Safe Reinforcement Learning for Sepsis Treatment

Yan Jia
Department of computer science
University of York
York, UK
yj914@york.ac.uk

John Burden
Department of computer science
University of York
York, UK
jjb531@york.ac.uk

Tom Lawton
Bradford Royal Infirmary and
Bradford Institute for Health Research
Bradford, UK
tom.lawton@bthft.nhs.uk

Ibrahim Habli
Department of computer science
University of York
York, UK
ibrahim.habli@york.ac.uk

***Abstract*—Sepsis, a life-threatening illness, is estimated to be the primary cause of death for 50,000 people a year in the UK and many more worldwide. Managing the treatment of sepsis is very challenging as it is frequently missed at an early stage and the optimal treatment is not yet clear. There are promising attempts to use Reinforcement Learning (RL) to learn optimal strategies to treat sepsis patients, especially for the administration of intravenous fluids and vasopressors. However, some RL agents only take the current state of patients into account when recommending the dosage of vasopressors. This is inconsistent with clinical safety practice in which the dosage of vasopressors is increased or decreased gradually. A sudden major change of the dosage might cause significant harm to patients and as such is considered unsafe in sepsis treatment. In this paper, we have adapted one of the deep RL methods published previously and evaluated whether the learned policy contains these sudden, major changes when recommending the vasopressor dosage. Then, we have modified this method to address the above safety constraint and learnt a safer policy by incorporating current clinical knowledge and practice.**

***Index Terms*—Sepsis treatment, Reinforcement learning, Safe policy**

I. Introduction

Sepsis is a life-threatening organ dysfunction caused by a dysregulated host response to infection [1]. A recent estimation is that one in five deaths worldwide is caused by sepsis [2] [3]. A major challenge is early detection of sepsis since the earlier the treatment begins the greater the chance of patient recovery. Once the condition has been detected, treatment normally involves administration of antibiotics and infection source control. When it turns into septic shock, administration of intravenous fluids and vasopressors will be necessary, but deciding on the treatment strategy for intravenous fluids and vasopressors is often difficult. Different fluid and vasopressor treatment strategies have been tested leading to quite different results in terms of patient mortality [4]. Further, many healthcare agencies and communities have devoted significant efforts to sepsis management, e.g. the Surviving Sepsis Campaign [5]. Despite such efforts, the optimal strategy for the administration of intravenous fluids and vasopressors remains unclear.

Some research groups have harnessed Reinforcement Learning (RL) to learn optimal strategies for recommending intravenous fluids and vasopressors. For example, the AI Clinician [6] has been developed using RL to dynamically recommend fluids and vasopressors for adult patients. RL is a very powerful machine learning (ML) technique which is widely used in complex decision making tasks to find an optimal policy [7]. It assumes that the environment can be viewed as a Markov Decision Process (MDP) in which an assumption is made that the future state of the process depends only on the current state; that is, given the current state, the future state does not depend on the cumulative history of past states.

However, for a complex intervention to be safe and effective, it is important for the technology to fit with accepted clinical practices and workflows. In the case of sepsis treatment, when recommending dosage for vasopressors, if we merely consider the optimal action based on the current state, it might cause a sudden major change in the dosage, which can be dangerous to some patients, e.g. resulting in acute hypotension (arising from rapidly decreasing doses), hypertension or cardiac arrhythmias (arising from rapidly increasing doses) [8] [9] [10]. Because the half life (the period of time for the concentration of a drug in the body to reduce by 50%) of Norepinephrine (a commonly used vasopressor) is measured in seconds or minutes [11], changes in Norepinephrine can have rapid effects on patients. The recommended dosage (and dosage changes) for intravenous fluids is less sensitive than for vasopressors as the half life of fluids is measured in hours [12], thus changes in fluid take a lot longer to take effect. Therefore, in this work, we focused on the safety of vasopressor administration.

In this paper, we have adapted a previously published deep RL method [13] used to learn the optimal treatment strategy, or policy, to investigate further the safety issues associated with sepsis treatment. The data used for learning the optimal policy is a large publicly available database - MIMIC III [14], collected from USA hospitals. As the MIMIC III data

This project is funded by Bradford Teaching Hospitals NHS Foundation Trust and supported by Assuring Autonomy International Programme.

set was generated by recording the real clinicians' actions, we refer to it as the clinician policy in comparison with the learnt optimal policy. We evaluated the learnt optimal policy and compared it against the clinician policy, i.e. the real patient trajectories in the test data set, including whether or not they show this kind of sudden major change when recommending vasopressor dosage for each patient. In doing so we identified that the learnt optimal policy has many more sudden major dosage changes than the clinician policy. As a consequence, we modified the model to capture the change between the current vasopressor dose and previous vasopressor dose in the state space. In addition, we modified the RL cost function to penalise the policy when this kind of behaviour is learnt. The result shows that what we have learnt has fewer sudden major changes and is therefore closer to the clinician's behaviour. Finally, we have evaluated both policies using a regression-based procedure for off-policy evaluation [15], which shows that the performance of our modified policy has higher value than the clinician policy. However, although the learnt optimal policy based on [13] seems to have higher value than our modified policy, in terms of vasopressor delivery, our modified policy is safer in the sense of being consistent with clinicians' knowledge, specifically in terms of avoiding sudden vasopressor changes and their harmful effects.

II. Background

RL consists of an agent interacting with its environment by producing actions and discovering errors or receiving rewards [7]. The environment is often represented by an MDP. An MDP is defined by $M = \langle S, A, P, R, \gamma \rangle$, where S is the state space, A is the action space, P is the transition function with $P(s'|s,a)$ denoting the probability of reaching state s' if taking action a in state s. R is the reward function with $R(s,a)$ being the expected immediate goodness of (s,a) and γ is the discount factor. A policy fully defines the agent's behaviour and maps the perceived states of the environment to actions for the agent to take. It is often denoted as π. If the agent is following policy π at time t, then $\pi(a|s)$ will be the probability that $A_t = a$ if $S_t = s$. The action-value function $Q_\pi(s,a) = E_\pi \left[\sum_{t'=t}^{T} \gamma^{t'-t} r_{t'} | S_t = s, A_t = a\right]$, is the expected discounted reward starting from state s, taking action a, when following policy π, where r_t is the reward received when transitions from the state s_t to the state s_{t+1} after taking action a_t and has mean $R(s_t, a_t)$ conditioned on (s_t, a_t), and T is the terminal time step.

A deep Q-Network (DQN) is a widely-used modern RL algorithm, which combines Q-learning [16] with a deep artificial Neural Network (NN) [17]. It learns the optimal policy by employing the same update rules and operating principles as Q-learning but using an NN as its action function representation. DQN uses the experiences or samples $\langle s, a, r, s' \rangle$ generated by interaction with the environment to train the NN. It uses a squared error loss function, which is the difference between the output of the network, $Q(s,a,\theta)$ and the desired target $Q_{target} = r + \gamma \max_{a'} Q(s', a', \theta)$ to update the NN. The parameters θ of the NN are updated as follows:

$$\theta_{k+1} \leftarrow \theta_k - \alpha \nabla E[(Q_{target} - Q(s,a;\theta))^2], \quad (1)$$

where k is the iteration step when training the NN. Simple DQNs have some shortcomings and there are various ways of refining them to improve their performance. One way to improve them is to use double DQNs which employ two NNs, one produced as in standard DQN, which is the main network with parameters θ, and the other a copy of the network from the last iteration used to obtain the Q-value, which is the target network with parameters θ'. The standard double Q-network loss is shown in (2).

$$L(\theta) = E[(Q_{double-target} - Q(s,a;\theta))^2], \quad (2)$$

where $Q_{double-target} = r + \gamma Q(s', argmax_{a'} Q(s', a'; \theta); \theta')$. In this work, we extended this equation to include a term which accounts for safety (see section IV).

III. Related Work

Much previous work in identifying the best treatment for sepsis patients has focused on clinical trials. In [18] they carried out a random trial to investigate the effect of the discontinuation of vasopressors in management of septic shock and found that tapering Norepinephrine rather than Vasopressin may be associated with a higher incidence of hypotension in patients recovering from septic shock who are on concomitant Norepinephrine and Vasopressin. However, the study was stopped early due to a significant difference in the incidence of hypotension between the control and experimental group, which also reveals the difficulty of conducting clinical trials to find the optimal treatment for sepsis patients. A recent review found that, in the last decade, the physiopathology of sepsis has become better understood. However, it concluded that clinical trials had yielded no satisfactory results [19].

More recently, researchers have utilised RL to learn the optimal treatment for sepsis patients. The application of an RL approach could reduce the time and cost to identify good treatment strategies by finding new insights in large patient-related clinical data sets, compared with clinical trials. For example, the AI Clinician [6] was built using RL on the MIMIC III database to explore the optimal treatment strategy for administering intravenous fluids and vasopressors. The state space included patients' demographics, Elixhauser premorbid status, vital signs, laboratory values, fluids and vasopressors received. The action space for the MDP is discretised into 25 possible actions with 5 possible choices for intravenous fluids and vasopressors respectively. This work used policy iteration to find an optimal policy. The group have also applied double DQN to determine the optimal policy, where they learned the treatment policy over continuous spaces using an NN with the same 25 actions [13]. They used SOFA (Sequential Organ Failure Assessment) score and Arterial Lactate (the level of lactate from arterial blood) to determine the intermediate reward and a terminal reward of +15 or -15 depending whether or not the patient survived their stay in hospital. The SOFA score is a measurement of organ failure with high values

associated with poor outcome; similarly, high levels of lactate are associated with poor outcomes in sepsis treatment. In detail the intermediate reward function is:

$$r(s_t, s_t+1) = C_0 \mathbb{1}\ (s_{t+1}^{SOFA} = s_t^{SOFA}\ \&\ s_{t+1}^{SOFA} > 0) + C_1(s_{t+1}^{SOFA} - s_t^{SOFA} + C_2 \tanh(s_{t+1}^{Lactate} - s_t^{Lactate}) \tag{3}$$

In this work, we have adapted the methods in [13] to train an agent to learn the optimal policy based on the same data set and the same patient cohort taken from MIMIC III and used this as a basis for evaluating and enhancing our approach. The patient cohorts are defined based on the sepsis-3 criteria – suspected infections combined with SOFA score ≥ 2. Patients who satisfy the sepsis-3 criteria, but with any one of the following conditions, 1. not adult, 2. intravenous fluid intake not documented, 3. possible withdrawal of treatment, 4. erroneous intake or output data, were excluded. The resulting patient cohorts were divided into a training dataset (80%, 20938), a validation dataset (10%, 2149) and a testing dataset (10%, 2160). For detailed patient features included in the state space, see the supplement to [6].

IV. Method and Results

For ease of presentation, we combine the description of our method and the results of the work.

A. Evaluation of the learned optimal policy

First, we have adapted the method in [13] to train the agent to learn the optimal policy. The main adaptation was considering 90-day mortality rather than deaths in hospital, as some patients might choose to be discharged in order to return home when they are unlikely to survive, so this is more appropriate when learning the optimal policy. We also used 47 features to represent the state space (as against 48 in the original work [13]) as one of the features is the time-steps to treat the patient, and we believe this is less relevant when clinicians are treating patients in reality, as they would not decide to stop treating the patient just because they have been treating the patient for a long period of time. Indeed, they have also made this kind of alterations in their later work [6]. The action space includes 25 possible actions with five discretised choices for the dose of intravenous fluids and five for vasopressors respectively, which is shown in Table I. Table I also shows the detailed dose range and dose median for the five vasopressor choices; this is important as this work is focused on the safety of vasopressor administration. Note that vasopressor dosage is shown in mcg/kg/min of Norepinephrine equivalent. Fig. 1 shows the comparison of the clinician policy and the learnt optimal policy on the test data set. As can be seen, the clinicians tended to prescribe less vasopressors and more intravenous fluids to patients than the learnt optimal policy (note the high frequencies in the clinician policy – 0,5,10,15,20 correspond to zero vasopressor dose). The proportion of time the clinicians prescribed vasopressors to patients was only 15% compared to 38% if the optimal policy recommendation was followed.

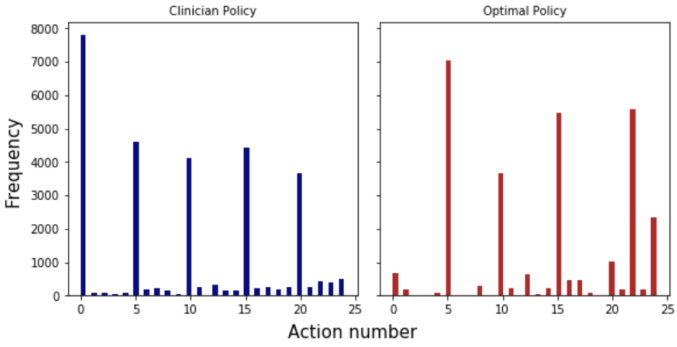

Fig. 1. Action frequency over all patient trajectories in test data set, where all actions are aggregated recommended by the clinician and optimal policies.

The results of this revision were generally consistent with the AI Clinician [6] and [13] in that the learnt optimal policy recommended more vasopressor than the clinicians' policy. Fig. 2 shows the correlations between the observed patient mortality and the difference between the doses suggested by the learnt optimal policy and the actual doses given by clinicians (i.e. the clinician policy). It shows that the minimum mortality rate is observed when there is no difference between the optimal policy and the clinician policy, which means patients who received doses similar to the doses recommended by the optimal policy have the lowest mortality. This implies that the learnt optimal policy is effective, which is the same as indicated in [13].

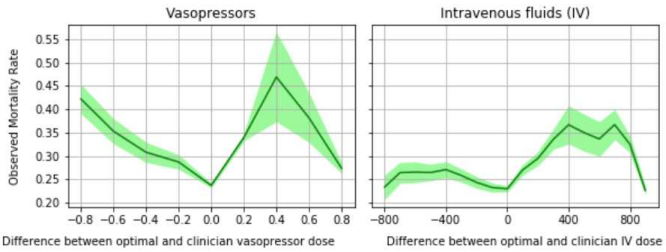

Fig. 2. Observed mortality rate variation with the difference between the doses recommended by the optimal policy and the actual doses, calculated by considering 90-day mortality.

Based on this comparison, the policy we have learnt can be seen to be similar to that in [13] in terms of optimality and validity. However, in contrast to [13], we have evaluated this policy from the safety perspective, specifically in terms of sudden changes in the recommended vasopressor dosage.

According to [20], doses of Norepinephrine over 0.5 mcg/kg/min are usually considered to be "high" and suggest the need for rescue or second-line therapy. Doses over 1.0 mcg/kg/min are rarely used. In the action space, shown in table I, the maximum dose change occurs when the recommendation changes from action 0 to action 4 in the following step for the same patient, or *vice versa*. This change is 0 to 0.786

TABLE I
DOSAGE ACTIONS

		Dose of vasopressor (mcg/kg/min)				
		No.: 0 **Range:** 0 **Median:** 0	1 (0.002, 0.079) 0.04	2 (0.08, 0.2) 0.135	3 (0.201,0.449) 0.27	4 (0.45, 1.005) 0.786
Dose of IV fluid	0	0	1	2	3	4
	1	5	6	7	8	9
	2	10	11	12	13	14
	3	15	16	17	18	19
	4	20	21	22	23	24

mcg/kg/min as 0.786 mcg/kg/min is the median of the fourth quartile and is clearly in a dangerous range.

We evaluated the maximum vasopressor dose change for the clinician policy and the learnt optimal policy on the test data set, which has 2,160 patients, by calculating the max absolute vasopressor dose change in one step for each patient during their treatment. In the clinician policy, we found 2.6% (57 patients) among 2,160 patients have this maximum dose change – 0.786 mcg/kg/min. In contrast, in the learnt optimal policy, we found 35% (756 patients) among 2,160 patients have this sudden change. Fig. 3 shows the comparison of max absolute vasopressor dose change between the clinician policy and the learnt optimal policy for these 2,160 patients. The max absolute vasopressor dose change following the learnt optimal policy is substantially higher than that of following the clinician policy. Further, many of the max absolute vasopressor dose changes in the learnt optimal policy reach 0.786, while there are only a few patients whose max absolute vasopressor dose change reaches 0.786 in the clinician policy. This implies that the learnt optimal policy may not be safe if used for treating sepsis patients, because of the prevalence of these sudden major dose changes and its harmful clinical effect. Fig. 3, also shows that there are a lot of patients whose max absolute vasopressor dose change was zero in the test data set (i.e. following the clinician policy). This is consistent with Fig. 1 in that clinicians tended to give less vasopressors, so some patients never got any vasopressors. However, if the learnt optimal policy was followed, many more patients would move out of the zero vasopressor "bucket". This also indicates that the learnt optimal policy tends to give more vasopressor to the patients than the clinician policy does. The apparent "noise" in Fig. 3 arises because the patients are sorted (ordered) first by the maximum change in the test data (i.e. the clinician policy), then by the maximum change in the optimal policy and, for some patients, the clinician policy gave a higher maximum change than the optimal policy.

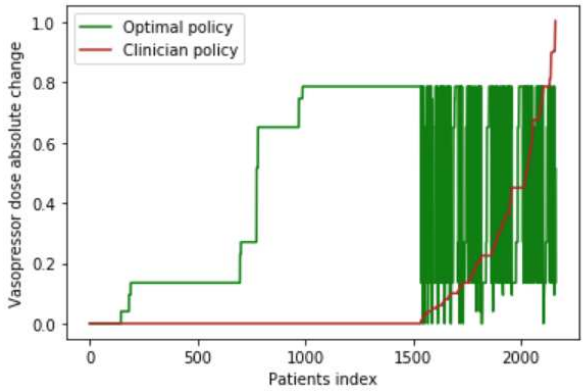

Fig. 3. Comparison of max absolute vasopressor dose change in one step for each patient in the test data set between the clinician and the learnt optimal policy.

B. Modification of the learnt optimal policy

In response to the above clinical safety concerns, we have modified the model to embrace the safety constraint, which is to reduce the rate of sudden major vasopressor dose changes, particularly to avoid the maximum change of 0.786 mcg/kg/min. We made two alterations to enable the agent to learn a safer policy.

Firstly, we added an extra feature in the state feature space, which is the relative dose change compared with the previous vasopressor dose for each patient. This enables the agent to take account of the difference between the current step and the previous step in terms of vasopressor dose while learning the optimal policy, rather than merely using the current step state features.

Secondly, we have also altered the cost function used for training. In [13], the authors added a regularisation term to the standard double Q-network loss (see equation (2)) to penalise output Q-values when it was outside the allowed thresholds ($|Q_{thresh}| = 20$). On this basis, we have added a second regularisation term to penalise the output Q-values when the recommended dose is higher or lower than the previous dose by 0.786 mcg/kg/min (i.e., a jump from action 0 to action 4 or *vice versa* in one step when recommending vasopressor doses for the patients). The altered loss function is as follows:

$$L(\theta) = E[(Q_{double-target} - Q(s,a;\theta))^2] + \lambda_1 max \\ (|Q(s,a;\theta)| - Q_{thresh}, 0) + \lambda_2 max(|V_{change}| - 0.75, 0) \tag{4}$$

V_{change} is the agent recommended dose (argmax of $Q(s,a;\theta)$) minus the vasopressor dose in the previous step.

After the implementation of these two alterations we have learnt a new modified policy. Fig. 4 shows the comparison of

the clinician policy and the learnt modified policy on the test data set. It shows that the modified policy also recommends more vasopressor than the clinician policy. However, in contrast to the learnt optimal policy in Fig. 1, the proportion of time the modified policy recommended vasopressor to patients was 24% compared to 38% in the learnt optimal policy.

Fig. 5 shows the relationship between observed mortality and the difference between the modified policy and the clinician policy, which also shows that when there is no difference between the modified policy and the clinician policy the observed mortality is reduced to a minimum. This also implies the validity of the modified policy.

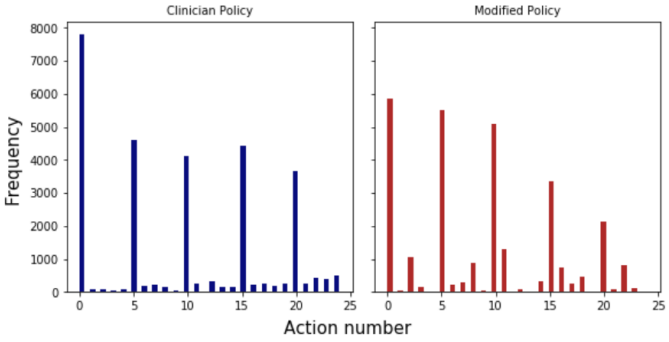

Fig. 4. Action frequency over all patient trajectories in test data set, where all actions are aggregated recommended by the clinician and modified policies.

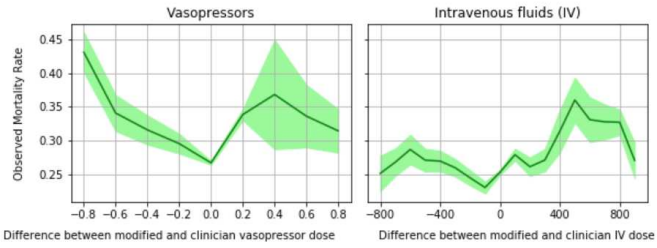

Fig. 5. Observed mortality rate variation with the difference between the doses recommended by the modified policy and the actual doses, calculated by considering 90-day mortality.

Having made these modifications we evaluated the new modified policy on the test data set to assess how many patients are subject to the sudden major change of vasopressor dosage, i.e. 0.786 mcg/kg/min. In total there are 7.87% (170 patients) amongst the 2,160 patients found with this change. Thus, the modified policy has reduced the rate of such sudden major changes of vasopressor dose by 77.5% when compared with the previous learnt optimal policy. Fig. 6 shows the maximum absolute vasopressor dose change in one step for each patient between the clinician policy and the modified policy. It shows a clear reduction in sudden major dose changes and the absolute change is much more reasonable compared to Fig. 3. However, it also shows an overall increase in the recommended vasopressor dose by comparison with the clinician policy, but the recommendations are much smoother. This is also consistent with the previous findings that it might be better to administer more vasopressor for sepsis patients but this modified policy behaves in a much safer way and is more consistent with clinical reality.

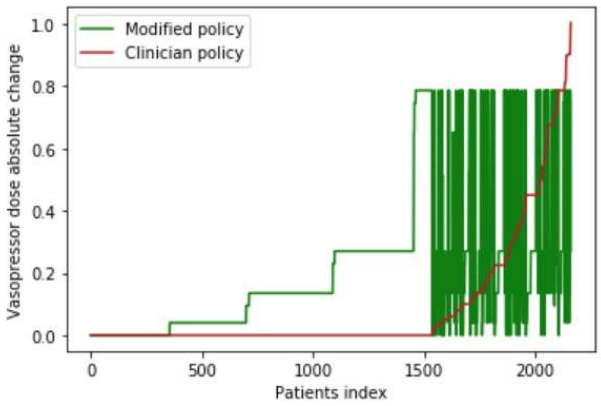

Fig. 6. Comparison of max absolute vasopressor dose change in one step for each patient in the test data set between the clinician and the modified policy.

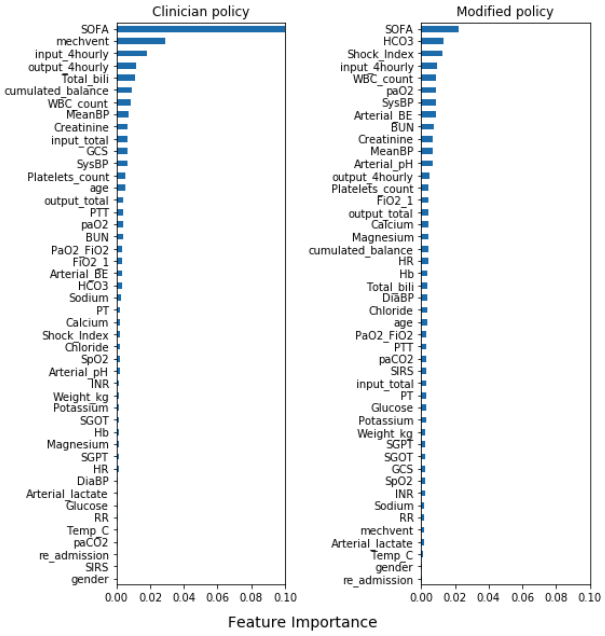

Fig. 7. Feature importance (from out of bag score) for clinician policy and the modified policy

In order to further assess the interpretability of the modified policy, we built classification random forest models to understand the relative importance assigned to the features in the state space when recommending vasopressor. Then, we compared the results with those of the clinician policy, see Fig.7. For both random forest models (clinician and modified policy), the current dose of vasopressor was discarded as the concern here is on what features influence whether or not vasopressor is recommended, not the size of the recommended dose. The relative importance of each feature was estimated using an out-of-bag score on the whole dataset, while we permuted the

values of each predictor, i.e. permutation feature importance [21]. In both policies, SOFA plays the most important role, which is as expected as SOFA describes sepsis-related organ failure. Shock index is also among the top rankings as it has been shown to indicate the need for vasopressor therapy. Gender and re-admission in both polices shows the least importance as these parameters are expected not to affect the decision of whether recommending vasopressor or not. This confirmed that the decisions suggested by the modified policy were clinically interpretable and relied primarily on sensible clinical parameters, such as SOFA, shock index, mean blood pressure, or white blood cells count (WBC_count) as shown in the figure.

C. Performance evaluation

It is not feasible to evaluate the policy on real patients because of ethical, legal and risk issues. Instead, we have carried out off-policy evaluation to assess the performance of the initial learnt optimal policy and the modified policy by fitting an MDP model $\widehat{M}$ from the current data to approximate the environment. Then the value function is computed using the estimated parameter transition probability $\widehat{P}$ and the reward $\widehat{R}$ recursively. The final estimated value averaged the resulting value function across all the observed trajectories in the test data set (refer to [15] [22] for a detailed description of the method). The average discounted reward of the chosen actions under the clinician policy across all of the trajectories in the test data set is also calculated as the benchmark, as shown in table II. It shows that the learnt optimal policy has a higher value than our modified policy. However, our modified policy is still higher than the clinician policy and in terms of vasopressor delivery, it is safer in the sense of avoiding sudden vasopressor changes and its dangerous effects on patients.

TABLE II
PERFORMANCE COMPARISON FOR DIFFERENT POLICIES

Policy	Estimated Discounted Reward
Clinician policy	7.16
Optimal policy	10.9
Modified policy	8.07

V. DISCUSSION

Effective treatment of sepsis is extremely important as it is one of the major causes of fatalities in hospitals. As noted above, clinical trials are expensive and are unlikely to identify effective treatment strategies quickly [19]. ML methods, such as RL, applied to clinical data have the potential to identify good treatment strategies more cost-effectively, but there are some potential pitfalls of ML which can give rise to unsafe outcomes, and some challenges in introducing ML into a clinical context.

First, the policies learnt in this case are produced by optimising the cost function. The cost function used in [13] does not address the rate of change of delivering vasopressor which is known to be potentially hazardous. The approach we have taken here enriches the cost function to take into account this important factor in clinical safety. There may be other factors that could beneficially be taken into account in identifying optimal treatment policies thus it is vital to combine knowledge of ML with clinical and patient safety expertise.

Second, if ML-based approaches are to be used in practice, in clinical settings, then it is necessary to assess their safety prior to deployment. A key element is to assure the safety of the ML-based system, potentially by producing a safety case [23] dealing with the specific challenges of ML, as illustrated in [24]. A second element is to engage with the clinicians in such a way that they are able to trust the recommendations produced by the system and are thus willing to use it in practice. We will address both these issues in conjunction with the team who developed the AI Clinician as part of the Assuring Autonomy International Programme [25].

Third, in settings as complex as sepsis care it is unrealistic to think that any treatment strategy learnt from historical clinical data will be "perfect" and also stable over time. Thus, it is important to learn from the behaviour of the system as observed, as opposed to the behaviour we anticipated (imagined), and to update our knowledge about the system as it evolves. The framework presented in [24] is intended to give a basis for monitoring and evaluating such complex systems to reflect the evolution of the system and its environment, and may therefore be helpful in this context.

VI. CONCLUSIONS

The work reported here illustrates both the potential value of ML in clinical settings, especially where clinical trials are costly and time-consuming, and the potential pitfalls of using ML to recommend treatment strategies. We have shown how to enrich the learning process with additional information about the safety of the treatment strategy and this is an important step forward in understanding how to integrate the notion of patient safety into the ML process. Further work will consider whether or not the learning process needs to be enriched further, taking into account other clinically relevant factors. It will also address the key challenge of assuring that the system is acceptably safe to use, prior to deployment, and how safety can continue to be improved by refining the treatment strategies in use.

The code for learning the modified policy and for producing the figures shown in the paper is available at: https://github.com/Yanjiayork/sepsisRL.

ACKNOWLEDGEMENT

This work is funded by Bradford Teaching Hospitals NHS Foundation Trust and supported by the Assuring Autonomy International Programme at the University of York. The views expressed in this paper are those of the authors and not necessarily those of the NHS, or the Department of Health and Social Care.

REFERENCES

[1] M. Singer et al.,"The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)", JAMA, vol. 315, no. 8, p. 801, 2016.

[2] J. Gallagher," 'Alarming' one in five deaths due to sepsis", BBC News, 2020. [Online]. Available: https://www.bbc.co.uk/news/health-51138859. [Accessed: 13- Feb- 2020].

[3] C. Davies, "NHS says sepsis monitoring system has saved hundreds of lives", the Guardian, 2019. [Online]. Available: https://www.theguardian.com/society/2019/aug/18/nhs-says-sepsis-monitoring-system-has-saved-hundreds-of-lives. [Accessed: 13- Feb- 2020].

[4] J. Waechter et al., "Interaction Between Fluids and Vasoactive Agents on Mortality in Septic Shock", Critical Care Medicine, vol. 42, no. 10, pp. 2158-2168, 2014.

[5] A. Rhodes et al."Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock: 2016." Critical Care Medicine, vol. 45, no. 3, pp. 486–552, 2017.

[6] M. Komorowski, L. Celi, O. Badawi, A. Gordon and A. Faisal, "The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care", Nature Medicine, vol. 24, no. 11, pp. 1716-1720, 2018.

[7] R. Sutton, and A. Barto. Reinforcement learning: An introduction. MIT press, 2018.

[8] K. Fadale, D. Tucker, J. Dungan and V. Sabol, "Improving Nurses' Vasopressor Titration Skills and Self-Efficacy via Simulation-Based Learning", Clinical Simulation in Nursing, vol. 10, no. 6, pp. e291-e299, 2014.

[9] "Noradrenaline (Norepinephrine) 1 mg/ml Concentrate for Solution for Infusion - Summary of Product Characteristics (SmPC) - (emc)", Medicines.org.uk. [Online]. Available: https://www.medicines.org.uk/emc/product/4115/smpc. [Accessed: 13- Feb- 2020].

[10] J. M. Allen, "Understanding vasoactive medications: focus on pharmacology and effective titration." Journal of Infusion Nursing, vol. 37, no. 2, pp. 82-86, 2014.

[11] H. Beloeil, J. Mazoit, D. Benhamou and J. Duranteau, "Norepinephrine kinetics and dynamics in septic shock and trauma patients", British Journal of Anaesthesia, vol. 95, no. 6, pp. 782-788, 2005.

[12] R. Hahn and G. Lyons, "The half-life of infusion fluids", European Journal of Anaesthesiology, vol. 33, no. 7, pp. 475-482, 2016.

[13] A. Raghu, M. Komorowski, I. Ahmed, L. Celi, P. Szolovits, and M. Ghassemi. "Deep reinforcement learning for sepsis treatment." arXiv preprint arXiv:1711.09602 (2017).

[14] A. Johnson et al., "MIMIC-III, a freely accessible critical care database", Scientific Data, vol. 3, no. 1, 2016.

[15] N. Jong, and P. Stone, "Model-based function approximation in reinforcement learning" In Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems, pp. 1-8. 2007.

[16] C. J. C. H. Watkins,"Learning from delayed rewards", Ph.D thesis, University of Cambridge, 1989.

[17] V. Mnih et al.,"Human-level control through deep reinforcement learning", Nature, vol. 518, no. 7540, pp. 529-533, 2015.

[18] K. Jeon, J. Song, C. Chung, J. Yang and G. Suh,"Incidence of hypotension according to the discontinuation order of vasopressors in the management of septic shock: a prospective randomized trial (DOVSS)", Critical Care, vol. 22, no. 1, 2018.

[19] G. Polat, R. Ugan, E. Cadirci and Z. Halici, "Sepsis and Septic Shock: Current Treatment Strategies and New Approaches", The Eurasian Journal of Medicine, vol. 49, no. 1, pp. 53-58, 2017.

[20] E. Bassi, M. Park, and L. Azevedo. "Therapeutic strategies for high-dose vasopressor-dependent shock." Critical care research and practice, 2013.

[21] L. Breiman, "Random Forests", Machine Learning, 45(1), 5-32, 2001. https://doi.org/10.1023/A:1010933404324

[22] N. Jiang, and L. Li. "Doubly robust off-policy value evaluation for reinforcement learning." arXiv preprint arXiv:1511.03722 (2015).

[23] I. Habli, S. White, M. Sujan, S. Harrison, and M. Ugarte, "What is the safety case for health IT? A study of assurance practices in England", Safety Science 110, pp.324-335, 2018.

[24] J.A. McDermid, Y. Jia, and I. Habli. "Towards a Framework for Safety Assurance of Autonomous Systems." In Artificial Intelligence Safety 2019, pp. 1-7. CEUR Workshop Proceedings, 2019.

[25] University of York, "Assuring Autonomy International Programme, University of York". [Online]. Available: https://www.york.ac.uk/assuring-autonomy/. [Accessed: 26- Feb- 2020].