RESEARCH ARTICLES

Open Access

Reinforcement learning for intensive care medicine: actionable clinical insights from novel approaches to reward shaping and off-policy model evaluation

Luca F. Roggeveen^{1,2}, Ali el Hassouni², Harm-Jan de Grooth¹, Armand R. J. Girbes¹, Mark Hoogendoorn² and Paul W. G. Elbers^{1*}[®] on behalf of the Dutch ICU Data Sharing Against COVID-19 Collaborators

Abstract

Background Reinforcement learning (RL) holds great promise for intensive care medicine given the abundant availability of data and frequent sequential decision-making. But despite the emergence of promising algorithms, RL driven bedside clinical decision support is still far from reality. Major challenges include trust and safety. To help address these issues, we introduce cross off-policy evaluation and policy restriction and show how detailed policy analysis may increase clinical interpretability. As an example, we apply these in the setting of RL to optimise ventilator settings in intubated covid-19 patients.

Methods With data from the Dutch ICU Data Warehouse and using an exhaustive hyperparameter grid search, we identified an optimal set of Dueling Double-Deep Q Network RL models. The state space comprised ventilator, medication, and clinical data. The action space focused on positive end-expiratory pressure (peep) and fraction of inspired oxygen (FiO2) concentration. We used gas exchange indices as interim rewards, and mortality and state duration as final rewards. We designed a novel evaluation method called cross off-policy evaluation (OPE) to assess the efficacy of models under varying weightings between the interim and terminal reward components. In addition, we implemented policy restriction to prevent potentially hazardous model actions. We introduce delta-Q to compare physician versus policy action quality and in-depth policy inspection using visualisations.

Results We created trajectories for 1118 intensive care unit (ICU) admissions and trained 69,120 models using 8 model architectures with 128 hyperparameter combinations. For each model, policy restrictions were applied. In the first evaluation step, 17,182/138,240 policies had good performance, but cross-OPE revealed suboptimal performance for 44% of those by varying the reward function used for evaluation. Clinical policy inspection facilitated assessment of action decisions for individual patients, including identification of action space regions that may benefit most from optimisation.

Conclusion Cross-OPE can serve as a robust evaluation framework for safe RL model implementation by identifying policies with good generalisability. Policy restriction helps prevent potentially unsafe model recommendations. Finally, the novel delta-Q metric can be used to operationalise RL models in clinical practice. Our findings offer a promising pathway towards application of RL in intensive care medicine and beyond.

*Correspondence: Paul W. G. Elbers p.elbers@amsterdamumc.nl

Full list of author information is available at the end of the article



© The Author(s) 2024. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Keywords Reinforcement learning, Explainability, Safety, Mechanical ventilation, Peep, fo2, Covid-19, Cross off-policy evaluation, Policy filtering, Clinical policy inspection

Introduction

The practice of intensive care medicine involves frequent and fast decision-making. To do so, intensive care professionals have access to large amounts of data as intensive care unit (ICU) patients are continuously monitored. Given the often sequential nature of the decision-making process, reinforcement learning (RL) [1] may be a particularly well-suited machine learning method to solve treatment challenges in the ICU. In RL, an agent interacts with an environment, and receives rewards or penalties based on the actions it takes. The agent's goal is to learn a policy that maximises the cumulative reward over time.

RL has enormous potential to improve critical care decision-making by providing a framework for learning from past experience and adapting to changing conditions. RL can be used to model the patient's condition and the potential outcomes of different treatment options, which can aid in the identification of high-risk patients, predict the risk of certain outcomes and find the best treatment pathway for a given patient. However, the majority of these models have been used to generate hypotheses and the use of RL models as bedside decision support aids is still a long way off. Open questions relate to model selection, evaluation metrics, and safe application in a clinical setting [2].

RL models are typically developed and evaluated using a single reward function. A reward function represents the goals of treatment and determines the agent's behaviour in the RL model. However, small changes in the reward function can have profound effects on the developed policies from these trained models [3]. Additionally, the reward signal may differ between patient populations or even individual patients due to operational differences in local guidelines as well as the availability of data used in the reward function. If a model uses a poorly defined reward function, it may generate policies that are not applicable to different patient groups or do not align with varying definitions of successful clinical outcomes.

Further, off-policy evaluation in RL is the process of evaluating the performance of a policy on historical actions with the environment rather than direct interaction of an RL algorithm with patients. Sparse reward signals, such as mortality, is a common challenge in off-policy evaluation as it limits the quality of the evaluation because there may be few or no rewards to base the evaluation on. This problem can be mitigated by adding a more frequently occurring signal to the reward such as incremental improvements of intermediary treatment goals such as restoration of organ function. This adds a new level of complexity as the different components of the reward function must now be weighed against each other. Additionally, if the action recommended by the model only occurs infrequently the quality of the off-policy evaluation degrades severely and can suffer from extreme bias and variance.

Lastly, RL algorithms can be difficult for healthcare providers to interpret and understand, reducing their acceptability. Current research into RL in the ICU mainly focuses on developing policies that are able to achieve better outcomes in terms of patient survival. However, RL can also be used as an analysis tool to evaluate the performance of past treatment strategies and gain insights into which strategies were most effective. This can be particularly useful for identifying patterns in patient care that are associated with better outcomes.

To address these challenges, particularly in regard to safe RL model deployment, we introduce the concept of cross off-policy evaluation. This is based on reward shaping, typically involves altering the reward function to guide the agent towards the desired outcome. Expanding on this, our approach evaluates models based on their performance under different weightings of reward components within a singular reward function. The premise being that a truly robust policy should exhibit its applicability across various reward structures. This ability to generalise is not just an indication of a model's robustness, but also a critical prerequisite for safe model deployment, ensuring that the model remains effective even when circumstances or priorities shift. We not only highlight the potential for this method to identify models that do not generalise effectively, but also showcase how it can be used to identify models that generalise well through the process of elimination. This provides a framework for model evaluation when the ideal reward is difficult to capture in a single number.

We also introduce a method of policy restriction that uses historical actions and the ordinal nature of the action space. This approach helps ensure that recommended actions align with previous treatment steps, thus promoting consistency in policy behaviour. In so doing, the strategy helps to mitigate the risk of unpredictable policy actions, enhancing the reliability of model deployment and contributing to a safer model operation.

Finally, we present different types of policy analysis and show how they can be applied to improve the interpretability from a clinical perspective. We illustrate our proposed solutions by applying them to the development and evaluation of RL models to optimise mechanical ventilation settings for critically ill patients with severe COVID-19 from the Dutch ICU Data Warehouse (DDW).

Methods

Clinical context

In this paper we focused on COVID-19 patients that required mechanical ventilation in the ICU to maintain adequate oxygenation and decarboxylation. Important ventilator parameters include respiratory rate, tidal volume, peak pressure, plateau pressure, positive endexpiratory pressure (PEEP) and fraction of inspired oxygen (FiO2). For most of these parameters, the ventilator mode determines whether they are controlled by the healthcare, controlled by the patient or only monitored. However, regardless of ventilator mode, PEEP and FiO2 are always controlled by healthcare professionals rather than only monitored. There is no consensus on optimal values for PEEP and FiO2 [4, 5]. If PEEP is too low, the lung may collapse causing decreased compliance and hypoxemia due to shunting. Overzealous application of PEEP may lead to reduced preload and hence reduced cardiac output, decreased compliance, increased deadspace ventilation causing hypercarbia and acidosis [6]. Similarly, if FiO2 is too low, hypoxia will ensue which will ultimately lead to organ failure and death. However, an FIO2 that is too high is associated with oxygen toxicity to the long and other organs [7]. Personalisation of PEEP and FiO2 may therefore be a valuable strategy. Some recent clinical trials in this direction have shown promising results [8, 9]. However, their approaches are labour intensive and require particular expertise in respiratory physiology as well as additional monitoring devices. Therefore, RL should be promising in this setting. To the best of our knowledge, no previous attempts have been made to use RL to optimise mechanical ventilation settings in COVID-19. There have been a limited number of attempts to use RL for optimising ventilator settings in patients with respiratory failure for reasons other than COVID-19. Prasad et al.'s application of Fitted Q-iteration focuses on optimising weaning protocols in ICUs, with the clinical goal of reducing reintubation occurrences and regulating patient physiological stability [10]. Peine et al. developed a RL-based model called VentAI to minimise 90-day mortality by optimising ventilatory settings, specifically tidal volume (TV), PEEP and FiO2, which are among the most commonly used settings on a ventilator under controlled ventilation [11]. Kondrup et al. proposed DeepVent, a Conservative Q-Learning algorithm (CQL), using a similar setup to Peine et al. and evaluated using Fitted Q Evaluation (FQE) [12]. Additionally, they introduced an intermediate reward function component based on the modified Apache II score.

Data extraction and preprocessing

The data for this study were sourced from the DDW, a database compiling information on critically ill COVID-19 patients from 25 ICUs in the Netherlands [13]. The DDW encompasses data on 3464 patients, covering two distinct periods of the pandemic in the Netherlands, often referred to as "wave 1" and "wave 2". These terms denote the first and the second major surge of COVID-19 cases, respectively, each of which saw a dramatic increase in infections. This database boasts more than 200 million individual clinical data points. At the time of this study, a snapshot of the DWWH, containing 3051 patients rather than the full 3464 were used for this experiment. The overall ICU mortality was 24.4%. Respiratory and haemodynamic parameters were among the most commonly recorded, including ventilation mode, prone position and ventilator settings. Medications administration and daily fluid balance were available for most patients. Lab records were widely available. Clinical features were derived using time aggregations applied at 4, 6, h, 24 h time intervals. A list of all available data parameters and features is available in Additional file 1: Appendix SA. Patient admissions were selected based on length of ICU stay, use of invasive mechanical ventilation, and data availability, for reasons described in Fig. 1.

Reinforcement learning problem definition

RL is a computational approach where a computer program, known as an 'agent', learns to make decisions by interacting with an environment, in this case, represented by ICU patient data. The goal of the agent is to maximise a 'reward', which in the ICU context, translates to optimal patient treatment outcomes. RL involves the agent interacting with the environment through 'states', which are snapshots of aggregated patient data over specific time intervals, and 'actions', the medical interventions or treatment decisions. A 'trajectory' pertains to the sequential events and decisions made throughout the entire course of a single patient's admission, encompassing the complete set of states and the corresponding actions executed during that individual's stay in the ICU.

Q-learning, an off-policy algorithm in RL, learns optimal decision strategies even from suboptimal actions. In the ICU, this means it can learn effective treatment strategies by analysing both optimal and non-optimal medical decisions made by healthcare professionals. This ability is crucial as it allows for learning from a wide range of historical real-world ICU scenarios without the need for experimental interventions on patients. In Q-learning, a



Fig. 1 Patient selection flowchart

function known as Q(s, a) estimates the utility of taking a particular action (a) in a given state (s), and then following the optimal strategy thereafter.

This section outlines the specific components of our model: the state space, action space, and reward structure, integral to our RL approach.

The state space defines the range of all possible conditions that an ICU patient can experience, as represented by a comprehensive combination of key clinical indicators and measurements. We used one-hour time steps and combined ventilator sensor data, medication records, clinical scores, and laboratory results into one state space. Missing values were imputed using the carryforward method, up to clinical cut-offs; see Additional file 1: Appendix SA for the full list of features.

The action space defines the set of all possible interventions and treatment adjustments that the RL model can select from, specifically centred around the settings of positive end-expiratory pressure (PEEP) and fraction of inspired oxygen (FiO2). Figure 2 shows the distribution of actions in the dataset. We used bins of PEEP and FiO2 based on clinical cut-offs, and specified one action for non-ventilation (NV) to allow for the entire ICU admission to be used as a trajectory for training RL models. Our decision to focus the action space on PEEP and FiO2, diverging from the broader dimensions of previous studies [peine '22] [Prassad '17], is strategic. It makes the RL model simpler, reducing the need for extensive training data and computational resources. Concentrating on these key ventilatory parameters allows for direct control, facilitating more straightforward and broadly applicable decision-making. It is worth noting that factors like tidal volume, ventilation mode, and others are incorporated into the state space, ensuring that the model still considers their influence while keeping the action space concise. This approach guarantees the model's adaptability to diverse ICU conditions while maintaining its capacity to provide clear, actionable guidance essential for effective patient care.

In the context of our RL approach, a "reward" serves as a quantitative measure of the quality of patient care, with multiple rewards provided throughout a single patient trajectory. In this study, the reward function encompasses various short-term treatment goals, including oxygenation and ventilation as well as long-term treatment goals such as mortality, length of stay, and discharge destination, collectively guiding the agent towards optimising patient treatment outcomes. For the intermediate reward, we included the P/F-ratio, the ratio of arterial partial pressure of oxygen (PaO2) to inspired oxygen concentration (FiO2), as a measure of oxygenation [14, 15]. The PF-ratio is also used to classify the severity of ARDS [16] and is confirmed as a risk factor for mortality in COVID-19 patients [17]. However, optimal P/F targets are not well-defined [18, 19]. We also included dead space, using Enghoff's modification of Bohr's equation [20] to estimate dead-space ventilation [18] from partial pressure of carbon dioxide (PaCO2) and the end-tidal carbon dioxide (ETCO2) levels: Vd/Vt = [PaCO2 - ETCO2] / PaCO2. Removal of CO2 is a primary goal of mechanical ventilation [21] and dead space is correlated with mortality in ARDS patients [22-25]. As exact targets are ill-defined, we primarily used changes between measurements as a



Physician Action distribution

Fig. 2 Action space density distribution of the historical actions of physicians in the dataset and illustration of the RL king-knight policy restriction. The red box shows under which actions the RL policy may recommend cessation of mechanical ventilation and the yellow box shows what actions a policy may next recommend if the current action (the small green box) is PEEP 6–10 cmH2O with FiO2 40–60%. NV stands for Non-invasively ventilated

delta target in the intermediate reward formulation. We defined a terminal reward component based on several factors. First, we included mortality, but also included the length of stay and the patient's discharge destination after the hospital admission. The primary objective is to improve the quality of life, which is conventionally measured using Quality-Adjusted Life Years [26]. However, due to insufficient post-discharge data, we employed the length of stay (LOS) in the ICU as a surrogate measure for quality of life. This choice was made to approximate the impact of healthcare interventions on patients' well-being. The exact formula for the reward function is provided in Additional file 2: Appendix SB.

Policy formulation

In RL models, while the primary goal is to learn optimal policies, these models inherently do not prescribe a specific action as optimal. The common practice is to employ a "greedy" policy, selecting actions with the highest expected reward. However, this approach might not always be suitable, particularly for unstable ARDS patients, where a significant deviation from previous actions could be detrimental. To mitigate this issue, we propose a policy restriction that confines the model to similar actions, allowing for one step up or down in either PEEP or FiO2. Illustrated in Fig. 2, our approach restricts the agent like a chess piece, with movements similar to a chessboard. The agent's deviations are akin to a king's movements, and the RL policy's advice on stopping mechanical ventilation is likewise constrained. We term this approach the "king-knight" policy that allows for structured flexibility, particularly when a patient is not on mechanical ventilation, where any action is permissible.

Model architecture, training and off-policy evaluation

The model architecture used in this paper is a Dueling Double-Deep Q Network (DDQN) [27, 28] as used in previous work [29]. We used an extensive hyperparameter grid to find a set of optimal model and training settings, including a variable amount of hidden layers (3 to 5) and nodes (32, 64, 128) per layer. Two learning rate (LR) decay (ReduceLRonPlateau and STEPLR) were explored for training. Training was performed using Prioritised Experience Replay [30] and all models were implemented with PyTorch [31]. We used the MAGIC [32]OPE estimator to assess policy performance. We defined the physician behavioural policy using K-nearest neighbour as in previous work [29].

Experimental setup

Given the unknown optional trade-off between intermediate and terminal rewards, our study involved training models using a diverse range of six weightings encompassing both reward components. To assess the generalisability of policies, we introduced cross off-policy evaluation, where policies trained under a specific set of reward weights were evaluated on the remaining five sets of reward weights. In our experiments, this weighting factor is varied across a set of predefined values: [0.25, 0.5, 1, 2, 4, 8]. This evaluation methodology necessitates that each individual reward component, namely the intermediate and terminal reward, inherently reflects clinically desirable outcomes in isolation. The experimental setup and methodology employed are depicted in Fig. 3. Best performing models were selected for further clinical policy inspection.

Policy evaluation and clinical policy inspection

To assess RL policies, the Off-Policy Policy Evaluation (OPE) method is utilised. OPE allows for the appraisal of a proposed AI policy by estimating its performance using real-world, historical ICU data, thus measuring its potential impact and effectiveness in past patient care scenarios without the need for actual policy execution.

To evaluate policies in a rigorous manner, we propose the introduction of a novel metric termed "delta-Q". Firstly, the Q-value, or q(s,a), represents the output of our model for a given state-action pair, indicating the



Fig. 3 Experiment design with a framework for off-policy evaluation and model and policy selection through cross-OPE evaluation and clinical policy inspection

expected utility of an action in a specific state. Delta-Q, on the other hand, is defined as the difference between the Q-value generated by our model and the Q-value representing the physician's action. Essentially, it measures the discrepancy in action quality between the AI's decision and the physician's choice. A delta-Q of zero implies that the model's decision aligns with the physician's, while a positive delta-Q suggests that the model's action might lead to improved treatment outcomes compared to historical decisions. By examining delta-Q values across all state-action pairs, we can identify specific areas of treatment that warrant further clinical scrutiny. Notably, delta-Q can be employed at both the dataset level and individual patient trajectories, enabling comprehensive analysis.

We also extend upon previous work on clinical policy inspection [29] and propose several model and policy visualisations for clinical validation and operationalisation. Due to the black-box nature of deep learning algorithms, we aim to provide clinical insight into policy behaviour for individual patients. For example, we can use delta-Q as a metric for alignment between physician and policy actions. Sudden changes in delta-Q, due to changes in state or physician action, over the course of the admissions can be used as a clinical alert. This allows physicians to evaluate model behaviour on a case-by-case basis and align with the clinical context in which the trained model and policy will be used.

Results

We extracted 1118 ICU admission from the DDW to create trajectories. 70% of trajectories were used for training and 30% for testing. The average duration of an admission was 12.8 days, resulting in a total of 34 years of data for training and 7.8 years for testing. Missing values were filled using the last value carry-forward method [33] and patients with excessive missing data were excluded, as shown in Fig. 1. A total of 158 features were used in the final state space. We trained 69,120 models using 8 model architectures and explored 18 hyperparameters. The total training time for this experiment was approximately 7 days on a virtual machine with 32 cores, 32 GB RAM, and 2 Nvidia 3070 Ti GPUs. Training was done using cuda [34].

Model and policy performance

For each model, we evaluated two policies, "greedy" and "king-knight", using the MAGIC OPE estimator. A total of 69,120 models were trained from which 138,240 policies were evaluated. From these, a set of 17,182 policies were selected and further evaluated using cross-OPE evaluation, as defined in the second evaluation round, see Fig. 3. Only 9.635 (56%) of cross-OPE evaluated models passed this evaluation phase. The results of this evaluation are shown in Fig. 4. Despite the successful performance of the models in the initial evaluation phase, a considerable number failed to exhibit comparable results during the cross-OPE evaluation utilising an alternate reward function shape. This may indicate the possibility of overfitting or the development of a policy that lacks generalisability. A final selection of top 100 models was based on cross-OPE performance and further explored in the clinical policy evaluation.

Model evaluation

We used the Q-value distribution of the state-action pairs of the dataset to investigate how the model ranks different actions on average. Figure 5 provides an example of a model with policies that had good quantitative (top ten percentile and only positive OPE and cross-OPE results) results, but shows a bias against one specific action, 'PEEP 10–14 and FiO2 80–100%.' Although the policies show clinically reasonable action recommendations, the bias in this model could disqualify it from clinical implementation, highlighting that OPE results alone are insufficient for a robust model and policy evaluation.

Policy evaluation

Figure 6 presents the delta-Q surface plot for a policy applied across the dataset, highlighting areas needing optimisation. The delta-Q is lower for actions in the 'PEEP 14 + cmH20 and FiO2 80–100%' range, and higher in the lower PEEP (0–6 cmH20, 6–10 cmH20) and lower FiO2 (21–40%, 40–60%) ranges, indicating these as critical improvement areas. This trend is consistent across models that passed the OPE and cross-OPE evaluations, directing physicians to prioritise these ranges in mechanical ventilation strategies for COVID ARDS patients.

Inherent to our safety-restricted policy design, clinical decisions can be more optimal than the RL-derived policy, as the safety-restricted policy limits the selection to less aggressive choices, barring potentially optimal but riskier actions. This includes situations where transitioning from a lower PEEP and FiO2 setting to a significantly higher one cannot be advised by the safety-restricted policy. These instances where the AI and clinical policy differ significantly are spread across the action space, suggesting a nuanced understanding of policy performance and indicating a need for future research, particularly in time series analysis. Figure 7's action distribution analysis for subgroups further reveals that the optimal policy typically favours treatments with higher FiO2 and PEEP values, though not in combination, reflecting a complex balance in treatment decision-making.



OPE performance (normalised) for selected (OPE>0) policies

Fig. 4 The histograms illustrate the evaluation outcomes with a detailing the Off-Policy Evaluation (OPE) for policies with positive OPE returns and b presenting the cross-OPE results. The vertical axis shows the density distribution, indicating data spread. The horizontal axis measures the relative performance or value of the target policy against the behaviour policy, using different reward functions without a specific unit. The "reward function version" corresponds to a series of weight factors: [0.25, 0.5, 1, 2, 4, 8], which are assigned to versions 1 through 6, respectively

Clinical policy inspection and RL model application

1.50 1.25 1.00 0.75 0.50 0.25 0.00 -

We introduce a method for applying RL to individual patient cases, exemplified in the log-Delta-Q trajectory 3D analysis, as presented in Fig. 8. This figure displays a 3D graph of PEEP, FiO2, and time, with log-Delta-Q values shown through varying hues. In specific instances, abrupt shifts in hue on the graph can signal physicians to reassess the patient's condition, suggesting potential changes in clinical state that might not be obvious through standard indicators. This method can assist physicians in prioritising their attention, aiding them in determining which patients may need more immediate and focused care.

Discussion and conclusion

In this study, we proposed a method to assess RL model performance using cross-OPE, which allows for assessment of generalisability of trained policies. We also introduced a novel policy filter, "king-knight", that restricts model recommendations to actions similar to those taken by physicians. Additionally, we proposed to operationalise the application of RL in ICU settings by implementing delta-Q based clinical alerts. These alerts prompt physicians to reassess treatment plans, particularly PEEP/FiO2 settings, in response to changes in patient or ventilator parameters. Additionally, Fig. 8 provides a graphical representation to aid physicians in tracking patient progress,



Fig. 5 Distribution of Q-values for all state-action pairs within the historical dataset. The x-axis denotes the Q-values, and the y-axis represents their frequency. The plot reveals a pronounced shift to the left in the Q-value distribution for 'PEEP 10–14 and FiO2 80–100%', implying a lower expected reward. A policy based on this model will likely infrequently suggest 'PEEP 10–14 and FiO2 80–100%'. Each curve in the plot is labelled according to the corresponding 'PEEP' and 'FiO2' action pair, as specified in the legend



Fig. 6 Displayed are three aggregate surface plots over the dataset: the first for physician-chosen actions, the second for policy-recommended actions, and the third, Delta-Q, which is calculated by subtracting physician Q-values from policy Q-values. The Delta-Q plot indicates smaller differences in high (14+) PEEP and FIO2 (80–100%) settings, suggesting this action aligns closely with the policy's guidance. In contrast, larger Delta-Q values in lower PEEP and FIO2 ranges suggest greater divergence, indicating that these areas may have more room for optimisation in alignment with policy recommendations



Survivors heatmap



Non-survivors heatmap







identifying unusual action choices, and exploring potential alternative treatment strategies. This approach may enhance real-time decision-making but also fosters a deeper understanding of patient-specific responses within the RL model's action space.

It is crucial to conduct a thorough clinical evaluation of trained RL models and policies before deployment in real-world settings. We showed how the use of off-policy evaluation alone may not be sufficient and may not identify models with clinically unexplainable behaviour. If a model's behaviour is not understandable or is at odds with clinical practice, it may not be trusted by physicians and could potentially lead to unsafe or suboptimal treatment decisions. Therefore, it is important to use a combination of methods, such as cross-OPE and clinical policy inspection, to ensure that the trained models and policies align with clinical expectations and have the desired performance.

It is important to note that there are limitations to these methods. One important caveat is that the proposed method relies heavily on domain expertise to define the components of the reward function and performing a



Fig. 8 An example of a trajectory 3D visualisation of the two components of the action space, PEEP and FiO2 over time with loq-Delta-Q values shown through varying hues

clinical evaluation is labour intensive and time consuming. Nonetheless, to ensure the safe deployment of RL models in clinical practice, we believe that despite the added labour and time required this step is essential for the safe and effective deployment of RL models in the healthcare setting. Additionally, it should be mentioned that our novel approaches to model evaluation can mitigate but not completely solve challenges related to offpolicy evaluation.

One limitation of this study is that it intentionally simplifies the action space, focusing primarily on optimising two key ventilatory settings. While this decision was made to make the model more efficient and interpretable, it raises the question of whether solely fine-tuning these parameters would significantly improve treatment outcomes without simultaneously optimising other crucial aspects of patient care, e.g. sedation, muscle blockers, haemodynamic stability or appropriate anticoagulation therapy. Additionally, it is important to note that this question may remain unanswered without empirical evidence, suggesting that an initial deployment of a simpler model may be practical, followed by iterative refinements to incorporate additional complexities as needed to enhance treatment efficacy.

Our proposed method also opens up new avenues for exploration, particularly in refining model selection through physician-patient consultations. Integrating patient preferences, like prioritising survival or quality of life, into the RL model's reward function is key to selecting a fitting reward function and corresponding optimal model, ensuring treatment advice aligns with each patient's specific preferences, thus enhancing patient-centric decision-making. The use of cross-OPE can assist researchers and physicians in creating models that perform well across a range of patient preferences. While there are limitations to these methods, they represent a step forward in the application of RL in healthcare, with the potential to improve treatment outcomes for critically ill patients.

Upcoming studies should focus on validating these methods across various clinical settings and adjusting RL models to align with evolving treatment standards. When planning for future localised applications of these models, it is also important to consider the centre-effect [35] variations in resources, practice patterns, and patient demographics may profoundly influence performance.

In conclusion, we have shown how cross-OPE can serve as a robust evaluation framework to identify policies with good generalisability. In addition, we demonstrated that policy restriction can help prevent potentially unsafe model recommendations. Finally, the novel delta-Q metric can be used to operationalise RL models in clinical practice. Our findings offer a promising pathway towards application of RL in intensive care medicine and beyond.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40635-024-00614-x.

Additional file 1. List of features.

Additional file 2: Reward function formulation for reinforcement learning model.

Acknowledgements

The Dutch ICU Data Sharing Against COVID-19 Collaborators have shared data. Partially funded by grants from ZonMw (Project 10430012010003, fle 50-55700-98-908), Zorgverzekeraars Nederland and the Corona Research Fund. The sponsors had no role in any part of the study.

Author contributions

LR and AH processed and analysed the data and drafted the manuscript. All authors contributed to interpretation of the results and critically reviewed the manuscript.

Funding

Partially funded by grants from ZonMw (Project 10430012010003, file 50-55700-98-908), Zorgverzekeraars Nederland and the Corona Research Fund. The sponsors had no role in any part of the study.

Availability of data and materials

All participating hospitals have access to the Dutch ICU Data Warehouse. External researchers can get access in collaboration with any of the participating hospitals. Contact details can be found on amsterdammedicaldatascience. nl. The code used for analysis is available upon request at gitlab.com.

Declarations

Ethics approval and consent to participate

The Medical Ethics Committee at Amsterdam UMC waived the need for patient informed consent and approved of an opt-out procedure for the collection of COVID-19 patient data during the COVID-19 crisis as documented under number 2020.156.

Consent for publication

Not applicable.

Competing interests

Not applicable.

Author details

¹Department of Intensive Care Medicine, Center for Critical Care Computational Intelligence (C4I), Amsterdam Medical Data Science (AMDS), Amsterdam Cardiovascular Science (ACS), Amsterdam Institute for Infection and Immunity (Al&II), Amsterdam Public Health (APH), Amsterdam UMC, Vrije Universiteit, Amsterdam, The Netherlands. ²Quantitative Data Analytics Group, Department of Computer Science, Faculty of Science, Vrije Universiteit, Amsterdam, The Netherlands.

Received: 15 June 2023 Accepted: 7 March 2024 Published online: 25 March 2024

References

- Sutton RS, Barto AG. Reinforcement Learning, second edition: An Introduction. MIT Press; 2018. Available: https://play.google.com/store/ books/details?id=uWV0DwAAQBAJ
- Gottesman O, Johansson F, Meier J, Dent J, Lee D, Srinivasan S, et al. Evaluating Reinforcement Learning Algorithms in Observational Health Settings. arXiv [cs.LG]. 2018. Available: http://arxiv.org/abs/1805.12298
- Lu M, Shahn Z, Sow D, Doshi-Velez F, Lehman L-WH. Is Deep Reinforcement Learning Ready for Practical Applications in Healthcare? A Sensitivity Analysis of Duel-DDQN for Hemodynamic Management in Sepsis Patients. AMIA Annu Symp Proc. 2020;2020: 773–782. Available: https://www.ncbi.nlm.nih.gov/pubmed/33936452
- Mercat A, Richard J-CM, Vielle B, Jaber S, Osman D, Diehl J-L, et al. Positive end-expiratory pressure setting in adults with acute lung injury and acute respiratory distress syndrome: a randomized controlled trial. JAMA. 2008;299: 646–655. doi:https://doi.org/10.1001/jama.299.6.646
- Higher versus Lower Positive End-Expiratory Pressures in Patients with the Acute Respiratory Distress Syndrome. N Engl J Med. 2004;351: 327–336. doi:https://doi.org/10.1056/NEJMoa032193
- Chikhani M, Das A, Haque M, Wang W, Bates DG, Hardman JG (2016) High PEEP in acute respiratory distress syndrome: quantitative evaluation between improved arterial oxygenation and decreased oxygen delivery. Br J Anaesth 117:650–658. https://doi.org/10.1093/bja/ aew314
- Cooper JS, Phuyal P, Shah N. Oxygen Toxicity. StatPearls. Treasure Island (FL): StatPearls Publishing; 2022. Available: https://www.ncbi.nlm.nih. gov/pubmed/28613494
- Somhorst P, van der Zee P, Endeman H, Gommers D (2022) PEEP-FiO2 table versus EIT to titrate PEEP in mechanically ventilated patients with COVID-19-related ARDS. Crit Care 26:272. https://doi.org/10.1186/ s13054-022-04135-5
- He H, Chi Y, Yang Y, Yuan S, Long Y, Zhao P et al (2021) Early individualized positive end-expiratory pressure guided by electrical impedance tomography in acute respiratory distress syndrome: a randomized controlled clinical trial. Crit Care 25:230. https://doi.org/10.1186/ s13054-021-03645-y
- Prasad N, Cheng L-F, Chivers C, Draugelis M, Engelhardt BE. A Reinforcement Learning Approach to Weaning of Mechanical Ventilation in Intensive Care Units. arXiv [cs.Al]. 2017. Available: http://arxiv.org/abs/ 1704.06300
- Peine A, Hallawa A, Bickenbach J, Dartmann G, Fazlic LB, Schmeink A et al (2021) Development and validation of a reinforcement learning algorithm to dynamically optimize mechanical ventilation in critical care. NPJ Digit Med 4:32. https://doi.org/10.1038/s41746-021-00388-6
- 12. Kondrup F, Jiralerspong T, Lau E, de Lara N, Shkrob J, Tran MD, et al. Towards Safe Mechanical Ventilation Treatment Using Deep Offline Reinforcement Learning. arXiv [cs.LG]. 2022. Available: http://arxiv.org/ abs/2210.02552
- Fleuren LM, Dam TA, Tonutti M, de Bruin DP, Lalisang RCA, Gommers D et al (2021) The Dutch Data Warehouse, a multicenter and fulladmission electronic health records database for critically ill COVID-19 patients. Crit Care 25:304. https://doi.org/10.1186/s13054-021-03733-z

- Horovitz JH, Carrico CJ, Shires GT (1974) Pulmonary response to major injury. Arch Surg 108:349–355. https://doi.org/10.1001/archsurg.1974. 01350270079014
- Bernard GR, Artigas A, Brigham KL, Carlet J, Falke K, Hudson L, et al. The American-European Consensus Conference on ARDS. Definitions, mechanisms, relevant outcomes, and clinical trial coordination. Am J Respir Crit Care Med. 1994;149: 818–824. doi:https://doi.org/10.1164/ ajrccm.149.3.7509706
- ARDS Definition Task Force, Ranieri VM, Rubenfeld GD, Thompson BT, Ferguson ND, Caldwell E, et al. Acute respiratory distress syndrome: the Berlin Definition. JAMA. 2012;307: 2526–2533. doi:https://doi.org/10. 1001/jama.2012.5669
- Gu Y, Wang D, Chen C, Lu W, Liu H, Lv T et al (2021) PaO2/FiO2 and IL-6 are risk factors of mortality for intensive care COVID-19 patients. Sci Rep 11:7334. https://doi.org/10.1038/s41598-021-86676-3
- Spinelli E, Mauri T. Why improved PF ratio should not be our target when treating ARDS. Minerva anestesiologica. 2021. pp. 752–754. doi:https://doi.org/10.23736/S0375-9393.21.15664-0
- Aggarwal NR, Brower RG (2014) Targeting normoxemia in acute respiratory distress syndrome may cause worse short-term outcomes because of oxygen toxicity. Ann Am Thorac Soc 11:1449–1453. https:// doi.org/10.1513/AnnalsATS.201407-297PS
- 20. Enghoff H. Volumen inefficax. Upsala Lakareforen Forh.
- 21. Robertson HT (2015) Dead space: the physiology of wasted ventilation. Eur Respir J 45:1704–1716. https://doi.org/10.1183/09031936.00137614
- Santus P, Radovanovic D, Saderi L, Marino P, Cogliati C, De Filippis G et al (2020) Severity of respiratory failure at admission and in-hospital mortality in patients with COVID-19: a prospective observational multicentre study. BMJ Open 10:e043651. https://doi.org/10.1136/bmjop en-2020-043651
- Nuckton TJ, Alonso JA, Kallet RH, Daniel BM, Pittet J-F, Eisner MD et al (2002) Pulmonary dead-space fraction as a risk factor for death in the acute respiratory distress syndrome. N Engl J Med 346:1281–1286. https://doi.org/10.1056/NEJMoa012835
- Sinha P, Calfee CS, Beitler JR, Soni N, Ho K, Matthay MA et al (2019) Physiologic Analysis and Clinical Performance of the Ventilatory Ratio in Acute Respiratory Distress Syndrome. Am J Respir Crit Care Med 199:333–341. https://doi.org/10.1164/rccm.201804-0692OC
- Villar J, Pérez-Méndez L, Blanco J, Añón JM, Blanch L, Belda J et al (2013) A universal definition of ARDS: the PaO2/FiO2 ratio under a standard ventilatory setting–a prospective, multicenter validation study. Intensive Care Med 39:583–592. https://doi.org/10.1007/ s00134-012-2803-x
- Räsänen P, Roine E, Sintonen H, Semberg-Konttinen V, Ryynänen O-P, Roine R (2006) Use of quality-adjusted life years for the estimation of effectiveness of health care: A systematic literature review. Int J Technol Assess Health Care 22:235–241. https://doi.org/10.1017/S0266 462306051051
- Wang Z, Schaul T, Hessel M, van Hasselt H, Lanctot M, de Freitas N. Dueling Network Architectures for Deep Reinforcement Learning. arXiv [cs.LG]. 2015. Available: http://arxiv.org/abs/1511.06581
- Sewak M. Deep Q Network (DQN), Double DQN, and Dueling DQN. In: Sewak M, editor. Deep Reinforcement Learning: Frontiers of Artificial Intelligence. Singapore: Springer Singapore; 2019. pp. 95–108. doi:https://doi.org/10.1007/978-981-13-8285-7_8
- Roggeveen L, El Hassouni A, Ahrendt J, Guo T, Fleuren L, Thoral P et al (2021) Transatlantic transferability of a new reinforcement learning model for optimizing haemodynamic treatment for critically ill patients with sepsis. Artif Intell Med 112:102003. https://doi.org/10.1016/j. artmed.2020.102003
- Schaul T, Quan J, Antonoglou I, Silver D. Prioritized Experience Replay. arXiv [cs.LG]. 2015. Available: http://arxiv.org/abs/1511.05952
- Paszke A, Gross S, Chintala S, Chanan G. Pytorch: Tensors and dynamic neural networks in python with strong gpu acceleration. PyTorch: Tensors and dynamic neural networks in Python with strong GPU acceleration. 2017;6.
- Thomas P, Brunskill E. Data-Efficient Off-Policy Policy Evaluation for Reinforcement Learning. International Conference on Machine Learning. jmlr.org; 2016. pp. 2139–2148. Available: http://www.jmlr.org/ proceedings/papers/v48/thomasa16.html

- Little RJA (1988) A Test of Missing Completely at Random for Multivariate Data with Missing Values. J Am Stat Assoc 83:1198–1202. https:// doi.org/10.1080/01621459.1988.10478722
- 34. NVIDIA and Vingelmann, Péter and Fitzek, Frank H.P. CUDA. 2020. Available: https://developer.nvidia.com/cuda-toolkit
- Localio AR, Berlin JA, Ten Have TR, Kimmel SE (2001) Adjustments for center in multicenter studies: an overview. Ann Intern Med 135:112–123. https://doi.org/10.7326/0003-4819-135-2-200107170-00012

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.