Reinforcement Learning Approach to Sedation and Delirium Management in the Intensive Care Unit

Niloufar Eghbali[1], Tuka Alhanai[2] and Mohammad M. Ghassemi[3]

***Abstract*— Common treatments in Intensive Care Units frequently involve prolonged sedation. Maintaining adequate sedation levels is challenging and prone to errors including: incorrect dosing, omission/delay in administration and, selecting a sub-optimal combination of sedatives. In this single-center retrospective study of 1,346 patients, we use a Deep Q Network approach to develop a multi-objective sedation management agent. The agent's objective was to achieve an adequate level of patient sedation without moving the patient's Mean Arterial Pressure (MAP) outside of a therapeutic range. To achieve this objective, the agent was allowed to periodically (every 4 hours) recommend how the dose of two commonly used sedatives (propofol, midazolam) and an opioid (fentanyl) should be adjusted: increased, decreased, or stay the same. To inform it's recommendations, the agent was provided with the patient's demographym and periodic measures including: vital signs, and depth of sedation. To mitigate the potential risk of delirium and the adverse effects of over sedation, a delirium control variable was integrated into the agent's reward function. We found that Physicians with dosing policies that agreed with our agent were 29% more likely to maintain the patient's sedation in a therapeutic range, compared to those that disagreed with our agent's policy.**

***Clinical relevance*— This study utilizes reinforcement learning to develop a sedation management agent, improving the ability to maintain target sedation levels by 29% compared to clinicians' policy, while considering optimal dosage regimens and delirium control in the ICU.**

I. INTRODUCTION

ICU patients often undergo treatments such as ventilation and incubation which may result in distressing side effects including: agitation, delirium, and pain [1]. To manage these side-effects, many patients require a combination of analgesic and sedative agents for several days ensure safe and effective treatment in the ICU [2], [3], [4].

Over- and under-sedation are among the most frequent adverse medication events related to prolonged administration of sedatives in the ICU [5]. Under-sedation typically results in untreated pain, which has several secondary consequences including: immunomodulation, and increased risk of post-traumatic stress disorder. In contrast, over-sedation results in longer ICU recovery times, increased risk of ventilator-associated pneumonia, higher likelihood of long-term brain dysfunction, increased morbidity and mortality [6], [7], [5], [1].

[1]Department of Computer Science, Michigan State University, East Lansing, MA, USA eghbaliz@msu.edu

[2]Division of Engineering, New York University Abu Dhabi, Abu Dhabi, UAE tuka.alhanai@nyu.edu

[3]Department of Computer Science, Michigan State University, East Lansing, MA, USA ghassem3@msu.edu

To complicate matters, critically ill patients may have changes in pharmacokinetics and pharmacodynamic characteristics due to pathophysiological variations or drug interactions, potentially increasing the chance of adverse medication errors. [3], [4]. Thus, it is particularly important to identify optimal sedative dosages to ensure patient safety in the ICU.

Sedation management in the ICU has been addressed as a sequential decision-making problem using reinforcement learning (RL) techniques. RL offers advantages by considering long-term outcomes, personalizing treatments, and operating without explicit mathematical models of biological systems [8], [9], [10].

Prior works have applied RL algorithms to regulate sedatives based on control variables such as bispectral index (BIS) and mean arterial pressure (MAP) [11], [12], [13], [14]. Padmanabhan et al. (2015) used a Q-learning RL algorithm to regulate propofol in simulated ICU patients [11]. In another work, Yu et al. (2019) applied Fitted Q-Iteration (FQI) and Bayesian inverse RL on retrospective data to infer latent reward functions for sedation regulation and weaning mechanical ventilation [13]. Further advancements have been made to incorporate both short-term and long-term treatment goals in RL-based sedation management. Yu et al. (2020) proposed a Supervised-Actor-Critic (SAC) RL algorithm that combined supervised learning with RL to minimize the deviation between the RL agent and clinician while aiming for effective patient treatment [15]. Eghbali et al. (2021) developed a deep deterministic policy gradient RL agent to regulate propofol dosages in a continuous action space, aiming to maintain adequate sedation levels while minimizing sedative dosages and considering patient health conditions [14].

While prior works demonstrate the potential of RL for sedation management, there remain clear opportunities for improvement. For instance, none of the prior works we surveyed account for the multi-objective, multi-action nature of the sedation management problem; in real clinical settings sedation management must also be balanced with delirium management, and the dose of one sedative (e.g. propofol) must be selected in light of the type and amount of other sedatives that a patient may be receiving.

Delirium is prevalent in ICU patients, with an estimated occurrence of 45%-87% [16], [17]. It is associated with increased mortality, cognitive impairment, extended ICU stays, and prolonged mechanical ventilation[18], [16]. Sedatives used in ICU protocols can contribute to delirium formation, prompting clinicians to include delirium screening as an integral component of active sedative management[1],

[16]. However, the interaction of delirium and sedation is neglected in studies that address sedation regimens. Most of the prior works concentrate on the regulation of a common sedative - propofol - while in real clinical settings, propofol is rarely used alone; it is typically used in conjunction with an opioid or other analgesics to manage ICU patients[19].

Our study extends earlier research and addresses the above limitations by: (1) training an agent to jointly regulate the two most commonly used sedatives (propofol and midazolam) along with a widely used opioid (Fentanyl) [20] and (2) accounting for both the level of sedation (as evaluated by the Richmond Sedation-Agitation Scale (RASS) [21]) as well as delirium in our formulation of the agent's reward function.

II. Method

A. Data Collection

All data in this study were sourced from the Medical Information Mart for Intensive Care (MIMIC-IV) dataset [22]. MIMIC-IV comprises de-identified clinical data from patients admitted to a tertiary academic medical center in Massachusetts, USA. We extracted all patients from MIMIC who received propofol, midazolam, and fentanyl during their ICU stay; this resulted in a sample of 1,346 patients. In Table I, we provide a summary of demographic characteristics for the patient sample.

B. Data Preparation

Each patient's ICU stay was divided into four-hour time windows, starting from the time they first received a sedative until the conclusion of their ICU stay. For each 4-hour interval of each patient, we extracted the amount of administered medications (dosages of propofol, midazolam, and fentanyl) and 18 clinical features based on the sedation monitoring guidelines by the American Society of Anesthesiologists, as well as studies conducted by [23], [15], [24] . The clinical features included: arterial pH, positive end-expiratory pressure (PEEP) set, inspired oxygen fraction (FiO2), arterial oxygen pressure, plateau pressure, average airway pressure, heart rate, respiratory rate, non-invasive blood pressure (mean, diastolic), oxygen saturation, mean arterial pressure (MAP), arterial blood pressure (diastolic), richmond sedation-agitation scale (RASS), delirium presence, age, and gender.

We utilized the sample-and-hold interpolation approach to deal with missing data: for a missing value in a given time window we kept the value of the previous time window. All remaining missing values (i.e. those without a predecessor) were imputed using k-nearest neighbor imputation.

TABLE I

Sample size and demographic characteristics of the patient sample extracted from MIMIC-IV.

Number of Patients in dataset	1,346
Gender proportion(F/M)	39.9% / 60.1%
Age(Years) Mean±Std	61.9 ± 16.8
ICU length of stay (hrs) Mean±Std	129.6 ± 114

TABLE II

Richmond agitation–sedation scale description and proportion of our patient sample available in each level

Score	Description	Proportion of data (%)
4	Combative	0.01
3	Very agitated	0.21
2	Agitated	0.66
1	Restless	3.89
0	Alert and calm	32.86
-1	Drowsy	20.75
-2	Light sedation	14.15
-3	Moderate sedation	10.24
-4	Deep sedation	8.12
-5	Unarousable	9.12

We randomly grouped patients into three sets: training (60%, 807 patients, 48,764 time windows), validation (20%, 270 patients, 18,023 time windows), and test set (20%, 269 patients, 17,193 time windows). All model development occured on the training and validation sets. Performance of the trained models was assessed on the test set.

C. Target Variables

In this sub-section we list the target variables considered by our reinformance learning agent's reward function.

Richmond Agitation–Sedation Scale (RASS): The "Richmond Agitation–Sedation Scale" (RASS) was selected as the target for the level of patient sedation. RASS is a 10-point scale including four "sedation" levels (-5 to -2), one "calm and alert" level (-1), and five levels of "anxiety or agitation" (0 to 4). In clinical practice, the goal is to maintain the sedation between -2 and 0[6], [21]. In Table II, we provide descriptions and proportions of each RASS level in the data.

Delirium: For each patient, the presence of delirium was encoded as a binary value where 1 denoted the presence of delirium. In clinical practice, the goal is to prevent the occurrence of delirium.

Mean Arterial Pressure (MAP): Sedatives, such as propofol, can have a detrimental influence on the hemodynamic stability of patients. Propofol, in particular, promotes vasodilation, which lowers MAP. To address this issue in our model, we utilized MAP as a hemodynamic parameter to refer to patients' physiological stability. In clinical practice, the goal is to keep the MAP in a therapeutic target range.

D. Model

Herein the sedation management problem is formulated as a Markov Decision Process (MDP) described by the tuple (S, A, P, R), where $s_t \in S$ represents the set of measurements that describes a patient status at a given time step t, $a_t \in A$ is the action and accounts for dosage of the drugs at time t, $P(s_{t+1}|s_t, a_t)$ is the probability of the next state given the current state, and $r(s_t, a_t) \in R$ is the observed reward following a transition at time step t. In each time step, the agent observes the current state and takes one action among the available set of actions for

which it receives a reward and moves to the next state. We apply the Deep Q-Network (DQN) [25] algorithm to train an agent that takes the state and returns the best possible action which maximizes the sum of rewards in an episode. Prior research largely used Q-learning as a reinforcement learning strategy, however, in this work, we employ DQN which utilizes deep neural networks and a replay of previous experience to enhance Q-learning algorithm. [25].

State: State $s_t \in S$ at each time step t is an 18-dimensional vector of measurements describing a patient's clinical and demographic metrics (described in data preparation section). We defined a finite number of states (1000) by clustering all patient's time series of states across train data. The Elbow method was applied to determine the optimal number of clusters.

Action: We utilized a discrete action space to decide between three alternative actions for each of the three medications: raising the dose, reducing the dose, or maintaining the dose. This pattern resulted in 27 distinct actions.

reward: $r(s_t, a_t) \in R$ is the observed reward of following action a_t at state s_t. We have defined the reward function based on three key variables discussed above. For each of the control parameters we define a maximum score of 1 that can be achieved when the parameter is in the target range. Therefore, in each timestep the maximum achievable immediate reward is 3. The Reward function is formulated below:

$$r_t = r_{RASS} + r_{MAP} + r_{delirium} \quad (1)$$

$$r_{MAP} = \frac{1}{1 + e^{-(MAP_t - 70)}} - \frac{1}{1 + e^{-(MAP_t - 100)}} \quad (2)$$

$$r_{RASS} = \frac{1}{1 + e^{-10(RASS_t + 2)}} - \frac{1}{1 + e^{-10(RASS_t - 0)}} \quad (3)$$

$$r_{delirium} = \begin{cases} 1, \text{if delirium does not occur at timestep t} \\ 0, \text{otherwise} \end{cases} \quad (4)$$

Policy: policy π is a function that maps the patient's state to drug doses: $a = \pi(s)$. In training, the RL agent uses a sequence of observed state-action pairs (s_t, a_t), called a trajectory (T), to learn the optimal policy π^* by maximizing the sum of discounted rewards defined as:

$$R(T) = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \gamma^3 r_{t+3} + \ldots + \gamma^T r_{t+T} \quad (5)$$

γ is a discount factor that reflects the relative importance of immediate and long-term rewards. If γ is close to 0, the agent merely takes immediate reward into account; if γ is close to 1, the agent is biased toward longer-term rewards. In our study, the value of γ was 0.8 and was determined by experimenting with several values of γ and keeping the value that maximized the model's performance on the validation set.

The DQN algorithm learns the optimal policy π^* by estimating the optimal Q-function. The Q-function of a policy π, $Q^\pi(s, a)$ estimates the expected return obtained from state s taking action a and following policy π thereafter. The highest return that may be achieved starting from state s, performing an action a, and following the optimal policy π afterward is known as the optimal Q-function $Q^*(s, a)$. $Q^*(s, a)$ satisfies bellman equation:

$$Q^*(s, a) = E[r + \gamma \max_{a'} Q^*(s', a')] \quad (6)$$

We used a neural network composed of three fully connected layers with ReLu activation functions in the first two layers and a linear activation function in the final layer to approximate Q^*. We save the collection of $< s, a, r, s' >$ transitions in the replay buffer and update the network using a batch of transitions sampled from the replay buffer. Sampling uncorrelated transitions in a batch enhances the overall stability. Our replay memory capacity was set at 2000 experiences, while our batch size was set to 256. The model was implemented in Pytorch 1.6.0 and used Adam optimization. To balance exploration and exploitation, we used ϵ-greedy approach where we chose a value of 0.001 for ϵ. We ran a total of 25,000 experiments to select hyper-parameters. For each hyper-parameter mentioned above, we uniformly sampled values from a random range and validated the performance on the validation set. The mean obtained reward during episodes was used as the validation criteria for choosing the hyper-parameters.

Baseline To evaluate our model, we compared its performance to the clinician's recorded performance in the MIMIC database making the reasonable assumption that clinical staff intends to keep patients in a therapeutic condition during their ICU stay. Therefore, we may use the accuracy of the clinician to compare against our model. Performance is defined for each trajectory (hours spent in ICU) as:

$$P_i^c = \frac{\text{duration in which control variable c is in target range}}{\text{ICU duration}_i} \quad (7)$$

where i denotes the i^{th} patient. For each control variable $c \in \{RASS, Delirium, MAP\}$, the performance metric presents the proportion of the total ICU stay hours that patient i spent in the therapeutic range. For our study, we specifically examined four types of performance measures: the Performance error (PE), root mean square error (RMSE), the mean performance error (MPE), and the median performance error (MDPE)[11]. Performance error is defined as:

$$PE_i^c = (1 - P_i^c) \quad (8)$$

MDPE gives the control bias observed for a single patient and is computed by:

$$MDPE_i^c = median(PE_i^c) \times 100 \quad (9)$$

$RMSE_i^c$ is the RMSE for each patient and control variable.

III. Results

Comparison between performance metrics of our model and that of clinicians are available in Table III. MDPE and RMSE for our model show that our learned sedation

TABLE III

PERFORMANCE METRICS FOR CONTROL VARIABLES FOLLOWING BOTH CLINICIANS' POLICY AND TRAINED POLICY

Performance Metrics	Control Variables					
	Delirium		RASS		MAP	
	Clinician's policy	Learned policy	Clinician's policy	Learned policy	Clinician's policy	Learned policy
Mean Performance Error (MPE)	73.09	**4.19**	29.84	**0.45**	29.24	**5.48**
Median Performance Error (MDPE)	81.82	3.39	20	1.3	25.64	0
Root Mean Square error (RMSE)	0.8	0.2	1.1	0.4	33.5	4.2

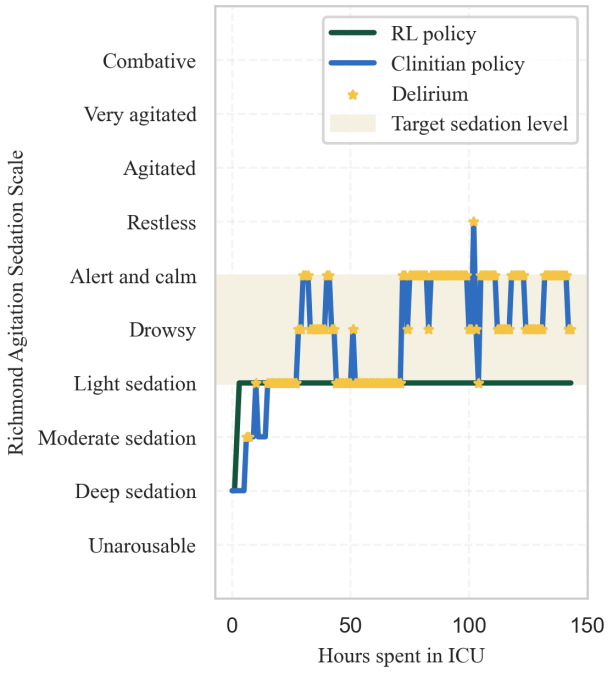

Fig. 1. Variation in sedation level for a randomly selected patient during ICU stay following both the clinician's policy(blue) and learned policy(green).

management approach results in less occurrences of out-of-therapeutic levels of sedation than seen among physicians. As demonstrated in Table III For 99.5% and 95.8% of the ICU stay, RASS and MAP measure values are in the desired range. MPE improved by 28.55% in RASS and 23.76% in MAP compared to the clinicians' policy.

Figure 1 presents the variation in sedation level for a randomly selected patient. The green line represents the sedation level as a consequence of our model's policy, whereas the blue line represents the sedation level resulting from clinician' policy. The existence of delirium is shown by * symbol. Highlighted section indicates the therapeutic range.

IV. DISCUSSION

Patients subjected to intensive interventions in the ICU, including ventilation and intubation, commonly experience associated complications, including pain, agitation, and delirium. As a result, there's a significant demand for prolonged sedative administration to guarantee the safety of these treatments. Inadequate sedation can precipitate a host of challenges, such as increased adverse incidents, diminished patient outcomes, an extended length of stay in the ICU, and augmented healthcare costs.

In this work, we developed an RL-based model based on the DQN algorithm to manage sedation and delirium in Intensive care units. We adopted RL because it is an effective framework for identifying the best sedative regimen for patients with different responses to the same medication and can learn the best sequence of decisions from retrospective data.

In the ICU, effective sedation management necessitates the combination of at least two distinct sedatives and analgesics. Building on prior research that focused solely on the frequently used sedative, propofol, we incorporated three of the predominant sedatives and analgesics. This approach aimed to formulate a more precise sedation strategy, considering the combined effects of these drugs. The agent was trained to determine the optimal sedative regimen by monitoring the patient's response to medication. To assess patients' levels of sedation, we utilized the Richmond Agitation and Sedation Scale (RASS) as a control variable. To mitigate the adverse effects of sedatives, we incorporated one of the most critical hemodynamic parameters, MAP, into the reward function. This approach gives credit to MAP values within the safe clinical range, thereby ensuring the patient's physiological stability. Compared to clinicians, our Method resulted in a 29% improvement in maintaining an adequate level of sedation, indicating the ability of the model to manage sedation. Furthermore, our model is able to control MAP during 94.5% of patients' length of ICU stay which is a 23.76% improvement in comparison to the clinician's policy. Delirium management is critical in ICU. Delirium is characterized by a rapidly changing cognitive state, inattention, disorganized thinking, and altered awareness. It occurs in up to 87% of ICU patients during their admission [16]. Sedative treatment that is administered incorrectly or insufficiently may trigger delirium or aggravate the symptoms. In practical clinical settings, sedation management protocols are accompanied by regular assessment of delirium, which is unaccounted for in prior works on sedation management. To overcome this limitation, we incorporated a delirium control variable in the reward function to prevent patients from developing delirium and minimizing the negative impact of sedatives. Following our trained policy, the incidence of delirium in patients will decrease by 68% , which is a notable improvement compared to the recorded performance of clinical staff.

Based on the assessments, we infer that our sedation management agent is an encouraging step toward automating

sedation in the ICU. Yet more work is required to make the approach presented here effective enough for practical implementation. Long-term administration of sedatives often leads to drug habituation, which means a patient's pharmacological response may change over the course of therapy. Future investigations should consider the impact of this habituation, guiding the direction of our subsequent research.

REFERENCES

[1] M. C. Reade and S. Finfer, "Sedation and delirium in the intensive care unit," *New England Journal of Medicine*, vol. 370, no. 5, pp. 444–454, 2014.

[2] J. W. Devlin, S. Mallow-Corbett, and R. R. Riker, "Adverse drug events associated with the use of analgesics, sedatives, and antipsychotics in the intensive care unit," *Critical care medicine*, vol. 38, pp. S231–S243, 2010.

[3] S. Liu, K. C. See, K. Y. Ngiam, L. A. Celi, X. Sun, and M. Feng, "Reinforcement learning for clinical decision support in critical care: comprehensive review," *Journal of medical Internet research*, vol. 22, no. 7, p. e18477, 2020.

[4] A. Wilmer, K. Louie, P. Dodek, H. Wong, and N. Ayas, "Incidence of medication errors and adverse drug events in the icu: a systematic review," *Quality and Safety in Health Care*, vol. 19, no. 5, pp. e7–e7, 2010.

[5] D. L. Jackson, C. W. Proudfoot, K. F. Cann, and T. S. Walsh, "The incidence of sub-optimal sedation in the icu: a systematic review," *Critical Care*, vol. 13, no. 6, p. R204, 2009.

[6] C. G. Hughes, S. McGrane, and P. P. Pandharipande, "Sedation in the intensive care setting," *Clinical pharmacology: advances and applications*, vol. 4, p. 53, 2012.

[7] R. Padmanabhan, N. Meskin, and W. M. Haddad, "Reinforcement learning-based control of drug dosing with applications to anesthesia and cancer therapy," in *Control Applications for Biomedical Engineering Systems*. Elsevier, 2020, pp. 251–297.

[8] A. Coronato, M. Naeem, G. De Pietro, and G. Paragliola, "Reinforcement learning for intelligent healthcare applications: A survey," *Artificial Intelligence in Medicine*, vol. 109, p. 101964, 2020.

[9] S. J. Bielinski, J. E. Olson, J. Pathak, R. M. Weinshilboum, L. Wang, K. J. Lyke, E. Ryu, P. V. Targonski, M. D. Van Norstrand, M. A. Hathcock *et al.*, "Preemptive genotyping for personalized medicine: design of the right drug, right dose, right time—using genomic data to individualize treatment protocol," in *Mayo Clinic Proceedings*, vol. 89, no. 1. Elsevier, 2014, pp. 25–33.

[10] C. Yu, J. Liu, and S. Nemati, "Reinforcement learning in healthcare: A survey," *arXiv preprint arXiv:1908.08796*, 2019.

[11] R. Padmanabhan, N. Meskin, and W. M. Haddad, "Closed-loop control of anesthesia and mean arterial pressure using reinforcement learning," *Biomedical Signal Processing and Control*, vol. 22, pp. 54–64, 2015.

[12] ——, "Optimal adaptive control of drug dosing using integral reinforcement learning," *Mathematical biosciences*, vol. 309, pp. 131–142, 2019.

[13] C. Yu, J. Liu, and H. Zhao, "Inverse reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units," *BMC medical informatics and decision making*, vol. 19, no. 2, pp. 111–120, 2019.

[14] N. Eghbali, T. Alhanai, and M. M. Ghassemi, "Patient-specific sedation management via deep reinforcement learning," *Frontiers in Digital Health*, vol. 3, p. 17, 2021.

[15] C. Yu, G. Ren, and Y. Dong, "Supervised-actor-critic reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units," *BMC medical informatics and decision making*, vol. 20, no. 3, pp. 1–8, 2020.

[16] R. Cavallazzi, M. Saad, and P. E. Marik, "Delirium in the icu: an overview," *Annals of intensive care*, vol. 2, no. 1, pp. 1–11, 2012.

[17] E. W. Ely, R. Margolin, J. Francis, L. May, B. Truman, R. Dittus, T. Speroff, S. Gautam, G. R. Bernard, and S. K. Inouye, "Evaluation of delirium in critically ill patients: validation of the confusion assessment method for the intensive care unit (cam-icu)," *Critical care medicine*, vol. 29, no. 7, pp. 1370–1379, 2001.

[18] C. N. Sessler and K. Varney, "Patient-focused sedation and analgesia in the icu," *Chest*, vol. 133, no. 2, pp. 552–565, 2008.

[19] J. Barr and A. Donner, "Optimal intravenous dosing strategies for sedatives and analgesics in the intensive care unit," *Critical care clinics*, vol. 11, no. 4, pp. 827–847, 1995.

[20] Y. Zhou, X. Jin, Y. Kang, G. Liang, T. Liu, and N. Deng, "Midazolam and propofol used alone or sequentially for long-term sedation in critically ill, mechanically ventilated patients: a prospective, randomized study," *Critical Care*, vol. 18, no. 3, pp. 1–9, 2014.

[21] C. N. Sessler, M. S. Gosnell, M. J. Grap, G. M. Brophy, P. V. O'Neal, K. A. Keane, E. P. Tesoro, and R. Elswick, "The richmond agitation–sedation scale: validity and reliability in adult intensive care unit patients," *American journal of respiratory and critical care medicine*, vol. 166, no. 10, pp. 1338–1344, 2002.

[22] A. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and R. Mark IV, "Mimic-iv (version 0.4)," *PhysioNet*, 2020.

[23] J. B. Gross, P. L. Bailey, R. T. Connis, C. J. Coté, F. Davis, B. Epstein, L. Gilbertson, D. Nickinovich, and J. Zerwas, "Practice guidelines for sedation and analgesia by non-anesthesiologists," *Anesthesiology*, vol. 96, no. 4, pp. 1004–1017, 2002.

[24] A. Jagannatha, P. Thomas, and H. Yu, "Towards high confidence off-policy reinforcement learning for clinical applications," in *CausalML Workshop, ICML*, 2018.

[25] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.