A RISK-BASED REINFORCEMENT LEARNING ALGORITHM TO IMPROVE ACUTE RENAL REPLACEMENT THERAPY

Brian McLaverty[1], Annabelle Lint[1], Gilles Clermont[1,2], Robert S. Parker[1,2*]
[1]Department of Chemical and Petroleum Engineering, University of Pittsburgh
[2]Department of Critical Care Medicine, University of Pittsburgh
Pittsburgh PA, 15260

Abstract

Kidney failure patients in the intensive care unit (ICU) require acute renal replacement therapy (ARRT). Hypotension commonly occurs during hemodialysis, a common ARRT, and is associated with premature termination of therapy and increased mortality. There is a critical need to develop personalized treatment support for hemodialysis that can avert hemodynamic instability and achieve individualized fluid removal goals. In this paper, a dynamic risk forecasting model was trained and validated on a population receiving hemodialysis at UPMC. The dynamic risk model demonstrated distinct differences in absolute risk and escalation in risk between hypotensive and normotensive hemodialysis sessions. A decision tree analysis derived a small set of risk states from the dynamic risk scores and predictive clinical variables. Q-learning was used to learn the optimal treatment policy for the finite state-action space, utilizing personalized risk as intermittent feedback to the agent for its actions. The learned policy could provide individualized treatment guidance to caretakers during hemodialysis and is a critical step towards improving renal replacement therapy outcomes.

Keywords

Reinforcement learning, Q-learning, machine learning, risk forecasting

Introduction

Intradialytic hypotension (IDH) occurs in 4%-30% of hemodialysis sessions (Kuipers et al., 2019). IDH is an independent predictor of mortality (Silversides et al., 2014), so it is important that clinicians can identify patients that are at high risk of hypotension and when hypotension will occur. Machine learning (ML) algorithms have been studied in several clinical settings involving early prediction of IDH. One such ML-derived early warning system for elective noncardiac patients under general anesthesia (Wijmberge et al., 2020) significantly decreased the incidence of hypotension during surgery. A different study (Yoon et al., 2020) built a hypotension forecasting model on ICU patients using a random forest algorithm. This risk model was coupled with a practical alert system to predict hypotension an hour before it occurs. The model predicted 80% of hypotensive events 15 minutes prior to occurrence and 60% of hypotensive events 60 minutes prior to occurrence (Yoon et al., 2020). Despite these advances in alert models, there is still a need for a treatment model that intelligently learns and suggests optimal treatment to clinicians to avoid IDH and other downstream consequences.

Reinforcement learning (RL) is one possible solution to fill this technology gap. A RL algorithm uses an agent to learn an optimal policy that maximizes process rewards. The application of reinforcement learning in a healthcare context has the advantage of not requiring a model to learn

* To whom all correspondence should be addressed: rparker@pitt.edu

an optimal treatment policy. RL algorithms have been applied to several healthcare situations, including sepsis (Komorowski et al., 2018), cancer (Padmanabhan et al., 2017), and anesthesia control (Padmanabhan et al., 2015), and have been successful in learning generalizable policies by optimizing a mixture of intermediate and terminal rewards. However, there are no existing reinforcement learning algorithms that have been developed specifically for dialysis or ARRT. Additionally, in modern RL-derived therapy and in clinical practice, the agent and clinician react to overt signs of patient instability. For example, during hemodialysis, clinicians take actions when a patient exhibits large changes in blood pressure from their individual baseline. We propose that an RL algorithm based on the projected risk of a future adverse clinical event could provide optimal treatment recommendations that improve patient outcomes by acting preemptively based on predicted patient instability.

In this paper, we develop an RL-based algorithm that recommends preemptive, personalized treatment for hemodialysis patients based on the patient's risk of future hypotension. First, a hypotension risk forecasting model was developed that was clinically interpretable and specifically applied to a hemodialysis population. Then, the risk model was used to construct a discrete state space that was coupled with a set of common clinical interventions to form a small, finite state-action space. The RL agent used risk-rewards as guidance to learn the optimal policy offline from a patient dataset. Thus, the agent learns to react to intradialytic hypotension preemptively, which could potentially lead to improved outcomes in the dialysis population.

Materials and Methods

Study Population and Collected Dataset

A cohort of 277 patients that underwent hemodialysis at UPMC were selected for analysis. 140 of these patients experienced dialysis hypotension, clinically defined as systolic blood pressure (SBP) < 90 mmHg and mean arterial blood pressure (MAP) < 65 mmHg, which was an independent predictor of mortality ($p < .05$, chi-squared test). Vitals, labs, and demographic information available for each hemodialysis session (n=1685) were collected from electronic health records (EHR). A dataset that included patient vitals, labs, underlying disease history, and indicators of prior hypotension was constructed and used as input to the risk prediction model. Additional vital signal features, such as minimum, maximum, slope, and linear weighted moving average (LWMA), were derived using the past 30 minutes of vitals time series data and appended to the constructed dataset. Sample features (n=89) were obtained or computed each minute; when data availability was limited, the last observation carried forward (LOCF) approach was used. Samples of the constructed dataset thus described candidate static and dynamic features available every minute for risk prediction.

Risk Model Training and Evaluation

A risk prediction model was developed by training and testing a random forest model using n=253 hypotensive hemodialysis sessions (HS) and n=1432 nonhypotensive hemodialysis sessions (NHS). Featurized data samples, available at each minute of hemodialysis, were randomly selected from NHS and at the time of hypotension from HS to train and test the model. The selected samples were split into a train (n=1128, 67%) and test set (n=557, 33%). Model parameters of the random forest were tuned by repeating stratified k-fold cross-validation (k=3) on the train dataset with a set of possible parameter values. Specifically, the model was trained using k-1 folds, and its prediction performance was evaluated on the remaining (validation) fold. Training and evaluation were repeated k times until each fold was used for validation. Tuned random-forest model parameters included: maximum tree depth (n=3), maximum features considered for node splitting (n=17), and number of trees in the forest (n=100). The model that maximized the average area under the receiver operating characteristic (ROC) curve across the validation folds was selected for further evaluation on the test dataset.

The trained random forest model was applied to minute-to-minute featurized raw data for hemodialysis sessions in the test dataset (n=557), providing dynamic absolute risk (probability) of hypotension. To further evaluate the clinical validity of the risk model, minute-to-minute risk trajectories in the timespan leading up to hypotension were produced for HS (n=84) and NHS (n=473) from the test dataset. The hypothetical hypotension onset time for NHS was chosen such that the distribution of hypotension onset times, relative to the start of hemodialysis, matched the distribution of real hypotension onset times from HS. Additional risk trajectories were produced and analyzed, beginning 30 minutes before the start of hemodialysis until the time of hypotension.

A Markov Decision Process Model of Hemodialysis

The hemodialysis process was modeled as a finite Markov decision process defined by a set of states, S, a set of actions, $A(s_k)$, a set of rewards R, and a probability transition function $P(s_{k+1}|s_k,a_k)$, which represents the probability of transitioning from state $s_k \in S$ to state s_{k+1} when action $a_k \in A$ is taken (Sutton and Barto, 2018). At each timestep k during hemodialysis, the agent (physician) takes an action a_k, and the patient transitions to a new state s_{k+1}, which is observed by the agent 15 minutes later. The agent is then rewarded $r_{k+1} \in R$ if taking action a_k in state s_k results in decreased personalized risk of hypotension.

MDP States (S)

The states of the MDP were derived using a data-driven approach, using the patients' physiological features and their projected absolute risk (AR) calculated by the random forest model. Input features (n=89) and the projected risk from the random forest model were extracted from each hemodialysis session every 15 minutes up to, but not including, the time of intradialytic hypotension. The

decision trees (n=100) of the random forest recursively partitioned subsamples of the dataset, using entropy reduction to measure the quality of the split by a feature at each decision tree node. The decrease in entropy resulting from partitioning the dataset with a given feature was calculated for each decision tree, and then averaged across the decision trees in the random forest. 8 of the 89 features with the highest average reduction in entropy were selected to be used for state space development. Features chosen include: current-time SBP, MAP, and diastolic blood pressure (DBP); minimum and past 30 minute LWMA of systolic blood pressure (SBP) and mean arterial pressure (MAP) measurements; and the slope of SBP over the past 30 minutes. Two additional risk measures, relative risk (*RR*) and personalized risk (*PR*), were calculated from the absolute risk projected from the model. *RR* describes hypotension risk relative to the average risk of NHS in the cohort (Eq. (1)).

$$RR = \frac{AR}{Average\ AR\ in\ NHS} \tag{1}$$

Personalized risk (*PR*) describes deviation in *RR* from patient's own *RR* at the start of the hemodialysis session (RR_0) (Eq. (2)).

$$PR = RR - RR_0 \tag{2}$$

The selected physiological features (n=8), *PR*, and *RR* were used to develop a finite state space. Principal component analysis was used to transform the original, correlated feature space to 4 principal components that describe >90% of data variance (Figure 1).

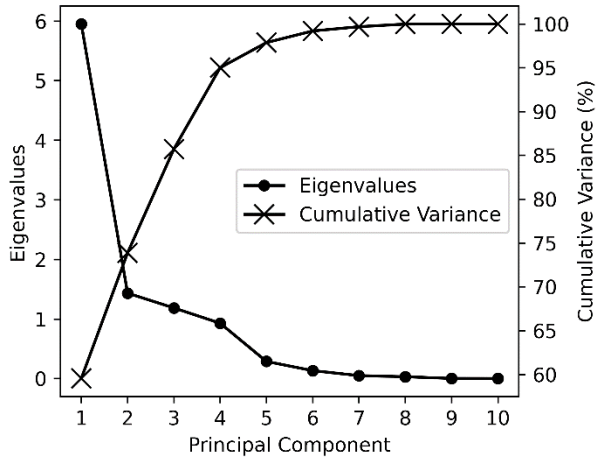

Figure 1. Eigenvalues and cumulative variance explained by the principal components.

Decision tree analysis, using entropy minimization as the splitting criterion, was applied to the reduced dimensionality dataset to separate high risk feature space, in which hypotension was experienced within 15 minutes, from lower risk space. The decision tree was applied to the dataset (max depth=3) and then pruned from the bottom up, removing split points that resulted in the smallest reduction in entropy. The remaining 6 leaves of the tree were used to define the risk states, which a hemodialysis patient could be assigned to in real time. Two additional terminal states were added to the state space, corresponding to hypotension and no hypotension.

MDP Actions (A)

The actions in set *A* were selected from clinician interventions during hemodialysis that would alter patient risk of hypotension. These included: i) increasing, decreasing, or no change in ultrafiltration rate, ii) administering mannitol or albumin bolus, and iii) initiating or increasing the rate of delivery of vasopressors. These actions were calculated every 15 minutes from start of dialysis. An increase or decrease in ultrafiltration or vasopressors was defined as at least a 25% increase or decrease in flow rate.

MDP Rewards (R)

The agent is rewarded at timestep k if a patient's personalized risk of hypotension decreased after taking action a_k on a patient in state s_k (Eq. (3)).

$$r_{k+1} = -RR_0[|PR_{k+1}| - |PR_k|] \tag{3}$$

In addition, the agent received a large penalty if an action a_k in s_k resulted in an immediate transition into a hypotensive terminal state.

Solving for the Optimal Policy

Reinforcement learning provided a solution to the MDP, where the agent was tasked to solve the MDP with the goal of maximizing expected cumulative discounted reward over time using suboptimal, stochastic transitions from real dialysis trajectories.

Q-learning is an algorithm that solves the Bellman optimality equation (Eq. (4-5)) without a model (Watkins and Dayan, 1992).

$$Q_o = r_{k+1} + \gamma \max_{a_{k+1}} Q^*(s_{k+1}, a_{k+1}) \tag{4}$$

$$Q^*(s,a) = \mathbb{E}[Q_o | s = s_k, a = a_k] \tag{5}$$

$Q^*(s_k,a_k)$ is the expected cumulative discounted reward for taking action a_k in state s_k and following the optimal policy afterwards. Specifically, real transitions were sampled from a fixed dataset $D = \{(s_k, a_k, s_{k+1}, r_{k+1})\}$ and the algorithm was updated (Eq. (6-8)) using current (Q_c) and new (Q_n) state-action value information (Watkins and Dayan, 1992; Lange et al., 2020; Padmanabhan et al., 2015).

$$Q_c = Q_{k-1}(s_k, a_k) \tag{6}$$

$$Q_n = r_{k+1} + \gamma \max_{a_{k+1}} Q_{k-1}(s_{k+1}, a_{k+1}) \tag{7}$$

$$Q_k(s_k, a_k) \leftarrow Q_c + \alpha[Q_n - Q_c] \quad (8)$$

The learning rate $\alpha \in [0,1]$ is the step size and the discount factor $\gamma \in [0,1]$ represents the present value of future rewards and acts as the agent's horizon. The discount factor was set to $\gamma = 0.10$ to reflect a short horizon and urgency to decrease personalized risk, and the learning rate was decayed according to a linear learning rate schedule (Even-Dar and Monsour, 2003). Theoretically, Q-learning converges to optimal Q^* if all state-action pairs are visited infinitely often and the learning rate is decayed appropriately. Q was initialized to zero and iteratively updated until $\Delta Q < 10^{-7}$ between successive iterations. After training, the optimal policy $\pi^*(s)$ for a given patient state was extracted from Q^* using Eq. (9):

$$\pi^*(s_k) = arg\max_{a_k} Q^*(s_k, a_k) \quad (9)$$

Results

Risk Trajectories

Minute-to-minute risk trajectories were generated by applying the trained random forest model on the hemodialysis sessions in the test dataset (n=557). Figure 2 depicts the evolution of relative risk for HS and NHS beginning two hours ahead of hypotension onset. There is distinct separation in average *RR* ($p<.05$, two-sample t-test) between the HS and NHS 2 hours ahead of event, where the average *RR* of HS is 4 times higher than NHS. This risk separation between the two groups increases leading up to the time of hypotension. On average, the relative risk of HS elevates beginning around 90 minutes prior to hypotension and escalates again around 30 minutes prior to hypotension. The risk of NHS remains low for the 2-hour period leading up to hypothetical hypotension onset.

Risk trajectories were generated beginning 30 minutes prior to hemodialysis initiation for HS and NHS in Figure 3. There is distinct separation in average risk ($p<.05$, two-sample t-test) between the HS and NHS throughout the observed period. HS and NHS experience distinct risk evolution from hemodialysis baseline. For HS, risk evolves from baseline and elevates over the observed period as continuous fluid removal induces hemodynamic instability on the average HS. Interestingly, NHS experience a slight elevation in risk at the start of hemodialysis; however, risk remains relatively stable afterward. This reflects the physiological stress induced on all patients upon ultrafiltration initiation. The risk trajectories in Figure 2 and Figure 3 are clinically relevant descriptions of the evolving risk of the average hemodialysis patient with and without impending hypotension. Hence, there is a distinct difference between the model generated risk trajectories for HS and NHS. These risk scores can be incorporated into an RL algorithm that could support clinician decision-making to avoid hypotension.

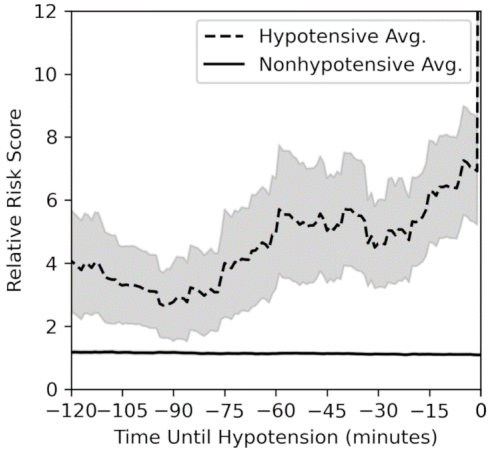

Figure 2. Mean RR trajectory before event shown for n=84 HS (dashed line) and n=473 NHS (solid line). Shaded gray: 95% confidence intervals.

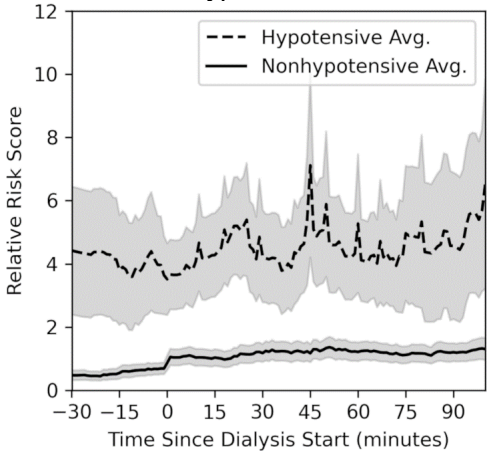

Figure 3. Mean RR trajectory beginning 30 minutes prior to start of dialysis shown for n=84 HS (dashed line) and n=473 NHS (solid line). Shaded gray: 95% confidence intervals.

Clinician vs. Optimal Policy

The actual clinician interventions at each timestep and across all hemodialysis sessions were collected. The optimal policy learned from the RL agent (Eq. (6)) was extracted from Q upon convergence and applied to patient risk state assignment at each timestep across all hemodialysis sessions. The clinician and RL policy were compared with respect to hemodialysis session outcome.

As depicted in Figure 4(a), clinicians most frequently made no change in ultrafiltration rate, followed by increasing and decreasing filtration rate in HS. Administration of albumin or mannitol bolus and initiation or escalation of vasopressor dose were infrequent interventions taken during hemodialysis

Conversely, Figure 4(b) demonstrates that the RL agent most frequently suggested to decrease ultrafiltration rate, followed by initiation or escalation of vasopressor dose and no change to ultrafiltration rate. This suggests that the RL agent was capable of learning that corrective clinical actions such as decreasing ultrafiltration rate and escalation of vasopressor dosage were beneficial interventions to decrease risk and avert hypotension.

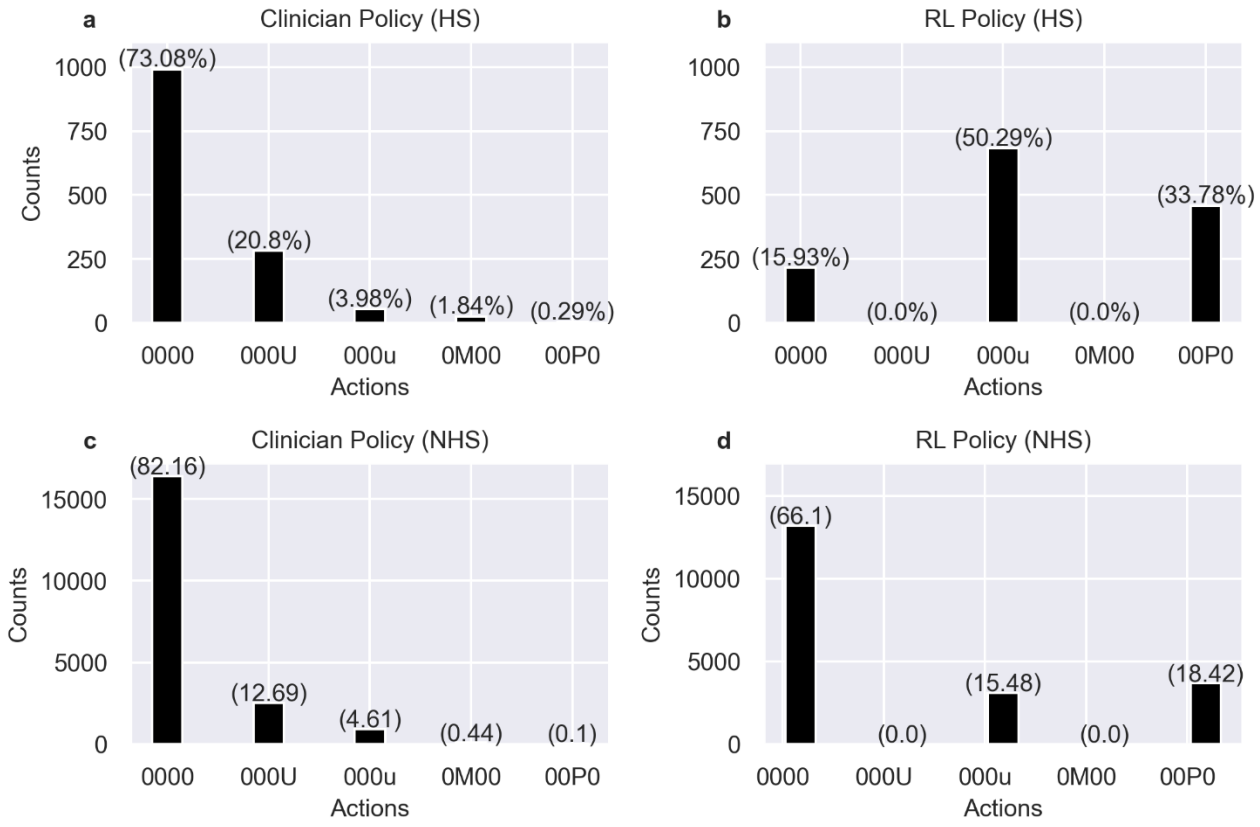

Figure 4. Clinician policy relative to hypotensive dialysis outcome (a). RL policy relative to hypotensive dialysis outcome (b). Clinician policy relative to nonhypotensive dialysis outcome (c). RL policy relative to nonhypotensive dialysis outcome (d). Action 0000: no change in ultrafiltration, action 000U: increase in ultrafiltration, action 000u: decrease in ultrafiltration, action 0M00: mannitol or albumin bolus, 00P0: initiation or increase in vasopressor dosage.

The clinician and RL policy with respect to NHS are shown in Figure 4(c) and Figure 4(d). The clinician policy for HS (Figure 4(c)) and NHS (Figure 4(a)) are similar in that clinicians most frequently made no change in ultrafiltration rate, followed by increasing ultrafiltration rate and decreasing ultrafiltration rate. Interestingly, clinicians increased ultrafiltration more frequently in HS versus NHS. The RL policy suggested corrective actions such as decrease ultrafiltration and increasing vasopressor dosage less frequently in NHS than in HS. These results support that the RL agent is capable of learning personalized, preemptive therapy using hypotension risk as feedback during the learning process.

Nevertheless, the RL agent suggested administration or increasing vasopressor dosage more frequently than clinicians during hemodialysis sessions (Figure 4). This intervention is only taken on hemodialysis patients in the intensive care unit (ICU) setting and is typically the final measure taken to prevent hypotension. Therefore, the agent-recommended interventions are aggressive and unlikely to be applied as frequently by a clinician. The optimal policy derived is deterministic and returns the single best action for a state. In cases where multiple actions may have similar rewards, this availability of similarly-valued options is not considered by the RL algorithm and may explain some of the differences between clinician and RL-recommended actions. In addition, the characteristics of the state must accurately reflect clinical reasoning for taking an intervention. To better reconcile RL and clinical actions may require extending the state to include history of clinical actions or underlying disease mechanism, such that recommended optimal actions are more consistent with clinician expectation. If a state has a rare, aggressive intervention and a less aggressive intervention with similarly high expected value, the less aggressive intervention could be chosen with greater frequency according to a user-defined probability distribution (Nanayakkara et al., 2022). Finally, although prevention of hypotension is a primary goal of hemodialysis, the ability to reach fluid and clearance goals needs to be represented in the state and rewards of the MDP.

Clinician vs. RL Treatment Outcomes

Microsimulations of the MDP using the RL-generated optimal policy were produced to evaluate its efficacy in reducing incidence of hypotension during dialysis. Specifically, 25,000 dialysis trajectories were produced *in silico* by initializing a set of risk states according to their distribution observed in the dataset. Then, state transitions were generated by applying optimal policy in Eq. 9 and producing subsequent states according to the probability transition function that was reconstructed from dataset

transitions. Each dialysis simulation was run until hypotension occurred or 3.5 hours of treatment time was reached. The RL-generated optimal policy resulted in a decrease in occurrence of intradialytic hypotension from 15.0% to 9.2% *in silico*. These results support that the agent-derived policy could potentially lead to improved patient outcomes at the expense of increased use of vasopressors.

Summary

A reinforcement learning agent was tasked to learn an optimal treatment policy to avoid hypotension in patients receiving hemodialysis. A hypotension risk model was trained and tested on a dialysis patient cohort from UPMC and produced risk score trajectories that captured clinically relevant evolution of intradialytic hypotension risk. The risk scores from the model were used both as feedback to the agent and to define the finite state space of the MDP. The agent learned an optimal policy that was clinically interpretable, and the suggested policy highlights the possible benefits of agent-recommended treatment. Simulation of dialysis treatments using the agent-suggested policy resulted in decreased incidence of hypotension *in silico*.

Acknowledgments

Financial support for this research was provided by the Department of Education Graduate Assistance in Areas of National Need fellowship program (DoEd GAAAN PP200A120195) as well as the National Institute of Health/ National Institute of Diabetes and Digestive and Kidney Diseases (R01-DK-131586).

References

Even-Dar, E., Mansour Y. (2003). Learning Rates for Q-Learning. *Journal o Machine Learning Research,* 5(1).

Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., & Faisal, A. A. (2018). The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine, 24*(11), 1716-1720.

Kuipers, J., Verboom, L. M., Ipema, K. J., Paans, W., Krijnen, W. P., Gaillard, C. A. J. M., Westerhuis, R., Franssen, C. F. M. (2019). The Prevalence of Intradialytic Hypotension in Patients on Conventional Hemodialysis: A Systematic Review with Meta-Analysis. *American Journal of Nephrology*, 49(6):497-506.

Lange S., Gabel T., Riedmiller, M. (2012). Batch Reinforcement Learning. (1992). *Springer*, 45-83.

Nanayakkara, T., Clermont, G., Langmead, C. J., & Swigon, D. (2022). Unifying cardiovascular modelling with deep reinforcement learning for uncertainty aware control of sepsis treatment. *PLOS Digital Health, 1*(2), e0000012.

Padmanabhan, R., Meskin, N., Haddad W. M. (2015). Closed-loop control of anesthesia and mean arterial pressure using reinforcement learning. *Biomedical Signal Processing and Control*, 22, 54-64.

Padmanabhan, R., Meskin, N., & Haddad, W. M. (2017). Reinforcement learning-based control of drug dosing for cancer chemotherapy treatment. *Mathematical Biosciences, 293*, 11-20.

Silversides, J. A., Pinto, R., Kuint, R., Wald, R., Hladunewich, M. A., Lapinsky, S. E., Adhikari, N. K. (2014). Fluid balance, intradialytic hypotension, and outcomes in critically ill patients undergoing renal replacement therapy: a cohort study. *Crit Care*, 18, 624.

Sutton, R. S., Barto, A. G. (2020). Reinforcement Learning: An Introduction. *MIT press*.

Watkins, C. J. C. H., Dayan, P. (1992). Q-learning. *Machine Learning*, 8, 279-292.

Wijmberge, M., Geerts, B. F., Hol, L., Lemmers, N., Mulder, M. P., Berge, P., Schenk, J., Terwindt, L. E., Hollmann, M. W., Vlaar, A. P., Veelo, D. P. (2020). Effects of a Machine Learning-Derived Early Warning System for Intraoperative Hypotension vs Standard Care on Depth and Duration of Intraoperative Hypotension During Elective Noncardiac Surgery: The HYPE Randomized Clinical Trial. *JAMA*, 323 (11): 1052-1060.

Yoon, J. H., Jeanselme, V., Dubrawski, A., Hravnak, M., Pinsky, M. R., Clermont, G. (2020). Prediction of hypotension events with physiological vital signatures in the intensive care unit. *Crit Care*. 24, 661.