high variance, we use a bootstrap procedure to estimate confidence intervals – for each policy, we sample 200 patient trajectories from a held-out dataset, compute the OPE estimates, and then repeat this process 500 times.

The policies we discover through these RL methods are determinstic in nature. Deterministic policies are hard to evaluate (Raghu et al., 2018a; Gottesman et al., 2018), due to importance sampling weights in the aforementioned estimators becoming zero, reducing the effective sample size of the OPE estimators. We follow a similar procedure to Komorowski et al. (2018) to soften the discovered policies, so they take the desired action 96% of the time and a random action the other 4% of the time.

4. Results

We now analyse the performance of the different methods, from both a quantitative and qualitative standpoint.

4.1. Quantitative analysis

We present numerical results for the performance of the different methods. We consider the mean and standard deviation (obtained via bootstrapping) under the PHWIS and PHWDR estimators. We consider the different methodologies mentioned above, as well as the clinical policy, a policy taking random actions at every timestep, and a policy taking zero drug at every timestep.

The results that follow are presented in the format: ***method name—state representation***. Methods considered are Q-

Value Iteration (QVI), Tabular Q-learning (TQL), Fitted Q-Iteration with Random Forest (FQIRF), Fitted Q-Iteration with Neural Network (FQINN), and Dueling Double Deep Q-Network (DQN). In state representations, 'O' refers to the original feature set, 'A' refers to autoencoder, and 'C' refers to concatenation of the current and previous three timesteps' states into a larger feature vector.

Table 1. Table showing off-policy evaluation performance (mean and standard deviation) of different methods and state representations under the PHWIS and PHWDR estimators. Also included are results for the clinician policy, a policy taking random actions, and a policy prescribing zero drug at every timestep.

Method	PHWIS	PHWDR
TQL-O	10.1 ±2.3	10.0 ± 1.55
TQL-A	10.0 ± 2.5	9.87 ± 1.48
TQL-C	**11.35 ± 1.56**	**10.4 ± 1.1**
QVI-O	7.75 ± 3.23	8.99 ± 1.81
QVI-A	9.11 ± 2.91	9.39 ± 1.88
QVI-C	8.48 ± 2.99	9.08 ± 1.77
FQIRF-O	9.53 ±2.56	9.29 ± 1.80
FQIRF-A	9.86 ±2.56	9.76 ± 1.63
FQIRF-C	10.2 ± 2.1	9.97 ± 1.45
FQINN-O	8.72 ± 3.1	9.52 ± 1.91
FQINN-A	9.79 ± 2.58	9.79 ± 1.72
FQINN-C	8.64 ± 2.95	9.19 ± 1.96
DQN-O	9.83 ± 2.30	9.73 ± 1.56
DQN-A	11.0 ± 2.0	**10.4 ±1.3**
DQN-C	10.4 ± 2.2	10.0 ± 1.4
Clinician	9.82 ± 0.67	9.62 ± 0.63
Random drug	6.83 ± 3.51	8.41 ± 2.00
Zero drug	10.7 ± 2.1	10.5 ± 1.6

Table 1 shows the performance of the different methods. Figure 1 shows boxplots representing the spread of performance across different bootstraps for selected models that performed best, for PHWIS and PHWDR respectively.

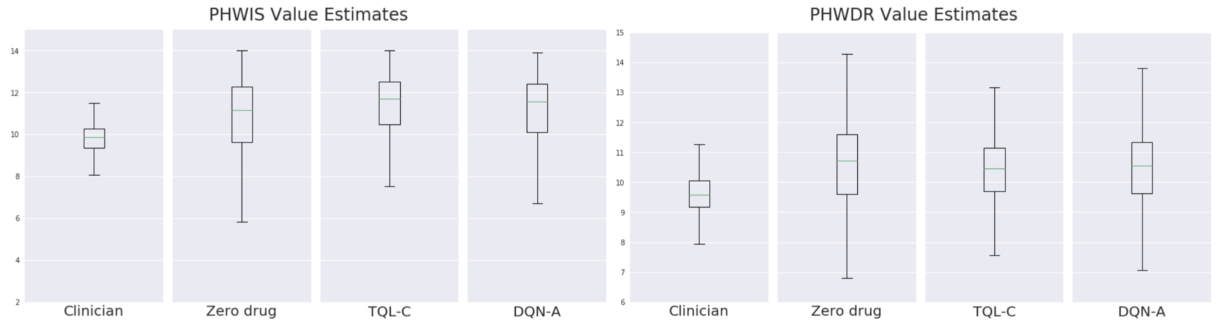
How do different RL methods perform? Considering PHWIS and PHWDR, we see that all policies have quite significant variance (a similar result seen in (Komorowski et al., 2018)), so it is challenging to assert that methods improve upon clinical performance. However, there seems to be some improvement over the base clinical policy by the better performing methods. TQL and DQN are the best-performing approaches, as evidenced by the quantitative results. QVI, at least with this framing of the problem, achieves poor results – this may be due to the difficulty in modelling the transition probability distribution in the underlying MDP. The FQINN methods perform quite poorly – the learned policies are quite different to the clinical policy (examples of learned policies are shown in Figure 3); this difference leads to poorer OPE results. FQIRF also suffers from this problem, but to a lesser degree.

What impact does state representation have? The autoencoded and concatenated state representations improve the average performance of most methods. Given that a patient's feature at a particular timestep is unlikely to be sufficient to capture their entire physiological state, this temporal encoding is well-motivated. The value of the state representation is diminished when we perform clustering, perhaps explaining the less-significant difference for TQL. The effect of state representation is presented in the box plot in Figure 2, comparing the PHWIS policy value for different state representations for the DQN model, which is more sensitive to the choice of state representation, being a continuous state-space model. We see more clearly the impact of the temporal representations.

Importantly, the state representations considered here are those that have been used in prior work; evaluating other state representations that are clinically-motivated is an important direction of future work.

Why does the zero-drug policy perform so well? Consider the two types of patient trajectories – those where patients survived (which will have positive returns), and those were patients did not survive (which will have negative returns). In cases where patients survived, clinicians may not prescribe any medication at particular times (the zero-drug policy), because patients are generally healthier and do not require medical interventions. Conversely, in those trajectories in which patients do not survive, clinicians rarely take the zero-drug policy, because these patients need medical interventions to manage their condition. The importance sampling weights associated with survival trajectories for the zero drug policy will be large, due to the general agreement between the evaluation policy (zero-drug policy) and the behaviour policy (clinician policy). The opposite is true for the non-survival trajectories. There are relatively few samples in the dataset with patients who did not survive *and* were treated with the zero-drug policy, as this is not clinically meaningful. Consequently, in the final estimator, for the zero-drug evaluation policy, the survival trajectories will be weighted highly, and the non-survival trajectories will be downweighted significantly, leading to a high estimated policy value. However, it is clear that not administering any drug is not a clinically feasible policy.

More generally, to summarize the problems with off-policy evaluation on this task, there are two key areas to consider. Firstly, there may be hidden confounders that influence the clinician's policy and patient outcome, in addition to the features that are present in the dataset. This is insufficient coverage of state representation. Secondly, as seen with the zero-drug policy, the behaviour policy (clinician's policy) does not cover the state space thoroughly enough. This



(a) PHWIS estimates

(b) PHWDR estimates

Figure 1. PHWIS and PHWDR estimates for 500 bootstraps for different evaluation policies.

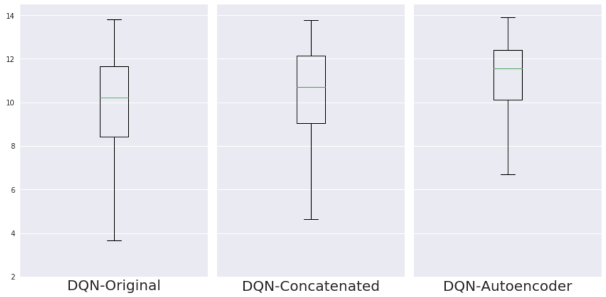


Figure 2. Impact of state representation on PHWIS estimates for the DQN model.

is insufficient coverage of the behaviour policy. Both of these issues severely affect the performance of importance sampling methods, leading to the unreliable performance observed. Related ideas have been expressed in Gottesman et al. (2018).

Given these issues, it is important to not treat quantitative evaluations as absolute, and also examine learned policies qualitatively to understand what has been discovered.

4.2. Qualitative analysis

As the quantitative evaluation methodologies suffer from high variance and specific failure cases (e.g. the zero-drug evaluation policy), we also examine the policies learned from a qualitative standpoint, aiming to understand which policies are also clinically meaningful.

We consider the overall cumulative distribution of actions for different severity patient states (based on a severity index, the SOFA score). We compare the average total variation distance between clinical and evaluation policies in these states to validate that the evaluation policies are not deviating too far from the base clinical policy – large deviations are (a) harder to evaluate (Gottesman et al., 2018) and (b) likely result in clinically meaningless policies.

Table 2. Average total variation distance between clinical policy and other policies, overall and stratified by patient severity (low, mid, and high). The clinician policy is formed using the kNN estimate used for evaluation.

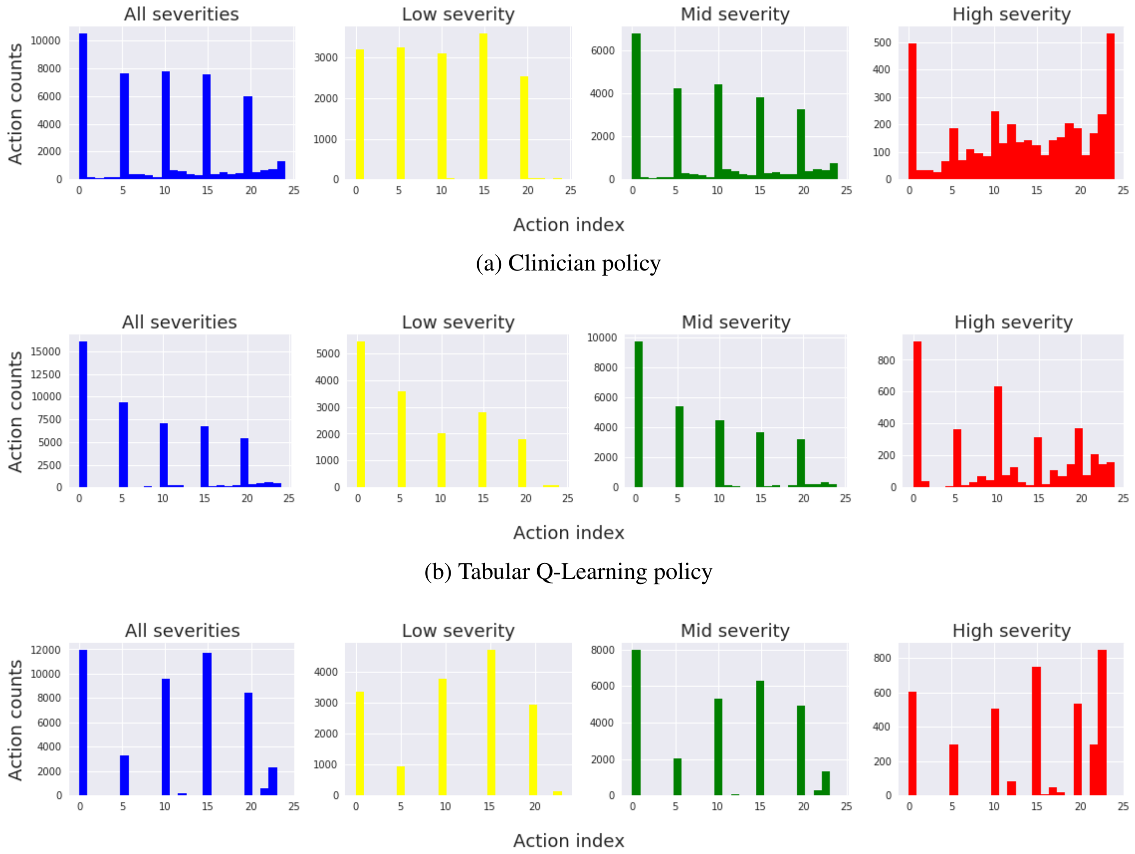
	Overall	Low	Mid	High
TQL-C	0.717	0.724	0.703	0.782
DQN-A	0.733	0.745	0.715	0.811
Random drug	0.934	0.938	0.933	0.928
Zero drug	0.755	0.775	0.739	0.786

Table 2 shows the average total variation distance between the clinical policy and selected learned policies for states with different severity of sepsis (noted as low, mid, and high severity). Figure 3 shows the distribution of actions for two policies that performed well, stratified by patient severity. Also shown is the clinical policy.

Comparing distributional distance and the policy histograms, we see that both the TQL-C and DQN-A policies are on average, closer to the clinician policy than the zero drug policy, although the difference is only slight in the high severity regime. They also both share some key characteristics overall with the clinician's policy, with both policies prescribing higher dosage amounts of drugs, demonstrated by large action counts for action indices 22-24. They appear to learn some similar features to the clinical policy, especially in the lower and medium severity regime.

5. Conclusion

In this work, we explored the task of developing baselines for the performance of several RL methodologies on the task of learning sepsis treatment strategies from observational data. We considered a range of different state representations and policy learning algorithms, and found that simple, tabular Q-learning can be used to learn quite effective policies, and is competitive with more complex continuous



(a) Clinician policy

(b) Tabular Q-Learning policy

(c) DQN policy

Figure 3. Comparing the cumulative action distributions under three policies: clinician, Tabular Q-Learning, and Deep Q-Network.

state-space methods, such as Deep Q-Learning. Developing such baselines has also reinforced the importance of informative state representations in learning good quality policies, especially in this medical scenario.

We also demonstrated the importance of considering both qualitative and quantitative evaluation when applying RL to real-world problems. These analyses revealed how both tabular Q-learning and Deep Q-learning could be used to find treatment strategies that potentially improve on what clinicians follow and advance the standard of patient care.

Acknowledgements

The author would like to thank the anonymous reviewers for their helpful comments and suggestions.

References

Cohen, J., Vincent, J.-L., Adhikari, N. K. J., Machado, F. R., Angus, D. C., Calandra, T., Jaton, K., Giulieri, S., J.Delaloye, Opal, S., Tracey, K., van der Poll, T., and Pelfrene, E. Sepsis: a roadmap for future research. *Lancet Infectious Diseases*, 15(5):581614, 2006.

Doroudi, S., Thomas, P. S., and Brunskill, E. Importance sampling for fair policy selection. 2017.

Ernst, D., Geurts, P., and Wehenkel, L. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6(4):503–556, 2005.

Gottesman, O., Johansson, F., Meier, J., Dent, J., Lee, D., Srinivasan, S., Zhang, L., Ding, Y., Wihl, D., Peng, X., et al. Evaluating reinforcement learning algorithms in observational health settings. *arXiv preprint arXiv:1805.12298*, 2018.

Indyk, P. and Motwani, R. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pp. 604–613. ACM, 1998.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pp. 448–456, 2015.

Jiang, N. and Li, L. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 652–661, 2016.

Komorowski, M., Gordon, A., Celi, L. A., and Faisal, A. 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*, December 2016.

Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., and Faisal, A. A. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716, 2018.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., D, W., Legg, S., and Hassabis, D. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015.

Nemati, S., Ghassemi, M. M., and Clifford, G. D. Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach. In *38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, August 2016.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Peng, X., Ding, Y., Wihl, D., Gottesman, O., Komorowski, M., Lehman, L.-w. H., Ross, A., Faisal, A., and Doshi-Velez, F. Improving sepsis treatment strategies by combining deep and kernel-based reinforcement learning. In *AMIA Annual Symposium Proceedings*, volume 2018, pp. 887. American Medical Informatics Association, 2018.

Prasad, N., Cheng, L.-F., Chivers, C., Draugelis, M., and Engelhardt, B. E. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. *arXiv preprint arXiv:1704.06300*, 2017.

Precup, D., Sutton, R. S., and Singh, S. P. Eligibility traces for off-policy policy evaluation. In *International Conference on Machine Learning*, pp. 759–766, 2000.

Raghu, A., Komorowski, M., Ahmed, I., Celi, L., Szolovits, P., and Ghassemi, M. Deep reinforcement learning for sepsis treatment. *arXiv preprint arXiv:1711.09602*, 2017a.

Raghu, A., Komorowski, M., Celi, L. A., Szolovits, P., and Ghassemi, M. Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach. In *Machine Learning for Healthcare Conference*, pp. 147–163, 2017b.

Raghu, A., Gottesman, O., Liu, Y., Komorowski, M., Faisal, A., Doshi-Velez, F., and Brunskill, E. Behaviour policy estimation in off-policy policy evaluation: Calibration matters. *arXiv preprint arXiv:1807.01066*, 2018a.

Raghu, A., Komorowski, M., and Singh, S. Model-based reinforcement learning for sepsis treatment. *arXiv preprint arXiv:1811.09602*, 2018b.

Schaul, T., Quan, J., Antonoglou, I., and Silver, D. Prioritized Experience Replay. *CoRR*, abs/1511.05952, 2015. URL http://arxiv.org/abs/1511.05952.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

van Hasselt, H., Guez, A., and Silver, D. Deep reinforcement learning with double q-learning. *CoRR*, abs/1509.06461, 2015. URL http://arxiv.org/abs/1509.06461.

Vincent, J.-L., Sakr, Y., Sprung, C. L., Ranieri, V. M., Reinhart, K., Gerlach, H., Moreno, R., Carlet, J., Gall, J.-R. L., and Payen, D. Sepsis in European intensive care units: results of the SOAP study. *Critical Care Medicine*, 34(2): 344–353, 2006.

Watkins, C. J. C. H. and Dayan, P. Q-learning. *Machine Learning*, 8(3):279–292, May 1992. ISSN 1573-0565. doi: 10.1007/BF00992698. URL https://doi.org/10.1007/BF00992698.