Discharge with Multiple Readmissions and Constraints

Maotong Sun

School of Management, Technical University of Munich
Jingui Xie

School of Management, Technical University of Munich
Munich Data Science Institute, Technical University of Munich

ARTICLE HISTORY
Compiled October 6, 2025

ABSTRACT
Effective discharge decision-making in Intensive Care Units (ICUs) is critical to optimizing patient outcomes and ensuring efficient utilization of limited resources, a challenge brought into sharper focus during the COVID-19 pandemic. Discharge decision-making involves a complex trade-off: discharging patients too early increases the risk of ICU readmission and mortality, while unnecessarily prolonged ICU stays elevate care costs and may adversely affect patient outcomes. We develop a data-driven framework that integrates Offline Reinforcement Learning (offline RL) with Constrained Reinforcement Learning (constrained RL) to derive personalized ICU discharge policies. We firstly include multiple ICU readmissions in the whole decision-making process. The approach also explicitly incorporates multiple clinical and operational constraints, including ICU length-of-stay (LOS) and readmission risk. Using real-world medical data from the MIMIC-IV database, we train and evaluate our model with Fitted Q-Iteration (FQI) and Fitted Q-Evaluation (FQE) under linear function approximation. The model distinguishes between initial and readmitted ICU cases by including readmission count in the state space. The proposed algorithm successfully learns discharge policies that reduce predicted mortality risk while satisfying thresholds on cumulative LOS and readmission risk. Numerical results show that the learned policies outperform observed clinical practices in mortality estimates and demonstrate conservative decision behavior for high-risk patient subgroups. By jointly optimizing discharge timing and multiple constraints, our approach yields more clinically aligned and efficient ICU discharge policies. This work demonstrates the feasibility and value of the combination of offline RL and constrained RL in supporting decisions in critical care environments.

KEYWORDS
Discharge decision optimization, constrained Markov decision process, offline reinforcement learning, primal-dual algorithm, off-policy evaluation

1. Introduction

Optimizing discharge decisions for patients in intensive care units (ICUs) represents a crucial challenge in healthcare management, given the substantial clinical and economic implications associated with these decisions. ICUs constitute resource-intensive settings, demanding timely and careful assessment of patient readiness for discharge. Effective discharge decision-making directly impacts patient recovery trajectories, hospital resource utilization, and overall healthcare expenditures. Moreover, the intricate

clinical conditions of ICU patients, coupled with intensive staffing requirements and elevated resource consumption, contribute significantly to the higher costs associated with ICU care compared to general medical-surgical wards (Bagshaw et al. 2020, Plotnikoff et al. 2021, Wu et al. 2025).

Consequently, healthcare providers face considerable pressure to discharge patients promptly once physiological stability is achieved and intensive care interventions are considered no longer necessary. On one hand, under the Affordable Care Act, it is in hospital's financial interest to discharge patients as soon as possible. However, premature discharge can result in unintended patient readmissions or rapid clinical deterioration, which exacerbate demands on hospital resources and are associated with substantially increased mortality risks and longer cumulative length-of-stay in the ICU (Kramer et al. 2013, Dam et al. 2025, Wu et al. 2025). Indeed, recent evidence suggests mortality rates among readmitted ICU patients can reach as high as 40%, markedly higher than the 3.6% to 8.4% mortality observed among patients not requiring readmission (Mcneill and Khairat 2020). Furthermore, hospitals in United States have a financial motivation to reduce the rate of premature discharge, as they can be penalized through the Hospital Readmission Reduction Program (HRRP) (Centers for Medicare & Medicaid Services 2024). HRRP penalizes the majority of hospitals for substandard readmission rates up to a maximum penalty rate of 1% of the medicare base payments to the hospital (Wasfy et al. 2017, James 2013). Therefore, elevated readmission rates due to premature patient discharge not only create health risks for patients and result in extended total ICU stays (with accompanying medical resource strain), but also subject hospitals to administrative penalties and economic losses. When a patient experiences multiple ICU readmissions within a given time-window, these readmission-related harms are naturally amplified. As such, ICU discharge decision-making should not only aim to limit the readmission risk but also account for patient's admission history - distinguishing among first-time ICU admission, initial readmission, and multiple readmissions.

On the other hand, prolonged ICU stay could harm patients. Studies show that prolonged ICU stays could lead to some mental health issues like anxiety or depression (Hatch et al. 2018, Garcez-Leme and Avelino-Silva 2023). Additionally, prolonged ICU stays can strain capacity, delay new admissions, and lead to inefficient use of critical resources such as staff time and equipment. This makes timely discharge decisions essential, particularly in high-demand hospital settings. One study notes that patients having 21+ days ICU length-of-stay (LOS) used 11.6% of bed-days, showing how much this impacts resource allocation (Moitra et al. 2016). The dilemma corresponding to the timing of discharge could become even more pressing during the public health crisis, e.g., COVID-19 pandemic, which has imposed a dramatically higher burden on critical care departments worldwide (Cohen et al. 2020).

Therefore, balancing the ICU LOS against the readmission risk represents a critical challenge in critical care management. This study presents a decision support framework that uses electronic health record (EHR) data, especially patients' physiological measurements collected during their ICU stay, to help clinicians determine the most appropriate time for discharge. By incorporating individual patient trajectories, our method generates personalized and risk adjusted recommendations for discharge that emphasize both safety and efficient use of resources. In this way, patients leave the ICU at a suitable time, improving individual outcomes and enhancing overall ICU efficiency.

Discharge decision-making in ICUs is frequently modeled as a sequential decision process using the Markov Decision Process (MDP) framework, which requires the

specification or estimation of key components such as transition probabilities and cost functions (Kreke et al. 2008, Chan et al. 2012, Ouyang et al. 2020, Shi et al. 2021, Lejarza et al. 2023). MDP also serves as the foundation for formulating reinforcement learning (RL) problems, which offers a model-free alternative that bypasses the need for explicit transition estimation. Due to its scalability and adaptability, RL has shown promise in addressing a range of complex problems in medical decision-making. For instance, Lee et al. (2015) develops and evaluates RL–based screening policies for allocating limited healthcare resources across at-risk patients, with a case study on Hepatocellular Carcinoma, demonstrating improved performance over current practices under varying resource constraints. Guo et al. (2022) introduce a RL framework for pandemic control, leveraging large-scale agent-based simulation and sequence learning to generate adaptive policies that effectively balance public health outcomes and economic costs. In the context of treatment optimization, Zhang et al. (2024) propose a goal-oriented hierarchical RL method that decomposes long-term therapeutic objectives into short-term subgoals, achieving improved drug dosing policies for sepsis patients and demonstrating a significant reduction in patient mortality. In these papers, researchers have used a variety of RL methods. In practical applications, RL has many variants and is often combined with other methods in machine learning. For the ICU discharge decision-making problem we are investigating, we will use offline constrained reinforcement learning (OCRL). OCRL aims to learning a decision making policy that performs well while satisfying constraints given a historical dataset.

It firstly enjoys the benefits of offline RL – not requiring interaction with the environment enables real-world application where collecting interaction data is expensive or dangerous, e.g., critical care decision-making. It could be both unethical and dangerous to try the policy recommended by the RL agent on patients to receive feedback. Also, it can enjoy the benefits of constrained RL – being able to specify constraints to the behavior of agents enables real-world applications with constrained concerns. As previously mentioned, clinicians must consider the safety requirements about readmission risk and the resource constraints in ICUs during the discharge decision-making process. Constrained RL could provide an effective approach for incorporating safety and resource constraints into the optimization problem. Moreover, we utilize the model-free structure, which offers a distinct advantage by circumventing the explicit estimation of transition probabilities and reward/cost functions typically required by traditional Dynamic Programming (DP) methods, making it especially valuable in healthcare settings where estimating these parameters is difficult due to the complexity and high dimensionality of the system dynamics.

As for constrained RL, it is common to model the decision problem under study using Constrained Markov Decision Process (CMDP). As a variant of the standard MDP, CMDP enables the incorporation of additional constraints in sequential decision-making problems, beyond the primary objective—such as minimizing mortality risk in the context of discharge decision-making. By introducing one or more constraints, this approach allows for a comprehensive consideration of multiple objectives during the decision-making process. For instance, in the discharge decision-making problem, it is essential to manage not only the readmission risk but also to ensure that the patient's LOS in the ICU remains within a specified range.

Our work builds on recent advances in OCRL, particularly the value-based frameworks proposed by Le et al. (2019) and Hong et al. (2024). While these methods offer strong theoretical guarantees, their practical application to complex, real-world settings such as ICU discharge decision-making problem remains limited due to both high computational demands and reliance on mixed policy formulations.

To address these challenges, we propose a streamlined synchronous multi-timescale architecture that integrates policy learning, evaluation, and constraints handling into a unified update process. This design reduces training overhead and eliminates the need to store historical policies, making the approach more practical for large-scale real-world medical decision-making applications. We validate the efficiency and effectiveness of our approach through experiments on a real-world medical dataset. This work makes the following contributions:

- We introduce a novel framework that incorporates multiple ICU readmissions into the discharge decision-making process, rather than treating each admission episode independently. To the best of our knowledge, this is the first study to explicitly model patient trajectories with repeated ICU stays in the context of sequential decision-making. While data sparsity remains a challenge—particularly for patients with higher readmission counts—our formulation enables a more holistic and temporally aware representation of patient health status and its impact on downstream decisions.
- We are the first to formulate the ICU discharge planning problem within an offline RL paradigm. Our approach models the task as a sequential decision process over a continuous state space constructed from patients' raw physiological measurements, without requiring discretization or transition probability estimation. This departs from traditional rule-based or score-based clinical tools such as SOFA or APACHE II/IV, enabling a more flexible, data-driven, and individualized representation of patient states.
- To accommodate multiple clinical objectives, we adopt the framework of constrained RL and formulate the problem as a CMDP. This formulation eliminates the need to encode competing clinical goals into a single composite cost function, which typically requires extensive hyperparameter tuning. Instead, CMDP offers a principled and interpretable mechanism for balancing objectives such as mortality risk, readmission risk, and ICU LOS.
- We validate our method using a real-world ICU dataset and benchmark its performance against historical clinical practices. Our experimental results show that the proposed approach effectively reduces mortality risk while simultaneously satisfying constraints on readmission risk and ICU LOS. These results highlight the potential of our method as a practical and clinically meaningful decision-support tool for guiding safe, efficient, and personalized ICU discharge decision-making.

The remainder of this paper is organized as follows. Section 2 reviews related work on ICU discharge decision-making and offline constrained reinforcement learning. Section 3 formalizes the ICU discharge decision-making problem as a CMDP. Section 4 presents the proposed offline constrained RL algorithm and its key components. In Section 5, we report the results of our numerical experiments and compare our approach with clinical baselines. Finally, Section 6 concludes the paper.

2. Literature Review

2.1. Hospital Discharge Decision-Making

Discharge decision-making is among the most crucial and challenging processes in intensive care units (ICUs) and hospitals. It has received significant attention in medical

research, with numerous empirical studies aiming to understand and improve discharge practices.

On one hand, a significant body of empirical research has thoroughly investigated discharge decision-making processes within hospital settings. For example, Long and Mathews (2018) demonstrate that ICU occupancy influences non-essential boarding time but does not affect the medically necessary length-of-stay (LOS). While early discharge can mitigate congestion and improve the availability of ICU beds for incoming patients, it often comes at the expense of patient outcomes. Using a large dataset of patients with congestive heart failure, Oh et al. (2018) find that inpatient stays shorter than the Centers for Medicare and Medicaid Services (CMS)-recommended LOS are associated with a 1.1% increase in readmission risk. Similarly, Kc and Terwiesch (2009, 2011) report that patients discharged prematurely face heightened risks of readmission, mortality, and other adverse outcomes. Medical literature further corroborates these findings, linking shorter LOS with poorer patient outcomes (Heggestad 2002, Kuo and Goodwin 2011). Conversely, Bartel et al. (2019) find that extending a patient's stay by just one additional day can reduce mortality risk by nearly 6.0%. Moreover, Carey (2015) and Oh et al. (2018) suggest that keeping patients longer during their initial hospital admission can significantly reduce the total readmission costs. These empirical findings form the foundation of our study, motivating an exploration of the trade-offs between LOS, patient outcomes, and resource utilization.

On the other hand, Markov Decision Process (MDP) and Dynamic Programming (DP) methods have been extensively applied to model discharge decision-making in healthcare. For instance, Kreke et al. (2008) firstly introduce a finite-horizon MDP framework aimed at determining optimal discharge strategies for patients with pneumonia-related sepsis to maximize their expected survival rates. Through structural and computational analysis, the study reveals that the optimal discharge strategy follows a non-stationary control-limit-type policy, implying that the level of illness at which it is optimal to discharge a patient changes over the course of the hospital stay. Chan et al. (2012) evaluates the effects of various ICU discharge policies under uncertainty, showing that incorporating predictive models for readmission risk with straightforward index-based policies can enhance patient throughput while maintaining or reducing mortality rates in ICUs with limited capacity. In another study, Ouyang et al. (2020) simultaneously considers ICU admission and discharge decisions in the MDP model to minimize the long-term average mortality. By allowing state transitions that reflect patient condition improvements or deteriorations, they characterize the optimal policy under scenarios where shorter ICU LOS can enhance patient outcomes, as well as circumstances where achieving better outcomes necessarily involves prolonged ICU LOS. Similarly, Shi et al. (2021) introduce a data-driven decision support framework for hospital discharge that integrates personalized readmission prediction with an MDP model to balance readmission risk and ward congestion. This approach demonstrates significant benefits, both through practical deployment in a partner hospital and extensive high-fidelity simulations. In turn, Alaeddini et al. (2019) sharpen the prediction side of this problem, and they propose a granular mixture KPCA Cox model that traces patient-specific readmission risk trajectories over the LOS. Lejarza et al. (2023) present a data-driven prescriptive framework that employs a clustering-based methodology to identify discrete patient health states from high-dimensional continuous electronic health record (EHR) data. Chuang et al. (2023) develops a MDP-based discharge decision model incorporating machine learning-driven predictive analytics to optimize long-term care (LTC) placements for older delayed discharge patients, reducing hospital costs by personalizing decisions based on patient health trajectories and

characteristics. Unlike the MDP-driven approaches above, Khatami et al. (2021) address discharge operations with a two-stage stochastic program that synchronizes bed releases and incoming admissions under uncertain processing times, thereby cutting discharge lateness and patient boarding.

Our approach employs a Constrained Markov Decision Process (CMDP) to model the ICU discharge decision-making process, exploring the trade-offs among patient's ICU LOS, patient health outcomes, and potential hospital payments. Moreover, our ICU discharge decision-making process accounts for the possibility of patients being readmitted to the ICU, meaning that discharge does not lead to an absorbing state. By incorporating the readmission count into the patient's state space, we enable the learning of distinct discharge policies tailored to different levels of readmission history. We further adopt a reinforcement learning (RL) approach to directly learn the discharge policy from various patient physiological data available in a real-world medical dataset. Consequently, this eliminates the necessity to construct a discrete state space using severity scoring systems (e.g., APACHE II, SAPS, SOFA) or clustering methods. In the next section, we will review the literature related to the RL techniques employed in our study.

2.2. Offline Constrained Reinforcement Learning

Reinforcement Learning (RL) has demonstrated significant potential for sequential decision-making across diverse domains, including robotics (Tang et al. 2024), safety assessment (Aghalari et al. 2021), livestream shopping (Liu 2022) and Large Language Models (DeepSeek-AI 2025). However, applying RL in high-stakes environments, such as ICU discharge decision-making, introduces two major challenges. First, clinical decisions must satisfy stringent constraints related to safety, resource utilization, or regulatory compliance. These requirements naturally lead to the use of Constrained Markov Decision Processes (CMDPs) (Altman 1999), which extend standard MDP frameworks by incorporating additional risk or cost functions that must remain below predefined thresholds. For smaller-scale problems, CMDPs can be solved using dynamic programming (Labbi and Berrospi 2007) or linear programming methods (Altman 1999). In contrast, large-scale applications typically necessitate model-free RL approaches that learn optimal policies directly from interactions with the environment, as explicitly modeling all state transitions is often infeasible.

The second challenge arises from the ethical and practical limitations of exploration in clinical settings, where the safety of patients is critical. Consequently, RL agents in these contexts must rely on pre-collected offline datasets rather than interactively sampling new data. This scenario has given rise to offline RL (Lange et al. 2012, Fujimoto et al. 2019, Kumar et al. 2019, 2020), where the objective is to learn an optimal policy solely from historical trajectories.

Although there is an extensive body of research on both offline RL and constrained RL individually, their integration into Offline Constrained Reinforcement Learning (OCRL) has received comparatively less attention. The Constrained Batch Policy Learning (CBPL) framework (Le et al. 2019) was among the first to tackle offline constrained policy learning by employing Fitted-Q Evaluation (FQE) to assess constraint violations and Fitted-Q Iteration (FQI) for policy improvement, all within a game-theoretic framework. Notably, CBPL is the only approach to date that offers provable sample efficiency for OCRL with function approximation. Building upon this, Hong et al. (2024) relax the Bellman completeness assumption while still relying on

mixed policy methods that necessitate storing all historical policies—a requirement that leads to memory costs growing linearly with the number of training steps, which is impractical in large-scale settings. Addressing this limitation, Cai et al. (2023) adapt the CBPL method for marketing budget allocation, significantly reducing the number of policies that need to be stored.

Beyond these theoretically grounded approaches, several practical algorithms for OCRL have emerged without formal guarantees. For instance, Lee et al. (2022) propose COptiDICE, an algorithm inspired by linear programming formulations of RL, while Liu et al. (2023) adapt the decision-transformer framework for the OCRL context through their CDT method. Also, Fang et al. (2024) propose offline Constraint Transformer (CT) for healthcare, using past patients' data to ensure safe medical decisions, like avoiding dangerous drug doses. Additionally, Xu et al. (2022) introduce CPQ, a Q-learning-based algorithm that penalizes out-of-distribution actions.

In summary, OCRL focuses on developing decision-making policies that perform effectively while satisfying essential constraints, using only historical trajectory data.

Our work builds on the foundational contributions of Le et al. (2019). Despite the theoretical advancements, directly applying its value-based OCRL algorithm to complex, high-stakes applications such as ICU discharge planning poses significant computational challenges. In particular, the combination of FQI for policy learning and FQE for off-policy evaluation (OPE) results in substantial training overhead, requiring extensive hyperparameter tuning and rendering the method less practical for real-world deployment. Moreover, both Le et al. (2019) and Hong et al. (2024) rely on a mixed policy formulation that necessitates maintaining and storing the full sequence of historical policies, incurring considerable memory and computational costs as training progresses.

To address these limitations, we introduce a synchronous architecture within the value-based OCRL framework, in which the parameters of FQI and FQE are updated jointly at each training step. The dual variable is also updated synchronously, eliminating the need for separate optimization schedules. This unified update scheme substantially reduces the computational burden associated with both training and hyperparameter tuning. Furthermore, by formulating the CMDP constraints in terms of cumulative costs, our method avoids the need for a mixed policy representation, both in theory and in implementation. We demonstrate the effectiveness and efficiency of our proposed approach through extensive numerical experiments using a real-world clinical dataset.

3. Model Formulation

In this section, we illustrate the details about formulating an indefinite-horizon discrete-time Constrained Markov Decision Process (CMDP) for the discharge decision-making problem in the ICU. Table 1 provides a summary of the mathematical notations used in this section.

State Space. To represent the physiological condition of a patient at each decision epoch t, let x_t denote the patient's physiological state. The physiological state representation incorporates variables from demographics (such as age, gender, and weight), vital signs (such as Glasgow Coma Scale, heart rate, and blood pressure), and laboratory measurements (such as blood urea nitrogen, partial thromboplastin time, and pH). Detailed descriptions of all physiological variables used in the analysis are provided in Table A1. In addition to capturing physiological status, the state

Table 1.: Mathematical notation used in the MDP and CMDP formulations.

Notation	**Definition**
t	Decision epoch
s_t	Patient state at decision epoch t
s_D	Terminal state indicating patient death after ICU discharge
s_H	Terminal state indicating successful ICU discharge without readmission or death
x_t	Physiological condition at decision epoch t
r_t	Readmission count at decision epoch t
a_t	Action taken at decision epoch t
$P^{a_t}(s_{t+1} \mid s_t)$	Transition probability from state s_t to s_{t+1} given action a_t
$c(s_t, a_t, s_{t+1})$	Instantaneous objective cost function
$\boldsymbol{g}(s_t, a_t, s_{t+1})$	Vector of instantaneous constraint costs
π	Policy over the decision-making process

space incorporates the number of times the patient has been readmitted, denoted as r_t (readmission count). Thus, the complete state of a patient at time t is represented by the vector $s_t = (x_t, r_t)$.

Furthermore, the state space incorporates two terminal states that mark the conclusion of the decision-making process. The first, denoted as s_D, signifies that the patient has died following a discharge decision made by the clinician. The second terminal state, denoted as s_H, corresponds to a successful discharge, defined as a scenario where the patient neither dies nor is readmitted to the ICU post-discharge. Upon reaching either of these terminal states, the decision-making process for that patient is terminated.

Action Space. At each decision epoch t, the clinicians must decide whether the patient is ready for discharge from the ICU. The action a_t is drawn from the action set $\mathcal{A} = \{0, 1\}$, where $a_t = 0$ signifies the decision to retain the patient in the ICU for at least one more period, ensuring their continued stay at decision epoch $t + 1$. Conversely, $a_t = 1$ denotes the decision to discharge the patient from ICU.

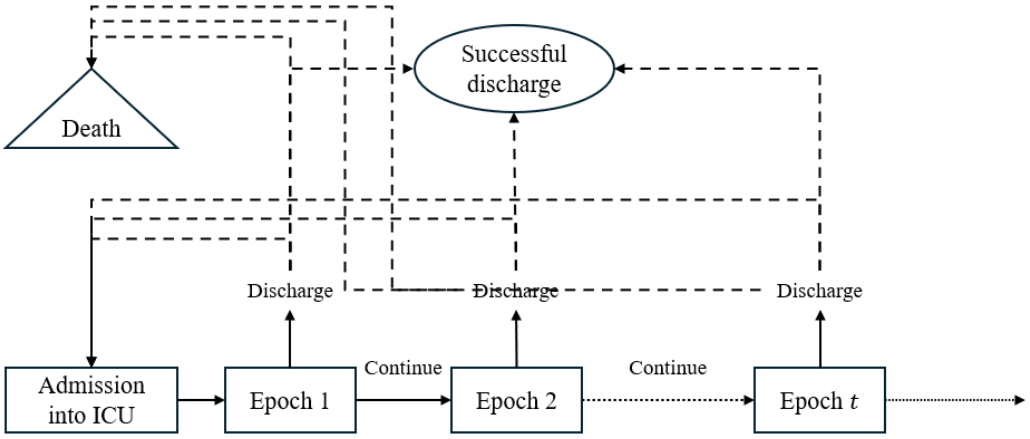


Figure 1.: An illustration of the discharge decision-making process in the ICU.

Transition Probability. The whole discharge decision-making process is shown as Figure 1. As illustrated in Figure 1, ICU discharge decision-making process ceases

once either of the two terminal state (death or successful discharge) is reached. Let $P^{a_t}(s_{t+1}|s_t)$ denote the probability that the patient will be in state $s_{t+1} \in \mathcal{S}$ at decision epoch $t+1$, given that the patient is in state s_t at decision epoch t and the chosen action is $a_t \in \mathcal{A}$. This probabilistic framework underpins the modeling of patient outcomes across successive decision points, capturing the dynamic nature of ICU care.

For each decision epoch $t = 1, 2, \ldots$, there are following four types of possible transitions:

- $P^{a_t=0}\Big(s_{t+1} = (x_{t+1}, r_{t+1} = r_t) \Big| s_t = (x_t, r_t)\Big)$ (the patient is staying in the ICU).
- $P^{a_t=1}\Big(s_{t+1} = s_H \Big| s_t = (x_t, r_t)\Big)$ (the patient is successfully discharged from the ICU).
- $P^{a_t=1}\Big(s_{t+1} = (x_{t+1}, r_{t+1} = r_t + 1) \Big| s_t = (x_t, r_t)\Big)$ (the patient is readmitted into the ICU).
- $P^{a_t=1}\Big(s_{t+1} = s_D \Big| s_t = (x_t, r_t)\Big)$ (the patient is deceased after being discharged from the ICU).

Objective Costs. In our study, we prioritize managing mortality risk as the primary objective cost when addressing ICU discharge decision-making. Mortality risk stands out as a vital clinical metric, steering clinicians toward optimal timing for patient discharge (Ouyang et al. 2020). We define the instantaneous objective cost function as follows

$$c(s_t, a_t, s_{t+1}) = \begin{cases} 1 & \text{if } a_t = 1 \text{ and } s_{t+1} = s_D, \\ 0 & \text{o.w.}, \end{cases} \tag{1}$$

where s_D denotes the terminal state of death, as established earlier. This function assigns a penalty of 1 exclusively when a discharge action ($a_t = 1$) results directly in the patient's death, encapsulating the critical outcome we strive to prevent. In all other cases, the cost would be 0, stating that our goal for minimizing mortality risk as the cornerstone of successful discharge decision-making.

Constraint Costs. In our decision-making framework for ICU discharge, we incorporate constraint costs to ensure that our policies align with essential quality and efficiency metrics. At each decision epoch t, the clinician's action a_t based on the patient's current state s_t and subsequent state s_{t+1}, generates a constraint cost vector $\boldsymbol{g}(s_t, a_t, s_{t+1}) \in \mathbb{R}^2$. This vector captures the impact of the decision on two critical constraints specific to this study: the readmission rate and the length-of-stay in the ICU.

The first constraint focuses on the readmission risk, a key indicator of performance-based payment scheme. Controlling this metric is important for preventing severe penalty for excessive readmission rates. The instantaneous readmission risk cost function is defined as

$$g_R(s_t, a_t, s_{t+1}) = \begin{cases} 1 & \text{if } a_t = 1 \text{ and } r_{t+1} = r_t + 1, \\ 0 & \text{o.w.}, \end{cases} \tag{2}$$

where $a_t = 1$ denotes a decision to discharge the patient from the ICU, and $r_{t+1} = r_t + 1$ indicates a readmission following that discharge. This formulation assigns a cost of 1

when a discharged patient is readmitted in to the ICU, reflecting the payment penalty, and 0 otherwise.

To complement the control of readmission risk, we also impose a constraint on the ICU length-of-stay (LOS), a critical indicator of operational efficiency in intensive care management. Constraining LOS promotes optimal utilization of ICU resources and supports improved patient throughput. This objective aligns with fast-track protocols in critical care, which encourage timely discharge decisions to reduce unnecessary ICU occupancy (Sato et al. 2009, Fitch et al. 2014). The instantaneous ICU LOS cost function is defined as

$$g_L\left(s_t, a_t, s_{t+1}\right) = \begin{cases} \Gamma_{r_t} & \text{if } a_t = 0, \\ 0 & \text{o.w.}, \end{cases} \tag{3}$$

where Γ_{r_t} denotes the continuation cost, which varies according to the patient's readmission count r_t. This cost captures the time spent in the ICU during a decision interval and is incurred only when the patient is kept in the ICU ($a_t = 0$). Once a discharge decision is made ($a_t = 1$), the instantaneous LOS cost for that interval drops to 0.

By integrating these two constraints – readmission risk and ICU LOS – our framework balances the dual objectives of patient safety and operational efficiency, adapting to the practical realities of data availability while aligning with established clinical management goals.

Objective. At each decision epoch t, the clinician could observe the patient's state $s_t = (x_t, r_t)$, where x_t denotes the current physiological condition and r_t is the cumulative readmission count. A stationary policy, represented by π, determines whether the patient is ready for discharge from the ICU. Thus, if we do not consider any constraints during the discharge decision-making process, the goal would be minimizing the expected mortality risk and the optimization problem is

$$\min_{\pi} \quad C\left(\pi\right), \tag{4}$$

where $C(\pi)$ is defined by

$$C\left(\pi\right) = \mathop{\mathbb{E}}_{s_0 \sim \chi, \pi}\left[\sum_{t=0}^{\infty} c\big(s_t, \pi(s_t), s_{t+1}\big)\right], \tag{5}$$

and χ denotes the distribution of initial states. The function $C(\pi)$ corresponds to the value (cost) function $C^{\pi}(s)$ averaged over the initial state distribution χ. The ∞ indicates the indefinite-horizon formulation of the MDP (and CMDP) for discharge decision-making process. To ensure that the cumulative cost is properly defined in the absence of discount factor, we impose a structural assumption (Assumption B.1) on the termination behavior of the decision-making process in Appendix B.

The analysis of the relationship between readmission count r and mortality risk cost function $C^{\pi}(s)$ is also included in Appendix B. The analysis indicates that if clinicians (the decision maker) aim to minimize the expected cumulative mortality risk $C(\pi)$ as the sole objective in ICU discharge decision-making, that is, solving optimization problem (4) without any constraints, then ICU readmission (increases in r) should be avoided. Because prolonged ICU LOS can reduce both mortality and readmission

risks (Shi et al. 2021), the patients should stay in the ICU as long as possible until the mortality risk (transiting into the terminal state s_D) is minimized.

However, in the CMDP framework we propose, minimizing expected mortality risk is not the sole objective. The goal is to learn an optimal policy that minimizes the expected cumulative objective cost while ensuring that the expected cumulative constraint costs related to both readmission risk and ICU LOS remain within their respective thresholds. This can be formulated as follows

$$\begin{aligned} \min_{\pi} \quad & C(\pi) \\ \text{s.t.} \quad & G_L(\pi) \leq l_{los} \\ & G_R(\pi) \leq l_{rr}, \end{aligned} \tag{6}$$

$$G_L(\pi) = \mathop{\mathbb{E}}_{s_0 \sim \chi, \pi} \left[\sum_{t=0}^{\infty} g_L\big(s_t, \pi(s_t), s_{t+1}\big) \right], \quad G_R(\pi) = \mathop{\mathbb{E}}_{s_0 \sim \chi, \pi} \left[\sum_{t=0}^{\infty} g_R\big(s_t, \pi(s_t), s_{t+1}\big) \right], \tag{7}$$

where $G_L(\pi)$ and $G_R(\pi)$ represent the expected cumulative constraint costs associated with ICU LOS and readmission risk, respectively, and l_{los} and l_{rr} are the corresponding given constraint thresholds. Similar to $C(\pi)$, $G_L(\pi)$ and $G_R(\pi)$ also correspond to the value functions averaged over the initial state distribution χ. $G_L(\pi)$ represents the expected cumulative ICU LOS throughout the entire decision-making process under policy π. This aligns with the definition of instantaneous ICU LOS cost in (3), where once a discharge action ($a = 1$) is taken and the patient leaves the ICU (temporarily or permanently), no instantaneous ICU LOS cost is incurred, whereas if the patient is kept in the ICU, the time interval until the next decision point is added into the cumulative ICU LOS. Similarly, $G_R(\pi)$ represents the expected cumulative readmission risk throughout the entire decision-making process under policy π. The interpretation is quite straightforward. Based on the definition of instantaneous readmission risk cost (2), each readmission event incurs a penalty cost of 1, incrementing the readmission counter. Therefore, $G_R(\pi)$ serves as a proxy for readmission risk by quantifying the expected number of readmission events under the policy π.

Firstly, consider the scenario in which only the ICU LOS constraint is enforced to manage critical care resource allocation. In this setting, certain patients may be discharged earlier, i.e., taking action ($a_t = 1$) in order to ensure that the expected cumulative LOS constraint cost $G_L(\pi)$ remains below the given threshold l_{los}. However, this earlier discharge may lead to a rise in the expected cumulative objective cost $C(\pi)$, which reflects the overall mortality risk across all initial states. Additionally, because a shorter ICU stay increases the likelihood of clinical deterioration post-discharge, the expected cumulative readmission risk cost $G_R(\pi)$ is also likely to increase under this constrained policy.

Secondly, motivated by institutional pressures such as the Hospital Readmissions Reduction Program (HRRP), it becomes necessary to simultaneously control the expected readmission risk by introducing a constraint threshold l_{rr}. While reducing readmission counts is often associated with improved long-term outcomes, this intuition does not universally apply. In certain cases, readmission offers critically ill patients a renewed opportunity to receive life-saving treatment. As previously defined, ICU readmission and transition to the terminal state s_D are mutually exclusive events –

they cannot occur simultaneously. Therefore, integrating constraints on both LOS and readmission risk may increase the expected mortality risk.

4. Algorithm Design

In this section, we present the algorithm framework underlying our approach, offering a detailed explanation of each step while emphasizing its key characteristics. The proposed method synthesizes fundamental principles from both Offline Reinforcement Learning (offline RL) and Constrained Reinforcement Learning (constrained RL), enabling effective policy learning under multiple operational and clinical constraints. We begin by providing a clear overview of the essential components of the framework, and then systematically analyze its design properties and advantages.

4.1. Primal-Dual Iteration Algorithm

We solve the constrained optimization problem (6) using Primal-Dual Iteration (PDI) based on Lagrangian relaxation (Altman 1999). The Lagrangian function is

$$L(\pi, \boldsymbol{\lambda}) = C(\pi) + \boldsymbol{\lambda}^\top \big(G(\pi) - \boldsymbol{l}\big), \tag{8}$$

where $\boldsymbol{\lambda} \in \mathbb{R}_+^d$ are the Lagrangian multipliers. Problem (6) then becomes equivalent to the saddle-point problem

$$\min_{\pi} \max_{\boldsymbol{\lambda}} L(\pi, \boldsymbol{\lambda}). \tag{9}$$

The dual function $d(\boldsymbol{\lambda}) = \min_\pi L(\pi, \boldsymbol{\lambda})$ provides a lower bound $P^* \geq d(\boldsymbol{\lambda})$ for all $\boldsymbol{\lambda} \in \mathbb{R}_+^d$ (Boyd and Vandenberghe 2004). The dual problem

$$D^* = \max_{\boldsymbol{\lambda}} d(\boldsymbol{\lambda}) \tag{10}$$

seeks the tightest bound. Since $d(\boldsymbol{\lambda})$ is concave (being the point-wise minimum of linear functions), gradient-based PDI can efficiently solve (10).

While PDI solves the dual problem, recovering the primal solution requires addressing the duality gap $P^* - D^*$. Strong duality (zero gap) would make problems (6) and (10) equivalent, but requires additional assumptions beyond our non-convex setting. We establish these assumptions in the following.

Assumption 4.1. *Slater's condition could be satisfied for all constraints, i.e.,* $\exists \pi$ *s.t.* $G(\pi) < \boldsymbol{l}$.

Assumption 4.2. *There exists constant* $g_{\max}$ *such that for any* $s \in \mathcal{S}$, $a \in \mathcal{A}$, $g(s, a) \leq g_{\max}$.

Assumption 4.3. *The constraint costs* $G(\pi)$ *of problem (6) are in the format of cumulative constraints.*

All three assumptions hold in our ICU discharge decision-making problem. First, Assumption 4.1 is satisfied as the constraint thresholds for both readmission risk and ICU length-of-stay (LOS) in our numerical experiments (Section 5) are set based on

values achievable in clinical practice within the dataset. Assumption 4.2 holds by the definitions of the instantaneous readmission risk cost function (2) and instantaneous ICU LOS cost function (3) in Section 3, where both instantaneous constraint costs are bounded. Finally, Assumption 4.3 is satisfied as shown in our definition of the two CMDP constraints ($G_L(\pi)$ and $G_R(\pi)$) in (7) of Section 3. While CMDPs can incorporate various constraint forms, both constraints studied here are expected cumulative constraints.

Theorem 4.4 (Theorem 3 in Paternain et al. 2023)**.** *Given Assumptions 4.1, 4.2, and 4.3, strong duality holds for (6), meaning that $P^* = D^*$.*

Therefore, we can interchange the order of max and min in (9):

$$\min_{\pi} \max_{\boldsymbol{\lambda}} L(\pi, \boldsymbol{\lambda}) = \max_{\boldsymbol{\lambda}} \min_{\pi} L(\pi, \boldsymbol{\lambda}). \tag{11}$$

With this equivalent unconstrained formulation, we apply PDI through alternating updates of the primal policy π and dual variables $\boldsymbol{\lambda}$. Unlike the mixed-policy approach of Le et al. (2019) (Algorithm 4), our method does not require primal problem convexity, enabling simpler policy representation and reduced memory usage. Offline RL algorithms approximate the policy π from historical datasets (detailed in Section 4.2), while Off-Policy Evaluation (OPE) methods evaluate constraint costs for updating dual variables (Section 4.3).

4.2. Offline Reinforcement Learning

The first critical component in our RL framework is the deployment of a value-based, model-free offline RL algorithm – Fitted Q-Iteration (FQI) (Ernst et al. 2005). FQI is a batch offline RL algorithm that iteratively fits a function approximator to Bellman backup targets. To begin, for both the objective and constraint costs ($c(\cdot)$ is used to represent these instantaneous costs) in the discharge decision-making process modeled in Section 3, the Bellman backup targets (bootstrapped targets) in the Q-Learning process are defined as

$$y = \begin{cases} c\left(s_t, a_t, s_{t+1}\right) & s_{t+1} = s_H \text{ or } s_D, \\ c\left(s_t, a_t, s_{t+1}\right) + \min\limits_{a \in \mathcal{A}} Q\left(s_{t+1}, a \middle| \theta\right) & \text{o.w.}, \end{cases} \tag{12}$$

where the objective is to minimize a mean squared error (MSE) loss function, expressed as

$$\mathcal{L}\left(\theta\right) = \frac{1}{n} \sum_{j=1}^{n} \left[y_j - Q\left(s_t^j, a_t^j \middle| \theta\right) \right]^2, \tag{13}$$

where n is the number of samples used for updating the parameters θ, and j denotes the index of the j-th sample. Conceptually, due to inherent distribution shift, the bootstrapped target y may suffer inaccuracies, particularly because the action minimizing the Q-value is itself bootstrapped. This can inadvertently drive the policy towards out-of-distribution (OOD) actions, where large positive extrapolation errors are more likely to occur.

Regulation techniques in machine learning could help prevent model overfitting given a finite amount of available data. Since offline RL bears greater similarity to supervised learning than online RL, it conceptually more aligns with the use of regulation techniques. As we are going to deploy the linear approximation method for approximating the Q-value in our offline RL part, we simply add the L2 regularization (weight decay) term to the loss function. The loss function now can be expressed as

$$\mathcal{L}_{reg}(\theta, \omega) = \mathcal{L}(\theta) + \omega \|\theta\|_2^2, \tag{14}$$

where ω controls the weight of penalty. Incorporating this term effectively reduces the susceptibility of our model to overfit outlier or extreme data points within the offline dataset. Additionally, we utilize mini-batch gradient descent updates rather than batch gradient descent, further enhancing computational efficiency and stability during training. The detailed procedure for the FQI algorithm incorporating L2 regularization for discharge decision-making is presented in Algorithm 2.

A fundamental challenge in offline RL is addressing the distribution shift that occurs during the training process. Fujimoto et al. (2019) (BCQ) and Kumar et al. (2019) (BEAR) tackled this challenge in the policy improvement phase by ensuring the learned policy π_{RL} remains close to the behavior policy π_D (i.e., clinical practices in the discharge decision-making context) using a distance metric $d(\pi_{RL}, \pi_D)$. Kumar et al. (2020) also developed a conservative method - Conservative Q-Learning (CQL), which is conservative in the estimation of state-action-values that are not in the dataset.

However, when modeling discharge decision-making as a CMDP, we face a more complex scenario. Our formulation incorporates two constraint costs (readmission risk and ICU LOS) in addition to the primary objective cost of mortality risk. These constraints inherently limit the RL agent's ability to select out-of-distribution actions. For instance, in an unconstrained MDP, an agent focused solely on minimizing mortality risk may invariably extend ICU stays, predominantly making decisions of staying in the ICU ($a = 0$). Therefore, the CMDP structure of discharge decision-making could mitigate the distribution shift challenge. Nevertheless, we still explore the application of a modified CQL approach rather than FQI for Q-value approximation and discharge policy derivation. The technical details of our modified CQL formulation and comprehensive numerical results are presented in Appendix I.

4.3. Fitted Q Evaluation

Fitted-Q-Evaluation (FQE) is a model-free Off-Policy Evaluation (OPE) method used in offline RL to estimate the Q-values for a specific policy π, and it can be treated as the substitution of agent-environment interaction part in online RL. In the setting of discharge decision-making problem, the Q-value represents the prediction of the expected cumulative cost (including bothe objective and constraint costs) from a given state-action pair $(s = (x, r), a)$ under the policy derived by offline RL agent (e.g., FQI algorithm). It works by using a historical dataset of trajectories to approximate the Q-value without needing to interact with the environment. Thus, FQE is particularly useful for safe policy evaluation in healthcare (Tang and Wiens 2021). While Le et al. (2019) introduced the original FQE algorithm, we present in Algorithm 3 a refined variant that employs mini-batch gradient descent updates instead of the batch gradient descent approach used in the original formulation — a modification paralleling our adaptations to the FQI algorithm.

The FQE algorithm finally outputs $\hat{Q}^{\pi}_{FQE}\left(s=(x,r),a\right)$, estimated Q-function of π. The final estimated value (constraint cost) can be calculated as the average value for the initial states s_0 in historical dataset $\mathcal{D}$

$$\hat{G}\left(\pi\right)=\frac{1}{n_e}\sum_{i=1}^{n_e}\hat{Q}^{\pi}_{FQE}\Big(s_0=(x_0,r_0),a\Big), \tag{15}$$

where n_e is the number of episodes in dataset $\mathcal{D}$, and $a=\pi\left(s_0\right)$ because we are considering deterministic policy. Similarly, we can estimate the objective cost $\hat{C}\left(\pi\right)$. Therefore, in offline RL, both $C\left(\pi\right)$ and $G\left(\pi\right)$ must be estimated using the FQE method as $\hat{C}\left(\pi\right)$ and $\hat{G}\left(\pi\right)$. Then, according to the optimization problem (6), in practice, we should have

$$\hat{G}\left(\pi\right)\leq l, \tag{16}$$

where l could be the given threshold for either readmission risk or ICU LOS. However, the estimation process in FQE often introduces fluctuations, which suggests the need for a more cautious approach to constraint formulation.

To address this, it becomes prudent to adopt a conservative constraint setting. We can reformulate it as a chance constraint offers a more robust solution. For example, if we aim to regulate the upper bound at a 98% confidence level, the constraint can be expressed as

$$\hat{G}\left(\pi\right)\leq l\Rightarrow\hat{G}\left(\pi\right)+\Delta\left(\pi\right)\leq l, \tag{17}$$

where $\Delta\left(\pi\right)$ is an upper confidence bound (UCB) that depends on both the sample variance and sample size. If we deploy normal approximation, it could be calculated as

$$\Delta\left(\pi\right)\approx z\cdot\hat{\sigma}, \tag{18}$$

where z-value is used to control the confidence interval (CI), and $\hat{\sigma}$ is the estimated standard error from samples. $\hat{\sigma}$ is always calculated as $\hat{\sigma}=\frac{s}{\sqrt{n_e}}$, with s being the sample standard deviation of the $\hat{Q}^{\pi}_{FQE}\left(s_0,a\right)$ values over the n_e initial states.

This transformation enhances the reliability of the constraint under estimation variability. $\Delta\left(\pi\right)$ represents the confidence margin or uncertainty adjustment added to the estimated constraint cost $\hat{G}\left(\pi\right)$ to account for estimation errors in FQE.

4.4. Solution Approach

Building upon the framework presented in earlier sections, we now detail Algorithm 1, which is adapted from the PDI-based OCRL algorithm - Algorithm 4, which is a modified version (without mixed-policy method) of the original one proposed by Le et al. (2019). Initially, we use a FQI agent to derive a policy that is optimal with respect to the instantaneous costs $c+\boldsymbol{\lambda}^{\top}\boldsymbol{g}$. Then, due to the offline setting, where no agent-environment interaction to evaluate the policy derived by the FQI agent, we use the FQE agents to evaluate both the objective cost and the constraint costs generated by the policy on validation set $\mathcal{D}_{val}$, and obtain $\hat{G}(\cdot)$ to complete the update of Lagrangian multipliers $\boldsymbol{\lambda}$.

Algorithm 1 Offline Policy Learning for Discharge Decision-Making with Constraints

1: Input one-step transitions $\mathcal{D}_{train} = \left\{\left(s_i, a_i, s_i', c_i, \boldsymbol{g}_i\right)_{i=1}^{N}\right\}$, and $\mathcal{D}_{val} = \left\{\left(s_i, a_i, s_i', c_i, \boldsymbol{g}_i\right)_{i=1}^{M}\right\}$, sample batch sizes n, number of training steps K, weight of penalty ω, learning rates α_{fqi}, α_{fqe} and $\alpha_{\boldsymbol{\lambda}}$, soft-update rate κ, soft-update frequency U

2: Randomly initialize the parametric Q-function estimator for FQI agent with θ_{fqi}^0

3: Randomly Initialize the target parametric Q-function estimator for FQI agent with $\theta'_{fqi} \leftarrow \theta_{fqi}^0$

4: Randomly initialize the parameters for FQE agents $\boldsymbol{\theta}_{fqe}^0$ and $\boldsymbol{\theta}'_{fqe}$. including one FQE agent for evaluating objective cost - mortality risk (θ_{mr}^0 and $\theta_{mr}^{'}$) and two for the constraint costs - readmission risk (θ_{rr}^0 and $\theta_{rr}^{'}$) and ICU LOS (θ_{los}^0 and $\theta_{los}^{'}$)

5: Initialize the dual variable λ_0

6: **for** step $k = 1 \rightarrow K$ **do**

7: Randomly sample n transitions from the training dataset $\mathcal{D}_{train}$

8: Compute the target values for all n samples

$$y_j = \begin{cases} c_j + \boldsymbol{\lambda}_{k-1}^{\top} \boldsymbol{g}_j & \text{if } s_j' = s_H \text{ or } s_D \\ \left(c_j + \boldsymbol{\lambda}_{k-1}^{\top} \boldsymbol{g}_j\right) + Q\left(s_j', \arg\min_{a'} Q\left(s_j', a' \middle| \theta_{fqi}^{k-1}\right) \middle| \theta'_{fqi}\right) & \text{o.w.} \end{cases}, \quad \forall j \in [1, n]$$

9: Update the parameters by doing mini-batch gradient descent on the L2-regularized loss function (14)

$$\theta_{fqi}^k \leftarrow \theta_{fqi}^{k-1} - \alpha_{fqi} \nabla_{\theta_{fqi}^{k-1}} \mathcal{L}_{reg}\left(\theta_{fqi}^{k-1}, \omega\right)$$

10: Then get $\pi_k(s) \leftarrow \arg\min_{a \in \mathcal{A}} Q\left(s, a \middle| \theta_{fqi}^k\right), \forall s \in \mathcal{S}$

11: Compute the target values based on the objective cost c, the constraint costs $\boldsymbol{g}$, and update the parameter vector $\boldsymbol{\theta}_{fqe}^k$ of the FQE agents (same as line 9)

12: Extract transitions with initial state from the validation dataset $\mathcal{D}_{val}$

13: Estimate $\hat{C}^{\pi_k}(s_0) = Q\left(s_0, \pi_k(s_0) \middle| \theta_{mr}^k\right)$, $\hat{G}_R^{\pi_k}(s_0) = Q\left(s_0, \pi_k(s_0) \middle| \theta_{rr}^k\right)$, and $\hat{G}_L^{\pi_k}(s_0) = Q\left(s_0, \pi_k(s_0) \middle| \theta_{los}^k\right)$

14: Update the dual variable $\boldsymbol{\lambda} \in \mathbb{R}_+^2$

$$\begin{aligned} \boldsymbol{\lambda}_k &= \left[\boldsymbol{\lambda}_{k-1} + \alpha_{\boldsymbol{\lambda}} \cdot \nabla_{\boldsymbol{\lambda}_{k-1}} L\left(\pi_k, \boldsymbol{\lambda}_{k-1}\right)\right]^+ \\ &= \left[\boldsymbol{\lambda}_{k-1} + \alpha_{\boldsymbol{\lambda}} \cdot \left(\hat{G}\left(\pi_k\right) - \boldsymbol{l}\right)\right]^+ \end{aligned}$$

15: Every U steps reset

$$\theta'_{fqi} \leftarrow \kappa \theta_{fqi}^k + (1-\kappa)\theta'_{fqi}; \quad \boldsymbol{\theta}'_{fqe} \leftarrow \kappa \boldsymbol{\theta}_{fqe}^k + (1-\kappa)\boldsymbol{\theta}'_{fqe}$$

16: **end for**

17: Output: $\pi_K(s) = \arg\min_{a \in \mathcal{A}} Q\left(s, a \middle| \theta_{fqi}^K\right), \forall s \in \mathcal{S}$

For FQI agent, besides the mini-batch gradient descent method and L2-regulated loss function shown in Section 4.2, different with the classical FQI introduced in Le et al. (2019), we utilize the target approximator used in Double Q-Learning (van Hasselt 2010), to enhance the stability of the training process and reduce overestimation. The Bellman backup targets in (12) become

$$y = \begin{cases} c\left(s_t, a_t, s_{t+1}\right) & \text{if } s_{t+1} = s_H \text{ or } s_D, \\ c\left(s_t, a_t, s_{t+1}\right) + Q\left(s_{t+1}, \arg\min_a Q\left(s_{t+1}, a \middle| \theta\right) \middle| \theta'\right) & \text{o.w.}, \end{cases} \tag{19}$$

where the selection of the action for next state is still due to the parameters θ, but a second set of parameters θ' is used to fairly evaluate the value of this policy. Similarly, for FQE agents, we also deploy the Double Q-Learning structure.

However, Algorithm 4 employs a batch gradient update scheme, requiring both the

FQI and FQE agents to achieve convergence prior to updating the Lagrangian multipliers. Such a sequential convergence requirement poses significant computational challenges in practical numerical experiments, substantially prolonging the overall training time, as the exact number of training steps needed for convergence is difficult to determine in advance. To mitigate this issue, we propose Algorithm 1, in which we adopt a synchronous updating structure that simultaneously updates the FQI agent, all FQE agents, and the Lagrangian multipliers through mini-batch gradient descent. This modification ensures more efficient training and reduces computational complexity, facilitating practical implementation.

We can analyze the computational complexity of Algorithm 1, and compare it with Algorithm 4. We use K represent the total number of training steps, while K_0 and K_m denote the number of training steps for FQI and m-th FQE agents (assuming that we have M FQE agents in total), respectively. In each training step, n samples from the training set $\mathcal{D}_{train}$ will be used, and the action space has a dimensionality of $|\mathcal{A}|$. Also, we assume that the FQI agent is implemented with a fully connected feedfoward neural network with the input dimension d, H_0 hidden layers, and e_h^0 neurons in each layer. Similarly, the M FQE agents are implemented with a structurally similar network with the input dimension d, H_m hidden layers for each FQE agent, e_h^m neurons in each layer, and its output dimension should be 1. With these assumptions and definitions, we can get Theorem 4.5 (the proof is provided in Appendix D).

Theorem 4.5 (Computational Complexity)**.** *The computational complexity of Algorithm 1 is*

$$\mathcal{O}\left(nK\left(P_0|\mathcal{A}| + \sum_{m=1}^{M} P_m\right)\right).$$

In contrast, the computational complexity of Algorithm 4 is

$$\mathcal{O}\left(nK\left(K_0P_0|\mathcal{A}| + \sum_{m=1}^{M} K_mP_m\right)\right).$$

P_0 represents the total number of trainable parameters in the FQI,

$$P_0 = de_0^0 + \sum_{h=1}^{H_0-1} e_{h-1}^0 e_h^0 + e_{H_1-1}^0|\mathcal{A}|,$$

and P_m represents the total number of trainable parameters in the m-th FQE agent,

$$P_m = de_0^m + \sum_{h=1}^{H_m-1} e_{h-1}^m e_h^m + e_{H_m-1}^m \times 1.$$

Remark 1. If we deploy the linear approximation method for FQI and FQE agents, then that can be treated as a simplified version of Theorem 4.5. We can treat d (input dimension) as the dimension of the feature vector $\phi(s,a) : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$, which could also represent the number of parameters need to be updated during the training process. Since there is no hidden layers and neuron now, FQI and all FQE agents can be treated as having the same number of parameters. Therefore, if we use

linear approximation method together with gradient descent update method, then we can have the computational complexity of Algorithm 1 is $\mathcal{O}(ndK|\mathcal{A}|)$, while that of Algorithm 4 is $\mathcal{O}\left(ndK\left(K_0|\mathcal{A}| + \sum_{m=1}^{M} K_m\right)\right)$.

Since we are using a synchronous updating process, we need to utilize the concept of multi-timescale stochastic approximation (as discussed in Chapter 6 of Borkar 2008). As outlined in Algorithm 1, the policy is first updated via the Q function approximation, followed by updates to the FQE parameters and the Lagrange multiplier λ, with the latter adjusted in the direction of the constraint slack. In this algorithm, the update of the approximator parameters of the FQI Q-value function, θ_{fqi}, occurs on the fastest timescale, while the update of the FQE parameters, θ_{fqe}, takes place on a slower scale. The slowest update is reserved for the Lagrange multiplier λ.

Intuitively, Algorithm 1 functions similarly to Algorithm 4, using the separation of timescales to treat θ_{fqe} and λ as quasi-static variables. The learning rates for θ_{fqi}, θ_{fqe}, and λ should satisfy the following conditions to ensure convergence

$$\begin{aligned} &\sum_k \alpha_\iota(k) = \infty, \sum_k \alpha_\iota(k)^2 < \infty, \forall \iota \in \{fqi, fqe, \lambda\}, \\ &\alpha_\lambda(k) = o\left(\alpha_{fqe}(k)\right), \alpha_{fqe}(k) = o\left(\alpha_{fqi}(k)\right). \end{aligned} \tag{20}$$

These conditions ensure that updates to λ occur at a slower rate than those of θ_{fqe}, which in turn are slower than updates of θ_{fqi}. In Section 5, we will demonstrate through numerical experiments that our algorithm can successfully converge.

5. Numerical Experiments

In this section, we assess the performance of our proposed approach (Algorithm 1) by applying it to a real-world medical dataset. This evaluation aims to examine the algorithm's effectiveness in optimizing ICU discharge decisions under multiple constraints.

5.1. Dataset and Clinical Settings

To evaluate the performance of our proposed Offline Constrained Reinforcement Learning (OCRL) framework, we conduct numerical experiments using the MIMIC-IV database (Johnson et al. 2024), a comprehensive repository of de-identified Electronic Health Records (EHR) and Intensive Care Unit (ICU) clinical information collected from the Beth Israel Deaconess Medical Center (BIDMC) in Boston between 2008 and 2022. The dataset encompasses clinical data from 65,366 patients spanning 94,458 ICU admissions, making it one of the largest publicly available resources for critical care research.

Before applying the OCRL framework, we implement a series of data preprocessing procedures to construct a clinically coherent and statistically robust cohort. Firstly, we restrict our analysis to patients with ICU length-of-stay (LOS) of fewer than 15 days, thereby excluding outlier cases with prolonged ICU stays that often reflect atypical clinical trajectories requiring individualized care (Shi et al. 2021). Secondly, we exclude patients who died during their initial ICU admission ($r = 0$), as attributing mortality in such cases poses substantial diagnostic ambiguity and potential confounding. In addition, we remove patients with more than six total ICU admissions to avoid data

sparsity in high-readmission strata, retaining only cases with up to five readmissions to ensure balanced representation across readmission levels.

A key characteristic of our OCRL-based policy learning framework (Algorithm 1) is its direct handling of continuous state space, and a critical component of continuous state space is the physiological state x_t. The physiological state x_t comprises 36 variables including their corresponding descriptions are listed in the Table A1. The selection of these physiological variables is based on two considerations. Firstly, they are all critically important physiological indicators for patients in the ICU. Secondly, we also referenced previous literature on prediction or decision-making problems in ICU settings to determine which physiological variables to choose (Lejarza et al. 2023, Kondrup et al. 2024, Cheng et al. 2024, Wu et al. 2025). Furthermore, for the remaining 36 physiological variables, the proportion of missing values is not particularly high, as we discuss later when addressing missing data.

Table A2 reports summary statistics for all the physiological variables used in the continuous state space including the mean, standard deviation, and Interquartile Range (IQR). Physiological measurements – most of which are recorded on an hourly basis – are grouped into time intervals that vary depending on the patient's readmission count (as detailed in the following paragraphs). For each time interval, the values of the physiological variables listed in Table A1 are aggregated by computing their mean.

Prior to handling the missing values introduced during the aggregation process, we first address anomalous entries in the physiological variables by applying the IQR outlier filtering method. This well-established, non-parametric approach has been widely recognized as effective for outlier detection in healthcare and clinical data settings (Nnamoko and Korkontzelos 2020, Mramba et al. 2024). For each variable, we compute the first (Q_1) and third (Q_3) quartiles, then calculated the IQR as $Q_3 - Q_1$. Observations falling below $Q_1 - 3 \cdot \texttt{IQR}$ or exceeding $Q_3 + 3 \cdot \texttt{IQR}$ are flagged as extreme outliers and excluded. The use of a 3.0 multiplier – more conservative than the conventional 1.5 – is intended to retain clinically meaningful extremes while eliminating values likely arising from data entry errors or sensor malfunctions. For data entries identified as outliers, we first treat them as missing values.

Missing values in physiological variables is addressed using a sequential combination of imputation methods based on the proportion of missing data. All variables are firstly imputed using the time-windowed forward-filling, i.e., forward-filling is performed only within the same ICU admission record. Variables with more than 75% missingness at this stage are excluded for further analysis. Next, for variables with the missingness below 75%, linear interpolation is deployed. If a variable still exhibits more than 30% missingness after the interpolation, then it would be removed. Finally, all remaining variable as shown in Table A1 and A2 are imputed via K-Nearest Neighbors (KNN) imputation with $k = 5$ (Salgado et al. 2016, Kondrup et al. 2024). This hierarchical imputation strategy leverages temporal continuity in ICU data while progressively applying more complex methods based on missingness severity.

Moreover, in numerical experiments, we need to provide clarity on the dynamics of the state space ($\mathcal{S}$). Let us assume a patient who dies within 30 days of their initial ICU admission. During this period, they may experience several times of ICU readmission, each affecting their state. However, to precisely define the terminal state s_D in our model, we specify that only the final ICU admission – immediately preceding death – transitions to s_D following a discharge action. This explicit characterization ensures that the framework precisely captures the patient's ultimate outcome, while accommodating the possibility of repeated ICU stays within the 30-day window. By focusing

on the last admission, the model maintains a rigorous and consistent representation of mortality as a terminal event.

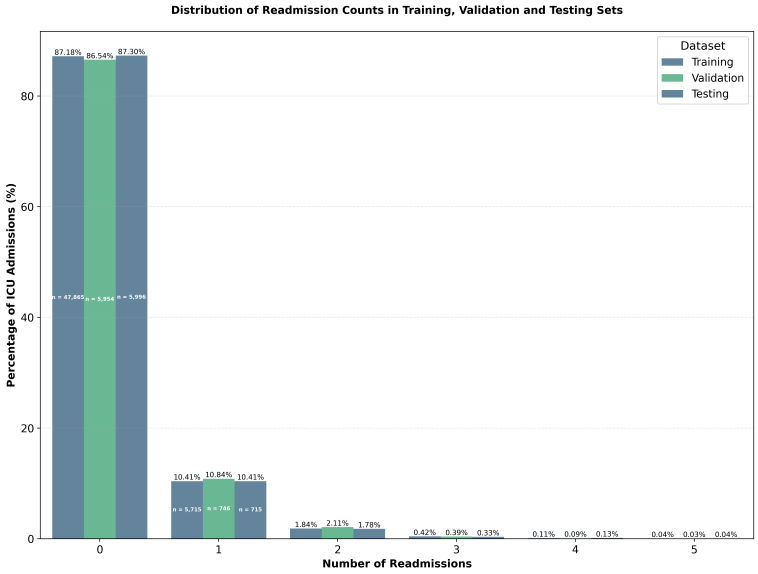


Figure 2.: Comparative distribution of readmission counts between training (54, 902 ICU admissions), validation (9, 135 ICU admissions) and testing (9, 143 ICU admissions) sets.

To model the varying monitoring density associated with different ICU readmission counts, we define the constant Γ_r which is used in the instantaneous constraint function for cumulative ICU LOS (3). It is changing with respect to the readmission count r. Specifically, for patients experiencing their first ICU admission ($r = 0$), the discharge decision is evaluated over a 12-hour assessment window. As the number of readmission increases, this window is progressively halved, following the schedule $12 \cdot 2^{-r}$ hours for $r = 1, 2, 3, 4$, after which it is held constant for $r \geq 4$. This design reflects the clinical intuition that patients with a history of ICU readmission typically require more frequent monitoring and reassessment — subject to the availability of critical care resources. To address differences in the scale of constraint costs during training, we normalize the values of Γ_r to the range $[0, 1]$, ensuring compatibility with the objective and other constraint terms. For instance, $\Gamma_0 = 1.0$, $\Gamma_1 = 0.5$, $\Gamma_2 = 0.25$, and so forth. This normalization facilitates stable optimization and interpretable trade-offs across multiple objectives in the constrained reinforcement learning framework.

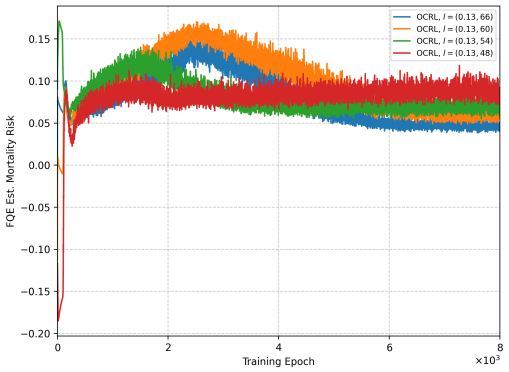
Finally, we partition the processed dataset $\mathcal{D}$ into training ($\mathcal{D}_{train}$), validation ($\mathcal{D}_{val}$), and testing ($\mathcal{D}_{test}$) subsets. After preprocessing and filtering, the final dataset comprises 491, 996 data samples drawn from 49, 581 patients, corresponding to a total of 68, 653 ICU admission episodes. To ensure fair and representative evaluation, patients are split into the three subsets in a 80% : 10% : 10% ratio, stratified at the patient level to preserve the statistical properties of admission trajectories. This allocation ensures that all ICU stays associated with a given patient are contained within a single partition, avoiding information leakage across sets. Figure 2 confirms that the distribution of readmission counts is well-aligned across the training, validation, and testing cohorts, indicating consistency in patient characteristics across splits. To further assess the robustness of Algorithm 1, we perform numerical experiments using temporal dataset partitioning. These experiments examine the algorithm's performance across different time periods, with comprehensive results presented in Ap-

pendix J.

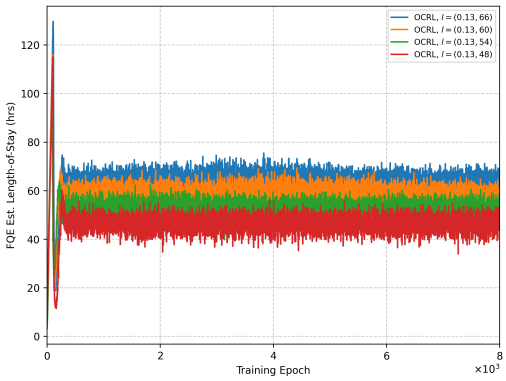
All experiments were performed on a high-performance Windows workstation equipped with dual AMD EPYC 7742 64-core processors (2.25 GHz) and an NVIDIA A40 GPU. The proposed algorithms were implemented using the `PyTorch` machine learning framework (Paszke et al. 2019), which enabled efficient training and evaluation of both linear-based and neural network–based approximators under the offline RL pipeline.

5.2. Algorithm Performance

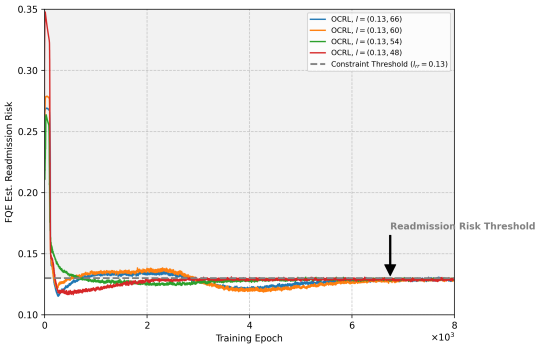
To evaluate the stability and convergence of the proposed Algorithm 1, we examine its training performance on both the training set $\mathcal{D}_{train}$ and the validation set $\mathcal{D}_{val}$. Given that the algorithm employs a multi-timescale update scheme, it is crucial to empirically assess whether its key components—specifically, the Lagrangian multipliers λ_{rr} and λ_{los}, as well as the embedded Fitted Q-Evaluation (FQE) agents—exhibit consistent and reliable convergence throughout the training process. This analysis provides foundational support for the algorithm's effectiveness in solving constrained offline reinforcement learning problems in high-stakes clinical settings.



(a) Training process of FQE agent for objective cost - mortality risk.



(b) Training process of FQE agent for constraint cost - ICU LOS.



(c) Training process of FQE agent for constraint cost - readmission risk.

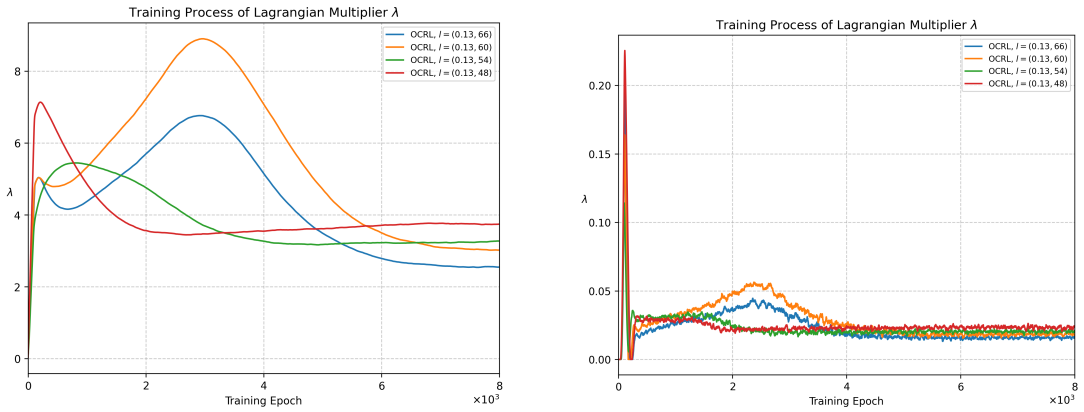
Figure 3.: The training process of FQE agents in Algorithm 1 with given thresholds $l_{los} \in \{48, 54, 60, 66\}$, and $l_{rr} = 0.13$.

Before presenting the results of the numerical experiments, we outline the hyperparameter settings used to train Algorithm 1. All numerical experiments in this section

are based on linear function approximation for both the FQI and FQE agents (the neural network approximation version is provided in Appendix H). Specifically, we employ one FQI agent to approximate the discharge policy and three FQE agents to estimate the value functions associated with the objective (mortality risk) and two constraints (readmission risk and ICU length-of-stay). Let α_{fqi} denote the learning rate for the FQI agent, and α_{fqe} the shared learning rate for all three FQE agents. To ensure stable and accurate policy evaluation during training, α_{fqe} is selected to be smaller than α_{fqi}. The candidate values considered for α_{fqi} include $\{3\times10^{-3}, 2\times10^{-3}, 1\times10^{-3}, 5\times10^{-4}, 1\times10^{-4}\}$, while those for α_{fqe} are $\{1\times10^{-3}, 5\times 10^{-4}, 1\times10^{-4}, 5\times10^{-5}, 1\times10^{-5}\}$. The Lagrangian multipliers associated with the constraints are updated using separate learning rates: α_λ^{rr} for the readmission risk constraint, and α_λ^{los} for the length-of-stay (LOS) constraint. Since the FQE agent used to estimate the LOS constraint typically exhibits a higher estimation variance (measured via the standard error $\hat{\sigma}$), we assign a smaller learning rate to α_λ^{los} to improve stability. Candidate values for α_λ^{rr} include $\{4\times10^{-4}, 3\times10^{-4}, 1\times10^{-4}, 1\times10^{-5}, 1\times10^{-6}\}$, and for α_λ^{los}, $\{1\times10^{-5}, 1\times10^{-6}, 1\times10^{-7}, 1\times10^{-8}\}$. In all cases, we ensure that both α_λ^{rr} and α_λ^{los} remain smaller than α_{fqe}. To penalize constraint violations, a penalty weight ω is introduced, with candidate values $\{1\times10^{-2}, 1\times10^{-3}, 1\times10^{-4}, 1\times10^{-5}\}$. Additional hyperparameters include the batch size n, chosen from $\{512, 256, 128, 64\}$, and the target network update frequency U, selected from $\{1\times10^{2}, 5\times10^{2}, 1\times10^{3}, 5\times10^{3}\}$. The soft-update rate κ is tuned over the set $\{1\times10^{-2}, 5\times10^{-3}, 1\times10^{-3}, 5\times10^{-4}, 1\times10^{-4}\}$. After extensive training and validation across the above configurations, the final selected hyperparameter values are: $\alpha_{fqi} = 2\times10^{-3}$, $\alpha_{fqe} = 5\times10^{-4}$, $\alpha_\lambda^{rr} = 3\times10^{-4}$, $\alpha_\lambda^{los} = 1\times10^{-6}$, $\omega = 1\times10^{-2}$, $n = 256$, $U = 1\times10^{2}$, and $\kappa = 1\times10^{-2}$. Moreover, we analyze the selection of key hyperparameters, including the learning rates α_λ^{rr} and α_λ^{los} for the two Lagrangian multipliers and the penalty weight ω. The results of this analysis are presented in Appendix F.

Figure 3 presents the training trajectories of the FQE agents incorporated into Algorithm 1, executed over 8×10^3 iterations with 1×10^3 training steps per iteration, resulting in a total of 8×10^6 training steps (i.e., $K = 8\times10^6$ in the Algorithm 1). We examine the convergence behavior under a fixed readmission risk constraint $l_{rr} = 0.13$ and four levels of ICU LOS constraints: $l_{los} \in \{48, 54, 60, 66\}$. As depicted in Figure 3b, the FQE estimations for LOS constraint costs converge early in training and remain close to the specified thresholds. Nonetheless, moderate fluctuations persist throughout the training process, which is characteristic of off-policy evaluation methods such as FQE (see Section 4.3 for discussion). A similar pattern of convergence with minor oscillations is observed in the FQE agents responsible for evaluating the objective cost, i.e., the mortality risk, as shown in Figure 3a. To enhance stability and enforce constraint satisfaction under uncertainty, we incorporate a probabilistic upper bound using a z-score of 2.33, corresponding to the upper bound of a 98% one-sided confidence interval under the standard normal distribution. This ensures that the upper confidence bound of the estimated constraint costs remains below the prescribed thresholds. Lastly, Figure 3c confirms that the FQE agents estimating the readmission risk constraint consistently converge toward the target value of $l_{rr} = 0.13$, validating the reliability of the proposed training framework across varying constraint conditions.

Figure 4 presents the convergence behavior of the Lagrangian multipliers λ_{rr} and λ_{los} during training. As shown in Figure 4b, under a fixed readmission risk threshold $l_{rr} = 0.13$, tighter ICU LOS constraints (i.e., smaller values of l_{los}) consistently correspond to larger values of the associated multiplier λ_{los}. This indicates that stricter LOS requirements exert greater influence on the FQI agent's policy updates, reflect-



(a) The training process of λ_{rr}.

(b) The training process of λ_{los}

Figure 4.: The training process of Lagrangian multipliers $\boldsymbol{\lambda}$ inside Algorithm 1.

ing an increased pressure to comply with resource constraints. Interestingly, a similar pattern is observed in Figure 4a, where the values of λ_{rr}—despite being associated with an identical readmission threshold $l_{rr} = 0.13$ across all cases—also increase as l_{los} becomes more restrictive. This suggests a coupling effect: when the LOS constraint tightens, the effective burden of satisfying the readmission constraint simultaneously becomes greater. In other words, even if the readmission threshold remains unchanged, the FQI agent perceives it as more restrictive in the presence of a more binding LOS constraint, resulting in a higher corresponding multiplier λ_{rr}. This highlights the interplay between multiple constraints in the CMDP formulation and underscores the necessity of jointly managing them during policy learning.

5.3. Policy Interpretation

In this section, we evaluate the performance of Algorithm 1 on the held-out testing set $\mathcal{D}_{test}$ and provide an interpretation of its effectiveness in a realistic, out-of-sample setting. Following the experimental setup described in Section 5.2, we train the algorithm using the training set $\mathcal{D}_{train}$ and validation set $\mathcal{D}_{val}$ under a grid of constraint thresholds: $l_{rr} \in \{0.12, 0.13, 0.14\}$ for the readmission risk and $l_{los} \in \{48.0, 54.0, 60.0, 66.0\}$ for the ICU LOS. These threshold values are set with reference to the average values observed in the training, validation, and testing sets (as shown in Table 2). Specifically, the tested values of l_{rr} and l_{los} are deliberately chosen to be lower than the dataset averages, allowing us to evaluate the performance of our offline constrained policy learning framework (Algorithm 1) under more stringent readmission risk and ICU LOS constraints compared to the average clinical practice in the MIMIC-IV. This design tests whether our approach can learn policies that improve upon current practice by achieving tighter constraints. The following analysis focuses on assessing the generalization performance of the learned policies across these constraint settings.

We begin by evaluating the accuracy and reliability of the FQE components within Algorithm 1. Specifically, we assess whether FQE can faithfully estimate the outcomes associated with clinician behavior as recorded in the training set $\mathcal{D}_{train}$, the validation set $\mathcal{D}_{val}$, and the testing set $\mathcal{D}_{test}$. To this end, FQE agents trained on the training set $\mathcal{D}_{train}$ are applied to evaluate the implicit policy underlying clinical practice. Table 2 reports the empirical averages of key outcome metrics—mortality risk, readmission

Table 2.: Comparison between observed clinical outcomes (average value) and FQE estimates across training ($\mathcal{D}_{train}$), validation ($\mathcal{D}_{val}$), and testing sets ($\mathcal{D}_{test}$).

Outcome Metric	**Training Set**		**Validation Set**		**Testing Set**	
	Observed	**FQE Est.**	**Observed**	**FQE Est.**	**Observed**	**FQE Est.**
Mortality Risk	0.082	0.082	0.088	0.079	0.081	0.081
Readmission Risk	0.147	0.147	0.155	0.146	0.145	0.147
Length-of-Stay (hrs)	66.39	66.42	66.16	65.39	66.18	66.73

risk, and ICU LOS — across initial states (s_0) in those three datasets, as well as the corresponding FQE estimates for the observed clinician policy. The close alignment between FQE estimations and empirical averages indicates that the FQE agents with linear function approximation are capable of accurately evaluating discharge policies, thus validating their utility in this decision-making context.

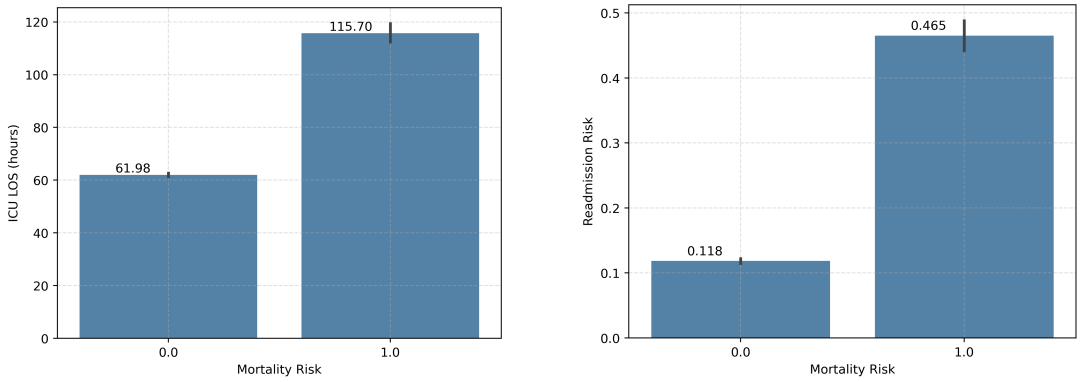
Table 3.: Comparison of mortality risks, readmission risks (RR), and ICU length-of-stay (LOS) estimated by FQE on testing set ($\mathcal{D}_{\text{test}}$) across RL agents with various constraint threshold parameters.

Cost Metric	**LOS constraint (hrs)**	**RR constraint threshold**		
		0.14	0.13	0.12
Mortality risk	66.0	**0.026**	**0.046**	0.097
	60.0	**0.033**	**0.064**	0.128
	54.0	**0.038**	**0.065**	0.130
	48.0	**0.045**	0.097	0.170

We next compare the FQE-estimated outcomes of Algorithm 1 under various constraint configurations on the testing set $\mathcal{D}_{test}$ against the clinical benchmarks reported in Table 2.

Table 3 reveals a clear trend: for any fixed readmission risk threshold l_{rr}, more stringent ICU LOS constraints (i.e., lower values of l_{los}) consistently lead to higher estimated mortality risks. This pattern aligns with findings in the clinical literature, which suggest that longer ICU stays can substantially improve survival rates (Bartel et al. 2019). However, as shown in Figure 5a, patient mortality risk exhibits a positive association with ICU LOS. This reflects the clinical reality that more critically ill patients tend to remain longer in the ICU and are at elevated risk of death. Despite this confounding relationship, Algorithm 1 is still able to learn reasonable and clinically meaningful policies. One possible reason is that the state representation includes a rich set of physiological variables, which help the algorithm distinguish between patients with different levels of severity.

Furthermore, Table 3 shows that, for a fixed LOS constraint l_{los}, tighter readmission risk thresholds (i.e., reducing l_{rr} from 0.14 to 0.12) paradoxically result in higher FQE-estimated mortality risk. This result goes against clinical understanding (see Figure 5b), which usually links lower readmission rates with better patient outcomes. The likely explanation is that, under the dual pressure of strict LOS constraints and reduced allowance for readmissions, the FQI agents are forced to make premature discharge



(a) (b)

Figure 5.: Relationship among the cumulative mortality risk costs, the cumulative length-of-stay (LOS) and the cumulative readmission risk costs for initial states s_0 in the training set $\mathcal{D}_{train}$.

decisions for a broad set of physiological states. In our framework, readmission and death are modeled as mutually exclusive outcomes. Thus, limiting readmission under early discharge increases the likelihood of mortality, which is reflected in the FQE estimates. Prior clinical research also highlights this trade-off, as premature ICU discharge decision has been associated with increased risk of readmission or post-discharge death (Badawi and Breslow 2012).

Table 4 provides a more detailed view of the policy's behavior. Although the OCRL policy results in more discharge actions overall ($a = 1$), these actions are mainly taken for patients during their first ICU admission ($r = 0$). In clinical practice, discharge decisions for first-time ICU admissions are often conservative. This caution stems from several factors: concerns about the elevated mortality risk associated with early readmissions (Bion and Dennis 2016), adherence to the medical ethical principle of "first, do no harm" (You and Ulrich 2024, Lighthall and Vazquez-Guillamet 2015), and the lack of sufficient empirical data to accurately assess the prognosis of patients during their initial ICU stay. Such conservatism is commonly referred to as watchful waiting in the ICU. As noted by Forster et al. (2020), discharge delays are frequent, with 15-31% of ICU patients experiencing such delays. While the OCRL policy demonstrates more assertive discharge behavior than clinical practice for first-time admissions (approximately 27% fewer "Stay in ICU" decisions), it exhibits similar caution for high-risk patients who subsequently died or required readmission after their initial ICU discharge. For this critical subgroup, the OCRL policy recommends continued ICU stay ($a = 0$) in 73.2% of cases, compared to 55.1% for the general cohort (Table 4). To further investigate the alignment between OCRL policy and clinical practices, we use the qSOFA score to assess patient severity and compare the score distributions under different discharge decisions made by the OCRL policy versus those observed in clinical practice. As a quick screening tool, qSOFA scores of 2 or above generally indicate greater clinical risk. Figure 6 shows that, although the OCRL agent recommends discharge more frequently for first-time ICU admissions, the qSOFA distribution of the discharge decision closely matches that observed in clinical practices. While real-world clinical judgment naturally incorporates a broader range of patient-specific factors, the qSOFA-based comparison offers a straightforward yet informative validation of the OCRL agent's decision-making rationale.

Table 4.: Discharge ($a = 1$) ratios for clinical practice vs. OCRL policy ($l_{rr} = 0.13$, $l_{los} = 54.0$) across various readmission counts (r) in $\mathcal{D}_{test}$.

Admission Type	**Policy**	**Stay in ICU** ($a = 0$)		**Discharge** ($a = 1$)	
		Count	Ratio	Count	Ratio
All Cases	Clinical Practice	42879	0.862	6868	0.138
	OCRL Policy	32903	0.661	16844	0.339
First Admission ($r = 0$)	Clinical Practice	27825	0.823	5996	0.177
	OCRL Policy	18644	0.551	15177	0.449
First Readmission ($r = 1$)	Clinical Practice	8044	0.918	715	0.082
	OCRL Policy	7249	0.828	1510	0.172
Second Readmission ($r = 2$)	Clinical Practice	3689	0.968	122	0.032
	OCRL Policy	3655	0.959	156	0.041
Third Readmission ($r = 3$)	Clinical Practice	1470	0.985	23	0.015
	OCRL Policy	1492	0.999	1	0.001

As for patients with multiple ICU admissions ($r \geq 2$), the policy becomes more conservative, suggesting that the algorithm adjusts its decisions based on higher clinical risk. Furthermore, among patients who were discharged in actual clinical practice ($a = 1$) but readmitted within 30 days, the OCRL agent instead recommends continued ICU stay ($a = 0$) in 55.8% of these cases. This indicates that the learned policy not only satisfies the global constraints but also recognizes high-risk individual cases where early discharge may be harmful. Also, the importance of including the readmission count r in the state space is highlighted, which enables the RL agent to better assess patient risk and adapt its decisions accordingly. A comparison with a variant of the method that does not incorporate readmission count r (OCRL-D) is provided in Appendix G. The results show that while OCRL-D does exhibit increasing conservativeness with larger r values, OCRL-R consistently better identifies high-risk cases and aligns its discharge behavior accordingly. These findings underscore the value of explicitly modeling readmission history when learning discharge policies in constrained, high-stakes clinical environments in the ICU.

Since linear function approximation is deployed in both FQI and FQE components, the interpretability of our proposed OCRL framework is enhanced. This modeling choice could ensure exact compatibility with SHAP (SHapley Additive exPlanations) values, a widely used framework for feature attribution. Unlike nonlinear approximators such as neural networks, linear approximation allow closed-form computation of SHAP values, enabling precise quantification of the marginal contribution of each physiological variable to the approximated Q-values. This interpretability is particularly valuable in clinical decision-making tasks, where transparency and feature-level reasoning are essential for building trust in algorithmic recommendations.

Figure 7 presents the feature importance analysis for the FQI agent in Algorithm 1, using SHAP values. Each subplot ranks features vertically in descending order of their aggregate influence, with the most impactful variables appearing at the top. The horizontal axis reflects the magnitude and direction of a feature's effect on the Q-value estimation. The left and right columns respectively display SHAP results for the two possible actions: staying in the ICU ($a = 0$) and discharging the patient

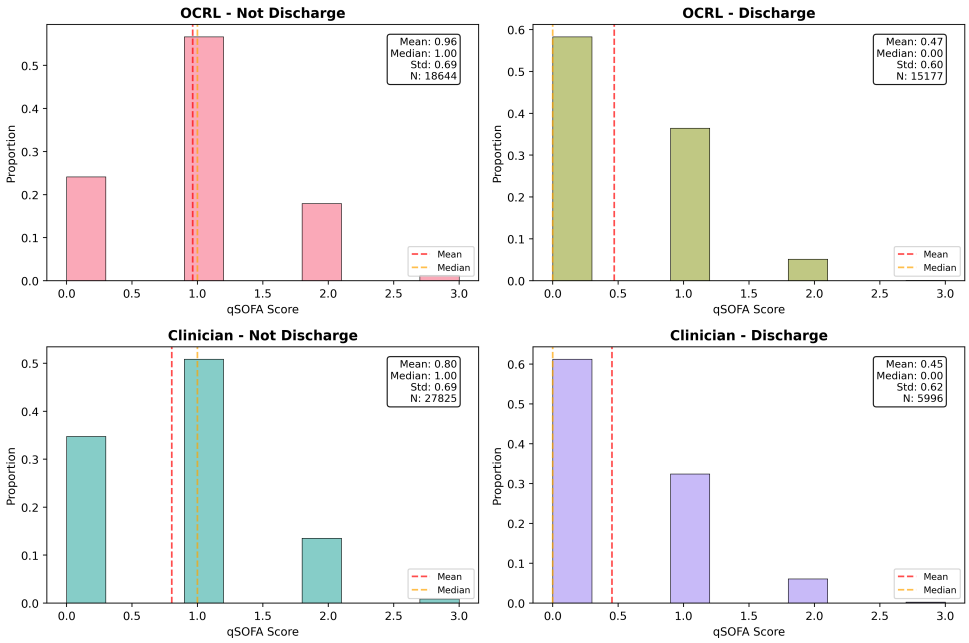


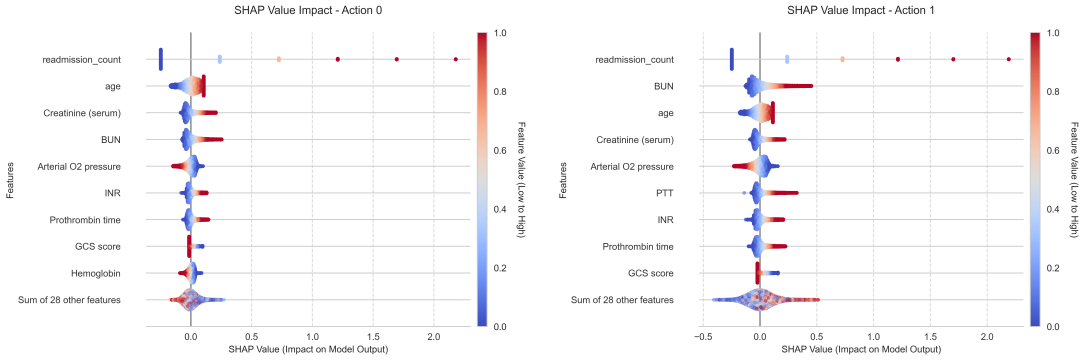
Figure 6.: Comparison of qSOFA score distribution analysis under different decision types.

($a = 1$). Subplots 7a and 7b correspond to a readmission risk threshold of $l_{rr} = 0.12$, while 7c and 7d reflect $l_{rr} = 0.13$; all are evaluated under a fixed ICU LOS constraint of $l_{los} = 54.0$ hours.

Across all subplots, the readmission count r—embedded in the state representation $s = (x, r)$—emerges as one of the most influential features in shaping discharge decisions. Notably, higher values of r (represented by red points) consistently exhibit positive SHAP values, indicating a strong association with increased approximated Q-values, and thus a greater estimated risk. This trend aligns with clinical expectations: patients with a history of multiple ICU readmissions are generally at higher risk.

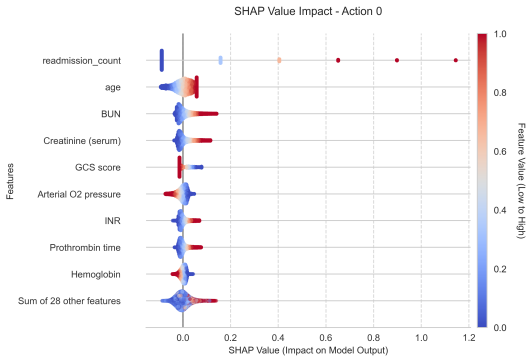
In addition to readmission count, several physiological features are consistently ranked as important across both actions and constraint settings. These include age, Glasgow Coma Scale (GCS) score, blood urea nitrogen (BUN), serum creatinine, and arterial oxygen pressure—each of which reflects a critical dimension of patient health. For example, age is positively associated with risk across all outcome metrics, including mortality, readmission, and ICU LOS. This is reflected in the SHAP analysis in Figure 7, where increased age corresponds to higher Q-values, especially under discharge decisions.

Each of the additional physiological features represents a critical aspect of the patient's clinical status: neurological condition (GCS score), respiratory function (arterial O_2 pressure), and renal function (BUN and serum creatinine levels). For instance, BUN measures the concentration of nitrogen in the blood and serves as a key indicator of kidney function; elevated BUN levels are commonly associated with increased mortality risk (Beier et al. 2011, Arihan et al. 2018). Similarly, the GCS score assesses a patient's level of consciousness, where lower scores typically signal higher mortality risk and a greater need for continued intensive care. As shown in Figure 7, the SHAP values corresponding to both BUN and GCS scores reflect their clinical relevance and confirm their importance in the FQI agent's decision-making process.

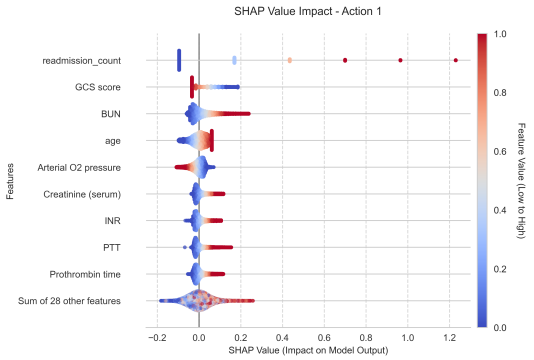


(a) SHAP Value Impact - Action 0, $l_{rr} = 0.12$

(b) SHAP Value Impact - Action 1, $l_{rr} = 0.12$



(c) SHAP Value Impact - Action 0, $l_{rr} = 0.13$



(d) SHAP Value Impact - Action 1, $l_{rr} = 0.13$

Figure 7.: SHAP values based feature importance analysis for the FQI agent in Algorithm 1 with different readmission penalty values: top row ($l_{rr} = 0.12$) and bottom row ($l_{rr} = 0.13$), both with $l_{los} = 54.0$. Action 0 (left column) represents staying in the ICU decision, while Action 1 (right column) represents discharge decision.

6. Conclusion

In this paper, we address the multifaceted challenges inherent in ICU discharge decision-making, encompassing safety concerns (mortality risk and readmission risk), financial considerations (readmission-related penalty payments), and operational efficiency (ICU length-of-stay). To tackle this multi-objective problem, we propose a policy learning framework that formulates the ICU discharge decision-making process as a discrete-time, indefinite-horizon Constrained Markov Decision Process (CMDP), thereby extending the classical MDP framework to accommodate multiple operational constraints simultaneously. Within this CMDP formulation, we incorporate the patient's ICU admission history through a readmission count variable in the state space, designate mortality risk minimization as the primary objective, and impose two additional constraints: readmission risk and ICU length-of-stay (LOS).

Our proposed policy learning framework integrates the constrained reinforcement learning formulation – specifically, Primal-Dual Iteration (PDI) – with a value-based offline reinforcement learning algorithm, Fitted Q-Iteration (FQI), to effectively solve a CMDP with multiple constraints without requiring any agent-environment interaction. Both the inherent properties of offline RL and the requirement for updating dual variables in the PDI algorithm necessitate the use of an Off-Policy Evaluation (OPE) method within our framework. To enable OPE during training, we incorporate

a modified Fitted Q-Evaluation (FQE) method as the core component for estimating the expected returns of policies derived by the FQI agent. Given the high computational cost associated with both FQI and FQE, we optimize the parameters of their respective function approximators simultaneously using different learning rates at each training step, thereby improving computational efficiency.

Using the real-world clinical dataset MIMIC-IV, we demonstrate through numerical experiments that our proposed multi-timescale offline constrained policy learning algorithm (Algorithm 1) reliably converges while satisfying all specified constraint thresholds. We evaluate the quality of the resulting policies by comparing FQE-estimated outcomes – mortality risk, readmission risk, and ICU LOS – under various constraint settings against clinical practices observed in the dataset. Notably, the learned policy exhibits adaptive conservativeness based on patients' readmission count, demonstrating the algorithm's capability to make appropriately cautious decisions for patients with multiple ICU admissions.

Our work has several limitations and directions for future extension. As more ICU databases worldwide become increasingly available and comprehensive, future research can pursue the following complementary directions. Firstly, evaluating the zero-shot transfer performance of Algorithm 1 (trained on MIMIC-IV) on external datasets to assess its out-of-distribution generalization. Secondly, employing transfer learning techniques – such as fine-tuning the MIMIC-IV-trained model on target datasets or using domain adaptation methods – to enhance the generalizability of our algorithmic framework across heterogeneous clinical settings. Moreover, in future research, we plan to refine our algorithm through collaboration with clinicians. Specifically, if non-physiological information from hospitals and rehabilitation departments can be incorporated – for example, bed availability that may necessitate earlier ICU discharge despite incomplete physiological recovery, staffing levels, post-ICU care capacity, or social support limitations that often delay discharge decisions – our framework could be further extended to account for these currently unmodeled non-physiological factors.

Funding

This work was partially supported by the Deutsche Forschungsgemeinschaft (DFG), Project No. 543063591.

Disclosure of Interest

The authors declare no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Consent and Approval

This study has been exempt from the requirement for approval by an institutional review board. This research study was conducted using a publicly available data that do not include personal information of patients.

The de-identified MIMIC-IV database (version 3.1) used in this study is publicly available through PhysioNet (`https://physionet.org/content/mimiciv/3.1/`).

Access requires completion of a training course in human subjects research and execution of a data use agreement.

We confirm that we have used Generative AI (ChatGPT-5, version August 2025) solely for language polishing and sentence-level editing of this manuscript. No AI tool was used to generate content, interpret data, formulate arguments, or draw conclusions; all substantive intellectual work—including study design, data analysis, interpretation, and writing of original text—was performed by the authors.

References

Aghalari A, Tajik N, Marufuzzaman M, Carruth D (2021) Inverse reinforcement learning to assess safety of a workplace under an active shooter incident. *IISE Transactions* 53(12):1337–1350.

Alaeddini A, Helm JE, Shi P, Faruqui SHA (2019) An integrated framework for reducing hospital readmissions using risk trajectories characterization and discharge timing optimization. *IISE Transactions on Healthcare Systems Engineering* 9(2):172–185.

Altman E (1999) *Constrained Markov Decision Processes* (Routledge), 1st edition.

Arihan O, Wernly B, Lichtenauer M, Franz M, Kabisch M, Muessig J, Masyuk M, Lauten A, Schulze PC, Hoppe UC, Kelm M, Jung C (2018) Blood urea nitrogen (bun) is independently associated with mortality in critically ill patients admitted to icu. *PLOS ONE* 13(1):e0191697.

Badawi O, Breslow MJ (2012) Readmissions and death after icu discharge: Development and validation of two predictive models. *PLOS ONE* 7(11):e48758.

Bagshaw SM, Tran DT, Opgenorth D, Wang X, Zuege DJ, Ingolfsson A, Stelfox HT, Thanh NX (2020) Assessment of costs of avoidable delays in intensive care unit discharge. *JAMA Network Open* 3(8):e2013913.

Bartel AP, Chan CW, Kim SH (2019) Should hospitals keep their patients longer? the role of inpatient care in reducing postdischarge mortality. *Management Science* 66(6):2326–2346.

Beier K, Eppanapally S, Bazick HS, Chang D, Mahadevappa K, Gibbons FK, Christopher KB (2011) Elevation of blood urea nitrogen is predictive of long-term mortality in critically ill patients independent of "normal" creatinine. *Critical Care Medicine* 39(2):305–313.

Bion J, Dennis A (2016) ICU Admission and Discharge Criteria. Webb A, et al., eds., *Oxford Textbook of Critical Care* (Oxford: Oxford University Press), 2 edition.

Borkar VS (2008) *Stochastic Approximation: A Dynamical Systems Viewpoint* (Cambridge University Press).

Boyd S, Vandenberghe L (2004) *Convex Optimization* (Cambridge: Cambridge University Press).

Cai T, Jiang J, Zhang W, Zhou S, Song X, Yu L, Gu L, Zeng X, Gu J, Zhang G (2023) Marketing budget allocation with offline constrained deep reinforcement learning.

Carey K (2015) Measuring the hospital length of stay/readmission cost trade-off under a bundled payment mechanism. *Health Economics* 24(7):790–802.

Centers for Medicare & Medicaid Services (2024) Hospital readmissions reduction program (hrrp). Available at: `https://www.cms.gov/medicare/payment/prospective-payment-systems/acute-inpatient-pps/hospital-readmissions-reduction-program-hrrp`, accessed: May 15, 2024.

Chan CW, Farias VF, Bambos N, Escobar GJ (2012) Optimizing intensive care unit discharge decisions with patient readmissions. *Operations Research* 60(6):1323–1341.

Cheng G, Xie J, Zheng Z, Luo H, Ooi OC (2024) Extubation decisions with predictive information for mechanically ventilated patients in the icu. *Management Science* 71(7):6069–6091.

Chuang Y, Zargoush M, Ghazalbash S, Samiedaluie S, Kuluski K, Guilcher S (2023) From prediction to decision: Optimizing long-term care placements among older delayed discharge patients. *Production and Operations Management* 32(4):1041–1058.

Cohen IG, Iltis AS, White DB (2020) Potential legal liability for withdrawing or withholding ventilators during COVID-19: Assessing the risks and identifying needed reforms. *Journal of the American Medical Association* .

Dam TA, de Bruin D, Cinà G, Thoral PJ, Elbers PW, den Uil CA, Crane RF (2025) Icu readmission and mortality risk prediction: Generalizability of a multi-hospital model. *Journal of Intensive Medicine* ISSN 2667-100X.

DeepSeek-AI (2025) Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Ernst D, Geurts P, Wehenkel L (2005) Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research* 6(Apr):503–556.

Fang N, Liu G, Gong W (2024) Offline inverse constrained reinforcement learning for safe-critical decision making in healthcare.

Fitch Z, Debesa O, Ohkuma R, Duquaine D, Steppan J, Schneider E, Whitman G (2014) A protocol-driven approach to early extubation after heart surgery. *The Journal of Thoracic and Cardiovascular Surgery* 147(4):1344–1350.

Forster GM, Bihari S, Tiruvoipati R, Bailey M, Pilcher D (2020) The association between discharge delay from intensive care and patient outcomes. *American Journal of Respiratory and Critical Care Medicine* 202(10):1399–1406.

Fujimoto S, Meger D, Precup D (2019) Off-policy deep reinforcement learning without exploration. *International Conference on Machine Learning*, 2052–2062.

Garcez-Leme LE, Avelino-Silva TJ (2023) Depression, delirium, and post-intensive care syndrome. *International Psychogeriatrics* 35(8):399–401, ISSN 1041-6102, issue Theme: Neurocognitive Markers Outside Dementia.

Guo X, Chen P, Liang S, Jiao Z, Li L, Yan J, Huang Y, Liu Y, Fan W (2022) Pacar: Covid-19 pandemic control decision making via large-scale agent-based modeling and deep reinforcement learning. *Medical Decision Making* 42(8):1064–1077.

Hatch R, Young D, Barber V, Griffiths J, Harrison DA, Watkinson P (2018) Anxiety, depression and post traumatic stress disorder after critical illness: a uk-wide prospective cohort study. *Critical Care* 22(1):310.

Heggestad T (2002) Do hospital length of stay and staffing ratio affect elderly patients' risk of readmission? a nation-wide study of norwegian hospitals. *Health Services Research* 37(3):647–665.

Hong K, Li Y, Tewari A (2024) A primal-dual-critic algorithm for offline constrained reinforcement learning. *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 280–288 (PMLR).

James J (2013) Medicare hospital readmissions reduction program. *Health Affairs Health Policy Brief* .

Johnson A, Bulgarelli L, Pollard T, Gow B, Moody B, Horng S, Celi LA, Mark R (2024) Mimic-iv (version 3.1). Accessed from PhysioNet.

Kc DS, Terwiesch C (2009) Impact of workload on service time and patient safety: An econometric analysis of hospital operations. *Management Science* 55(9):1486–1498.

Kc DS, Terwiesch C (2011) An econometric analysis of patient flows in the cardiac intensive care unit. *Manufacturing & Service Operations Management* 14(1):50–65.

Khatami M, Alvarado M, Kong N, Parikh PJ, Lawley MA (2021) Inpatient discharge planning under uncertainty. *IISE Transactions* 54(4):332–347.

Kondrup F, Jiralerspong T, Lau E, de Lara N, Shkrob J, Tran MD, Precup D, Basu S (2024) Towards safe mechanical ventilation treatment using deep offline reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 15696–15702.

Kramer AA, Higgins TL, Zimmerman JE (2013) The association between icu readmission rate and patient outcomes. *Critical Care Medicine* 41(1):24–33.

Kreke JE, Bailey MD, Schaefer AJ, Angus DC, Roberts MS (2008) Modeling hospital discharge policies for patients with pneumonia-related sepsis. *IIE Transactions* 40(9):853–860.

Kumar A, Fu J, Tucker G, Levine S (2019) Stabilizing off-policy q-learning via bootstrapping error reduction.

Kumar A, Zhou A, Tucker G, Levine S (2020) Conservative q-learning for offline reinforcement learning.

Kuo YF, Goodwin JS (2011) Association of hospitalist care with medical utilization after discharge: Evidence of cost shift from a cohort study. *Annals of Internal Medicine* 155(3):152–159.

Labbi A, Berrospi C (2007) Optimizing marketing planning and budgeting using markov decision processes: An airline case study. *IBM Journal of Research and Development* 51(3.4):421–431.

Lange S, Gabel T, Riedmiller M (2012) Batch reinforcement learning. *Reinforcement Learning*, 45–73 (Springer).

Le HM, Voloshin C, Yue Y (2019) Batch policy learning under constraints.

Lee E, Lavieri MS, Volk ML, Xu Y (2015) Applying reinforcement learning techniques to detect hepatocellular carcinoma under limited screening capacity. *Health Care Management Science* 18:363–375.

Lee J, Paduraru C, Mankowitz DJ, Heess N, Precup D, Kim KE, Guez A (2022) Coptidice: Offline constrained reinforcement learning via stationary distribution correction estimation.

Lejarza F, Calvert J, Attwood MM, Evans D, Mao Q (2023) Optimal discharge of patients from intensive care via a data-driven policy learning framework. *Operations Research for Health Care* 38:100400.

Lighthall GK, Vazquez-Guillamet C (2015) Understanding decision making in critical care. *Clinical Medicine & Research* 13(3-4):156–168.

Liu X (2022) Dynamic coupon targeting using batch deep reinforcement learning: An application to livestream shopping. *Marketing Science* 42(4):637–658.

Liu Z, Guo Z, Yao Y, Cen Z, Yu W, Zhang T, Zhao D (2023) Constrained decision transformer for offline safe reinforcement learning.

Long EF, Mathews KS (2018) The boarding patient: Effects of icu and hospital occupancy surges on patient flow. *Production and Operations Management* 27(12):2122–2143.

Mcneill H, Khairat S (2020) Impact of intensive care unit readmissions on patient outcomes and the evaluation of the national early warning score to prevent readmissions: Literature review. *JMIR Perioperative Medicine* 3(1).

Moitra V, Guerra C, Linde-Zwirble W, Wunsch H (2016) Relationship between icu length of stay and long-term mortality for elderly icu survivors. *Critical Care Medicine* 44(4):655–662.

Mramba LK, Liu X, Lynch KF, Yang J, Aronsson CA, Hummel S, Norris JM, Virtanen SM, Hakola L, Uusitalo UM, Krischer JP (2024) Detecting potential outliers in longitudinal data with time-dependent covariates. *European Journal of Clinical Nutrition* 78(4):344–350.

Nnamoko N, Korkontzelos I (2020) Efficient treatment of outliers and class imbalance for diabetes prediction. *Artificial Intelligence in Medicine* 104:101815.

Oh Jh, Zheng Z, Bardhan IR (2018) Sooner or later? health information technology, length of stay, and readmission risk. *Production and Operations Management* 27(12):2038–2053.

Ouyang H, Argon NT, Ziya S (2020) Allocation of intensive care unit beds in periods of high demand. *Operations Research* 68(2):591–608.

Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Köpf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S (2019) Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems 32*, 8024–8035.

Paternain S, Calvo-Fullana M, Chamon LFO, Ribeiro A (2023) Safe policies for reinforcement learning via primal-dual methods. *IEEE Transactions on Automatic Control* 68(3):1321–1336.

Plotnikoff KM, Krewulak KD, Hernández L, Spence K, Foster N, Longmore S, Straus SE, Niven DJ, Parsons Leigh J, Stelfox HT, Fiest KM (2021) Patient discharge from intensive care: an updated scoping review to identify tools and practices to inform high-quality care. *Critical Care* 25(1):438.

Salgado C, Azevedo C, Proença H, Vieira S (2016) Missing data. *Secondary Analysis of Electronic Health Records* (Cham: Springer).

Sato M, Suenaga E, Koga S, Matsuyama S, Kawasaki H, Maki F (2009) Early tracheal extubation after on-pump coronary artery bypass grafting. *Ann Thorac Cardiovasc Surg* 15(4):239–242.

Shi P, Helm JE, Deglise-Hawkinson J, Pan J (2021) Timing it right: Balancing inpatient congestion vs. readmission risk at discharge. *Operations Research* 69(6):1842–1865.

Tang C, Abbatematteo B, Hu J, Chandra R, Martín-Martín R, Stone P (2024) Deep reinforcement learning for robotics: A survey of real-world successes.

Tang S, Wiens J (2021) Model selection for offline reinforcement learning: Practical considerations for healthcare settings. *The 6th Machine Learning for Healthcare Conference*, volume 149 of *Proceedings of Machine Learning Research*, 2–35 (PMLR).

van Hasselt H (2010) Double q-learning. *Advances in Neural Information Processing Systems*, volume 23, 2613–2621.

Wasfy JH, Zigler CM, Choirat C, Wang Y, Dominici F, Yeh RW (2017) Readmission rates after passage of the hospital readmissions reduction program: a pre–post analysis. *Annals of Internal Medicine* 166(5):324–331.

Wu CP, Shirley RB, Milinovich A, Liu K, Mireles-Cabodevila E, Khouli H, Duggal A, Bhattacharyya A (2025) Exploring timely and safe discharge from icu: a comparative study of machine learning predictions and clinical practices. *Intensive Care Medicine Experimental* 13(1):10.

Xu H, Zhan X, Zhu X (2022) Constraints penalized q-learning for safe offline reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*.

You SB, Ulrich CM (2024) Ethical considerations in evaluating discharge readiness from the intensive care unit. *Nursing Ethics* 31(5):896–906.

Zhang Q, Li T, Li D, Lu W (2024) A goal-oriented reinforcement learning for optimal drug dosage control. *Annals of Operations Research* 338:1403–1423.

Appendix A. Summary Statistics in the Dataset

Table A1.: Definitions of physiological variables used in the numerical experiments.

Variable	Definition
Age	Patient's chronological age in years.
Gender (Male) (%)	Percentage of male patients in the cohort.
Weight	Patient's body weight measured in kilograms.
Heart Rate	Number of heartbeats per minute.
Arterial O_2 Pressure	Partial pressure of oxygen in arterial blood, reflecting oxygenation status.
Hemoglobin	Concentration of hemoglobin in blood, responsible for oxygen transport.
Arterial CO_2 Pressure	Partial pressure of carbon dioxide in arterial blood, indicating ventilatory status.
Hematocrit (Serum)	Percentage of red blood cells in serum by volume.
WBC	White blood cell count, indicating immune response and infection status.
Chloride (Serum)	Serum chloride concentration, important for maintaining acid-base balance.
Creatinine (Serum)	Serum concentration of creatinine, a marker of renal function.
Glucose (Serum)	Serum glucose concentration, indicative of blood sugar regulation.
Glucose (Whole Blood)	Glucose concentration measured in whole blood, reflecting real-time glycemic status.
Magnesium	Serum magnesium level, essential for neuromuscular and enzymatic functions.
Sodium (Serum)	Serum sodium concentration, critical for fluid and electrolyte balance.
pH (Venous)	Acidity or alkalinity of venous blood, used to assess metabolic status.
pH (Arterial)	Acidity or alkalinity of arterial blood, important for evaluating respiratory and metabolic conditions.
FiO_2	Fraction of inspired oxygen in inhaled air, used in oxygen therapy management.
Tidal Volume	Volume of air moved into or out of the lungs during normal breathing.
Respiratory Rate	Number of breaths taken per minute.
Arterial Base Excess	Amount of excess or insufficient base in the blood, indicating metabolic disturbances.
BUN	Blood urea nitrogen level, a key indicator of renal function.
Ionized Calcium	Serum ionized calcium concentration, critical for cardiac and neuromuscular activity.
Total Bilirubin	Total concentration of bilirubin in blood, used to assess liver function.
Potassium (Serum)	Serum potassium concentration, essential for cellular function and cardiac activity.
HCO_3 (Serum)	Serum bicarbonate concentration, contributing to acid-base homeostasis.
Platelet Count	Number of platelets per unit volume of blood, critical for blood clotting.
Prothrombin Time	Time taken for blood to clot via the extrinsic pathway, indicative of coagulation status.
PTT (Partial Thromboplastin Time)	Time taken for blood to clot via the intrinsic pathway, assessing coagulation function.
INR (International Normalized Ratio)	Standardized ratio of prothrombin time, used to monitor anticoagulation therapy.
Blood Pressure Systolic	Maximum arterial pressure during heart contraction.
Blood Pressure Diastolic	Minimum arterial pressure between heartbeats.
Blood Pressure Mean	Average arterial pressure during a cardiac cycle.
Temperature (°C)	Core body temperature measured in degrees Celsius.
SaO_2	Arterial oxygen saturation, reflecting the percentage of oxygenated hemoglobin.
GCS Score	Glasgow Coma Scale score, assessing a patient's level of consciousness.

Table A2.: Summary statistics of physiological variables. IQR = interquartile range.

Variable	**Mean**	**SD**	**Q1**	**Q3**	**IQR**
Age	63.91	16.04	54.00	76.00	22.00
Weight	82.75	35.78	66.54	94.98	28.44
Heart Rate	85.49	16.35	73.73	96.00	22.27
Arterial O_2 Pressure	123.90	46.20	90.60	147.00	56.40
Arterial CO_2 Pressure	41.20	7.74	36.61	44.82	8.21
Hemoglobin	9.90	1.89	8.50	11.10	2.60
PH (Venous)	7.38	0.06	7.35	7.41	0.06
PH (Arterial)	7.40	0.06	7.37	7.44	0.07
Hematocrit (serum)	30.22	5.46	26.20	33.50	7.30
WBC	11.27	5.48	7.60	13.80	6.20
Chloride (serum)	103.25	6.39	99.20	107.00	7.80
Creatinine (serum)	1.33	0.97	0.70	1.58	0.88
Glucose (whole blood)	135.72	31.72	115.80	150.00	34.20
Glucose (serum)	132.44	43.49	103.00	150.00	47.00
Magnesium	2.09	0.32	1.90	2.30	0.40
Sodium (serum)	138.58	4.94	136.00	141.00	5.00
Inspired O_2 Fraction	49.80	15.23	40.00	56.67	16.67
Arterial Base Excess	0.77	3.92	-1.00	2.67	3.67
BUN	28.72	22.12	14.00	36.40	22.40
Ionized Calcium	1.13	0.07	1.09	1.16	0.07
Total Bilirubin	1.14	1.36	0.44	1.20	0.76
Potassium (serum)	4.09	0.53	3.70	4.40	0.70
HCO_3 (serum)	24.73	4.84	22.00	27.00	5.00
Platelet Count	205.33	112.48	128.00	260.00	132.00
Prothrombin time	15.15	4.07	12.60	16.20	3.60
PTT	36.98	15.30	27.90	39.50	11.60
INR	1.37	0.37	1.10	1.50	0.40
Blood Pressure Systolic	119.13	18.70	105.83	130.83	25.00
Blood Pressure Diastolic	63.19	12.36	54.60	70.77	16.17
Blood Pressure Mean	78.71	12.66	70.00	86.33	16.33
Temperature C	36.86	0.50	36.58	37.11	0.53
SaO_2	96.58	2.19	95.17	98.17	3.00
GCS Score	12.86	3.19	11.00	15.00	4.00
Respiratory Rate	14.99	7.23	9.67	19.90	10.23
Tidal Volume	0.48	0.13	0.40	0.55	0.15

Appendix B. Relationship between Readmission Count and Mortality Risk

This section analyzes the relationship between readmission count r and the optimal cost function $C^*(s=(x,r))$ related to mortality risk under the unconstrained optimization problem (4) presented in Section 3, particularly examining their relationship when the physiological conditions are identical.

Assumption B.1. *For each ICU admission, from any initial state $s=(x,r)$, the policy π would discharge the patient — i.e., selects the action $a=1$ — within a finite number of steps with probability one. Formally,*

$$P^{\pi}\big(\exists\, t<\infty : a_t = 1 \;\big|\; s=(x,r)\big) = 1, \qquad \forall\, x, \forall\, r.$$

In particular, no trajectory governed by π leaves a patient in the ICU infinitely.

In clinical settings, every ICU admission must eventually result in either successful recovery or death. Empirically, ICU stays longer than one month are rare (Flaws et al. 2024, Lefering et al. 2024), and the primary goal of ICU care is to stabilize and improve the patient's condition. Organizational rules require discharging patients from the ICU within 4-24 h once stability criteria are met (National Institute for Health and Care Excellence 2017). Also, ethical guidelines mandate withdrawal of futile life support rather than infinite ICU care (Dabi and Rahman 2023). Assumption B.1 formalizes these clinical and administrative constraints by ruling out pathological or unrealistic trajectories. It ensures that each decision episode ends in finite time, allowing the total expected cost to be meaningfully defined even without a discount factor. Moreover, this assumption guarantees the convergence of dynamic programming algorithms such as value iteration and backward induction, since the process almost surely reaches a terminal state.

In the formulation where the state space incorporates the readmission count r, we further investigate an important monotonicity property of the optimal cost function $C^*(s=(x,r))$ with respect to r, which reflects the compounding risk associated with multiple ICU admissions. We need one general assumption used in medical decision-making process about the state transition probability $P^a(s_{t+1}|s_t)$.

Assumption B.2. *For the fixed decision/action $a\in\mathcal{A}$, the transition probabilities $P^a(s'=(x',r') \mid s=(x,r))$ satisfy the **First-Order Stochastic Dominance (FOSD)** condition with respect to the physiological condition x, and the readmission count r. Specifically, for any $f_h(x_1)\le f_h(x_2)$ and $r_1\le r_2$, the following holds:*

$$\sum_{s'\in\mathcal{S}:f_h(x')\le f_h(x)} P^a\Big(s'\Big|s_1=(x_1,r_1)\Big) \ge \sum_{s'\in\mathcal{S}:f_h(x')\le f_h(x)} P^a\Big(s'\Big|s_2=(x_2,r_2)\Big), \quad \forall x, \tag{B1}$$

where $f_h(x)$ represents a severity scoring system regarding to the physiological states, and lower values of $f_h(x)$ correspond to a more favorable health condition and less mortality risk (e.g., APACHE II score, SAPS score, and SOFA score). As for the terminal state s_H and s_D, we could define $f_h(s_H)=-\infty$ and $f_h(s_D)=+\infty$ to indicate the most favorable and least favorable outcomes, respectively.

This assumption formalizes that a lower readmission count r leads to stochastically

more favorable state transitions. For the primary objective cost – mortality risk – it is essential to ensure that the following condition holds: if two patients exhibit identical physiological conditions (i.e., the same x or severity score) but differ in their number of readmission counts (i.e., different r), the patient with fewer readmission counts should have a comparatively lower mortality risk. Readmissions can expose patients to additional risks of complications, infections, and other adverse events associated with hospital stays (Shah et al. 2018). Therefore, this condition reflects the intuition that fewer readmission counts generally correlate with better health stability and reduced risk of adverse health outcomes. We conclude this in the Lemma B.3, with the proof is provided in Appendix C.

Lemma B.3. *For the optimal policy π^* of the unconstrained problem (4), the optimal cost function $C^*(s)$, i.e., $C^*(x, r)$, is non-decreasing with respect to the number of readmission times r for any given physiological condition x.*

Therefore, if clinicians aim to minimize the expected cumulative mortality risk $C(\pi)$ as the sole objective during the ICU discharge decision-making process (i.e., the optimization problem (4) without any constraints included), then the readmission (r increases) should be avoided. Since longer ICU LOS should be able to reduce both the mortality risk and readmission risk (Shi et al. 2021), the patients should stay in the ICU as long as possible until the mortality risk (transiting into the terminal state s_D) is minimized.

Appendix C. Proof of Lemma B.3

Proof. Assume that we are using value-iteration to derive the optimal cost function in (4), then for iteration $n = 0, 1, 2, \ldots$, we can define the cost function

$$\begin{aligned} C_{n+1}(s) &= \min_{a\in\mathcal{A}} \left\{ c(s,a) + \sum_{s'\in\mathcal{S}} P^a\left(s'|s\right) C_n\left(s'\right) \right\} \\ &= \min \left\{ c\Big(s, a=1\Big) + \sum_{s'\in\mathcal{S}} P^{a=1}\Big(s'\Big|s\Big) C_n\left(s'\right), \right. \\ &\qquad \left. c\Big(s, a=0\Big) + \sum_{s'\in\mathcal{S}} P^{a=0}\Big(s'\Big|s\Big) C_n\left(s'\right) \right\}, \quad \forall s \in \mathcal{S}. \end{aligned} \tag{C1}$$

Since the instantaneous objective cost function (1) also depends on the next state s', we could have

$$\begin{aligned} c\Big(s, a=1\Big) &= \sum_{s'\in\mathcal{S}} P^{a=1}\Big(s'\Big|s\Big) c\left(s, a=1, s'\right), \\ c\Big(s, a=0\Big) &= \sum_{s'\in\mathcal{S}} P^{a=0}\Big(s'\Big|s\Big) c\left(s, a=0, s'\right). \end{aligned} \tag{C2}$$

Now, we assume that there are two states $s_1 = (x_1, r_1)$ and $s_2 = (x_2, r_2)$, where $x_1 = x_2 = x$ (i.e., $f_h(x_1) = f_h(x_2)$) and $r_1 \leq r_2$. By the definition of the instantaneous objective cost function (1) corresponding to the mortality risk, for $a = 0$, we could get

$$c\Big(s_1 = (x, r_1), a = 0\Big) = c\Big(s_2 = (x, r_2), a = 0\Big) = 0. \tag{C3}$$

Next, for $a = 1$, the possible next state could be $s' = (x', r' = r + 1)$, s_H, and s_D. By the FOSD assumption (Assumption B.2) and instantaneous objective cost function (1), the state s_2 possesses higher probability transiting into the terminal state s_D, which is considered as the worst health condition ($f_h(s_D) = +\infty$), we can get that

$$c\Big(s_1 = (x, r_1), a = 1\Big) \leq c\Big(s_2 = (x, r_2), a = 1\Big). \tag{C4}$$

Therefore, both $c\big(s = (x, r), a = 0\big)$ and $c\big(s = (x, r), a = 1\big)$ are non-decreasing in r for a fixed physiological condition x.

Assuming that $C_n\Big(s' = (x', r')\Big)$ is non-decreasing in r' for the fixed x', then by Assumption B.2, we can get

$$\sum_{s'\in\mathcal{S}} P^a\Big(s'|s = (x, r_1)\Big) C_n\Big(s' = (x', r_1')\Big) \leq \sum_{s'\in\mathcal{S}} P^a\Big(s'|s = (x, r_2)\Big) C_n\Big(s' = (x', r_2')\Big), \quad \forall a. \tag{C5}$$

Therefore, we can derive that, if for a fixed physiological condition x, $C_n(s = (x, r))$ is non decreasing in r, then so is $C_{n+1}(s = (x, r))$. Any bounded function C_0 that is

already non-decreasing in r can be chosen. In our case, according to (C3) and (C4), we can let $C_0(s)$ equal to either $c(s, a=0)$ or $c(s, a=1)$.

According to Assumption B.1, we can prove that the decision-making process terminates almost surely in finite time steps. We can firstly define

$$\tau := \inf\{t \geq 0 : s_t \in \{s_H, s_D\}\}. \tag{C6}$$

Then, we let k denote the k-th time of taking the discharge action, and we can have

$$P^{a_t^k=1}\left(s_{t+1}^k \in \{s_H, s_D\} \middle| s_t^k\right) \geq \epsilon, \quad \exists \epsilon \in (0, 1] \tag{C7}$$

Hence we could have

$$\begin{aligned} P^\pi(\tau = \infty) &= P^\pi\left(\bigcap_{k=1}^{\infty}\left\{s_{t+1}^k \notin \{s_H, s_D\}\right\}\right) \\ &= \mathbb{E}^\pi\left[\prod_{k=1}^{\infty} P^{a_t^k=1}\left(s_{t+1}^k \notin \{s_H, s_D\} \middle| s_t^k\right)\right] \\ &\leq \mathbb{E}^\pi\left[\prod_{k=1}^{\infty}(1-\epsilon)\right] \\ &= (1-\epsilon)^\infty = 0. \end{aligned} \tag{C8}$$

Therefore, the discharge decision-making process will reach one of the two terminal states in finite time with probability 1.

Next, we need to prove that $\lim_{n\to\infty} C_n(s) = C^*(s)$. Let us consider the Bellman recursion process over a finite-horizon. For $n = 0, 1, 2, \ldots$, we can have

$$\begin{aligned} V^{(n)}(s) &= \min_\pi \mathbb{E}\left[\sum_{t=0}^{n} c(s_t, \pi(s_t))\right] \\ &= \min_{a\in\mathcal{A}}\left\{c(s,a) + \sum_{s'\in\mathcal{S}} P^a(s'|s) \min_\pi \mathbb{E}\left[\sum_{t=1}^{n} c(s_t, \pi(s_t))\right]\right\} \\ &= \min_{a\in\mathcal{A}}\left\{c(s,a) + \sum_{s'\in\mathcal{S}} P^a(s'|s) V^{(n-1)}(s')\right\}, \end{aligned} \tag{C9}$$

and $V^{(0)}(s) = \min_{a\in\mathcal{A}} c(s,a)$.

Then, we can prove the step-to-step equality $C_n = V^{(n)}$. If $n = 0$, we can let $C_0(s) = V^{(0)}(s) = \min_{a\in\mathcal{A}} c(s,a)$, which satisfies the requirement for $C_0(s)$. Then, We

can assume that $C_n(s) = V^{(n)}(s)$ for any state s, and we can have

$$\begin{aligned} C_{n+1}(s) &= \min_{a\in\mathcal{A}} \left\{ c(s,a) + \sum_{s'\in\mathcal{S}} P^a\left(s'|s\right) C_n\left(s'\right) \right\} \\ &= \min_{a\in\mathcal{A}} \left\{ c(s,a) + \sum_{s'\in\mathcal{S}} P^a\left(s'|s\right) V^{(n)}\left(s'\right) \right\} \\ &= V^{(n+1)}(s). \end{aligned} \tag{C10}$$

Since the decision process stops almost surely in finite time according to the Assumption B.1, we have for every s

$$\lim_{n\to\infty} C_n(s) = \lim_{n\to\infty} V^{(n)}(s) := C^*(s). \tag{C11}$$

Because each $V^{(n)}$ is non-decreasing in readmission count r for a fixed physiological condition x (same as C_n), the point-wise limit inherits the same property

$$C^*\Big(s_1 = (x, r_1)\Big) \leq C^*\Big(s_2 = (x, r_2)\Big), \quad r_1 \leq r_2. \tag{C12}$$

□

Appendix D. Proof of Theorem 4.5

Proof. We break through the arguments into three parts, analyzing the computational costs of FQI, FQE, and then assembling them in the two algorithm frameworks (Algorithm 1 and Algorithm 4).

For FQI with the neural network approximation method, we assume that each gradient-descent update (including one forward pass and one backward pass) costs $\Theta(P_0)$ for a network with P_0 parameters. Then, for the target computation (calculating the values of y), it required $n|\mathcal{A}|$ forward pass, each costing $\Theta(P_0)$. Hence,

$$C_{target} = \mathcal{O}\left(n|\mathcal{A}|P_0\right).$$

Next, we need to perform one forward pass and one backward pass on the same n samples. This costs $\Theta(nP_0)$. Thus,

$$C_{update} = \mathcal{O}\left(nP_0\right).$$

Lastly, through the Bellman-Backup iterations, FQI repeats the above two phases K_0 times to minimize the loss function (e.g., MSE loss). We should have

$$C_{FQI} = K_0 \times \left(C_{target} + C_{update}\right) = \mathcal{O}\left(nK_0P_0|\mathcal{A}|\right).$$

As for the FQE agent, it estimate the action-value function of a fixed policy (as the input), so it does not enumeration actions when computing target values. Similarly, we can get that for each FQE agent, the computational complexity should be

$$C_{FQE} = \mathcal{O}\left(nK_mP_m\right).$$

Thus, we can derive that the computational complexity of Algorithm 4 should be

$$\mathcal{O}\left(K\left(nK_0P_0|\mathcal{A}| + n\sum_{m=1}^{M}K_mP_m\right)\right) \rightarrow \mathcal{O}\left(nK\left(K_0P_0|\mathcal{A}| + \sum_{m=1}^{M}K_mP_m\right)\right).$$

For our proposed approach – Algorithm 1, all parameters are now updated within the same training step. We avoid performing policy optimization at each training step, nor is there a complete FQE performed on the updated policy at each training step. Thus, the computational complexity can be concluded as

$$\mathcal{O}\left(K\left(nP_0|\mathcal{A}| + n\sum_{m=1}^{M}P_m\right)\right) \rightarrow \mathcal{O}\left(nK\left(P_0|\mathcal{A}| + \sum_{m=1}^{M}P_m\right)\right).$$

□

Appendix E. Algorithm

In this section, we list the critical algorithms referenced in the main body of the paper.

Algorithm 2 FQI with Regulated Loss Function for Discharge Decision-Making

1: Input the dataset of one-step transitions $\mathcal{D} = \left\{(s_i, a_i, s'_i, c_i)_{i=1}^{N}\right\}$, batch size n, number of training steps K, learning rate α_{fqi}, weight of penalty ω, soft-update rate κ, soft-update frequency U

2: Initialize the parametric Q-function estimator with random parameters θ_0

3: **for** step $k = 1 \rightarrow K$ **do**

4: Randomly sample n transitions from $\mathcal{D}$

5: Compute the target values for all n samples $(\forall j \in [1, n])$

$$y_j = \begin{cases} c_j & \text{if } s'_j = s_H \text{ or } s_D \\ c_j + \min\limits_{a' \in \mathcal{A}} Q\left(s'_j, a' | \theta_{k-1}\right) & \text{o.w.} \end{cases}$$

6: Update the parameter by

$$\theta_k \leftarrow \theta_{k-1} - \alpha \nabla_{\theta_{k-1}} \mathcal{L}(\theta_{k-1}),$$

where the loss function is

$$\mathcal{L}_{reg}\left(\theta_{k-1}, \omega\right) \triangleq \frac{1}{n}\left[\sum_{j=1}^{n}\left(y_j - Q\left(s_j, a_j | \theta_{k-1}\right)\right)^2\right] + \omega \left\|\theta_{k-1}\right\|^2.$$

7: **end for**

8: Output: θ_K

Algorithm 3 FQE with Sampling for Discharge Decision-Making

Input dataset $\mathcal{D} = \left\{(s_i, a_i, s'_i, c_i)_{i=1}^{N}\right\}$, policy $\tilde{\pi}$ to be evaluated, batch size n, number of training steps K, learning rate α_{fqe}, and the soft-update rate κ, soft-update frequency U

Initialize the first parametric Q-function estimator with random parameters θ_0

for step $k = 1 \rightarrow K$ **do**

Randomly sample n transitions from $\mathcal{D}$

Compute the target values for all n samples

$$y_j = \begin{cases} c_j & \text{if } s'_j = s_H \text{ or } s_D \\ c_j + Q\left(s'_j, \tilde{\pi}(s'_j)\middle|\theta_{k-1}\right) & \text{o.w.} \end{cases}, \quad \forall j \in [1, n]$$

Update the parameter by

$$\theta_k \leftarrow \theta_{k-1} - \alpha \nabla_{\theta_{k-1}} \mathcal{L}(\theta_{k-1})$$

end for

Output $\hat{C}^{\tilde{\pi}}(s) = Q\left(s, \tilde{\pi}(s)|\theta_K\right) \quad \forall s$

Algorithm 4 Iterative Primal-Dual Update with FQI & FQE

Input learning rate α_λ for Lagrangian multiplier

Initialize $\lambda_0 = 0$

for step $k = 1, 2, \ldots, K$ **do**

Update policy π via FQI algorithm (until convergence)

$$\pi_k \leftarrow \text{FQI}(\lambda_{k-1})$$

Update the dual variable via using FQE to evaluate $G(\pi_k)$, and another FQE agent need to be trained

for evaluating $C(\pi_k)$

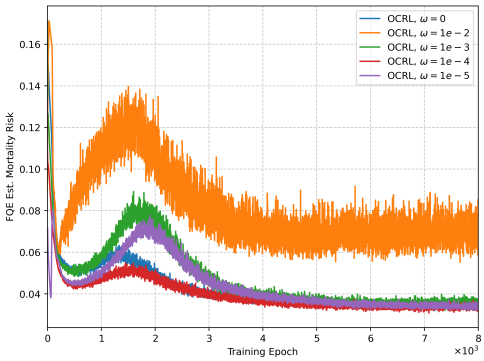
$$\lambda_k = \left[\lambda_{k-1} + \alpha_\lambda \cdot (\hat{G}(\pi_k) - l)\right]^+$$

end for

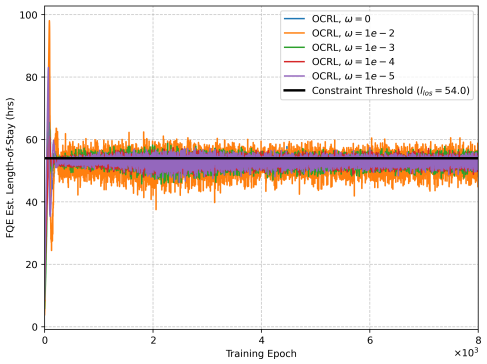
Appendix F. Hyperparameter Sensitivity Analysis

This section examines the influence of key hyperparameters in training our proposed multi-timescale offline policy learning framework (Algorithm 1). The first part focuses on the weight of the L2 regularization term in (14), denoted by ω, while the second part analyzes the learning rates of the two Lagrangian multipliers, α_λ^{rr} and α_λ^{los}.

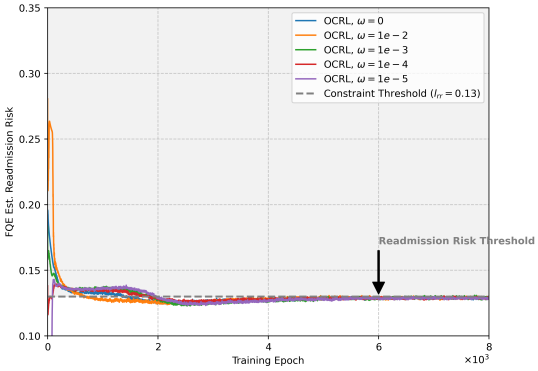
F.1. L2 Regularization



(a) Training process of FQE agents for objective cost - mortality risk.



(b) Training process of FQE agents for constraint cost - ICU LOS.

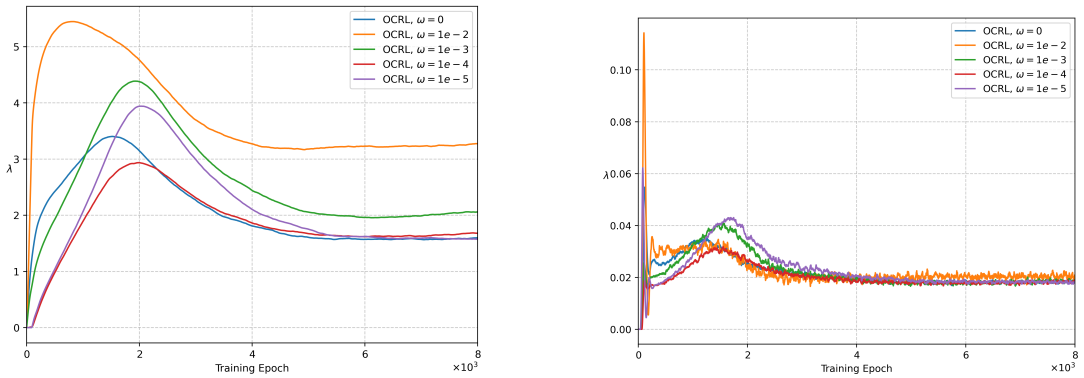


(c) Training process of FQE agents for constraint cost - readmission risk.

Figure F1.: Training processes of FQE agents in Algorithm 1 with varying weights of the L2 regularization term, under the fixed thresholds $l_{los} = 54$ and $l_{rr} = 0.13$.

The complete training processes for five different weights ω of the L2 regularization term in (14) are shown in Figures F1 and F2. With different ω settings, all configurations successfully converge within 8×10^6 steps. We observe no significant differences among the training processes for $\omega \in \{0, 1 \times 10^{-3}, 1 \times 10^{-4}, 1 \times 10^{-5}\}$. As shown in Figure F1a, all settings achieve significantly low FQE-estimated mortality risk ($\hat{C}(\pi_K)$). This demonstrates that our proposed algorithmic framework for multi-timescale offline constrained policy learning can improve overall performance in ICU discharge decision-making, as evaluated by FQE-estimated average mortality risk, while satisfying constraints on average readmission risk and ICU length of stay (LOS), both also estimated via FQE. However, since FQE is only one OPE method that provides a perspective on policy value estimation (Tang and Wiens 2021, Kondrup et al. 2024), we

must further examine the performance of these policies in the ICU discharge decision-making process across different readmission counts r. Table F1 presents the action distributions for different regularization settings. To evaluate their performance on a larger set of samples, we use data entries from both the validation set and the testing set. OCRL-v0 Policy denotes the final output policy π_K obtained from Algorithm 1 with $\omega = 0$. Similarly, OCRL-v2, OCRL-v3, and OCRL-v4 Policies correspond to π_K obtained using $\omega = 1 \times 10^{-3}$, 1×10^{-4}, and 1×10^{-5}, respectively.



(a) The training process of λ_{rr}.

(b) The training process of λ_{los}.

Figure F2.: The training process of Lagrangian multipliers $\boldsymbol{\lambda}$ inside Algorithm 1.

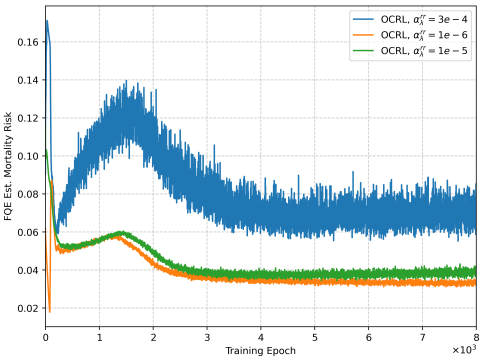
Across all four ω settings, we observe substantial divergence between the trained policies and clinical practices in MIMIC-IV for all readmitted cases ($r \geq 1$), particularly for cases with readmission counts of two or more ($r \geq 2$), where sample sizes in the data are limited. In these relatively rare but clinically more severe and high-risk readmitted cases, the learned policies exhibit behavior similar to their approach for first-time ICU admissions—namely, they adopt a considerably more aggressive discharge strategy than clinical practice, with a disproportionately high frequency of discharge actions ($a = 1$). However, readmitted cases, particularly those with $r = 1$ and $r = 2$, constitute a non-negligible portion of the dataset, and clinicians typically employ more conservative discharge policies for these patients, recognizing their elevated risk profiles and history of potentially premature discharge (Ofoma et al. 2018).

Therefore, the OCRL-v1 Policy presented in the main text employs $\omega = 1 \times 10^{-2}$. While the FQE-estimated mortality risk for $\omega = 1 \times 10^{-2}$ is higher than that achieved with other ω settings, Table F1 demonstrates that this configuration yields a more clinically reasonable action distribution. When the L2 regularization weight in (14) is set to $\omega = 1 \times 10^{-2}$, the resulting policy exhibits more conservative behavior across all readmitted cases ($r \geq 1$) compared to policies derived with alternative ω values. Notably, for cases with multiple readmissions ($r \geq 3$), which are characterized by limited sample sizes but heightened clinical risk, the OCRL-v1 Policy demonstrates even greater conservatism than the clinical practices observed in MIMIC-IV, applying discharge actions ($a = 1$) with particular caution in these sensitive scenarios. Similar to Conservative Q-Learning (CQL) and other conservative offline RL methods, the L2-regularized policy ($\omega = 1 \times 10^{-2}$) exhibits more pessimistic FQE-estimated returns but demonstrates more clinically appropriate behavior, particularly in high-risk scenarios. This trade-off between pessimistic value estimates and practical safety is a well-documented characteristic of conservative algorithms in offline RL (Kumar et al. 2020).

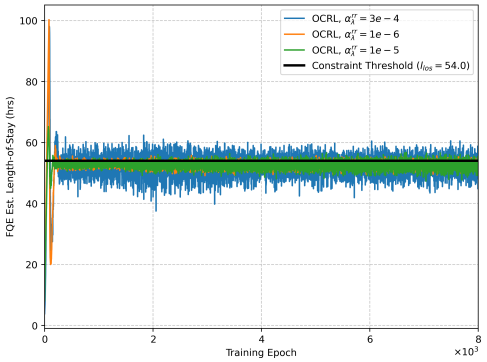
Table F1.: Comparison of discharge action ratios ($a = 1$) across admission types categorized by readmission count (r) in both $\mathcal{D}_{val}$ and $\mathcal{D}_{test}$ among clinical practice and OCRL policy ($l_{rr} = 0.13$, $l_{los} = 54.0$) with various weights of regularization term (ω).

Admission Type	**Policy**	**Stay in ICU** ($a = 0$)		**Discharge** ($a = 1$)	
		Count	Ratio	Count	Ratio
All Cases	Clinical Practice	84392	0.860	13748	0.140
	OCRL-v0 Policy	65779	0.670	32361	0.330
	OCRL-v1 Policy	63953	0.652	34187	0.348
	OCRL-v2 Policy	63330	0.645	34810	0.355
	OCRL-v3 Policy	68168	0.695	29972	0.305
	OCRL-v4 Policy	69128	0.704	29012	0.296
First Admission ($r = 0$)	Clinical Practice	55456	0.823	11950	0.177
	OCRL-v0 Policy	41506	0.616	25900	0.384
	OCRL-v1 Policy	36875	0.547	30531	0.453
	OCRL-v2 Policy	38584	0.572	28822	0.428
	OCRL-v3 Policy	43029	0.638	24377	0.362
	OCRL-v4 Policy	44561	0.661	22845	0.339
First Readmission ($r = 1$)	Clinical Practice	16166	0.917	1461	0.083
	OCRL-v0 Policy	13021	0.739	4606	0.261
	OCRL-v1 Policy	14341	**0.814**	3286	**0.186**
	OCRL-v2 Policy	12987	0.737	4640	0.263
	OCRL-v3 Policy	13493	0.765	4134	0.235
	OCRL-v4 Policy	13365	0.758	4262	0.242
Second Readmission ($r = 2$)	Clinical Practice	7362	0.965	267	0.035
	OCRL-v0 Policy	6188	0.811	1441	0.189
	OCRL-v1 Policy	7265	**0.952**	364	**0.048**
	OCRL-v2 Policy	6460	0.847	1169	0.153
	OCRL-v3 Policy	6426	0.842	1205	0.158
	OCRL-v4 Policy	6209	0.814	1420	0.186
Third Readmission ($r = 3$)	Clinical Practice	2893	0.983	50	0.017
	OCRL-v0 Policy	2650	0.900	293	0.100
	OCRL-v1 Policy	2937	**0.998**	6	**0.002**
	OCRL-v2 Policy	2799	0.951	144	0.049
	OCRL-v3 Policy	2754	0.936	189	0.064
	OCRL-v4 Policy	2626	0.892	317	0.108
Fourth Readmission ($r = 4$)	Clinical Practice	1952	0.992	15	0.008
	OCRL-v0 Policy	1860	0.946	107	0.054
	OCRL-v1 Policy	1967	**1.000**	0	**0.000**
	OCRL-v2 Policy	1932	0.982	35	0.018
	OCRL-v3 Policy	1908	0.970	59	0.030
	OCRL-v4 Policy	1832	0.931	135	0.069
Fifth Readmission ($r = 5$)	Clinical Practice	563	0.991	5	0.009
	OCRL-v0 Policy	554	0.975	14	0.025
	OCRL-v1 Policy	568	**1.000**	0	**0.000**
	OCRL-v2 Policy	568	1.000	0	0.000
	OCRL-v3 Policy	560	0.986	8	0.014
	OCRL-v4 Policy	535	0.942	33	0.058

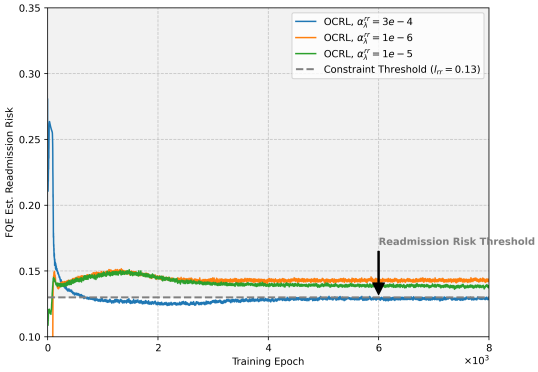
F.2. Learning Rates for Lagrangian Multipliers



(a) Training process of FQE agents for objective cost - mortality risk.



(b) Training process of FQE agents for constraint cost - ICU LOS.



(c) Training process of FQE agents for constraint cost - readmission risk.

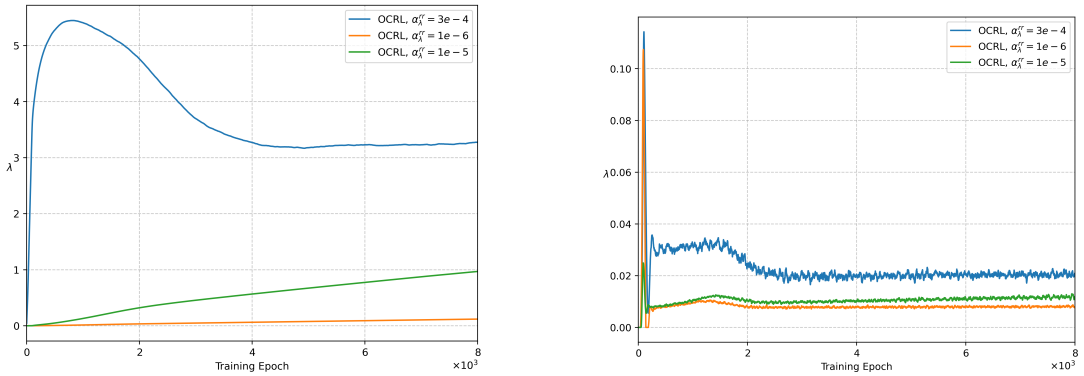
Figure F3.: The training process of FQE agents in Algorithm 1 with varying learning rates α_λ^{rr}, under the fixed thresholds $l_{los} = 54$ and $l_{rr} = 0.13$.

In our indefinite-horizon CMDP formulation of the ICU discharge decision-making problem, we incorporate two constraints – readmission risk and ICU length-of-stay (LOS). Accordingly, we introduce two learning rates, α_λ^{rr} and α_λ^{los}, for updating the corresponding Lagrangian multipliers (dual variables) λ_{rr} and λ_{los}. The primary objective in choosing these learning rates is to ensure the stability of the resulting policy π_K, thereby enabling the multi-timescale offline constrained policy learning framework (Algorithm 1) to converge both reliably and efficiently.

The learning rate combination used in our main analysis is $\alpha_\lambda^{rr} = 3 \times 10^{-4}$ and $\alpha_\lambda^{los} = 1 \times 10^{-6}$. This setting not only satisfies the condition given by (20), but also successfully converges to stable policies π_K across various given thresholds l_{rr} and l_{los} in our numerical experiments.

We first fix $\alpha_\lambda^{los} = 1 \times 10^{-6}$ and use our most commonly employed thresholds $l_{rr} = 0.13$ and $l_{los} = 54.0$ to examine what occurs when we adopt α_λ^{rr} values that are significantly smaller than 3×10^{-4}. Figure F3 and F4 present the complete training processes, where we compare $\alpha_\lambda^{rr} = 3 \times 10^{-4}$ against $\alpha_\lambda^{rr} = 1 \times 10^{-5}$ and $\alpha_\lambda^{rr} = 1 \times 10^{-6}$. As shown in Figure F3c, when α_λ^{rr} is reduced to 1×10^{-5} or 1×10^{-6}, the OCRL agent cannot meet the readmission risk constraint of $l_{rr} = 0.13$. Correspondingly, Figure F4a demonstrates that the Lagrangian multiplier λ_{rr} fails to converge within the allocated

8×10^6 training steps under these lower learning rates.



(a) The training process of λ_{rr}.

(b) The training process of λ_{los}.

Figure F4.: The training process of Lagrangian multipliers $\boldsymbol{\lambda}$ inside Algorithm 1 with varying learning rates α_λ^{rr}, under the fixed thresholds $l_{los} = 54$ and $l_{rr} = 0.13$.

Moreover, we maintain $\alpha_\lambda^{los} = 1 \times 10^{-6}$ but impose a stricter constraint threshold of $l_{rr} = 0.12$ to compare the performance of $\alpha_\lambda^{rr} = 3 \times 10^{-4}$ against $\alpha_\lambda^{rr} = 4 \times 10^{-4}$. Figure F5 reveals that with the higher learning rate ($\alpha_\lambda^{rr} = 4 \times 10^{-4}$, orange line), the Lagrangian multiplier λ_{rr} undergoes wild oscillations of increasing magnitude, showing no signs of approaching convergence throughout 8×10^6 training steps.

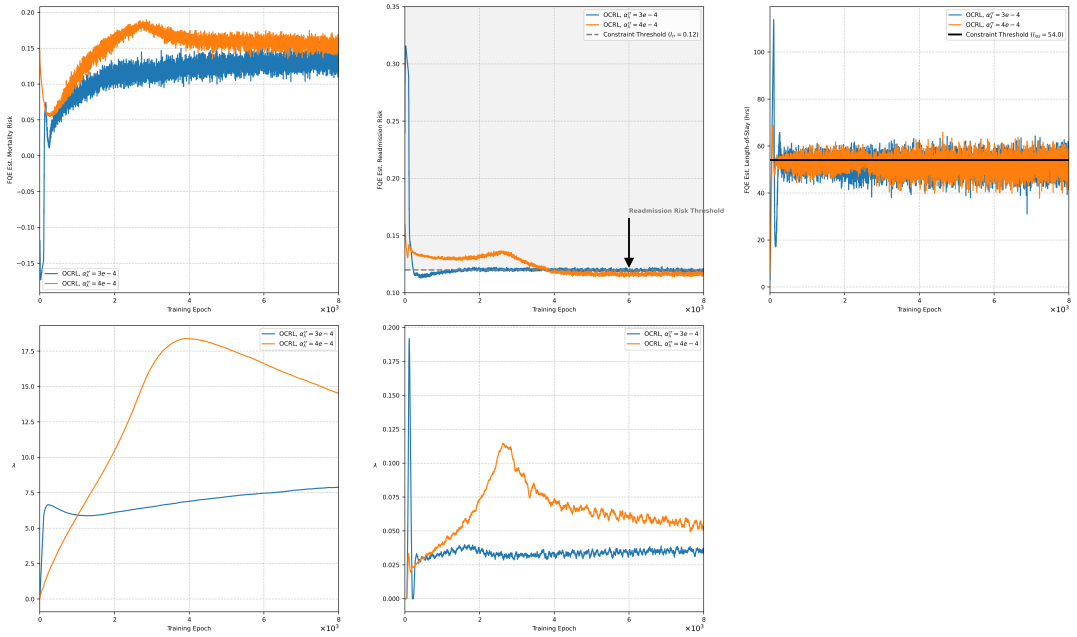
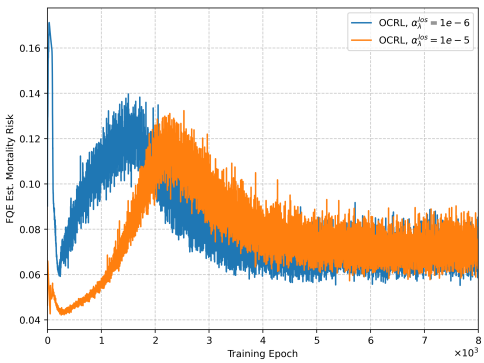
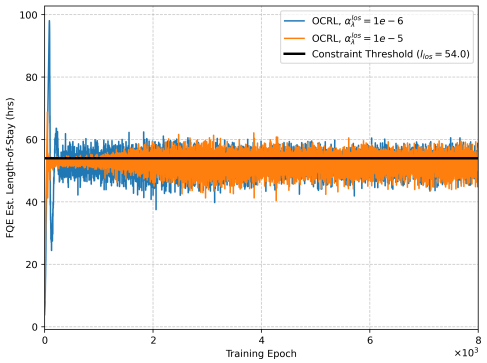


Figure F5.: The training dynamics of FQE agents and Lagrangian multipliers inside Algorithm 1 with varying learning rates α_λ^{rr}, under the fixed thresholds $l_{los} = 54$ and $l_{rr} = 0.12$.

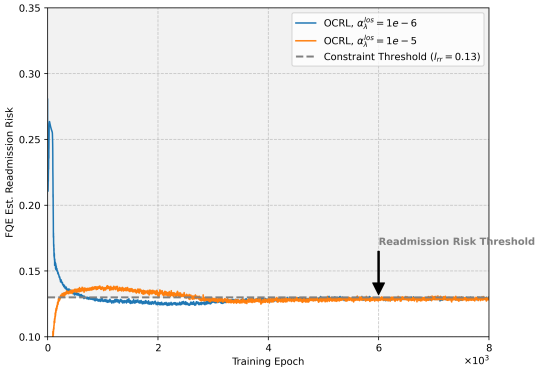
We next examine the learning rate α_λ^{los} for the Lagrangian multiplier λ_{los} by comparing our baseline value of $\alpha_\lambda^{los} = 1 \times 10^{-6}$ with an order-of-magnitude increase to $\alpha_\lambda^{los} = 1 \times 10^{-5}$. While both learning rates achieve convergence within 8×10^6 steps and satisfy the constraint thresholds $l_{los} = 54.0$ and $l_{rr} = 0.13$ (Figures F6 and F7),



(a) Training process of FQE agents for objective cost - mortality risk.



(b) Training process of FQE agents for constraint cost - ICU LOS.



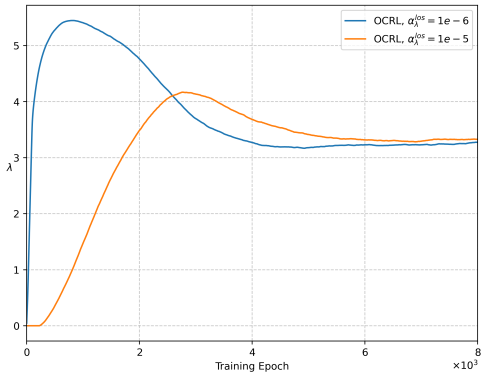
(c) Training process of FQE agents for constraint cost - readmission risk.

Figure F6.: The training process of FQE agents in Algorithm 1 with varying learning rates α_λ^{los}, under the fixed thresholds $l_{los} = 54$ and $l_{rr} = 0.13$.

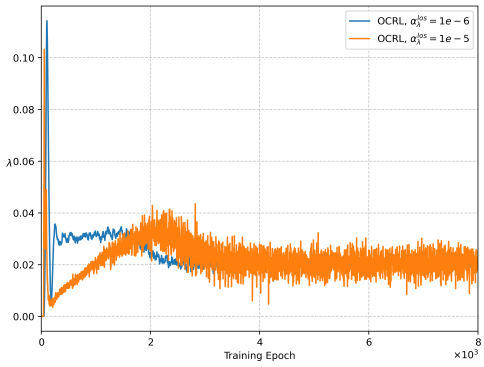
the higher learning rate ($\alpha_\lambda^{los} = 1 \times 10^{-5}$) introduces significant oscillations in λ_{los} updates, as evident in Figure F7b.

We conduct additional experiments using alternative constraint threshold settings, ($l_{los} = 48.0, , l_{rr} = 0.13$) and ($l_{los} = 60.0, , l_{rr} = 0.13$), to compare the performance of $\alpha_\lambda^{los} = 1 \times 10^{-6}$ with that of a lower learning rate, $\alpha_\lambda^{los} = 1 \times 10^{-7}$. As shown in Figures F8 and F9, both learning rates produce comparable training performance. For $l_{los} = 48.0$, the adopted rate of 1×10^{-6} results in a more stable overall training trajectory, with notably faster convergence of λ_{los}. The smaller rate, 1×10^{-7}, also performs well – although it converges more slowly, it ultimately produces smoother and more stable convergence curves. Overall, under both threshold settings, each learning rate yields stable policy outputs by the end of training.

In conclusion, Algorithm 1 demonstrates robust performance across a range of learning rate selections for both α_λ^{rr} and α_λ^{los}. The selected learning rate combination ensures stable policy convergence under diverse constraint threshold configurations, validating the algorithmic framework's reliability in practical applications.



(a) The training process of λ_{rr}.



(b) The training process of λ_{los}.

Figure F7.: The training process of Lagrangian multipliers $\boldsymbol{\lambda}$ inside Algorithm 1 with varying learning rates α_λ^{los}, under the fixed thresholds $l_{los} = 54$ and $l_{rr} = 0.13$.

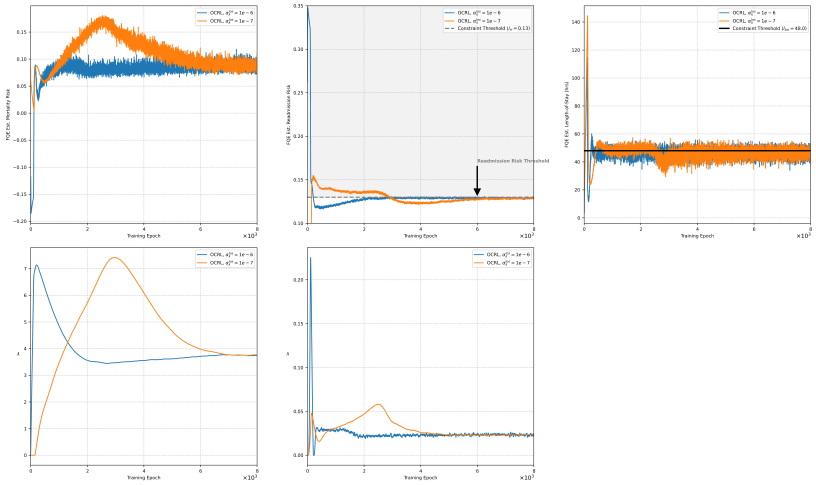


Figure F8.: The training dynamics of FQE agents and Lagrangian multipliers inside Algorithm 1 with varying learning rates α_λ^{los}, under the fixed thresholds $l_{los} = 48$ and $l_{rr} = 0.13$.

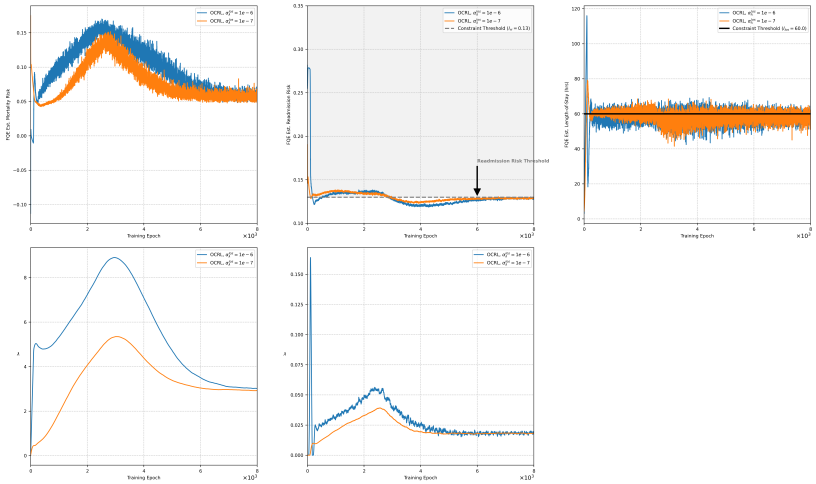


Figure F9.: The training dynamics of FQE agents and Lagrangian multipliers inside Algorithm 1 with varying learning rates α_λ^{los}, under the fixed thresholds $l_{los} = 60$ and $l_{rr} = 0.13$.

Appendix G. Policy Performance without Considering Readmission Status

To assess the importance of explicitly modeling readmission status, we evaluate a baseline formulation in which the ICU discharge decision-making process is modeled as a Constrained Markov Decision Process (CMDP) without incorporating the readmission count r into the state representation. In this simplified setting, the state space $\mathcal{S}$ consists solely of physiological measurements x, and the decision process terminates once the discharge action ($a = 1$) is taken.

Under this formulation, the transition dynamics are as follows

- $P^{a_t=0}\Big(s_{t+1} = x_{t+1}\Big|s_t = x_t\Big)$ (the patient is staying in the ICU).
- $P^{a_t=1}\Big(s_{t+1} = s_H\Big|s_t = x_t\Big)$ (the patient is successfully discharged from the ICU).
- $P^{a_t=1}\Big(s_{t+1} = s_R\Big|s_t = x_t\Big)$ (the patient is readmitted into the ICU after being discharged from the ICU).
- $P^{a_t=1}\Big(s_{t+1} = s_D\Big|s_t = x_t\Big)$ (the patient is deceased after being discharged from the ICU).

In this setting, all post-discharge trajectories lead to absorbing terminal states, with s_R representing readmission as a one-time outcome rather than part of a recurrent sequence. The same constraint thresholds (e.g., $l_{rr} = 0.13$, $l_{los} = 54.0$) are applied as in the original model that includes r in the state space. However, since the expected cumulative readmission risk is now bounded within $[0, 1]$—reflecting a single possible readmission event—it becomes easier to satisfy the readmission constraint in this reduced model. Similarly, cumulative ICU LOS is naturally shorter because it does not account for repeated admissions.

Table G1.: Discharge action ratios ($a = 1$) under clinical practice, OCRL-R policy ($l_{rr} = 0.13$, $l_{los} = 54.0$), and OCRL-D policy ($l_{rr} = 0.13$, $l_{los} = 54.0$) across admission types categorized by readmission count (r) in $\mathcal{D}_{test}$.

Admission Type	**Policy**	**Stay in ICU** ($a = 0$)		**Discharge** ($a = 1$)	
		Count	Ratio	Count	Ratio
All Cases	Clinical Practice	42879	0.862	6868	0.138
	OCRL-R Policy	32903	0.661	16844	0.339
	OCRL-D Policy	33918	0.682	15829	0.318
First Admission ($r = 0$)	Clinical Practice	27825	0.823	5996	0.177
	OCRL-R Policy	18644	0.551	15177	0.449
	OCRL-D Policy	21534	0.637	12287	0.363
First Readmission ($r = 1$)	Clinical Practice	8044	0.918	715	0.082
	OCRL-R Policy	7249	0.828	1510	0.172
	OCRL-D Policy	6449	0.736	2310	0.264
Second Readmission ($r = 2$)	Clinical Practice	3689	0.968	122	0.032
	OCRL-R Policy	3655	0.959	156	0.041
	OCRL-D Policy	2843	0.746	968	0.254
Third Readmission ($r = 3$)	Clinical Practice	1470	0.985	23	0.015
	OCRL-R Policy	1492	0.999	1	0.001
	OCRL-D Policy	1365	0.914	128	0.086

To compare the effects of including versus excluding readmission count r in the state representation, we examine policy performance under three scenarios: (1) observed clinical decisions from the dataset, (2) the OCRL-R policy, which incorporates readmission count in the state space, and (3) the OCRL-D policy, which omits it. Table G1 presents a breakdown of ICU discharge ($a = 1$) and retention ($a = 0$) actions under each policy, stratified by readmission count r.

As shown in Table G1, the OCRL-D policy exhibits a more conservative discharge behavior than OCRL-R for first admissions ($r = 0$), but tends to be more aggressive for patients reamitted in to ICU ($r \geq 1$). This is problematic, as patients with multiple times of readmission are at significantly higher risk of mortality and adverse health outcomes, and should be retained in the ICU for longer periods. While OCRL-D does exhibit increasing conservativeness with rising r values (e.g., discharge ratio decreasing from 0.363 to 0.086), OCRL-R consistently better identifies high-risk cases and aligns its discharge behavior accordingly, even under stricter constraint thresholds. These findings underscore the value of explicitly modeling readmission history when learning discharge policies in constrained, high-stakes clinical environments in the ICU.

Appendix H. Neural Network - based Function Approximator

H.1. Training Details and Hyperparameter Configuration

Table H1 summarizes the hyperparameter settings used in Algorithm 1, where both the FQI and FQE agents are implemented using neural network - based function approximators. The model is trained over 4 million iterations using an adaptive gradient method – AdamW optimizer (Loshchilov and Hutter 2017), and mean squared error (MSE) as the loss function. The FQI model is implemented with a three-layer feedforward neural network comprising 128, 64, and 32 neurons in the first, second, and third hidden layers, respectively. As for the FQE agents, we use one hidden layer with 500 neurons. For both FQI and FQE, Leaky ReLU activation function is used with its negative slope equal to 0.1. Similar to the numerical experiment of linear approximation method, separate learning rates are employed for the FQI agent, FQE agents, and dual variables to facilitate stable training dynamics under constraint optimization. Same as the training process of linear approximation, a soft-update strategy is adopted for the target Q-networks, with parameters updated every 100 steps at a rate of $\kappa = 10^{-2}$.

Table H1.: Hyperparameter settings for Algorithm 1 using neural network-based function approximators.

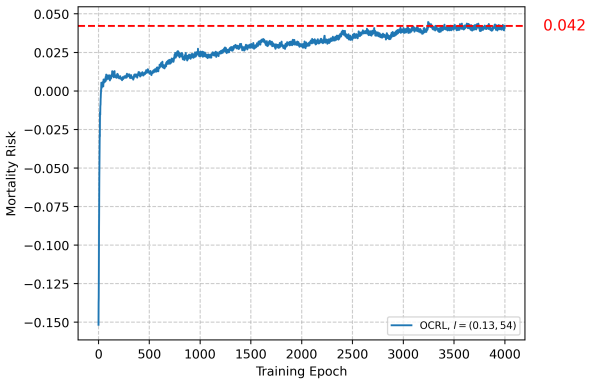
Hyperparameter	**Value**
Training steps	4×10^6
Batch size n	256
Learning rate (FQI agent) α_{fqi}	2×10^{-3}
Learning rate (FQE agents) α_{fqe}	5×10^{-4}
Learning rate (dual variable) $\alpha_{\lambda_{rr}}$	3×10^{-4}
Learning rate (dual variable) $\alpha_{\lambda_{los}}$	1×10^{-7}
Loss function $\mathcal{L}(\cdot)$	MSE
Optimizer	`AdamW`
Weight decay	1×10^{-2}
Activation function	`Leaky ReLU`
Soft-update rate κ	1×10^{-2}
Target update frequency U	100

H.2. Policy Performance

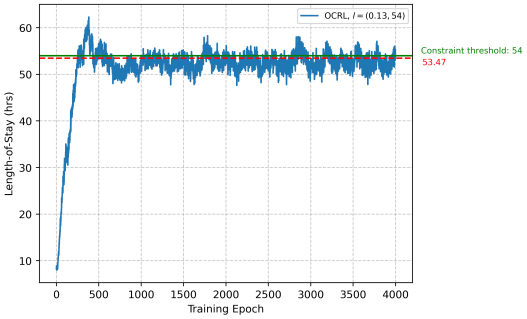
The numerical results presented below correspond to the constraint thresholds $l_{rr} = 0.13$ and $l_{los} = 54$. Figure H1 demonstrates the convergence behavior of the FQE-estimated objective and constraint costs, confirming the stability and reliability of our proposed approach with neural network approximation method.

Table H2.: FQE-estimated outcomes between OCRL-R policy ($l_{rr} = 0.13$, $l_{los} = 54$) and OCRL-NN policy ($l_{rr} = 0.13$, $l_{los} = 54$) on testing set $\mathcal{D}_{test}$.

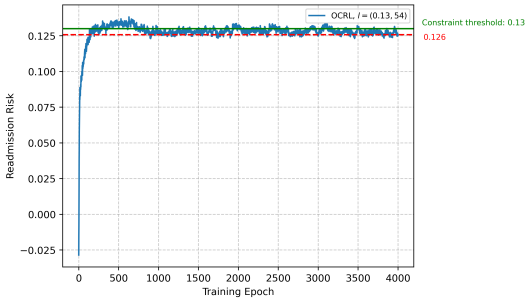
Outcome Metrics	**OCRL-R policy**	**OCRL-NN policy**
Mortality Risk	0.065	0.042
Percentage of the same action	68.45%	68.94%



(a) Training process of FQE agent for mortality risk.



(b) Training process of FQE agent for ICU LOS.



(c) Training process of FQE agent for readmission risk.

Figure H1.: The training process of FQE agents using neural network approximation in Algorithm 1 with given thresholds $l_{los} = 54$, and $l_{rr} = 0.13$.

Appendix I. Conservative Q-Learning (CQL) - based OCRL Algorithm Framework

I.1. Training Details and Hyperparameter Configuration

In addition to employing Fitted Q-Iteration (FQI), we explore an advanced value-based offline Reinforcement Learning (RL) algorithm, Conservative Q-Learning (CQL) (Kumar et al. 2020), to enhance policy learning within Algorithm 1, particularly employing linear approximation methods. Although, as discussed in Section 4.2, the constrained optimization framework intrinsic to the discharge decision-making problem inherently mitigates distribution shift issues characteristic of offline RL, integrating CQL further strengthens the conservativeness of our approach. Consequently, this integration effectively limits the frequency of discharge actions selected by the RL agent.

Table I1.: Hyperparameter settings for Algorithm 1 using CQL instead of FQI.

Hyperparameter	**Value**
Training steps	8×10^6
Batch size n	256
Learning rate (CQL agent) α_{cql}	2×10^{-3}
Learning rate (FQE agents) α_{fqe}	5×10^{-4}
Learning rate (dual variable) $\alpha_{\lambda_{rr}}$	2×10^{-4}
Learning rate (dual variable) $\alpha_{\lambda_{los}}$	1×10^{-7}
CQL penalty coefficient β	0.2
Loss function $\mathcal{L}(\cdot)$	CQL loss (L2-regularized MSE + penalty)
Optimizer	`SGD`
Weight decay	1×10^{-4}
Soft-update rate κ	1×10^{-2}
Target update frequency U	100

Specifically, to apply CQL, we modify the L2-regularized Mean Squared Error (MSE) loss function (14) into the following form

$$\mathcal{L}_{cql}(\theta) = \beta \cdot (\mathbb{E}_{a\sim\pi_{\mathcal{D}}}\left[Q(s,a|\theta)\right] - \mathbb{E}_{a\sim\pi_{RL}}\left[Q(s,a|\theta)\right]) + \mathcal{L}_{reg}(\theta), \tag{I1}$$

where $\pi_{\mathcal{D}}$ denotes clinical practices from the dataset and π_{RL} represents the learned policy. CQL enforces pessimism in value learning by maximizing the Q values under π_{RL}, which is characterized by the low Q values, and balancing by minimizing the Q-values of actions from $\pi_{\mathcal{D}}$. Note that CQL implicitly constrains the learned policy π_{RL} over $\pi_{\mathcal{D}}$. The hyperparameter β controls the degree of conservativeness. The hyperparameter β determines the level of conservatism; higher values enforce greater conservatism, whereas lower values yield more exploratory policies.

I.2. Policy Performance

We present numerical results of Algorithm 1 employing CQL under constraints $l_{rr} = 0.14$, $l_{los} = 54$, and penalty coefficient $\beta = 0.2$. We define OCRL-R as the policy using FQI (as presented in the main body) and OCRL-C as the policy using CQL.

Table I2 compares the FQE-estimated outcomes, specifically mortality risk, on the testing dataset $\mathcal{D}_{test}$. The estimated mortality risk for OCRL-C increases from 0.038

(OCRL-R) to 0.058, thus becoming closer to the observed average mortality risk (0.081) in the test set. This occurs due to the CQL loss function (I1) prompting the RL agent to select actions more aligned with clinical practice.

Table I2.: Comparison of FQE-estimated outcomes between OCRL-R policy ($l_{rr} = 0.14$, $l_{los} = 54.0$) and OCRL-C policy ($l_{rr} = 0.14$, $l_{los} = 54.0$, and $\beta = 0.2$) on testing set $\mathcal{D}_{test}$.

Outcome Metrics	**OCRL-R policy**	**OCRL-C policy**
Mortality Risk	0.038	0.058
Percentage of the same action	67.04%	82.06%

In Table I3, the OCRL-C policy demonstrates substantially increased conservativeness, exhibiting a higher proportion of decisions to keep patients in the ICU ($a = 0$) across all readmission counts. Notably, the OCRL-C policy selects no discharge actions ($a = 0$) for patients experiencing multiple readmissions ($r = 2, 3$). While this increased conservatism enhances patient safety, overly conservative policies may restrict the discovery of potentially more efficient discharge policies.

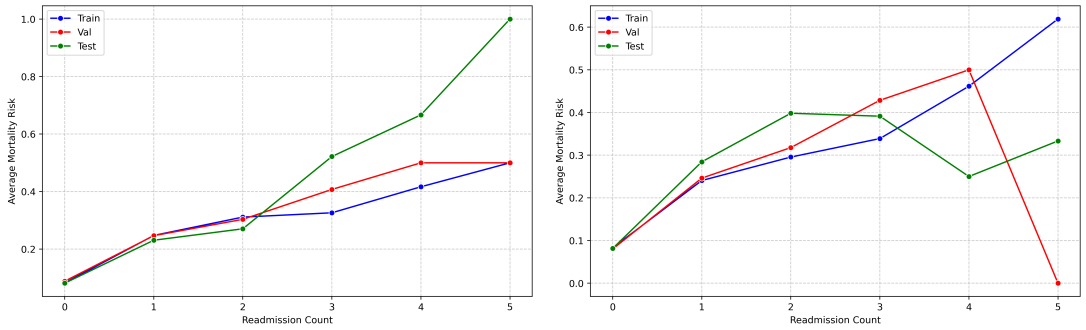
Table I3.: Discharge action ratios ($a = 1$) under clinical practice, OCRL-R policy ($l_{rr} = 0.14$, $l_{los} = 54.0$), and OCRL-C policy ($l_{rr} = 0.14$, $l_{los} = 54.0$, and $\beta = 0.2$) across admission types categorized by readmission count (r) in $\mathcal{D}_{test}$.

Admission Type	**Policy**	**Stay in ICU** ($a = 0$)		**Discharge** ($a = 1$)	
		Count	Ratio	Count	Ratio
All Cases	Clinical Practice	42879	0.862	6868	0.138
	OCRL-R Policy	31185	0.627	18562	0.373
	OCRL-C Policy	44314	0.891	5433	0.109
First Admission ($r = 0$)	Clinical Practice	27825	0.823	5996	0.177
	OCRL-R Policy	18051	0.534	15770	0.466
	OCRL-C Policy	28474	0.842	5347	0.158
First Readmission ($r = 1$)	Clinical Practice	8044	0.918	715	0.082
	OCRL-R Policy	6599	0.753	2160	0.247
	OCRL-C Policy	8673	0.990	86	0.010
Second Readmission ($r = 2$)	Clinical Practice	3689	0.968	122	0.032
	OCRL-R Policy	3242	0.851	569	0.149
	OCRL-C Policy	3811	1.000	0	0.000
Third Readmission ($r = 3$)	Clinical Practice	1470	0.985	23	0.015
	OCRL-R Policy	1476	0.989	17	0.011
	OCRL-C Policy	1493	1.000	0	0.000

Appendix J. Temporal Dataset Partitioning Experiments

In order to investigate the robustness of the algorithmic framework - Algorithm 1, we conduct numerical experiments regarding to temporal MIMIC-IV dataset partitioning. For previous numerical experiments, the patients' data is divided into training, validation, and testing set randomly. We now partition the dataset into training, validation, and testing sets based on the recorded temporal information.

The MIMIC-IV dataset has been processed to anonymize specific patient temporal information, leaving only the *anchor_year_group* column to indicate the time period of each record. Given that COVID-19 pandemic significantly impacted hospital operations, we would like to use post-pandemic data for validation and testing to evaluate the learned policies from pre-pandemic training data. However, to ensure sufficient data volume, we also include 2017-2019 data in the validation and test sets. Thus, our training set comprises data from 2008-2016, while the validation and test set span 2017-2022.



(a) Random dataset partitioning.

(b) Temporal dataset partitioning.

Figure J1.: The relationship between patient mortality risk $C(\pi_{\mathcal{D}})$ and readmission count r across training, validation, and testing sets.

For the random dataset partitioning used in the main text, the relationship between the readmission count and mortality risk in training, validation, and testing set is shown in Figure J1a. We can observe that while the average mortality risk ($C(\pi_{\mathcal{D}})$) shows some variation across the three sets when readmission count is high (≥ 3) owing to insufficient sample size, the mortality risk consistently increases with readmission count across all three sets. However, as shown in Figure J1, after adopting temporal dataset partitioning, the relationship between average mortality risk and readmission count no longer consistently exhibits monotonic increase across the three sets. Specifically, while this relationship is maintained in the training set, it does not hold for the validation ans testing set. We now proceed to analyze the performance of the proposed OCRL-based policy learning framework (Algorithm 1) under the temporal dataset partitioning.

As shown by Figure J2, under the temporal dataset partitioning, the multi-timescale offline constrained learning framework (Algorithm 1) still achieves successful convergence during the training process. Two different L2 regularization weights are tested: OCRL-R-v0 without L2 regularization term ($\omega = 0$), and OCRL-R-v1 with $\omega = 1e-3$. Despite their similar convergence behavior shown in Figure J2, where both achieve successful convergence, the resulting action distributions vary across different readmission counts.

Firstly, as illustrated in Figure J1b, the validation set under temporal dataset par-

titioning (red line) shows that the first four readmission counts ($r \leq 4$) still exhibit increasing average increasing mortality risk. The mortality risk dropping to 0 for the fifth readmission ($r = 5$) is understandable given that only one patient falls into this category. Concerning OCRL agents' performance, Table J1 demonstrates that both OCRL-R-v0 and OCRL-R-v1 successfully capture the trend of decreasing discharge action ratios with increasing readmission count. However, OCRL-R-v1 adopts a more conservative discharge policy for readmitted patients. This conservative approach is particularly pronounced for cases with higher readmission counts ($r \geq 4$), where limited sample sizes exist, allowing OCRL-R-v1 to be more cautious than the clinical practices in the MIMIC-IV dataset.

Table J1.: Discharge action ratios ($a = 1$) under clinical practice, OCRL-R-v0 policy ($l_{rr} = 0.13$, $l_{los} = 54.0$, and $\omega = 0$) and OCRL-R-v1 policy ($l_{rr} = 0.13$, $l_{los} = 54.0$, and $\omega = 1e-3$) across admission types categorized by readmission count (r) in $\mathcal{D}_{val}$.

Admission Type	**Policy**	**Stay in ICU** ($a = 0$)		**Discharge** ($a = 1$)	
		Count	Ratio	Count	Ratio
All Cases	Clinical Practice	45291	0.859	7454	0.141
	OCRL-R-v0 Policy	37066	0.703	15679	0.297
	OCRL-R-v1 Policy	37018	0.702	15727	0.298
First Admission ($r = 0$)	Clinical Practice	31789	0.828	6596	0.172
	OCRL-R-v0 Policy	25309	0.659	13076	0.341
	OCRL-R-v1 Policy	24796	0.646	13589	0.354
First Readmission ($r = 1$)	Clinical Practice	8594	0.922	723	0.078
	OCRL-R-v0 Policy	7332	0.787	1985	0.213
	OCRL-R-v1 Policy	7538	0.809	1779	0.191
Second Readmission ($r = 2$)	Clinical Practice	3012	0.966	107	0.034
	OCRL-R-v0 Policy	2683	0.860	436	0.140
	OCRL-R-v1 Policy	2853	0.915	266	0.085
Third Readmission ($r = 3$)	Clinical Practice	1233	0.983	21	0.017
	OCRL-R-v0 Policy	1107	0.883	147	0.117
	OCRL-R-v1 Policy	1165	0.929	89	0.071
Fourth Readmission ($r = 4$)	Clinical Practice	373	**0.984**	6	0.016
	OCRL-R-v0 Policy	354	0.934	25	0.066
	OCRL-R-v1 Policy	375	**0.989**	4	0.011
Fifth Readmission ($r = 5$)	Clinical Practice	290	**0.997**	1	0.003
	OCRL-R-v0 Policy	281	0.966	10	0.034
	OCRL-R-v1 Policy	291	**1.000**	0	0.000

Secondly, for testing set under temporal dataset partitioning, the scenario becomes more complicated. As shown in Figure J1b, in the testing set (green line), the monotonically increasing relationship between mortality risk and readmission count breaks down when readmission count $r \geq 3$. In Table J2, we observe that OCRL-R-v1 continues to exhibit greater conservatism compared to OCRL-R-v0. While both agents are influenced by the altered relationship between mortality risk and readmission count in the $r = 3$ and $r = 4$ cases, the L2 regularization term helps alleviate this effect. Furthermore, for the most challenging scenario with minimal sample size ($r = 5$), OCRL-R-v1 can still maintain conservatism similar to or even higher that clinical practices

in MIMIC-IV dataset, making more cautious ICU discharge decision-making.

Table J2.: Discharge action ratios ($a = 1$) under clinical practice, OCRL-R-v0 policy ($l_{rr} = 0.13$, $l_{los} = 54.0$, and $\omega = 0$) and OCRL-R-v1 policy ($l_{rr} = 0.13$, $l_{los} = 54.0$, and $\omega = 1e-3$) across admission types categorized by readmission count (r) in $\mathcal{D}_{test}$.

Admission Type	**Policy**	**Stay in ICU** ($a = 0$)		**Discharge** ($a = 1$)	
		Count	Ratio	Count	Ratio
All Cases	Clinical Practice	45863	0.860	7438	0.140
	OCRL-R-v0 Policy	37390	0.701	15911	0.299
	OCRL-R-v1 Policy	37542	0.704	15759	0.296
First Admission ($r = 0$)	Clinical Practice	31466	0.828	6553	0.172
	OCRL-R-v0 Policy	24993	0.657	13026	0.343
	OCRL-R-v1 Policy	24498	0.644	13521	0.356
First Readmission ($r = 1$)	Clinical Practice	9050	0.924	742	0.076
	OCRL-R-v0 Policy	7794	0.796	1998	0.204
	OCRL-R-v1 Policy	8116	0.829	1676	0.171
Second Readmission ($r = 2$)	Clinical Practice	3562	0.969	113	0.031
	OCRL-R-v0 Policy	3150	0.857	525	0.143
	OCRL-R-v1 Policy	3353	0.912	322	0.088
Third Readmission ($r = 3$)	Clinical Practice	989	0.977	23	0.023
	OCRL-R-v0 Policy	735	0.731	271	0.269
	OCRL-R-v1 Policy	812	0.807	194	0.193
Fourth Readmission ($r = 4$)	Clinical Practice	311	0.987	4	0.013
	OCRL-R-v0 Policy	226	0.717	89	0.283
	OCRL-R-v1 Policy	270	0.857	45	0.143
Fifth Readmission ($r = 5$)	Clinical Practice	491	0.994	3	0.006
	OCRL-R-v0 Policy	492	0.996	2	0.004
	OCRL-R-v1 Policy	493	0.998	1	0.002

The following analysis evaluates the performance of OCRL agents trained on the 2008–2016 dataset, using the 2017–2022 data for validation and testing. As shown in Figure J3, the average mortality risk increases monotonically with readmission count r across the three training-year groups. However, this pattern does not fully hold for the 2017–2019 and 2019–2022 cohorts, particularly when $r \geq 3$.

Next, we examine the performance of agents trained on 2008–2016 data when evaluated on the 2017–2019 cohort, as shown in Table J3. Clinical practice during this period continues to exhibit a decreasing discharge action ratios ($a = 1$) with increasing readmission count r, reflecting heightened conservatism. However, the average mortality risk at $r = 4$ and $r = 5$ is lower than at $r = 3$, which may influence the behavior of OCRL agents. The OCRL-R-v0 agent, trained without L2 regularization, is notably affected, showing higher discharge action ratios at $r = 3$ and $r = 4$ than at $r = 2$. In contrast, OCRL-R-v1, which incorporates L2 regularization, is less sensitive to this anomaly, showing a deviation only at $r = 3$ while maintaining cautious behavior at $r = 4$ and $r = 5$, consistent with clinical practices in MIMIC-IV.

Finally, we examine the performance of agents trained on 2008-2016 data when evaluated on 2020-2022 data. Since the 2020-2022 dataset contains no cases with $r = 5$, comparisons are limited to $r = 0$ through $r = 4$. While clinical practice discharge

Table J3.: Discharge action ratios ($a = 1$) under clinical practice, OCRL-R-v0 policy ($l_{rr} = 0.13$, $l_{los} = 54.0$, and $\omega = 0$) and OCRL-R-v1 policy ($l_{rr} = 0.13$, $l_{los} = 54.0$, and $\omega = 1e-3$) across admission types categorized by readmission count (r) **from 2017 to 2019** in both $\mathcal{D}_{val}$ and $\mathcal{D}_{test}$.

Admission Type	**Policy**	**Stay in ICU** ($a = 0$)		**Discharge** ($a = 1$)	
		Count	Ratio	Count	Ratio
All Cases	Clinical Practice	63046	0.863	10000	0.137
	OCRL-R-v0 Policy	51384	0.703	21662	0.297
	OCRL-R-v1 Policy	52182	0.714	20864	0.286
First Admission ($r = 0$)	Clinical Practice	41310	0.826	8724	0.174
	OCRL-R-v0 Policy	32711	0.654	17323	0.346
	OCRL-R-v1 Policy	32495	0.649	17539	0.351
First Readmission ($r = 1$)	Clinical Practice	12892	0.925	1043	0.075
	OCRL-R-v0 Policy	10939	0.785	2996	0.215
	OCRL-R-v1 Policy	11414	0.819	2521	0.181
Second Readmission ($r = 2$)	Clinical Practice	5513	0.968	183	0.032
	OCRL-R-v0 Policy	4851	**0.852**	845	**0.148**
	OCRL-R-v1 Policy	5197	**0.912**	499	**0.088**
Third Readmission ($r = 3$)	Clinical Practice	1985	0.981	38	0.019
	OCRL-R-v0 Policy	1629	**0.805**	394	**0.195**
	OCRL-R-v1 Policy	1764	**0.872**	259	**0.128**
Fourth Readmission ($r = 4$)	Clinical Practice	565	0.986	8	0.014
	OCRL-R-v0 Policy	481	**0.839**	92	**0.161**
	OCRL-R-v1 Policy	528	**0.921**	45	**0.079**
Fifth Readmission ($r = 5$)	Clinical Practice	781	0.995	4	0.005
	OCRL-R-v0 Policy	773	**0.985**	12	**0.015**
	OCRL-R-v1 Policy	784	**0.999**	1	**0.001**

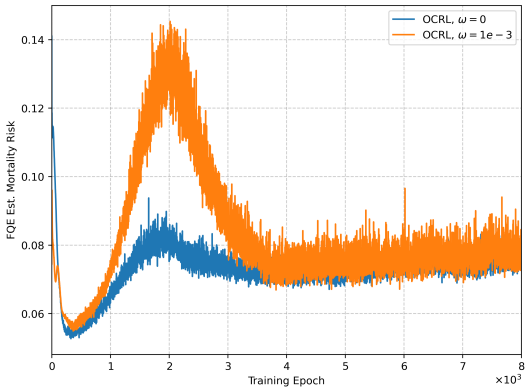
action ratios continue to decrease with increasing readmission count r, the relationship between average mortality risk and readmission count at $r = 3$ proves highly misleading (as shown in Figure J3). Specifically, the average mortality risk at $r = 3$ is substantially lower than at $r = 2$, causing OCRL-R-v1 with L2 regularization to be slightly affected, exhibiting marginally higher discharge action ratios at $r = 3$ versus $r = 2$ (0.101 vs. 0.081). Nevertheless, OCRL-R-v1 maintains a cautious discharge policy for $r = 4$ cases. While OCRL-R-v0 successfully maintains gradually decreasing discharge action ratios from $r = 0$ to $r = 3$, it exhibits two concerning behaviors: first, its discharge action ratios consistently exceed those of both clinical practice and OCRL-R-v1, reflecting a more assertive discharge policy; second, it behaves particularly aggressively at $r = 4$, precisely when sample sizes are limited yet patient conditions are most critical.

In conclusion, the proposed multi-timescale offline constrained policy learning algorithmic framework - Algorithm 1 exhibits adequate robustness when subjected to temporal dataset partitioning. The COVID-19 pandemic from 2019-2022 had a substantial impact on ICU data. Nevertheless, our algorithmic framework demonstrates robustness in two key aspects. First, the three FQE agents responsible for evaluating training effectiveness successfully converge during training. Second, when evaluated

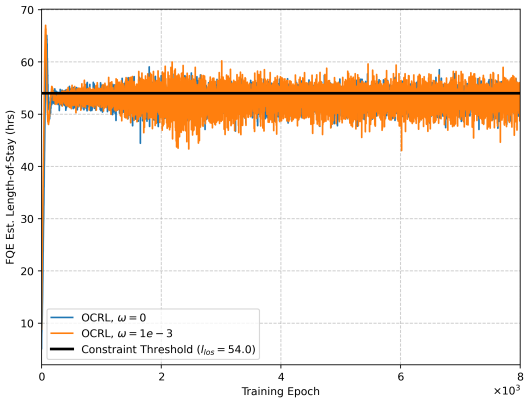
Table J4.: Discharge action ratios ($a = 1$) under clinical practice, OCRL-R-v0 policy ($l_{rr} = 0.13$, $l_{los} = 54.0$, and $\omega = 0$) and OCRL-R-v1 policy ($l_{rr} = 0.13$, $l_{los} = 54.0$, and $\omega = 1e-3$) across admission types categorized by readmission count (r) **from 2020 to 2022** in both $\mathcal{D}_{val}$ and $\mathcal{D}_{test}$.

Admission Type	**Policy**	**Stay in ICU** ($a = 0$)		**Discharge** ($a = 1$)	
		Count	Ratio	Count	Ratio
All Cases	Clinical Practice	28108	0.852	4892	0.148
	OCRL-R-v0 Policy	23072	0.699	9928	0.301
	OCRL-R-v1 Policy	22378	0.678	10622	0.322
First Admission ($r = 0$)	Clinical Practice	21945	0.832	4425	0.168
	OCRL-R-v0 Policy	17591	0.667	8779	0.333
	OCRL-R-v1 Policy	16799	0.637	9571	0.363
First Readmission ($r = 1$)	Clinical Practice	4752	0.918	422	0.082
	OCRL-R-v0 Policy	4187	0.809	987	0.191
	OCRL-R-v1 Policy	4240	0.819	934	0.181
Second Readmission ($r = 2$)	Clinical Practice	1061	0.966	37	0.034
	OCRL-R-v0 Policy	982	**0.894**	116	**0.106**
	OCRL-R-v1 Policy	1009	**0.919**	89	**0.081**
Third Readmission ($r = 3$)	Clinical Practice	231	0.975	6	0.025
	OCRL-R-v0 Policy	213	**0.899**	24	**0.101**
	OCRL-R-v1 Policy	213	**0.899**	24	**0.101**
Fourth Readmission ($r = 4$)	Clinical Practice	119	0.983	2	0.017
	OCRL-R-v0 Policy	99	**0.818**	22	**0.182**
	OCRL-R-v1 Policy	117	**0.967**	4	**0.033**

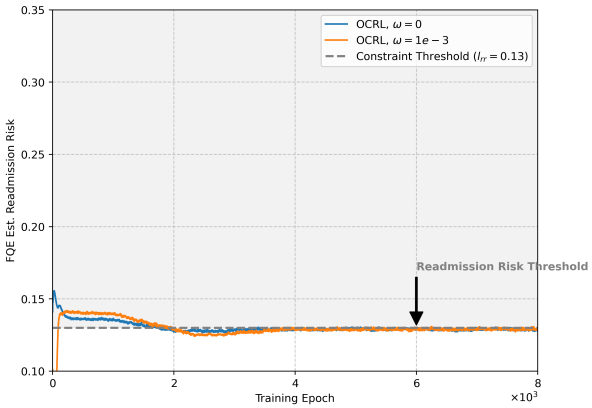
on validation and testing sets containing 2017-2022 data, the OCRL agents trained on 2008-2016 data maintain cautious ICU discharge decision-making comparable to clinical practices for high readmission count cases, despite the uncertain relationship between average mortality risk and readmission count. Furthermore, given the limited sample sizes for high readmission count cases and the inclusion of other important physiological features beyond readmission count in the state space, minor fluctuations in OCRL policy performance for these cases are expected and reasonable.



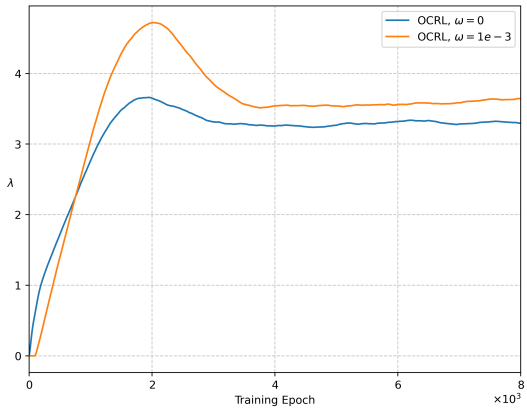
(a) Training process of FQE agents for objective cost - mortality risk.



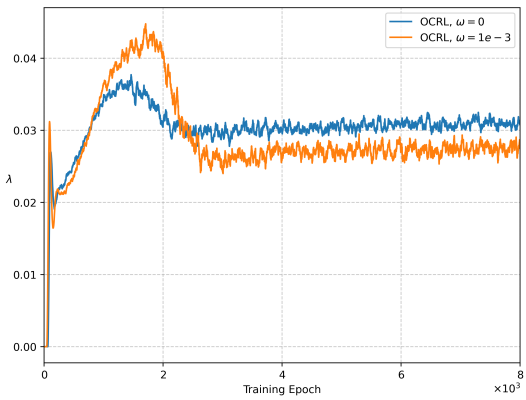
(b) Training process of FQE agents for constraint cost - ICU LOS.



(c) Training process of FQE agents for constraint cost - readmission risk.



(d) The training process of λ_{rr}.



(e) The training process of λ_{los}.

Figure J2.: The training process of FQE agents and Lagrangian multipliers in Algorithm 1 with varying L2 regularization weights under temporal dataset partitioning. Constraint threshold setting: $l_{los} = 54$, and $l_{rr} = 0.13$.

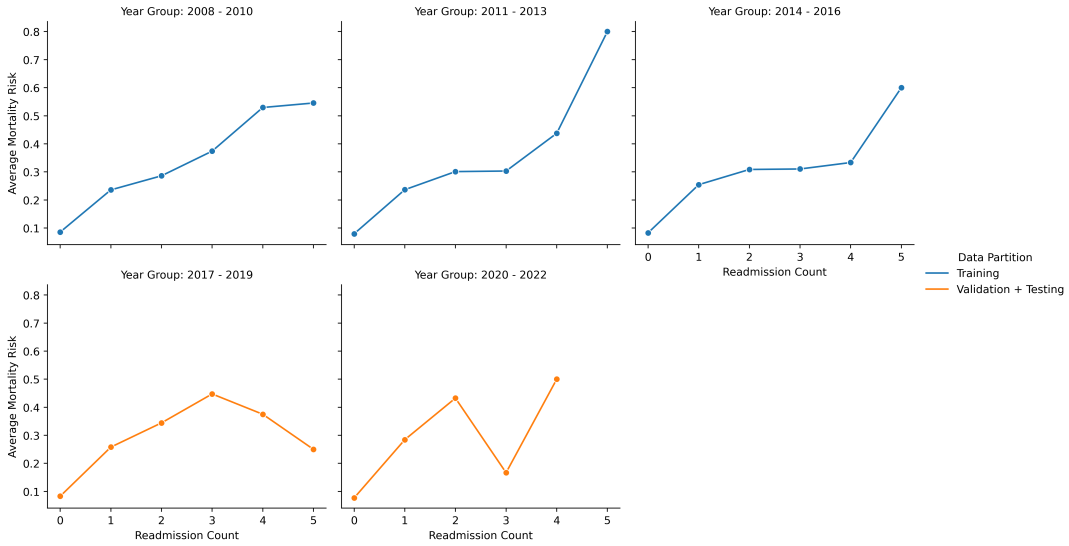


Figure J3.: The average mortality risk $C\left(\pi_{\mathcal{D}}\right)$ by readmission count r across dataset partitions (training v.s. validation + testing) and year groups (from 2008 to 2022).

References

Dabi A, Rahman O (2023) Termination of life support. *StatPearls [Internet]* (StatPearls Publishing), pMID: 32644460.

Flaws D, Fraser JF, Laupland K, et al. (2024) Time in icu and post-intensive care syndrome: how long is long enough? *Critical Care* 28:34.

Kondrup F, Jiralerspong T, Lau E, de Lara N, Shkrob J, Tran MD, Precup D, Basu S (2024) Towards safe mechanical ventilation treatment using deep offline reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 15696–15702.

Kumar A, Zhou A, Tucker G, Levine S (2020) Conservative q-learning for offline reinforcement learning.

Lefering R, Waydhas C, DGU T (2024) Prediction of prolonged length of stay on the intensive care unit in severely injured patients—a registry-based multivariable analysis. *Frontiers in Medicine* 11:1358205.

Loshchilov I, Hutter F (2017) Decoupled weight decay regularization. *International Conference on Learning Representations*.

National Institute for Health and Care Excellence (2017) Rehabilitation after critical illness in adults. Quality standard QS158, URL `https://www.nice.org.uk/guidance/qs158/resources/rehabilitation-after-critical-illness-in-adults-pdf-75545546693317`, accessed: Accessed 6 August 2025.

Ofoma UR, Dong Y, Gajic O, Pickering BW (2018) A qualitative exploration of the discharge process and factors predisposing to readmissions to the intensive care unit. *BMC Health Services Research* 18(1):6.

Shah M, Patil S, Patel B, Agarwal M, Davila CD, Garg L, Agrawal S, Kapur NK, Jorde UP (2018) Causes and predictors of 30-day readmission in patients with acute myocardial infarction and cardiogenic shock. *Circulation: Heart Failure* 11(4):e004310.

Shi P, Helm JE, Deglise-Hawkinson J, Pan J (2021) Timing it right: Balancing inpatient congestion vs. readmission risk at discharge. *Operations Research* 69(6):1842–1865.

Tang S, Wiens J (2021) Model selection for offline reinforcement learning: Practical considerations for healthcare settings. *The 6th Machine Learning for Healthcare Conference*, volume 149 of *Proceedings of Machine Learning Research*, 2–35 (PMLR).