Deep Conservative Reinforcement Learning for Personalization of Mechanical Ventilation Treatment

Flemming Kondrup[*]
McGill University
Montreal, Canada
flemming.kondrup@mail.mcgill.ca

Thomas Jiralerspong[*]
McGill University
Montreal, Canada
thomas.jiralerspong@mail.mcgill.ca

Elaine Lau[*]
McGill University
Montreal, Canada
tsoi.lau@mail.mcgill.ca

Nathan de Lara
McGill University
Montreal, Canada
nathan.delara@mail.mcgill.ca

Jacob Shkrob
McGill University
Montreal, Canada
jacob.shkrob@mail.mcgill.ca

My Duc Tran
McGill University
Montreal, Canada
my.d.tran@mail.mcgill.ca

Doina Precup
Mila, McGill University, DeepMind
Montreal, Canada
dprecup@cs.mcgill.ca

Sumana Basu
Mila, McGill University
Montreal, Canada
sumana.basu@mail.mcgill.ca

Abstract

Mechanical ventilation is a key form of life support for patients with pulmonary impairment. An important challenge faced by physicians is the difficulty of personalizing treatment and thus to offer the best ventilation settings for each patient. This leads to sub-optimal care which further leads to complications such as permanent lung injury, diaphragm dysfunction, pneumonia and potentially death. It is therefore essential to develop a decision support tool to optimize and personalize ventilation treatment.

We present DeepVent, the first deep reinforcement learning model to address ventilation settings optimization. Given a patient, DeepVent learns to predict the optimal values for the ventilator parameters Adjusted Tidal Volume (V_t), FiO_2 (Fraction of inspired O_2) and PEEP (Positive End-Expiratory Pressure) with the final objective of promoting 90 day survival. We use the MIMIC-III dataset, comprised of 19,780 patients under ventilation. We show that our use of Conservative Q-Learning addresses the challenge of overestimation of the values of out-of-distribution states/actions and that it leads to recommendations within safe ranges, as outlined in recent clinical trials. We evaluate our model using Fitted Q Evaluation, and show that it is predicted to outperform physicians. Furthermore, we design a clinically relevant intermediate reward to address the challenge of sparse reward. Specifically, we employ the Apache II score, a widely used score by physicians to assess the severity of a patient's condition, and show that it leads to improved performance.

Keywords: Reinforcement Learning, Conservative Q-Learning, Mechanical Ventilation, Personalized Healthcare

We would like to thank Bogdan Mazoure (Mila) and Eyal de Lara (University of Toronto) for reviewing the paper and sharing constructive feedback, Adam Oberman (Mila) for their advice on the intermediate reward and Andrew Bogecho (McGill) for helping us getting access to the McGill RL Lab compute.

[*] These authors contributed equally

1 Introduction

The COVID-19 pandemic has put enormous pressure on the healthcare system, particularly on intensive care units (ICUs). In cases of severe pulmonary impairment, mechanical ventilation assists breathing in patients and acts as the key form of life support. However, the optimal ventilator settings is individual specific and often unknown [1], leading to ventilator induced lung injury (VILI), diaphragm dysfunction, pneumonia and oxygen toxicity [2]. Previous work has approached ventilation optimization with RL using a tabular approach [3]. Our work makes three key contributions:

- We propose the first deep reinforcement learning approach to personalize mechanical ventilation settings
- We demonstrate the potential of Conservative Q-Learning [4], a recently proposed offline deep reinforcement learning algorithm, to address overestimation of the values of out-of-distribution states/actions, which is very important in a healthcare context, where data is limited and risk in decision making must be avoided
- We introduce an intermediate reward based on the Apache II mortality prediction score [5] to address the challenge of sparse reward

We compare DeepVent's decisions to those of physicians, as recorded in an existing standard dataset, as well as to those of an agent trained with Double Deep Q-Learning (DDQN) [6]. DeepVent is predicted to outperform physicians while avoiding the overestimation problems of DDQN, thus making safe recommendations.

2 Preliminaries

2.1 Double Deep Q-Networks (DDQN)

Overestimation occurs when the estimated value of a random variable is higher than its true value. Deep Q-Networks are known to overestimate the values of unseen state-action pairs. DDQNs were introduced as a solution by modifying the calculation of the target value [6]. At any point in time, one of DDQN's networks, chosen at random, is updated, by using as target the estimate from the other network. Although this partially solves overestimation, DDQNs can still suffer from it [6], particularly in offline RL where exploration is limited to the dataset used in training. This can lead to important overestimation in state-action pairs underrepresented in the dataset, or out-of-distribution (OOD), leading to sub-optimal action choices [4] which may translate to unsafe recommendations, putting patients at risk.

2.2 Conservative Q-Learning (CQL)

To address the challenge of overestimation in an offline setting, Conservative Q-Learning (CQL) was proposed [4]. It learns a conservative estimate on the Q-function by incorporating a regularizer $E_{\mathbf{s_t} \sim \mathcal{D}, \mathbf{a_t} \sim A}[Q(\mathbf{s_t}, \mathbf{a_t})]$ on top of the standard TD error to minimize the overestimated Q-values of unseen actions. In addition, the term $-E_{\mathbf{s_t}, \mathbf{a_t} \sim \mathcal{D}}[Q(\mathbf{s_t}, \mathbf{a_t})]$ is added to maximize the Q-values in the dataset, providing a lower bound in expectation of the policy. CQL minimizes the estimated Q-values for all actions while simultaneously maximizing the estimated Q-values for the actions appearing in the dataset. This prevents overestimation of OOD state-action pairs which are underrepresented in the dataset.

3 Datasets

We used the MIMIC-III database [7] containing data of 61,532 ICU stays at the Beth Israel Deaconess Medical Center. Standardized Query Language (SQL) was used to extract data for a total of 19,780 patients under ventilation. For features with $<$ 30% of the data was missing, KNN imputation was used [8]. If 30% to 95% of the data was missing, time-windowed sample-and-hold was used [8]. If $>$ 95% was missing, the variable was removed. An out-of-distribution (OOD) set of outlier patients was created with patients having at least one feature in the top or bottom 1% of the distribution at the start of their ICU stay.

4 Proposed Approach

Our MDP was defined similarly to [3], with episodes lasting from the time of the patient's intubation to 72 hours after.

State Space The state space $\mathcal{S}$ is composed of 37 variables[1]:

- Demographics: Age, gender, weight, readmission to the ICU, Elixhauser score
- Vital Signs: SOFA, SIRS, GCS, heart rate, sysBP, diaBP, meanBP, shock index, respiratory rate, temperature, spO2

[1] For variable definitions refer to the MIMIC-III paper [7]

- Lab Values: Potassium, sodium, chloride, glucose, bun, creatinine, magnesium, carbon dioxide, Hb, WBC count, platelet count, ptt, pt, inr, pH, partial pressure of carbon dioxide, base excess, bicarbonate
- Fluids: Urine output, vasopressors, intravenous fluids, cumulative fluid balance

Action Space The 3 ventilator settings of interest are:

- Adjusted tidal volume *or* Vt (Volume of air in and out of the lungs with each breath adjusted by ideal weight)
- PEEP (Positive End Expiratory Pressure)
- FiO2 (Fraction of inspired oxygen)

The action space $\mathcal{A}$ is the Cartesian product of the set of these three settings. Each setting can take one of seven values corresponding to ranges. We can therefore represent an action as the tuple $a = (v, o, p)$ with $v \in Vt, o \in FiO_2, p \in PEEP$.

Reward Function The main objective of our agent is to keep a patient alive in the long-term. Therefore, even if DeepVent only treats patients for 72 hours, it learns how to maximize their 90 day survival. We thus define a terminal reward $r(s_t, a_t, s_{t+1})$, which takes at the final state the value -1 if the patient passes away within 90 days and $+1$ otherwise.

The sole use of a sparse terminal reward is known to cause poor performance [9] in RL tasks. We therefore developed an intermediate reward based on the Apache II score [5], a widely used score in ICUs to assess the severity of a patient's disease. Apache II takes various physiological variables such as temperature, blood pressure etc. as input and returns a score based on how far each variable is from the healthy range. A higher score is associated with variables being far from the normal range, and thus a more severe state. The score was adapted to the variables present in MIMIC-III. In order to not simply define reward based on how well a patient was doing but rather their evolution through time, our intermediate reward consists of the change in Apache II score between s_{t+1} and s_t, which is normalized by dividing it by the total range of the score. Combining our intermediate and terminal rewards, we obtain our final reward function:

$$r(s_t^i, a_t^i, s_{t+1}^i) = \begin{cases} +1 & \text{if } t+1 = l_i \text{ and } m_{t+1}^i = 1 \\ -1 & \text{if } t+1 = l_i \text{ and } m_{t+1}^i = 0 \\ \frac{(A_{t+1}^i - A_t^i)}{\max_A - \min_A} & \text{otherwise} \end{cases}$$

where:
$\mathbf{A}_t^i$ is the modified Apache II score of patient i at t
$\mathbf{m}_t^i$ = 0 if patient i is dead at 90 days and 1 otherwise
$\mathbf{l}_i$ is the length of patient i's stay at the ICU
$\mathbf{max}_A, \mathbf{min}_A$ are the max. and min. values of our modified Apache II score

Off-Policy Evaluation (OPE) The performance of various OPE methods was recently evaluated in healthcare [10], and Fitted Q Evaluation (FQE) was found to consistently provide the most accurate results. We thus use FQE [11], which takes as input a dataset D and a policy π, and outputs a value estimate for each state in D, corresponding to an approximation of the cumulative discounted reward received by following a policy π starting at a given state. It is important to note that FQE outputs an approximation of true performance and not its exact value. Clinical trials would thus be required to confirm the results in section 5.1.

5 Results & Discussion

5.1 DeepVent Overall Performance

To begin, we compare the performance of DeepVent- (CQL without intermediate reward), DeepVent (CQL with intermediate reward), and the physician when applied to the patients in our test set (see Table 1).

Table 1: Mean initial state value estimates for physician, DeepVent- and DeepVent, with std. errors. DeepVent- significantly outperforms the physician. Adding the Apache II intermediate reward (DeepVent) further improves the estimate.

PHYSICIAN	DEEPVENT-	DEEPVENT
$\mathbf{0.502} \pm 0.00709$	$\mathbf{0.762} \pm 0.00402$	$\mathbf{0.797} \pm 0.00670$

The initial state of an episode represents the state of a given patient when ventilation is initiated. The performance of DeepVent- or DeepVent can be approximated by the value estimation output by FQE for the initial state of a patient. Although DeepVent was trained with intermediate rewards, FQE's value estimation only depends on the dataset $\mathcal{D}$ and the actions chosen by the policy π used to train FQE. Because we trained FQE using the dataset without intermediate rewards for both DeepVent- and DeepVent, the estimates are solely based on the terminal reward and can thus be used as a fair comparison between the two models. Since the physician policy effectively generates the episodes in our dataset, its value estimates for each initial state can be computed by taking the cumulative discounted reward for the episode starting at that state. We observe that DeepVent outperforms physicians by a factor of 1.52. The addition of the intermediate reward increases this factor to 1.59.

5.2 DeepVent and Safe Recommendations

We next evaluate DeepVent's action distributions and compare it with DDQN and the physician (see Figure 1).

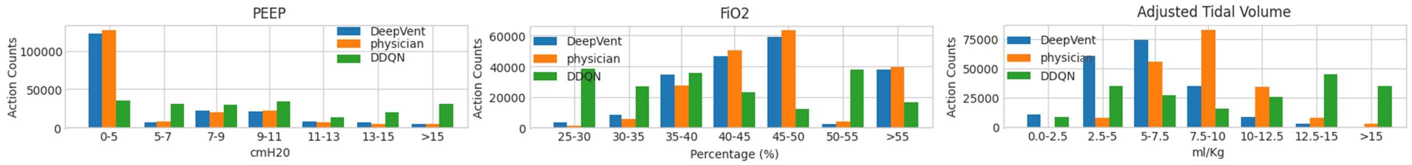

Figure 1: Distribution of actions across ventilator settings. Unlike DDQN, DeepVent makes recommendations in safe and clinically relevant ranges for each setting.

DeepVent was observed to suggest safe setting recommendations. The standard of care in terms of PEEP setting is commonly initiated at 5 cmH2O [12] which is supported by the high number of recommendations by physicians being in the range of 0-5 cmH2O in our dataset. DeepVent spontaneously chose to adopt this strategy by making most recommendations in the range of 0-5 cmH2O. In contrast, DDQN chose settings distributed along all the options, ranging up to 15 cmH2O, where physicians rarely went. High PEEP settings have been associated with higher incidence of pneumothorax [13], inflammation [14] and impaired hemodynamics [15], and should therefore be avoided.

In terms of FiO2 setting, DeepVent was once again found to follow clinical standards of care. More specifically, we observe that DeepVent often chose actions in the same ranges as physicians in our dataset, with many recommendations in the ranges of 35-50% and >55% and few recommendations below 35% and between 50-55%. In contrast, DDQN made few recommendations in ranges often suggested by physicians, and many in those that were rarely employed.

Finally, for the adjusted tidal volume, the optimal tidal volume is usually found in the 4-8 ml/kg range [16, 17]. DeepVent made a majority of recommendations in the range of 2.5-7.5 ml/kg, with an important amount of these being concentrated in the 5-7.5 ml/kg range. In contrast, DDQN made many recommendations in higher ranges, often even going above 15 ml/kg, a range rarely observed in clinical practice and associated with increased lung injury and mortality [18].

5.3 DeepVent in Out-Of-Distribution Samples

We next investigated whether the sub-optimal recommendations made by DDQN might be caused by value overestimation. To do so, we investigated the mean initial values for DeepVent and DDQN (as estimated by FQE). It is interesting to not only understand how well the model performs on data similar to that on which it was trained, but also on outlier data. We thus consider both an in-distribution (ID) and an out-of-distribution (OOD) setting (see Figure 2).

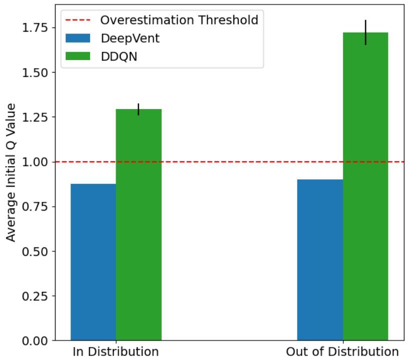

Figure 2: Mean initial Q-values for both in and out of distribution settings for DeepVent and DDQN (with variances - DeepVent's variance is not visible because of its small value). The horizontal line is the maximum expected return per episode. In contrast to DeepVent, DDQN clearly suffers from overestimation, which is aggravated in the OOD setting

Since the maximal expected return for an episode in our dataset is set at 1, values above this threshold should be considered as overestimated. We observe that DDQN overestimates policy values in both the ID and OOD settings. In addition, DDQN's overestimation is exacerbated in the OOD setting. This failure to accurately assess these OOD states may be the cause of the unsafe recommendations discussed above. DeepVent seems to avoid these problems, as its average initial state value estimate stays below the overestimation threshold of 1 in both settings, and barely changes in OOD.

6 Conclusion

In this work, we developed DeepVent, a decision support tool for personalizing mechanical ventilation treatment using offline deep reinforcement learning. We showed that our use of Conservative Q-Learning leads to settings in clinically relevant and safe ranges by addressing the problem of overestimation of the values of out-of-distribution state-action pairs. Furthermore, we showed using FQE that DeepVent achieves a higher estimated performance when compared to physicians, which can be further improved by implementing our Apache II based intermediate reward. We conclude that DeepVent intuitively learns to pick actions that a physician would agree with, while using its capacity to overview vast amounts of clinical data at once and understand the long-term consequences of its actions to improve outcomes for patients. Moreover, the fact that DeepVent is associated with low overestimation in out-of-distribution settings makes it much more reliable, and thus closes the gap between research and real-world implementation. Future work should aim to investigate the potential of the DeepVent methodology in other healthcare applications.

References

[1] H. Zein, A. Baratloo, A. Negida, and S. Safari. Ventilator Weaning and Spontaneous Breathing Trials; an Educational Review. *Emerg (Tehran)*, 4(2):65–71, 2016.

[2] T. Pham, L. J. Brochard, and A. S. Slutsky. Mechanical Ventilation: State of the Art. *Mayo Clin Proc*, 92(9):1382–1400, 09 2017.

[3] A. Peine, A. Hallawa, J. Bickenbach, G. Dartmann, L. B. Fazlic, A. Schmeink, G. Ascheid, C. Thiemermann, A. Schuppert, R. Kindle, L. Celi, G. Marx, and L. Martin. Development and validation of a reinforcement learning algorithm to dynamically optimize mechanical ventilation in critical care. *NPJ Digit Med*, 4(1):32, Feb 2021.

[4] A. Kumar, A. Zhou, G. Tucker, and S. Levine. Conservative q-learning for offline reinforcement learning, 2020.

[5] W.A. Knaus, E.A. Draper, D.P. Wagner, and J.E. Zimmerman. APACHE II: a severity of disease classification system. *Crit Care Med.*, 1985.

[6] H. van Hasselt, A. Guez, and D. Silver. Deep reinforcement learning with double q-learning, 2015.

[7] A. E. W. Johnson, T. J. Pollard, L. Shen, L. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 2016.

[8] C.M. Salgado, C. Azevedo, H. Proença, and S.M. Vieira. Missing data. in: Secondary analysis of electronic health records. *Springer, Cham.*, 2016.

[9] Maja J. Mataric. Reward functions for accelerated learning. 1994.

[10] S. Tang and J. Wiens. Model selection for offline reinforcement learning: Practical considerations for healthcare settings, 2021.

[11] Michita Imai Takuma Seno. d3rlpy: An offline deep reinforcement library. In *NeurIPS 2021 Offline Reinforcement Learning Workshop*, December 2021.

[12] GF Nieman, J Satalin, P Andrews, H Aiash, NM Habashi, and LA Gatto. Personalizing mechanical ventilation according to physiologic parameters to stabilize alveoli and minimize ventilator induced lung injury (vili). *Intensive Care Med Exp.*, 2017.

[13] J Zhou, Z Lin, X Deng, B Liu, Y Zhang, Y Zheng, H Zheng, Y Wang, Y Lai, W Huang, and X Liu et al. Optimal Positive End Expiratory Pressure Levels in Ventilated Patients Without Acute Respiratory Distress Syndrome: A Bayesian Network Meta-Analysis and Systematic Review of Randomized Controlled Trials. *Front. Med.*, 2021.

[14] A. Güldner, A. Braune, L. Ball, P. L Silva, C. Samary, A. Insorsi, R. Huhle, I. Rentzsch, C. Becker, L. Oehme, M. Andreeff, M. F Vidal Melo, T. Winkler, and P. Pelosi et al. Comparative Effects of Volutrauma and Atelectrauma on Lung Inflammation in Experimental Acute Respiratory Distress Syndrome. *Critical care medicine*, 2016.

[15] S. N. PROVE Network Investigators for the Clinical Trial Network of the European Society of Anaesthesiology, Hemmes, M. Gama de Abreu, P. Pelosi, and M. J. Schultz. High versus low positive end-expiratory pressure during general anaesthesia for open abdominal surgery (PROVHILO trial): a multicentre randomised controlled trial. *Lancet (London, England)*, 2014.

[16] A.M. Luks. Ventilatory strategies and supportive care in acute respiratory distress syndrome. *Influenza and other respiratory viruses, 7 Suppl 3*, 2013.

[17] O. Kilickaya and O. Gajic. Initial ventilator settings for critically ill patients. *Critical care*, 2013.

[18] A. Serpa Neto, S. O. Cardoso, J. A. Manetta, V. G. Pereira, D. C. Espósito, M. Pasqualucci, M. C. Damasceno, and M. J. Schultz. Association between use of lung-protective ventilation with lower tidal volumes and clinical outcomes among patients without acute respiratory distress syndrome: a meta-analysis. *JAMA*, 2012.