Position: Reinforcement Learning in Dynamic Treatment Regimes Needs Critical Reexamination

Zhiyao Luo [1] **Yangchen Pan** [1] **Peter Watkinson** [2] **Tingting Zhu** [1]

Abstract

In the rapidly changing healthcare landscape, the implementation of offline reinforcement learning (RL) in dynamic treatment regimes (DTRs) presents a mix of unprecedented opportunities and challenges. This position paper offers a critical examination of the current status of offline RL in the context of DTRs. We argue for a reassessment of applying RL in DTRs, citing concerns such as inconsistent and potentially inconclusive evaluation metrics, the absence of naive and supervised learning baselines, and the diverse choice of RL formulation in existing research. Through a case study with more than 17,000 evaluation experiments using a publicly available Sepsis dataset, we demonstrate that the performance of RL algorithms can significantly vary with changes in evaluation metrics and Markov Decision Process (MDP) formulations. Surprisingly, it is observed that in some instances, RL algorithms can be surpassed by random baselines subjected to policy evaluation methods and reward design. This calls for more careful policy evaluation and algorithm development in future DTR works. Additionally, we discussed potential enhancements toward more reliable development of RL-based dynamic treatment regimes and invited further discussion within the community. Code is available at https://github.com/GilesLuo/ReassessDTR.

1. Introduction

The advent of machine learning in the medical field has opened new avenues for treatment optimization (Chernozhukov et al., 2018; Myszczynska et al., 2020). Among various machine learning techniques, reinforcement learning (RL) has steadily gained recognition as a transformative tool in healthcare (Coronato et al., 2020; Yu et al., 2021), particularly within dynamic treatment regimes (DTRs)(Chakraborty & Murphy, 2014). The core strength of RL-DTRs is their ability to learn from patient responses to treatments and adapt treatment plans accordingly, leading to better patient outcomes. This adaptability is especially valuable in healthcare, where patient conditions and treatment responses are often diverse and can change over time.

Recent advancements in RL applied to DTRs, especially in offline RL(Agarwal et al., 2020), have demonstrated promising potential to guide future treatment decisions without direct environment interaction. For instance, Wu et al. introduced a weighted dueling double deep Q-network with embedded human expertise (WD3QNE), addressing domain knowledge embedding. Luckett et al. explored V-learning to estimate optimal DTRs in mobile health, addressing the challenges of indefinite time horizons and high-frequency decision-making. These studies highlight the diverse applications of offline RL in healthcare, from improving personalized treatment in chronic diseases to optimizing real-time interventions in mobile platforms.

The application of RL in DTRs is not without its criticisms (Jeter et al., 2019) (Gottesman et al., 2018), prompting a thorough reassessment of its utility and necessity in this field. One major issue is the lack of standardized evaluation metrics. For example, Komorowski et al. use weighted importance sampling; Raghu et al.; Raghu et al.; Peng et al. use different variants of doubly robust off-policy evaluation methods (Jiang & Li, 2016). Another approach involves using direct methods (Mannor et al., 2007), yet this, too, lacks uniformity across studies. The diversity of evaluation techniques underscores the need for a consensus on benchmark metrics to facilitate meaningful comparisons in this field. Furthermore, the diverse formulations of Markov decision processes (MDPs) and the lack of established baselines make it challenging to gauge the treatment improvement offered by RL approaches over traditional methods. Regarding the state space, Komorowski et al. converted continuous variables into discrete clusters, in contrast to Liu et al.; Wang et al., who used continuous variables directly. In ac-

[1]Department of Engineering Science, University of Oxford, Parks Road, Oxford OX1 3PJ, United Kingdom [2]Nuffield Department of Population Health (NDPH), University of Oxford, Richard Doll Building, Old Road Campus, Headington, Oxford OX3 7LF, United Kingdom . Correspondence to: Zhiyao Luo <zhiyao.luo@eng.ox.ac.uk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

tion space design, although discrete actions are commonly employed, some studies, such as (Wang et al., 2022; Huang et al., 2022) have explored continuous action spaces. For reward, Komorowski et al. chose simplistic 90-day mortality for patient outcomes, whereas Raghu et al.; Raghu et al.; Peng et al. integrated clinically validated risk scores into the reward function, but with varying weights on reward components. This diversity highlights systematic reevaluation to enhance RL's reliability and effectiveness in developing treatment plans.

The position of this paper is that **while RL holds significant promise for optimizing DTRs in healthcare, there is a critical need for a comprehensive reassessment of its application.** Previous studies (Gottesman et al., 2018; Tang & Wiens, 2021) focused mainly on practical healthcare concerns with respect to policy evaluation. However, we call for a more thorough understanding of methodological inconsistencies, including variations in policy evaluation metrics, the absence of standardized baselines for comparison, and the diverse formulations of MDPs. The paper advocates for the establishment of uniform standards and methodologies to ensure that RL's application is both scientifically rigorous and practically applicable to healthcare.

Our paper presents a critical perspective on the effectiveness and necessity of using RL in DTRs. Initially, we review fundamental concepts in offline RL and its application in DTRs. We then review the literature in this domain, highlighting the various policy evaluation methods used in different studies and pointing out that the diversity often leads to significant variability in performance among algorithms. Furthermore, we compare basic baselines (e.g., random policies and supervised learning) that a notable number of existing studies omitted and found that RL algorithms can underperform compared to these simple baselines, which raises a cause for concern. Moving forward, we standardize our evaluation method to focus on the impact of different reward designs, demonstrating that varying rewards can lead to relatively disparate performance. Through our critical analysis, we offer several considerations to improve the reliability of model development and evaluations in this field.

2. Background

This section provides the basic notation and conceptual background necessary for applying offline RL to DTRs, along with a common offline RL formulation for addressing problems in DTR.

2.1. RL and Offline RL

RL is building upon Markov decision process (MDP) denoted as $M = \{\mathcal{S}, \mathcal{A}, P, r, \gamma\}$ (Puterman, 2014), where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\gamma \in [0, 1)$ is the discount factor, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and $P : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$) are the reward and transition functions, respectively. The *value function* is the expectation of future discounted total reward obtained by following a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, $v^\pi(s) = \mathbb{E}^\pi[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)|s_0 = s]$ where $\mathbb{E}^\pi$ means the expectation under the policy π and the transition probability. The corresponding *action-value* function is $q^\pi(s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[v^\pi(s')]$. The goal is to find an *optimal policy* π^* that maximizes the values $\forall s \in \mathcal{S}$.

In an *offline RL* setting, our focus is on learning an optimal policy for decision-making based on a pre-gathered dataset, denoted as $\mathcal{D} = \{s_i, a_i, r_i, s'_i\}_{i=0}^{n-1}$. This dataset is assumed to be the result of actions taken according to a specific behavior policy $\pi_\mathcal{D}$. One primary challenge in offline RL is that $\pi_\mathcal{D}$ may not thoroughly explore all possible actions, leading to potential overestimation of those out-of-distribution actions. Acting greedily with respect to such actions could be problematic (Fujimoto et al., 2019).

To address the challenge, a widely adopted method involves restricting the learned policy π to be close to a baseline policy $\pi_\mathcal{D}$, by incorporating a KL-divergence term into the optimization objective: $\max_\pi \mathbb{E}_{s\sim\rho}[\sum_a \pi(a|s)q(s, a) - \tau D_{\mathrm{KL}}(\pi(\cdot|s)||\pi_\mathcal{D}(\cdot|s))]$ with $\tau > 0$. This formulation ensures that the optimized policy π is only supported where $\pi_\mathcal{D}$ is non-zero, effectively setting $\pi(a|s) = 0$ wherever $\pi_\mathcal{D}(a|s) = 0$. This principle underlies numerous offline RL strategies (Wu et al., 2019; Peng et al., 2020; Nair et al., 2021; Brandfonbrener et al., 2021; Fujimoto & Gu, 2021). Fujimoto & Gu introduce a behaviour cloning regularization term, $(\pi(s) - a)^2$. An alternative approach is in-sample policy optimization, which aims to prevent the selection of actions outside $\pi_\mathcal{D}$'s distribution. A more recent work (Xiao et al., 2023) proposes an in-sample softmax to directly sample from support. We will later provide more details about the algorithms used in our experiments.

2.2. Dynamic Treatment Regime

A Dynamic Treatment Regime (DTR) represents a sequence of decision rules that guide treatment adaptations over time based on an individual patient's evolving conditions and responses. An action (i.e., treatment) should take into account the patient's current state, including medical history and previous treatment responses, to recommend the next best action. Offline RL is particularly suitable for DTR because it allows the use of extensive historical healthcare data to learn optimal treatment strategies without the need for real-time experimentation in patients.

In this work, we use the intravenous vasopressor fluid (IV) and vasopressor dosage regime task (Komorowski et al., 2018) for sepsis treatment in the intensive care unit (ICU) as a proof-of-position. The dataset is derived from the Medical Information Mart for Intensive Care III (MIMIC-

III) database (Johnson et al., 2016). The selection of patients follows the sepsis-3 criteria (Singer et al., 2016), focusing on the early stage of sepsis management up to 24 hours prior to and 48 hours after the estimated onset of sepsis.

3. Problem Formulation in Offline RL for DTR

A DTR problem is typically modeled as a finite-time MDP as introduced in Section 2.1:

States ($s \in \mathcal{S}$): time-varying clinical variables that represent the medical condition of a patient. Particularly, this paper considers *continuous states* to avoid introducing variance from clustering-based state discretization algorithms.

Actions ($a \in \mathcal{A}$): actionable treatment decisions. This work primarily focuses on *discrete actions*, as they are not only more extensively studied in literature but also have a more established theoretical foundation of policy evaluation.

Rewards (r): the outcome of an action taken in a particular state, reflecting improvements or deterioration in the patient's condition. Various factors can be taken into account in the reward design, including the outcome of treatment, the clinical risk score, and the abnormality of vital signs.

In our sepsis cohort, the curation and preprocessing of raw data are mostly reproduced from Komorowski et al.'s work. Data are binned into 4-hour intervals, with 46 observational variables at each step. The dataset categorizes IV fluid and vasopressor dosages into five classes (See Appendix B), resulting in a 5×5 discrete action space. Additionally, we employed a series of important modifications. First, we change the treatment outcome from 90-day to in-hospital mortality under clinical guidance to strengthen the correlation between actions and outcomes. Secondly, we exclude patients who have inconsistent time-series data[1]. Furthermore, we removed the 'input 4-hourly' feature from the observation space, as the feature is part of the action.

4. Diversity and Inconsistency of Policy Evaluation Methods in RL-DTR

4.1. Challenges of Policy Evaluation in RL

Evaluating offline RL algorithms for DTRs is challenging for several reasons: **a)** The dataset is fixed and observational, meaning RL cannot be evaluated by interacting with the environment. **b)** Medical decision-making is complex, as the effects of treatments may not be immediately apparent and can be influenced by many confounding factors. **c)** Patients' responses to treatments are inherently uncertain and variable, making it difficult to assess the true effectiveness of proposed treatment policies. These challenges make it harder to evaluate offline RL algorithms in the context of DTRs, compared to traditional RL settings.

Recent literature addresses the evaluation challenges in DTR by a range of policy evaluation techniques. Notable among these are Inverse Probability Weighting (IPW) (Liu et al., 2017), Weighted Importance Sampling (WIS) (Kidambi et al., 2020; Nambiar et al., 2023), the Direct Method (DM)(Huang et al., 2022; Kondrup et al., 2023), and Doubly Robust (DR)(Raghu et al., 2017; Wu et al., 2023; Wang et al., 2018) estimators. These methods tried to tackle the confounding variables and create a counterfactual estimation based on historical data.

4.2. Existing Evaluation Methods in RL-DTR

An offline policy evaluation (OPE) aims to estimate a policy's value $\hat{V}^{\pi}$ using a behavior policy $\pi_{\mathcal{D}}$. The Direct Method (DM) directly approximates the value function by constructing a model to predict the expected reward for each state-action pair under the target policy. The DM estimator for the policy value is formulated as:

$$\hat{V}^{\pi}_{DM} = \mathbb{E}^{\pi}\hat{Q}^{\pi_{\mathcal{D}}}(s,a) \tag{1}$$

where $\hat{Q}^{\pi_{\mathcal{D}}}(s,a)$ is the estimated action-value function under the behavior policy. Despite its straightforward approach, DM relies heavily on the model's accuracy for estimating $\hat{Q}^{\pi_{\mathcal{D}}}$, making it susceptible to model misspecification and bias due to the imbalanced nature of medical data. Importance Sampling (IS)(Tokdar & Kass, 2010) adjusts returns from the behavior policy to approximate those under the target policy. WIS mitigates IS's high variance by normalizing each weight by the sum of all weights before applying them to the returns. The WIS estimator is expressed as:

$$\hat{V}^{\pi}_{WIS} = \frac{\sum_{i=1}^{N}[\rho^{i}_{1:T_i}G^{i}]}{\sum_{i=1}^{N}\rho^{i}_{1:T_i}} \tag{2}$$

where $\rho^{i}_{1:T_i} = \prod_{t=1}^{T_i}\frac{\pi(a^i_t|s^i_t)}{\pi_{\mathcal{D}}(a^i_t|s^i_t)}$, and $G^i = \sum_{t=1}^{T_i}\gamma^{t-1}r^i_t$. Modifications such as bootstrapping ($\hat{V}^{\pi}_{WIS_b}$), ratio truncation ($\hat{V}^{\pi}_{WIS_t}$), or a combination of both ($\hat{V}^{\pi}_{WIS_{bt}}$) can further decrease variance. A comprehensive introduction is available in Appendix C. The Doubly Robust (DR) estimator(Jiang & Li, 2016) combines the IS with DM, iteratively calculated per-trajectory estimation as:

$$\hat{V}^{T_i+1-t}_{DR} = \hat{V}^{\pi}(s^i_t) + \eta_t(r^i_t + \gamma\hat{V}^{H-t}_{DR} - \hat{Q}^{\pi_e}(s^i_t, a^i_t))$$

Starting from $t = H$ with $V^0_{DR} := 0$, this calculation proceeds in reverse order to derive $V^{T_i}_{DR}$ for patient i. The overall estimate is then $\hat{V}^{\pi}_{DR} := \sum_{i=1}^{N}\hat{V}^{T_i}_{DR}$. Refer to Appendix C for our implementation of the DR estimator.

[1]The original patient cohort included patients who lost track of information in the middle of treatment trajectories. This means that they omit the irregular sampling frequency and consider them as consecutive time steps, which is not a common practice in RL. Therefore, we remove patients who have missing information within 4 hours during admissions.

5. Reward Design Choices

This section examines reward design choices in RL for DTR literature. We aim to underscore the importance of reward structures in RL's effectiveness and comparability in DTR.

5.1. Outcome-Based Reward

The initial reward design, (Komorowski et al., 2018), employs a straightforward reward setting: $r = 0$ for non-terminal steps, $r = +100$ for patient survival or $r = -100$ for death at the final step. This binary approach oversimplifies the complexity of medical scenarios and omits critical factors such as the risk of deterioration, abnormality of the critical signs, and the progression rate of the disease, all of which significantly affect patient mortality. From an RL perspective, this approach might lead to more challenging issues such as credit assignment, sampling inefficiency and learning variance due to sparse reward.

5.2. Risk-based Reward

Incorporating intermediate reward is shown to be beneficial in goal-reaching RL (Zhai et al., 2022). In DTR applications, intermediate rewards are generally represented by clinical risk scores. Here we introduce two typical risk score-based rewards in the RL-DTR literature.

5.2.1. ICU Risk-Based Reward

The Sequential Organ Failure Assessment (SOFA) (Kajdacsy-Balla Amaral et al., 2005; Jones et al., 2009) score, a commonly used critical care metric to quantify the severity of a patient's organ function or rate of failure, has been used as a reward design in the literature (Raghu et al.; Wang et al.). In addition to the SOFA score, lactate levels are included in the reward calculation, as they are biomarkers(Nguyen et al., 2004) for tissue hypoxia and metabolic dysfunction. Represented SOFA score as $\kappa \in [0, 1, 2, ..., 24]$ and the value of lactate (mmol/L) as v, the reward is intricately defined as:

$$\begin{aligned} r_t^i = c_0(\mathbf{1}_{\kappa_t^i=\kappa_{t+1}^i} \cdot \mathbf{1}_{\kappa_{t+1}^i>0}) + c_1(\kappa_{t+1}^i - \kappa_t^i) \\ + c_2 \tanh(v_{t+1}^i - v_t^i) + \mathbf{1}_{t=T_i} r_{\text{outcome}} \end{aligned} \tag{3}$$

where c_0, c_1, c_2 is -0.025, -0.125, and -2, respectively. $r_{outcome}$ is 15 for an alive patient; otherwise -15. This formula integrates both stability and changes in organ function (using the SOFA score) and metabolic alterations (using lactate levels).

5.2.2. Early Warning Risk-Based Reward

We also present a reward function based on the National Early Warning Score 2 (NEWS2) (Inada-Kim & Nsutebu, 2018) system to keep in line with the latest medical applications [2]. We normalize NEWS2 to a range of [0,1], representing the probability of mortality. r_{outcome} is set to -1 in the event of death and to 0 otherwise. The reward is then

$$r_t^i = -r_{\text{NEWS2}} + \mathbf{1}_{t=T_i} r_{\text{outcome}} \tag{4}$$

This normalization creates a consistent focus on mortality and eliminates the need for tuning weights between different reward components.

6. Baselines Comparisons

This section addresses the selection of baselines. Upon reviewing the literature, we observed several inconsistencies: **1)** the use of different baseline sets across studies, with some lacking state-of-the-art (SOTA) offline RL algorithms; **2)** the absence of naive baselines, such as random policy, for essential sanity checks; **3)** the omission of supervised learning baselines. Consequently, this section outlines a set of baselines that we consider appropriate for comparison.

Supervised learning baselines: Using Supervised Learning (SL) algorithms as a benchmark is crucial to determining RL's benefits in DTR. Comparing RL with simpler SL algorithms helps us understand if the complexities of RL lead to better results. Although offline RL may not always perfectly match clinicians' actual decisions, its performance should be close to real-world outcomes (as indicated by clinician decisions in test data). We employ a Long-Short-Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997) network to minimize cross-entropy loss, a standard loss function for classification tasks (see the Appendix C.4).

Naive baselines: We include random π_r, zero-drug π_{min}, max-drug π_{max}, and alternating policy π_{alt} (elaborated in the section D.1). While often neglected in RL-DTR research, naive baseline comparison is vital for two reasons: **1)** to assess if naive baselines inadvertently outperform clinicians, indicating potential flaws in evaluation metrics; **2)** to establish a lower performance bond, providing a worst-case scenario benchmark for algorithmic assessment.

Deep Q-Network (DQN): DQN(Mnih et al., 2015) integrates neural networks to approximate the Q-function with a target network and Experience Replay Buffer. Although DQN was originally designed for online RL, one can easily adapt it to the offline case where the replay buffer is fixed. The objective function for DQN can be expressed as:

$$\arg\min_{\theta} \mathbb{E}_{\mathcal{D}} \left[\left(r + \gamma \max_{a'} Q_{\theta'}(s', a') - Q_{\theta}(s, a) \right)^2 \right] \tag{5}$$

where θ and θ' are the parameters of the current and target Q-networks, respectively.

Conservative Q-Learning (CQL): The primary objective

[2] Recent clinical studies (Sivayoham et al., 2021; Mellhammar et al., 2019) indicates the possible advantage of using NEWS2 against SOFA score in managing sepsis.

of CQL(Kumar et al., 2020) is to construct a conservative estimate of the Q function, intentionally underestimating the Q values. This is achieved by incorporating an additional regularization term into the standard Bellman update equation. The loss function $L_{CQL}(\theta)$ for CQL is given by:

$$L_{\text{DQN}}(\theta) - \lambda \mathbb{E}_{\mathcal{D}}\left[\log \sum_{a} \exp(Q_\theta(s,a)) - \mathbb{E}_{a\sim\pi(\cdot|s)}[Q_\theta(s,a)]\right]$$

Here, λ is a regularization parameter, and $L_{DQN}(\theta)$ is DQN's loss function.

Batch Constrained Q-Learning (BCQ): BCQ(Fujimoto et al., 2019) ensures the action is in-distribution by introducing a Varational Auto-encoder (Sohn et al., 2015) structure. The objective function for BCQ can be formulated as:

$$\arg\min_{\theta} \mathbb{E}_{\mathcal{D}}\left[\left(r+\gamma \max_{a'\in\mathcal{A}_\phi(s')} Q_{\theta'}(s',a') - Q_\theta(s,a)\right)^2\right] \quad (6)$$

where $\mathcal{A}_\phi(s')$ denotes the set of actions similar to the training batch, generated by a VAE with parameters ϕ.

Implicit Q-Learning (IQL): IQL focuses on directly optimizing the Q-function without explicitly defining a policy(Kostrikov et al., 2022). This approach aims to improve learning efficiency and stability. The update equation for IQL can be written as:

$$\arg\min_{\theta} \mathbb{E}_{\mathcal{D}}\left[\left(r + \gamma \mathbb{E}_{a'\sim\pi_\theta}[Q_{\theta'}(s',a')] - Q_\theta(s,a)\right)^2\right] \quad (7)$$

Here, π_θ denotes the implicit policy derived from the Q-function parameterized by θ.

As the domain of offline RL rapidly evolves, newer algorithms(Fujimoto & Gu, 2021; Xiao et al., 2023) are emerging as more advanced alternatives but are not reviewed here.

7. Experiments

This section presents empirical findings using the sepsis dataset, illustrating the significant impact of varying policy evaluation algorithms, metrics, and reward design. The data set is divided into training, validation, and testing sets, comprising 70%, 15%, and 15% of the data, respectively. This partitioning adheres to the patient stratification method detailed in Appendix B. For hyperparameter optimization, grid search is performed in a unified search space. Please see Appendix E.2 for any missing details. In addition to the main result below, Appendix H provides the test set performance using Outcome and SOFA reward in Table 10 and 11, with additional 35 experiments on subsets.

7.1. Overall Comparison Results

We compare naive baselines, supervised learning, and RL (10 policies in total) on 9 metrics with 13 patient groups and 3 different reward designs, resulting in 17,550 evaluation experiments. In addition to the performance comparison in the overall test set, we stratified the test set according to clinical outcome and risks, producing 12 subsets. Due to space limitations, only selected figures and tables are shown here, with complete results in Appendix H.

To compare the performance across different reward designs and policy evaluation metrics in a straightforward way, we used a measure called "Number of Wins" (No. Wins). This measure counts how many times one algorithm outperforms all others. A win is counted when the algorithm is the best on a metric for a particular patient group with a specific reward setting. From our experiments, it is evident that the best algorithm varies across settings; moreover, although some algorithms achieve wins under policy evaluation methods such as WIS and DR, they may still behave unreasonably, i.e., deviating significantly from a doctor's policy, as indicated by metrics like RMSE or F1. We summarize the key observations below.

Effectiveness of Naive Baselines: Surprisingly, the naive baselines have wins over RL, SL, and even the doctor returns in the overall test set, as shown in Figure 1a and Figure 1b. Intuitively, naive baselines should not win even once. Table 1 reveals that the weight policy performed better than all RL, SL, and even the doctor return in the overall test set on WIS and WIS_t. However, this trend is inconsistent with the result in Table 10, where RL algorithms generally surpass naive baselines. These findings again support our recommendation for including naive baselines as an easy check to the reliability of OPE methods.

RL Performances Across Rewards: We observed significant differences in how RL algorithms perform with different rewards. For instance, DQN won the most on the NEWS2 reward, while CQL won the most on the outcome reward (see Figure 1b). Yet, on the SOFA reward, all RL policies won less than SL, making it difficult to identify a consistently performing superior algorithm across various reward settings.

RL vs SL: When RL models outperform SL on OPE metrics, it is reasonable to anticipate that RL should also demonstrate comparable performance to SL on supervised learning metrics. However, our findings present a more complex scenario. There are instances where RL's superiority in OPE metrics does not translate to better performance like RMSE or F1 score (see Table 1's DQN column), while we also observe the opposite cases (see Table 1's CQL column, WIS, WIS_b and WIS_t rows). This discrepancy raises critical questions about the effectiveness and reliability of OPE methods in evaluating RL models.

Comparison of DR and IS: Our analysis indicated that DR either overestimates or underestimates IS (see Table 1 , and Table 10, 11 in Appendix, and compare the DR row with other importance sampling rows.), subject to the reward setting. Instead of being doubly robust, our experiments

metric	alt	max	min	random	weight	SL	DQN	CQL	IQL	BCQ
RMSE_{IV}	763.89	861.51	645.83	671.39	645.83	557.81 ± 9.27	638.51 ± 8.63	541.67 ± 5.74	578.96 ± 10.06	626.2 ± 9.56
$\text{RMSE}_{\text{vaso}}$	0.67	0.89	0.32	0.5	0.59	0.31	0.44 ± 0.07	0.3 ± 0.01	0.31 ± 0.01	0.31
WIS	−4.58	−4.62	−4.58	−3.84	−3.78	−4.22 ± 0.41	−3.79 ± 0.01	−4.1 ± 1.43	−5.83	−4.58
WIS_b	−5.43	−4.81	−5.76	−4.4	−4.73	−4.62 ± 0.17	−3.88 ± 0.73	−4.48 ± 0.77	−5.31 ± 0.06	−5.41 ± 0.17
WIS_t	−4.58	−4.62	−4.58	−3.97	−3.78	−4.57 ± 0.62	−3.84 ± 0.11	−4.1 ± 1.43	−5.83	−4.58
WIS_{bt}	−5.64	−4.69	−5.61	−4.5	−4.5	−4.68 ± 0.3	−3.87 ± 0.67	−4.38 ± 0.98	−5.27 ± 0.05	−5.55 ± 0.19
DR	−0.54	−0.19	−1.55	−0.35	−0.3	−0.36	−0.14 ± 0.04	−0.71 ± 0.05	−0.51 ± 0.04	−1.54 ± 0.01
P.F1	0.2	0.02	0.2	0.2	0.0	0.31 ± 0.01	0.06 ± 0.02	0.33 ± 0.01	0.34 ± 0.01	0.23 ± 0.01
S.F1	0.19	0.02	0.19	0.19	0.0	0.3 ± 0.01	0.06 ± 0.02	0.32 ± 0.01	0.33 ± 0.01	0.22 ± 0.01
$G_{\mathcal{D}}$							-4.39			

Table 1. **Comparison across policies on the test set using NEWS2 reward.** The best and second-best algorithms are highlighted in red and blue, respectively. RMSE_{IV} and $\text{RMSE}_{\text{vaso}}$ mean the RMSE loss for the IV fluid treatment and vasopressor treatment. The predicted action class is mapped to continuous values by taking the median of the bin, as detailed in Appendix B. P.F1 and S.F1 denote the patient-wise F1 and sample-wise F1. The above denotation applies to all the following tables.

metric	alt	max	min	random	weight	SL	DQN	CQL	IQL	BCQ
RMSE_{IV}	788.91	880.8	756.41	774.78	749.58	637.22 ± 10.04	734.32 ± 11.14	609.62 ± 11.74	645.67 ± 6.1	719.77 ± 12.57
$\text{RMSE}_{\text{vaso}}$	0.54	0.85	0.27	0.47	0.56	0.25	0.4 ± 0.07	0.26 ± 0.01	0.25 ± 0.01	0.26
WIS	−3.23	−3.51	−3.44	−3.14	−3.38	−3.01 ± 0.07	−2.85 ± 0.72	−1.9 ± 0.36	−2.98 ± 0.13	−3.44
WIS_b	−3.33	−3.49	−3.41	−3.23	−3.4	−2.89 ± 0.06	−2.85 ± 0.5	−2.09 ± 0.19	−2.84 ± 0.02	−3.34 ± 0.02
WIS_t	−3.23	−3.51	−3.44	−3.14	−3.38	−3.01 ± 0.07	−2.85 ± 0.72	−1.9 ± 0.36	−2.98 ± 0.13	−3.44
WIS_{bt}	−3.29	−3.47	−3.42	−3.21	−3.41	−2.93 ± 0.05	−2.81 ± 0.49	−2.1 ± 0.24	−2.86 ± 0.05	−3.34 ± 0.05
DR	−0.26	−0.05	−1.46	−0.39	−0.46	−0.41 ± 0.02	−0.13 ± 0.09	−0.48 ± 0.13	−0.43 ± 0.02	−1.27 ± 0.03
P.F1	0.25	0.02	0.25	0.25	0.0	0.31 ± 0.01	0.07 ± 0.02	0.33 ± 0.02	0.34 ± 0.02	0.27 ± 0.01
S.F1	0.24	0.02	0.24	0.24	0.0	0.29 ± 0.01	0.07 ± 0.02	0.31 ± 0.02	0.33 ± 0.02	0.25
$G_{\mathcal{D}}$							-3.41			

Table 2. **Comparison across policies on subset "rate $[-\infty, -0.4]$ high variance" set using the NEWS2 reward.** "rate $[-\infty, -0.4]$" refers the subgroup where patients are quickly out of risk of death, i.e., a large negative change rate of NEWS2 score. "high variance" means that patient in this subgroup was observed with high fluctuation in terms of change rate of the risk of death.

suggest that the estimators in DR tend to dual 'unrobust' due to behavioral and value approximation error (further discussion in Section 7.2.). This finding calls for a reassessment of the robustness claims of DR methods in healthcare contexts.

RL outperforms SL on particular patient cohorts: We also found RL outperforms SL on both OPE metrics and supervised learning metrics for specific patient subgroups. A notable illustration can be seen in Table 2, particularly when comparing the performance of SL with CQL. In this comparison, CQL outperforms SL across nearly all metrics and achieves similar results to SL in the DR metric. This finding highlights RL's capacity to develop more effective treatment strategies than SL for certain patient groups, demonstrating its potential for individualized treatment solutions.

7.2. Understanding the Variance of Policy Evaluation

This section aims to offer an in-depth study of the causes of variance in policy evaluation (or at least a portion of these causes). Intuitively, the policy evaluation can be unreliable when the behavioral policy or target value function does not approximate well with the real one. We illustrate this point of view by visualizing the inference probability/value of the behavioral ($\hat{\pi}_{\mathcal{D}}$)/value ($\hat{Q}^{\pi}$) estimators against their sample losses.

7.2.1. Variance from $\hat{\pi}_{\mathcal{D}}$

A common assumption in OPE research is $\hat{\pi}_{\mathcal{D}} \simeq \pi_{\mathcal{D}}$. We challenged this assumption within the DTR setting by analyzing testing samples with significant loss, as shown in Figure 3. The variance in $\hat{\pi}_{\mathcal{D}}$ can be understood through two critical observations:

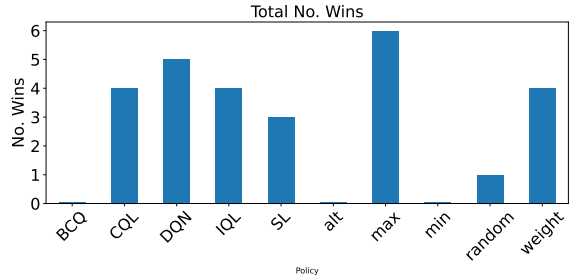
Trajectories with small behavior probabilities determine importance sampling: Importance sampling gives more weight to the trajectory with large ratios, and large ratios are more likely to be produced by trajectories with small behavior probabilities on the denominator. Therefore, it is critical for a $\hat{\pi}_{\mathcal{D}}$ to produce small probabilities correctly since such large errors can severely distort OPE estimates when paired with small probabilities.

Negative correlation between errors and inference probabilities: Smaller probabilities, which contribute more to the importance sampling, are often associated with higher loss, indicating greater deviation (see Figure 3a).

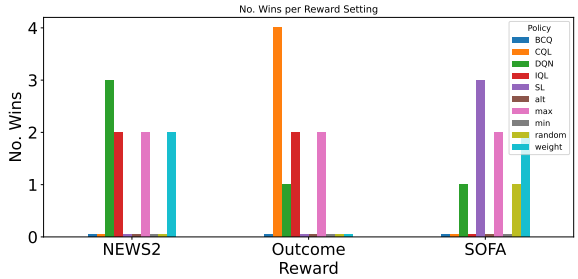
7.2.2. Variance from $\hat{Q}^{\pi}$

Quality of value approximation: The quality of learning $\hat{Q}^{\pi}$ depends on reward settings, which lead to inconsistent DR policy estimation results (as discussed in Section 7.1) This insight can be validated by Fig 3b, 3c and 3d across three reward settings.

Large errors center around high Q estimates: Large

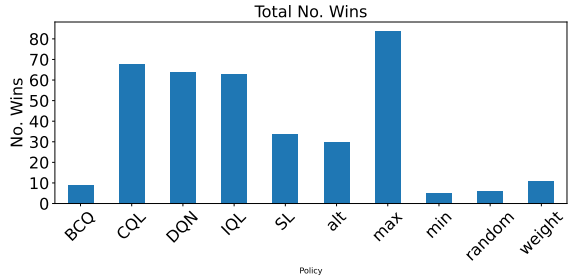


(a) Overall No. wins.

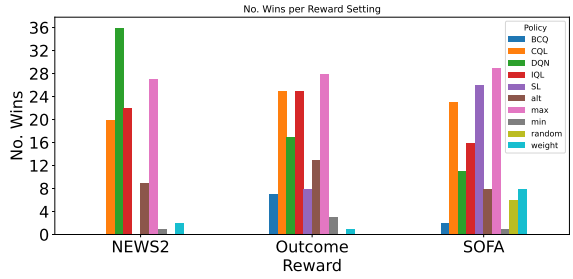


(b) No. wins under the three respective reward settings.

Figure 1. **Number of wins for each policy in the (overall) test set.** Wins are calculated based on the mean performance of 5 random seeds. Alt, min, max, random, and weight policies are naive baselines. This denotation applies to all the following figures.



(a) Sum of No. wins across patient subgroups.



(b) Sum of No. wins on stratification groups under the three respective reward settings.

Figure 2. **A summed number of wins across patient subgroups stratified by mortality risk rate of change.** This figure presents the cumulative performance of each algorithm, measured by the No. win across 12 stratified subsets derived from the test set. Wins are calculated for each algorithm within each subset across all metrics and subsequently aggregated to reflect overall performance. This approach allows for an assessment of the average algorithmic efficacy in various subgroups of patients, stratified by changes in mortality risk.

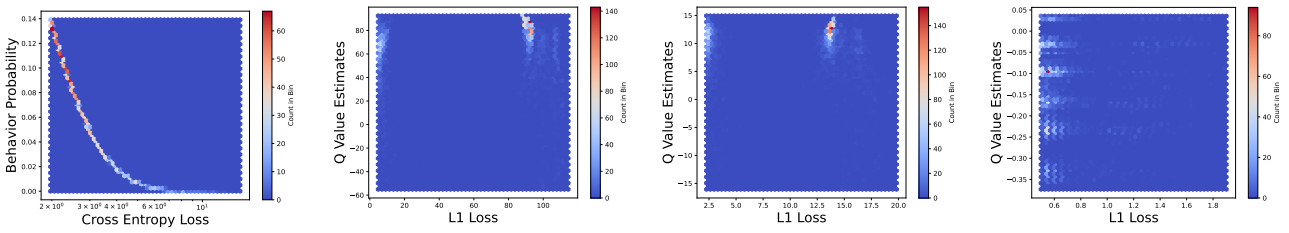
errors are concentrated around high Q estimates for both the Outcome reward ($r_{\text{alive/death}} = \pm 100$) and SOFA reward($r_{\text{alive/death}} = \pm 15$) as shown in Figure 3 (b) and (c), respectively). This suggests that the value estimator does not adequately capture the termination reward (i.e., whether the patient survives or dies). In contrast, the NEWS2 reward does not exhibit this pattern, as its reward function is smoother, ranging from -1 to 0, with -1 indicating death and 0 otherwise. The new experiments support our position of reevaluating RL-DTR. Specifically, careful design of the reward function is crucial not only for clinical meaningfulness but also for facilitating learning value estimators for OPE.

7.2.3. Investigating Model Calibration for $\hat{\pi}_{\mathcal{D}}$

Model calibration refines a model's predictive probabilities to match the actual outcome likelihoods, thereby ensuring the model's confidence reflects its empirical accuracy. We investigated the influence of model calibration on behavior policy and performance change in OPE. In this study, the temperature scale (Guo et al., 2017) is used as the model calibrator. Implementation details and hyperparameter choice for temperature scaling are provided in Appendix G.

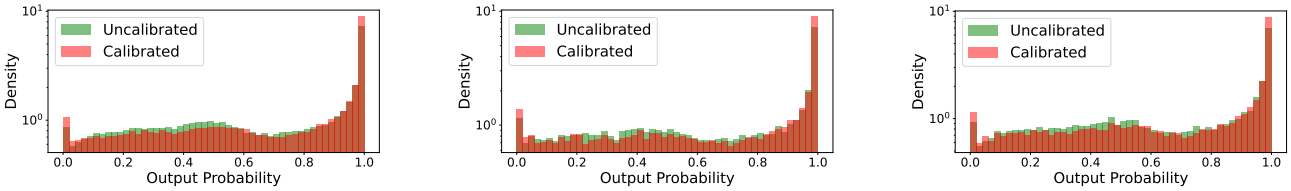
Figure 4 shows the probability distribution before and after calibration. It can be seen that more probabilities are distributed near 0 and 1 after calibration. As discussed in Section 7.2.1, importance sampling gives small probabilities more weight, and small probabilities can lead to extremely high losses, indicating that calibration may increase OPE variance in this case. To further illustrate this point, we plotted the importance ratio (i.e., $\rho^i_{1:T_i} = \prod_{t=1}^{T_i} \frac{\pi(a^i_t|s^i_t)}{\pi_{\mathcal{D}}(a^i_t|s^i_t)}$) of the random policy π_r in Figure 5. Large ratios became even larger after calibration, which implies that model calibration may increase the variance of OPE and should be applied with caution. An additional 14 figures for baseline ratio comparison are reported in the Appendix I.

To further investigate the impact of model calibration on OPE, we ran OPE for all naive baselines using the calibrated behavior policy. The results are presented in Tables 7, 8 and 9 in the Appendix G. Since we do not have access to the ground truth reward estimates for the naive baselines, we use the criterion 'higher than $G_{\mathcal{D}}$' as a sanity check: **If a naive baseline can surpass the performance of clinicians according to the OPE results, it suggests that the OPE method may not be reliable.** This is because we expect clinicians who have extensive domain knowledge to generally outperform naive baselines. The results show that some naive baselines can still surpass clinical experts, regardless of reward design. This result again supports our position of reevaluating DTR and indicates that model calibration may



(a) $\hat{\pi}_{\mathcal{D}}$ (b) $\hat{Q}^{\pi}$ on the outcome reward (c) $\hat{Q}^{\pi}$ on the SOFA reward (d) $\hat{Q}^{\pi}$ on the NEWS2 reward

Figure 3. **Behavioral and value estimator versus their losses on the testing set.** The count in each bin is indicated by a colour bar, transitioning from blue to red as the number increases. **(a)** depicts the behavioral loss (samples with a cross-entropy loss > 90th percentile) versus the inference probability. **(b)**, **(c)**, and **(d)** show the direct method estimator loss (samples with L1 loss >90th percentile) on the outcome, SOFA, and NEWS2 reward, respectively.



(a) Training set (b) Validation set (c) Test set

Figure 4. **Comparison of output probability between calibrated and uncalibrated $\hat{\pi_{\mathcal{D}}}$.** The plot shows a histogram of output probability and the number of counts in the dataset with a logarithm scale on the y-axis on training, validation and test set, respectively. It is observed that the frequencies of extreme probabilities (i.e., probabilities near 0 and 1) are higher after calibration.

not be universally helpful in DTR.

8. Consideration of RL for DTR

Building on our main findings from the previous sections, here we outline several practical considerations that are crucial when conducting experiments in RL for DTR.

SL as a baseline: Including SL as a baseline is crucial for evaluating whether RL can outperform a supervised approach using both OPE and supervised learning metrics. While comparing RL to SL is standard practice in the offline RL community, it may not be common knowledge to a broader audience. Therefore, we emphasize the importance of including SL as a baseline when assessing RL performance.

Benchmarking against simplicity: Naive baselines provide a clear and straightforward benchmark that any advanced model should surpass to justify its complexity. According to our experiments, these simple strategies can surprisingly outperform complex RL models under certain conditions. This highlights the importance of including these simple strategies for sanity checks.

Data stratification towards equitable DTR: Data stratification reveals the effectiveness of a policy on different patient subgroups and helps to identify treatment learning bias, ensuring equitable healthcare outcomes across patient populations. A comparative analysis of Figure 1 and Figure 2 support this insight: On one hand, algorithms that excel in the overall test set may not necessarily maintain their superiority in stratified patient groups. On the other hand, RL has the potential to derive improved policies from suboptimal offline data, and its effectiveness may be particularly pronounced in specific patient groups. Stratifying data by patient groups can be a valuable strategy to swiftly pinpoint where RL provides the most benefit. This targeted approach not only facilitates the identification of these advantages but also allows for detailed examination and validation by medical experts. Such a nuanced analysis could lead to more personalized and effective treatment strategies, demonstrating the true potential of RL in healthcare. Our study used episodic stratification based on the rate of change in NEWS2 scores. However, alternative stratification approaches, such as initial state stratification, could also be considered for future research.

Alternative OPE methods: Recent studies have introduced more advanced methods for quantifying the dispersion or variance of OPE (Thomas et al., 2015a;b; Gottesman et al., 2020) and avoiding overfitting the importance-weighted return as an RL agent (Liu et al., 2022). While these methods provide valuable insights into RL treatment decisions and are encouraged to be used in future RL-DTR works, they do

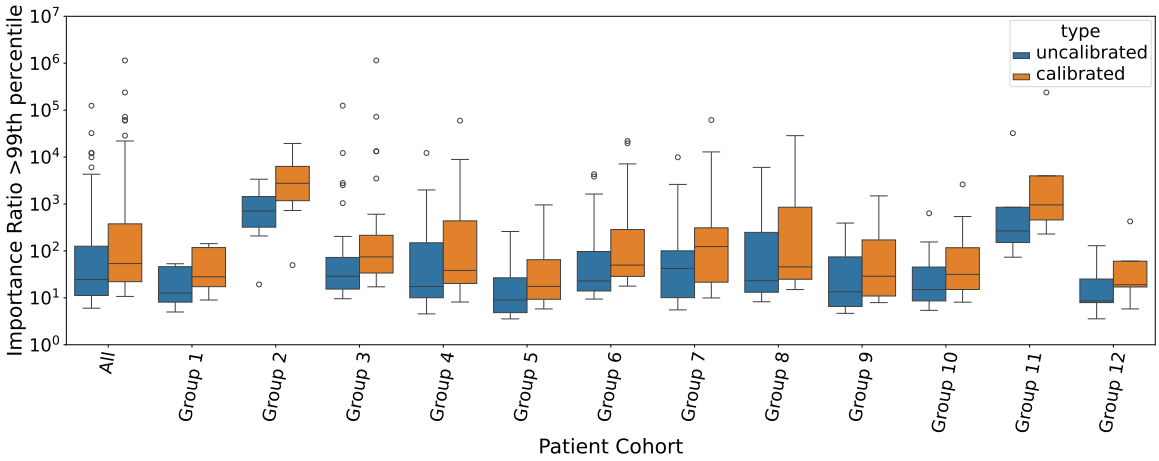


Figure 5. **Importance ratio histogram of random policy > 99th percentile.** The horizontal axis includes different datasets, where 'All' means the test set and the rest are NEWS2 risk-stratified subsets, indexed by the ascending order of NEWS2 change rate. The calibrated model contains more extremely large ratios > 99th percentile. Only ratio outliers (i.e., > 99th percentile) are plotted for visualization convenience. To view the other 14 ratio plots for 5 baseline policies in 3 different reward settings, please see Appendix I.

not eliminate the impact of variance or reduce the variance of policy evaluation. This highlights the need to develop OPE methods specifically tailored to DTR. Another category of OPE methods, that were proposed for resolving some theoretical challenges such as convergence or stability, can also be adapted for use in DTR environments. This group includes algorithms such as least squared TD (Bradtke & Barto, 1996), emphatic TD (Sutton et al., 2016), gradient TD (Maei, 2011), and accelerated TD (Pan et al., 2017b;a). These methods present opportunities for further investigation and development.

Behavioral model selection and calibration: Previous work by Raghu et al. suggests that judicious selection and calibration of the behavioral model can help mitigate variance in OPE. However, our experiments indicated that the effectiveness of calibration may not be universal. We suggest exploring a spectrum of models and calibration techniques to identify the most suitable approach for the specific DTR setting.

9. Discussions

Our work critically examines the application of offline RL algorithms in DTRs, focusing on three key aspects: diversity in policy evaluation methods, variability in reward definitions, and the absence of informative baselines. We demonstrate that the comparative performance of RL algorithms can vary significantly depending on these three factors through extensive empirical analysis using the medical sepsis dataset. Additionally, we offer practical suggestions to guide future research in this field. Moving forward, it is essential to address these complexities with a more structured and standardized approach to fully realize the benefits of RL in DTRs.

Future work and limitations. Our study has several limitations that should be acknowledged and could be important future work directions. First, while we validated our position using the Sepsis dataset, future research could explore the generalizability of our findings across other relevant datasets. Second, our work focuses on linear function approximation and does not explore other representation learning methods. Investigating locality-encouraging representations could be beneficial (Engel et al., 2004; Gomes & Krause, 2010; Schlegel et al., 2017; Pan et al., 2021), as patients with highly similar conditions are likely to benefit from similar treatments. Future research might also consider exploring neural network-based representation learning methods, such as recurrent neural networks (RNN), or studying causal effects (Raghu et al., 2018b; Zhang, 2020).

Third, this paper does not consider that multiple optimal treatments might exist for a patient's condition. Although one might expect this to have a minor effect on policy evaluation, as it is counterintuitive for vastly different doses to be simultaneously optimal for the same patient. However, it could still impact algorithm comparison if an algorithm consistently favors a certain type of mode. For instance, an SL algorithm might fit the mode presented in the training data, whereas an offline RL agent might learn a different mode. To address this issue, specialized algorithms that can capture multiple modes may be employed (Bishop, 1994; Pan et al., 2020) to model a policy. Finally, we performed episodic stratification based on the rate of change in NEWS2 scores, while there are other interesting stratification approaches, such as initial state stratification, to be explored.

Acknowledgements

Zhiyao Luo was supported by the Tang Oxford Scholarship from the China Oxford Scholar Fund (COSF) and the SBFT fellowship from the Sino-British Fellowship Trust (SBFT). Tingting Zhu was supported by the Royal Academy of Engineering under the Research Fellowship scheme.

Impact Statement

Applying RL in high-stakes medical decision-making like Sepsis carries significant implications for patient care and treatment outcomes. While RL has the potential to revolutionize personalized medicine by learning from patient responses and adapting treatment plans accordingly, its deployment in clinical settings must be approached with utmost caution.

This position paper critically examines the application of RL in DTRs. Particularly, our critique highlights significant challenges that must be addressed to prevent potential harm. The inconsistent and potentially inconclusive evaluation metrics present a clear risk: without robust assessment frameworks, the deployment of RL in medical decision-making could lead to suboptimal or even harmful treatment recommendations. Furthermore, our call for incorporating more baselines into the research of RL in DTRs underlines the necessity for benchmarking against simpler baselines. This is not only a matter of scientific rigor but also of ethical responsibility, ensuring that the adoption of more complex models is justified by demonstrable benefits to patient outcomes.

It is the research community's collective responsibility to ensure that these technologies are introduced in a manner that is safe, ethical, and ultimately beneficial to patient care. The impact of RL in healthcare could be profound, but it must be guided by a commitment not to harm, ensuring that the leap forward does not come at the expense of patient trust or safety.

References

Agarwal, R., Schuurmans, D., and Norouzi, M. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*, pp. 104–114. PMLR, 2020.

Bishop, C. M. Mixture density networks. In *Technical Report NCRG/94/001*, 1994.

Bradtke, S. J. and Barto, A. G. Linear least-squares algorithms for temporal difference learning. *Machine Learning*, 1996.

Brandfonbrener, D., Whitney, W. F., Ranganath, R., and Bruna, J. Offline RL without off-policy evaluation. In *Advances in Neural Information Processing Systems*, 2021.

Chakraborty, B. and Murphy, S. A. Dynamic treatment regimes. *Annual review of statistics and its application*, 1:447–464, 2014.

Chen, C.-C., Chong, C.-F., Liu, Y.-L., Chen, K.-C., and Wang, T.-L. Risk stratification of severe sepsis patients in the emergency department. *Emergency medicine journal: EMJ*, 23(4):281, 2006.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. Double/debiased machine learning for treatment and structural parameters, 2018.

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

Engel, Y., Mannor, S., and Meir, R. The kernel recursive least-squares algorithm. *IEEE Transactions on Signal Processing*, 52:2275–2285, 2004.

Fujimoto, S. and Gu, S. S. A minimalist approach to offline reinforcement learning. *Advances in neural information processing systems*, 34:20132–20145, 2021.

Fujimoto, S., Meger, D., and Precup, D. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, 2019.

Gomes, R. and Krause, A. Budgeted nonparametric learning from data streams. In *International Conference on Machine Learning*, pp. 391–398, 2010.

Gottesman, O., Johansson, F., Meier, J., Dent, J., Lee, D., Srinivasan, S., Zhang, L., Ding, Y., Wihl, D., Peng, X., et al. Evaluating reinforcement learning algorithms in observational health settings. *arXiv preprint arXiv:1805.12298*, 2018.

Gottesman, O., Futoma, J., Liu, Y., Parbhoo, S., Celi, L., Brunskill, E., and Doshi-Velez, F. Interpretable off-policy evaluation in reinforcement learning by highlighting influential transitions. In *International Conference on Machine Learning*, pp. 3658–3667. PMLR, 2020.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Huang, Y., Cao, R., and Rahmani, A. Reinforcement learning for sepsis treatment: A continuous action space solution. In *Machine Learning for Healthcare Conference*, pp. 631–647. PMLR, 2022.

Inada-Kim, M. and Nsutebu, E. News 2: an opportunity to standardise the management of deterioration and sepsis. *Bmj*, 360, 2018.

Jeter, R., Josef, C., Shashikumar, S., and Nemati, S. Does the" artificial intelligence clinician" learn optimal treatment strategies for sepsis in intensive care? *arXiv preprint arXiv:1902.03271*, 2019.

Jiang, N. and Li, L. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 652–661. PMLR, 2016.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Jones, A. E., Trzeciak, S., and Kline, J. A. The sequential organ failure assessment score for predicting outcome in patients with severe sepsis and evidence of hypoperfusion at the time of emergency department presentation. *Critical care medicine*, 37(5):1649, 2009.

Kajdacsy-Balla Amaral, A. C., Andrade, F. M., Moreno, R., Artigas, A., Cantraine, F., and Vincent, J.-L. Use of the sequential organ failure assessment score as a severity score. *Intensive care medicine*, 31:243–249, 2005.

Kidambi, R., Rajeswaran, A., Netrapalli, P., and Joachims, T. Morel: Model-based offline reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020.

Klein Klouwenberg, P. M., Cremer, O. L., van Vught, L. A., Ong, D. S., Frencken, J. F., Schultz, M. J., Bonten, M. J., and van der Poll, T. Likelihood of infection in patients with presumed sepsis at the time of intensive care unit admission: a cohort study. *Critical care*, 19:1–8, 2015.

Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., and Faisal, A. A. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018.

Kondrup, F., Jiralerspong, T., Lau, E., de Lara, N., Shkrob, J., Tran, M. D., Precup, D., and Basu, S. Towards safe mechanical ventilation treatment using deep offline reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 15696–15702, 2023.

Kostrikov, I., Nair, A., and Levine, S. Offline reinforcement learning with implicit q-learning. In *International Conference on Learning Representations*, 2022.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 2020.

Liu, Y., Logan, B., Liu, N., Xu, Z., Tang, J., and Wang, Y. Deep reinforcement learning for dynamic treatment regimes on medical registry data. In *2017 IEEE international conference on healthcare informatics (ICHI)*, pp. 380–385. IEEE, 2017.

Liu, Y., Flet-Berliac, Y., and Brunskill, E. Offline policy optimization with eligible actions. In *Uncertainty in Artificial Intelligence*, pp. 1253–1263. PMLR, 2022.

Luckett, D. J., Laber, E. B., Kahkoska, A. R., Maahs, D. M., Mayer-Davis, E., and Kosorok, M. R. Estimating dynamic treatment regimes in mobile health using v-learning. *Journal of the American Statistical Association*, 2019.

Maei, H. R. Gradient temporal-difference learning algorithms. *PhD thesis, University of Alberta Education and Research Archive*, 2011.

Mannor, S., Simester, D., Sun, P., and Tsitsiklis, J. N. Bias and variance approximation in value function estimates. *Management Science*, 53(2):308–322, 2007.

Mellhammar, L., Linder, A., Tverring, J., Christensson, B., Boyd, J. H., Sendi, P., Åkesson, P., and Kahn, F. News2 is superior to qsofa in detecting sepsis with organ dysfunction in the emergency department. *Journal of clinical medicine*, 8(8):1128, 2019.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *nature*, 518(7540): 529–533, 2015.

Myszczynska, M. A., Ojamies, P. N., Lacoste, A. M., Neil, D., Saffari, A., Mead, R., Hautbergue, G. M., Holbrook, J. D., and Ferraiuolo, L. Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nature Reviews Neurology*, 16(8):440–456, 2020.

Nair, A., Gupta, A., Dalal, M., and Levine, S. Awac: Accelerating online reinforcement learning with offline datasets. *International Conference on Learning Representations*, 2021.

Nambiar, M., Ghosh, S., Ong, P., Chan, Y. E., Bee, Y. M., and Krishnaswamy, P. Deep offline reinforcement learning for real-world treatment optimization applications. In

Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 4673–4684, 2023.

Nguyen, H. B., Rivers, E. P., Knoblich, B. P., Jacobsen, G., Muzzin, A., Ressler, J. A., and Tomlanovich, M. C. Early lactate clearance is associated with improved outcome in severe sepsis and septic shock. *Critical care medicine*, 32(8):1637–1642, 2004.

Pan, Y., Azer, E. S., and White, M. Effective sketching methods for value function approximation. *Conference on Uncertainty in Artificial Intelligence*, 2017a.

Pan, Y., White, A., and White, M. Accelerated gradient temporal difference learning. *AAAI Conference on Artificial Intelligence*, pp. 2464–2470, 2017b.

Pan, Y., Imani, E., Farahmand, A.-m., and White, M. An implicit function learning approach for parametric modal regression. *Advances in Neural Information Processing Systems*, 33:11442–11452, 2020.

Pan, Y., Banman, K., and White, M. Fuzzy tiling activations: A simple approach to learning sparse representations online. In *International Conference on Learning Representations*, 2021.

Peng, X., Ding, Y., Wihl, D., Gottesman, O., Komorowski, M., Li-wei, H. L., Ross, A., Faisal, A., and Doshi-Velez, F. Improving sepsis treatment strategies by combining deep and kernel-based reinforcement learning. In *AMIA Annual Symposium Proceedings*, volume 2018, pp. 887. American Medical Informatics Association, 2018.

Peng, X. B., Kumar, A., Zhang, G., and Levine, S. Advantage weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2020.

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Raghu, A., Komorowski, M., Ahmed, I., Celi, L., Szolovits, P., and Ghassemi, M. Deep reinforcement learning for sepsis treatment. *arXiv preprint arXiv:1711.09602*, 2017.

Raghu, A., Gottesman, O., Liu, Y., Komorowski, M., Faisal, A., Doshi-Velez, F., and Brunskill, E. Behaviour policy estimation in off-policy policy evaluation: Calibration matters. *arXiv preprint arXiv:1807.01066*, 2018a.

Raghu, A., Komorowski, M., and Singh, S. Model-based reinforcement learning for sepsis treatment. *arXiv preprint arXiv:1811.09602*, 2018b.

Schlegel, M., Pan, Y., Chen, J., and White, M. Adapting kernel representations online using submodular maximization. In *International Conference on Machine Learning*, pp. 3037–3046, 2017.

Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G. R., Chiche, J.-D., Coopersmith, C. M., et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama*, 315(8):801–810, 2016.

Sivayoham, N., Hussain, A. N., Shabbo, L., and Christie, D. An observational cohort study of the performance of the reds score compared to the sirs criteria, news2, curb65, sofa, meds and piro scores to risk-stratify emergency department suspected sepsis. *Annals of medicine*, 53(1): 1863–1874, 2021.

Sohn, K., Lee, H., and Yan, X. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

Sutton, R. S., Mahmood, A. R., and White, M. An emphatic approach to the problem of off-policy temporal-difference learning. *Journal of Machine Learning Research*, 2016.

Tang, S. and Wiens, J. Model selection for offline reinforcement learning: Practical considerations for healthcare settings. In *Machine Learning for Healthcare Conference*, pp. 2–35. PMLR, 2021.

Thomas, P., Theocharous, G., and Ghavamzadeh, M. High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015a.

Thomas, P., Theocharous, G., and Ghavamzadeh, M. High confidence policy improvement. In *International Conference on Machine Learning*, pp. 2380–2388. PMLR, 2015b.

Tokdar, S. T. and Kass, R. E. Importance sampling: a review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):54–60, 2010.

Wang, L., Zhang, W., He, X., and Zha, H. Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2447–2456, 2018.

Wang, Z., Zhao, H., Ren, P., Zhou, Y., and Sheng, M. Learning optimal treatment strategies for sepsis using offline reinforcement learning in continuous space. In *International Conference on Health Information Science*, pp. 113–124. Springer, 2022.

Wu, X., Li, R., He, Z., Yu, T., and Cheng, C. A value-based deep reinforcement learning model with human expertise in optimal treatment of sepsis. *NPJ Digital Medicine*, 6 (1):15, 2023.

Wu, Y., Tucker, G., and Nachum, O. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

Xiao, C., Wang, H., Pan, Y., White, A., and White, M. The in-sample softmax for offline reinforcement learning. In *International Conference on Learning Representations*, 2023.

Yu, C., Liu, J., Nemati, S., and Yin, G. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

Zhai, Y., Baek, C., Zhou, Z., Jiao, J., and Ma, Y. Computational benefits of intermediate rewards for goal-reaching policy learning. *Journal of Artificial Intelligence Research*, 73:847–896, 2022.

Zhang, J. Designing optimal dynamic treatment regimes: A causal reinforcement learning approach. In *International conference on machine learning*, pp. 11012–11022. PMLR, 2020.

A. Appendix Content

- Appendix B: Data Description and Stratification.
- Appendix C: Details of off-policy evaluation methods and our implementation.
- Appendix D: Details of naive baselines.
- Appendix E: Model Implementation and training details.
- Appendix F: Supplementary materials for Section 7.2.
- Appendix G: Details of model calibration and the comparison between calibrated and uncalibrated models.
- Appendix H: Full results, including the testing result on Outcome and SOFA reward settings (2 tables), and the respective performance on stratified patient groups (36 tables).
- Appendix I: Importance ratio histogram comparison between calibrated and uncalibrated behavior model on 3 reward settings across 5 naive baselines.

B. Data Description

B.1. Patient Stratification

Patient stratification is a pivotal process in healthcare(Chen et al., 2006; Klein Klouwenberg et al., 2015), primarily due to the heterogeneity of patient responses and outcomes. This practice is critical for ensuring that medical interventions are tailored to the unique characteristics and needs of different patient groups. Despite its importance, many existing studies in the field of RL-DTR overlook this crucial step prior to model development.

In our work, we address this gap by incorporating a simple yet effective patient stratification into our RL framework. We use the NEWS2 score as a basis for stratifying patients, specifically focusing on their risk of mortality. This approach allows us to train and evaluate our RL models on a cohort that is balanced in terms of mortality risk and outcome, thereby reducing data selection bias.

We define our stratification process by considering the rate of change in NEWS2 scores during admission, which reflects the speed of patient deterioration or recovery. Additionally, patients are categorized into two groups based on the standard deviation (SD) of their NEWS2 score changes: those with high variance indicating a more fluctuated state and those with low variance indicating a more stable deterioration/recovery, split by the SD median in the subgroup. The NEWS2 scores are segmented into distinct brackets [-0.4, -0.15, 0, 0.15, 0.4], and the variance is classified as either low or high. A comprehensive data distribution plot is shown in Fig6.

For dataset split, we first divided all data into subgroups (2 outcomes, 6 NEWS2 score bins and 2 SD levels) and make sure each group is (near) evenly distributed in the training, validation and test set for fair evaluations.

B.2. Action Discretization

The 2 drugs of interest is binned into 5 classes each to formulate a discrete aciton space. We keep the binning method aligned with Komorowski et al.:

Action	IV Fluids		Vasopressor	
	Range	Median	Range	Median
0	0	0	0	0
1	(0, 50]	40.0	(0, 0.08]	0.044
2	(50, 180]	93.75	(0.08, 0.22]	0.15
3	(180, 530]	315.35	(0.22, 0.45]	0.301
4	(530, ∞)	949.8	(0.45, ∞)	0.9

Table 3. Action Discretization Range and Median

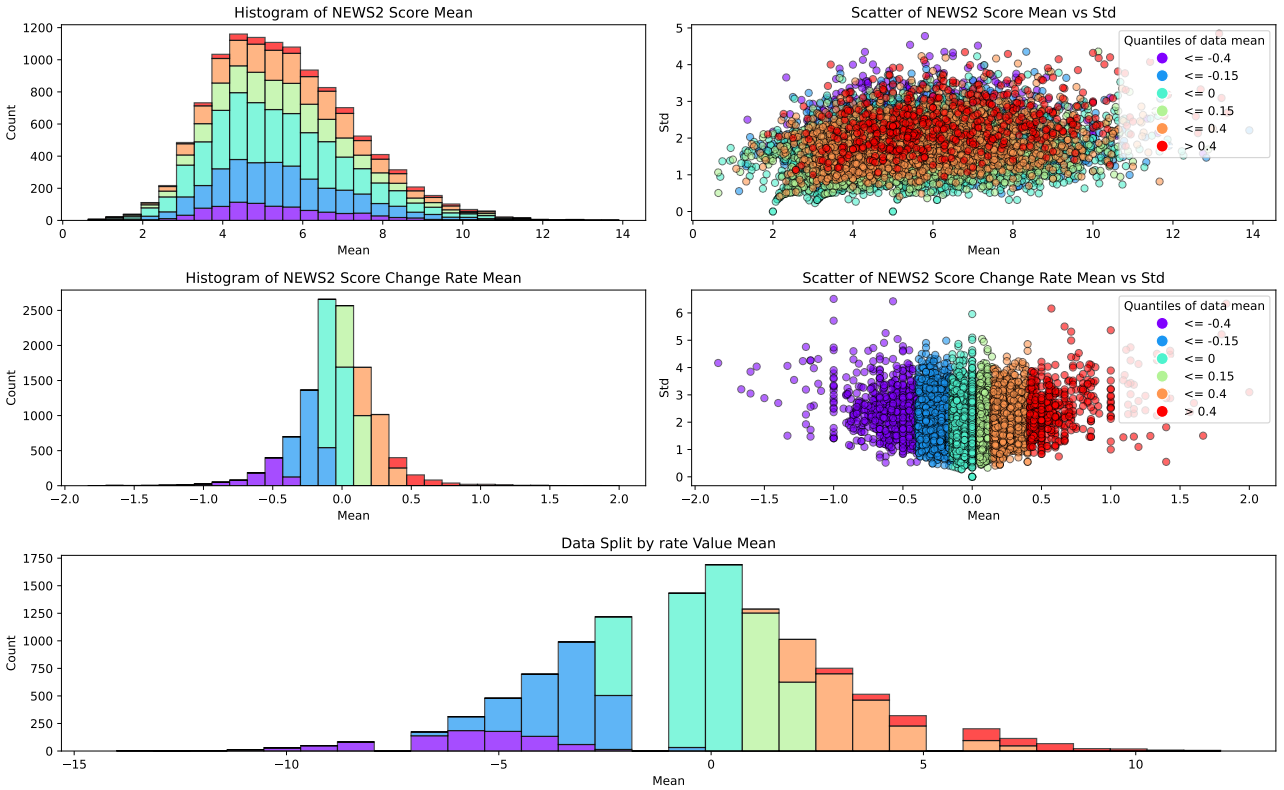


Figure 6. Data Stratification using change rate of NEWS2 score and admission outcome. The plot includes **(top left)** the distribution of mean NEWS2 score, **(top right)** mean NEWS2 against SD , **(middle left)** distribution of NEWS2 rate of change , **(middle right)** NEWS2 rate of change mean against its SD , and **(bottom)** the distribution of mean NEWS2 score for each stratified fold. Stratified by the rate of change of NEWS2, it is observed that patient of different deterioration/recovery speed distributed across all risk levels. The plot also shows that the number of patients can significantly differ across risk stratification, supporting the need of RL evaluation on all subsets.

Please note that the median for each bin is different from the material provided in https://static-content.springer.com/esm/art%3A10.1038%2Fs41591-018-0213-5/MediaObjects/41591_2018_213_MOESM1_ESM.pdf since we employed a series of modifications (See section 3) to improve medical soundness.

C. Equations and Implementations of Off-policy Evaluation Methods

C.1. Weighted Importance Sampling with Bootstrapping

The bootstrapped estimator in the context of RL involves resampling the dataset to generate multiple bootstrap samples. The value estimate for each sample is calculated, with the final bootstrapped estimate being the average of these individual estimates. Mathematically, the bootstrapped version of the Weighted Importance Sampling (WIS) estimator can be expressed as:

$$\hat{V}^{\pi}_{WIS_b} = \frac{1}{B}\sum_{b=1}^{B}\left(\frac{\sum_{i=1}^{N_b}\rho_i^{(b)}G_i^{(b)}}{\sum_{i=1}^{N_b}\rho_i^{(b)}}\right) \tag{8}$$

Here, B represents the number of bootstrap samples, N_b is the number of trajectories in the b-th bootstrap sample, $\rho_i^{(b)}$ denotes the importance sampling ratio for the i-th trajectory in the b-th sample, and $G_i^{(b)}$ is the return for the i-th trajectory. This can be simplified as:

$$\hat{V}^{\pi}_{WIS_b} = \mathbb{E}_{b\sim B}\left[\hat{V}^{\pi}_{WIS_{(b)}}\right] \tag{9}$$

For our implementation, we set $B = 100$ and $N_b = N$ for all $b \in B$.

C.2. Truncated Weighted Importance Sampling

Truncated Weighted Importance Sampling (WIS) involves limiting the influence of trajectories with disproportionately high importance weights. This is achieved by truncating the importance weights at a specified threshold τ. The truncated WIS estimator is given by:

$$\hat{V}^{\pi}_{WIS_{tr}} = \frac{\sum_{i=1}^{N}\min(\rho_i,\tau)G_i}{\sum_{i=1}^{N}\min(\rho_i,\tau)} \tag{10}$$

In our application, we select a truncation threshold of $\tau = 1$.

C.3. Combined Weighted Importance Sampling with Bootstrapping and Truncation

This approach integrates the robustness of bootstrapping with the stability of truncation, offering a more reliable estimate. The combined estimator is defined as:

$$\hat{V}^{\pi}_{WIS_{tr,B}} = \mathbb{E}_{b\sim B}\left[\hat{V}^{\pi}_{WIS_{tr(b)}}\right] \tag{11}$$

This revised section succinctly explains the implementation of different off-policy evaluation methods, ensuring the mathematical expressions are consistent with the defined symbols and terms, and providing clarity in the methodology.

To complete the paragraph with the appropriate equations and coherent explanations, here is the revised section:

C.4. Implementation of Behavioral Cloning Policy

We implement a supervised learning approach using a Long Short-Term Memory (LSTM) neural network to approximate the doctor's policy, referred to as the behavioral policy. The neural network is trained on a combined set of training and test data, and model selection is performed based on performance on a separate validation set. The training objective is to

minimize the cross-entropy loss between the model's predicted actions and the actions taken by the doctor over time. The loss function can be formulated as:

$$L_{\text{CE}}(\hat{\pi}_{\mathbb{D}}, \pi_{\mathbb{D}}) = -\mathbb{E}_{a\sim\mathcal{D}} \log \hat{\pi}_{\mathbb{D}}(a|s;\theta) \cdot a \tag{12}$$

Here, θ represents the parameters of the LSTM network and y_t^i is the actual action taken by the doctor at time t for patient i. The model with the highest patient-wise F1 score on the validation set is selected as the best model. We prefer the F1 score over the Area Under the Receiver Operating Characteristic (AUROC) due to the latter's potential to be misleading in imbalanced datasets.

C.5. Doubly Robust Estimator

The doubly robust estimator combines elements of both the direct method and importance sampling. For the direct method component, we use a training dataset comprising both the training and test sets, with the validation set used for model selection. The estimator for the direct method is updated using an offline 'SARSA' approach, which can be represented as:

$$L_{\text{DM}} = L_{\text{MSE}}\left(r + \gamma Q(s', a'), Q(s, a)\right) \tag{13}$$

The best model is chosen based on the minimum Temporal Difference (TD) error observed on the validation set. The behavioral policy utilized is consistent with the one described in the previous section.

D. Baseline Details

D.1. Naive Baselines

The naive baselines apply action based on simple rules without considering the difference of states. To avoid zero in the importance ratio, we set a small value to zeros when $\pi(a_i) = 0$. For convenience, we denote this small value as ϵ_1. The probability of the rest action will receive a reduction of ϵ_2 to guarantee that $\sum_a \pi(a) = 1$. Assume there are M actions in the action space. We denote the first action a_0 as zero drug, and the last action a_{M-1}as max drug. The equations for each baseline are given in Table 4.

Table 4. Equations for Naive Baselines

Policy	Equation
Alt Policy	$\pi_{alt}(a) = \begin{cases} 0.5 - \epsilon_2 & \text{if } a = a_0 \\ 0.5 - \epsilon_2 & \text{if } a = a_{M-1} \\ \epsilon_1 & \text{otherwise} \end{cases}$
Max Policy	$\pi_{max}(a) = \begin{cases} 1 - \epsilon_2 & \text{if } a = a_{M-1} \\ \epsilon_1 & \text{otherwise} \end{cases}$
Min Policy	$\pi_{min}(a) = \begin{cases} 1 - \epsilon_2 & \text{if } a = a_0 \\ \epsilon_1 & \text{otherwise} \end{cases}$
Random Policy	$\pi_r(a) = \frac{1}{M}$
Weight Policy	$\pi_w(a = a_j) = p_j, \quad \text{for } j = 1, 2, ..., M$

where p is the occurrence probability vector of all 25 actions present in the training set.

$$
\begin{aligned}
p = [&0.00144, 0.07288, 0.12563, 0.13930, 0.10696, \\
&0.00178, 0.03407, 0.04142, 0.06409, 0.08265, \\
&0.00173, 0.02256, 0.02194, 0.04831, 0.04838, \\
&0.00184, 0.03228, 0.02171, 0.02993, 0.02659, \\
&0.00221, 0.02637, 0.01584, 0.01834, 0.01162]
\end{aligned} \tag{14}
$$

E. Training and Hyperparameter Search

E.1. Network Structure

We employ a simplified model structure consisting of a single linear layer to minimize the influence of network architecture on the convergence of our RL algorithms. This approach also eliminates the influence of training tricks for deep learning. The input to this model is a flattened 3-frame observation window, which effectively utilizes patient data from the past 12 hours of admission, ensuring that the model has access to a relevant and recent history of patient states.

E.2. Hyperparameter Search

We conducted a comprehensive search for optimal hyperparameters, common to all RL algorithms used in this study. The key hyperparameters and their respective search ranges are presented in Table 5.

Owner	Hyperparameter	Values
All	Seed	[6311, 6890, 663, 4242, 8376]
	learning rate	[0.01, 0.001, 0.0001, 0.00001]
	batch size	[256]
	n step	[1]
	γ	[0.99]
BCQ	unlikely action threshold	[0.3, 0.5]
	imitation logits penalty	[0.02, 0.1, 0.5]
IQL	actor update frequency	[1, 5]
	quantile	[0.7, 0.9]
	β	[0.7, 1.0]
	τ	[0.001]
CQL	α	[0.1, 0.5, 1.0]

Table 5. **Summary of Hyperparameters and Their Values.** The owner column shows where the hyperparameter was used. β is the temperature parameter for policy loss calculation, τ is the coefficient for soft update of target networks, and α is the weight for the conservative loss.

The seeds for the random number generator were chosen to avoid any intentional bias and ensure reproducibility. They were generated using a standard random number generation process in Python.

F. Supplementary of 'Understanding the Variance of Policy Evaluation'

F.1. Behavioral Model and Value Estimation Error

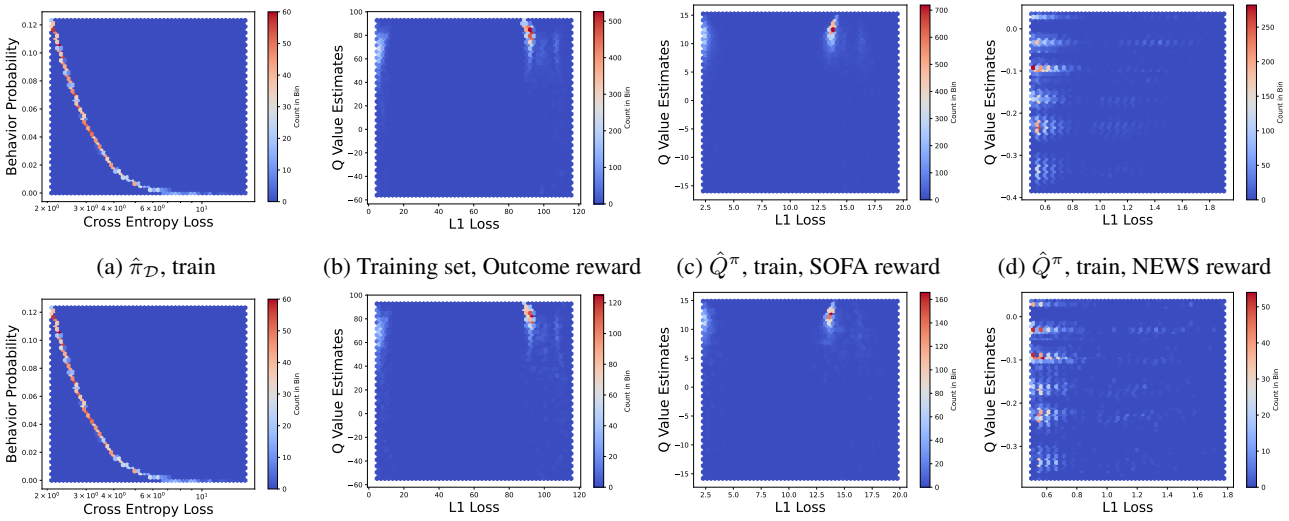
Fig 7 shows the correlation between behavioral cloning loss w.r.t the inference probability for $\hat{\pi}_{\mathcal{D}}$, and the correlation between value approximation loss w.r.t the inference state-action value for $\hat{Q}^{\pi}$. It is supplementary to Section 7.2, showing that the correlations exists not only in the testing set but in both the training and validation sets, independent of the reward settings.

We want to explain further that the error of $\hat{Q}^{\pi}$ and $\hat{\pi}_{\mathcal{D}}$ is not a consequence of ill model training. Our selected $\hat{\pi}_{\mathcal{D}}$ reaches an F1 score of 0.7 on both OPE training set and validation set[3]. Similarly, we select the $\hat{Q}^{\pi}$ that can best minimize the TD error. Your provided text describes loss distribution trends in a model's performance, contrasting behavior models and value

[3]"OPE training set' here means a combined dataset of training and test set. 'Validation set' still follows its original meaning. We avoid

functions. To enhance clarity and logical flow, the text can be revised for more precise language, structured presentation of the data, and improved coherence. Here is a refined version:

Figure 8 provides a visualization of loss distributions across samples, highlighting the predictive accuracy of both the behavior model, denoted as $\hat{\pi}_{\mathcal{D}}$, and the value functions, $\hat{Q}^{\pi}$. The behavior model generally predicts well across the majority of samples but exhibits notably higher losses in a minority of cases. This pattern follows a logarithmic trend on a linear exponential (y) scale, a characteristic often observed in imbalanced learning scenarios. For the value functions, the losses distribute differently. Specifically, the loss distribution for the Outcome and SOFA rewards form Gaussian-shaped peaks centered at 15 and 100, respectively. This distribution correlates with their termination rewards of ± 15 and ± 100. Such peaks suggest that the value functions struggle to accurately predict outcomes when there is a significant discrepancy between intermediate and termination rewards. In contrast, the NEWS2 score does not show this pattern, likely due to its more gradual reward function.



(a) $\hat{\pi}_{\mathcal{D}}$, train (b) Training set, Outcome reward (c) $\hat{Q}^{\pi}$, train, SOFA reward (d) $\hat{Q}^{\pi}$, train, NEWS reward

(e) $\hat{\pi}_{\mathcal{D}}$, val (f) $\hat{Q}^{\pi}$, val, Outcome reward (g) $\hat{Q}^{\pi}$, val, SOFA reward (h) $\hat{Q}^{\pi}$, val, NEWS2 reward

Figure 7. **a)**, **e)** Cross entropy losses against inference probabilities in the training and validation set for $\hat{\pi}_{\mathcal{D}}$. **b)**, **c)**, **d)**, **f)**, **g)**, **h)** L1 losses against Q estimates in the training and validation set on the 3 reward settings for $\hat{Q}^{\pi}$.

G. Model Calibration

We implemented temperature scaling (Guo et al., 2017) to calibrate the behavior model. Temperature scaling is a post-processing technique that adjusts the model's predicted probabilities to better match the observed probabilities, thereby (potentially) improving the model's calibration. We present the implementation details, OPE results, and a visual comparison between the calibrated and uncalibrated models.

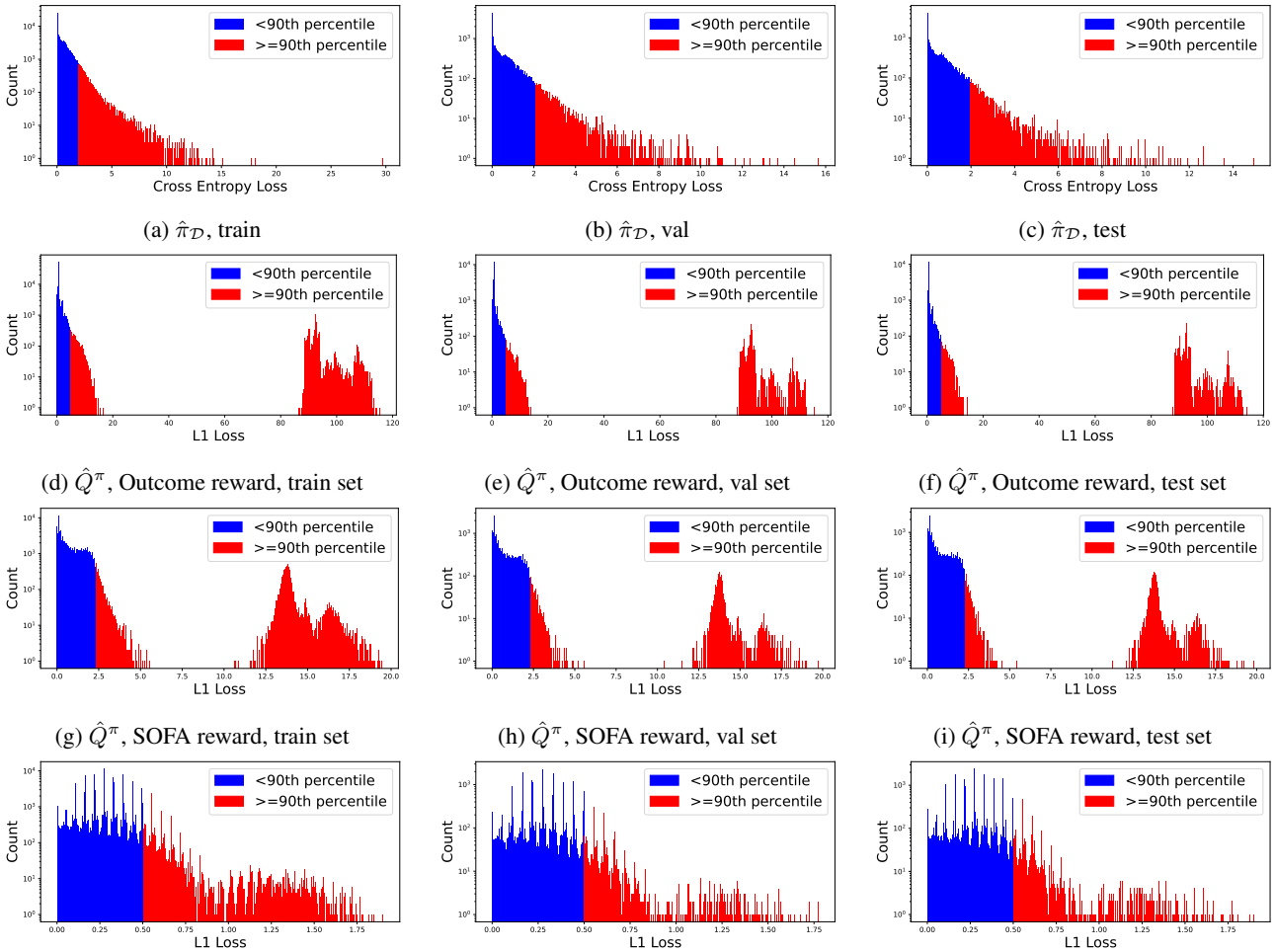
G.1. A brief Introduction to Temperature Scaling

Formally, let $f(x)$ represent the logits output by a neural network for a given input x, and let $P(y|x)$ denote the predicted probability distribution over classes y, obtained by applying the softmax function. Temperature scaling introduces a temperature parameter $T > 0$ to adjust this distribution as follows:

$$P_T(y|x) = \text{softmax}\left(\frac{f(x)}{T}\right) \tag{15}$$

Here, the softmax function is defined as $\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}}$ for logits z_i, where the summation in the denominator spans all class logits for the instance. The temperature T serves to "soften" ($T > 1$) or "sharpen" ($T < 1$) the probability distribution, with $T = 1$ leaving the original predictions unchanged.

using the original training, validation and test set as RL for decoupling



(a) $\hat{\pi}_{\mathcal{D}}$, train (b) $\hat{\pi}_{\mathcal{D}}$, val (c) $\hat{\pi}_{\mathcal{D}}$, test

(d) $\hat{Q}^{\pi}$, Outcome reward, train set (e) $\hat{Q}^{\pi}$, Outcome reward, val set (f) $\hat{Q}^{\pi}$, Outcome reward, test set

(g) $\hat{Q}^{\pi}$, SOFA reward, train set (h) $\hat{Q}^{\pi}$, SOFA reward, val set (i) $\hat{Q}^{\pi}$, SOFA reward, test set

(j) $\hat{Q}^{\pi}$, NEWS2 reward, train set (k) $\hat{Q}^{\pi}$, NEWS2 reward, val set (l) $\hat{Q}^{\pi}$, NEWS2 reward, test set

Figure 8. **a)**, **b)**, **c)** $\hat{\pi}_{\mathcal{D}}$ cross-entropy loss histogram on training, validation and test set. **d)**, **e)**, **f)** $\hat{Q}^{\pi}$ L1 loss histogram on Outcome reward, **g)**, **h)**, **i)**, $\hat{Q}^{\pi}$ L1 loss histogram on SOFA reward,**j)**, **k)**, **l)** $\hat{Q}^{\pi}$ L1 loss histogram on NEWS2 reward.

The optimal value of T is typically determined through a calibration process on the validation dataset, aiming to minimize a calibration-specific loss function. Here, we choose the Negative Log Likelihood (NLL). This optimization can be succinctly expressed as:

$$T^* = \arg\min_{T} \; L_{\text{calib}}(P_T(y|x), y_{\text{true}}) \tag{16}$$

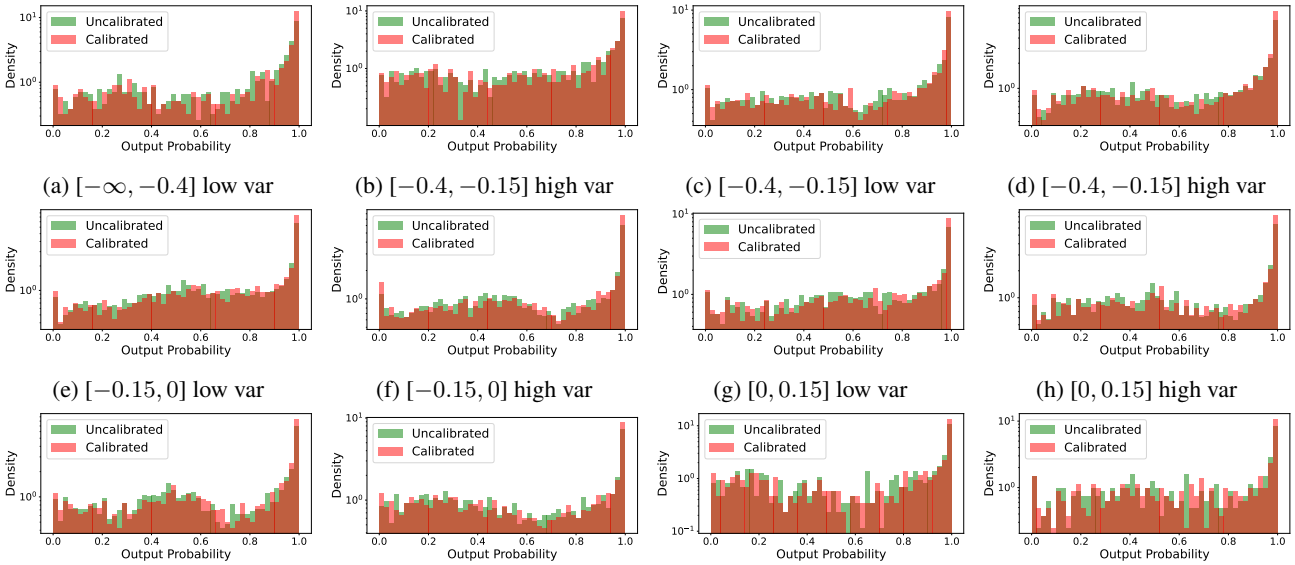
where y_{true} represents the true class labels in the validation dataset. We search hyperparameters to determine the best-calibrated model (See Table 6).

learning rate	[0.1, 0.05, 0.02, 0.01, 0.001, 5e-4, 1e-4, 5e-5, 1e-5]
batch size	[512]

Table 6. Hyperparameter search for temperature scaling calibration on the behavior policy.

G.2. Comparing Calibrated and Uncalibrated Behavior Models on OPE

To further investigate the impact of model calibration on OPE, we ran OPE for all naive baselines using the calibrated behavior policy. The results are presented in Tables 7, 8 and 9. Since we do not have access to the ground truth reward



(a) $[-\infty, -0.4]$ low var (b) $[-0.4, -0.15]$ high var (c) $[-0.4, -0.15]$ low var (d) $[-0.4, -0.15]$ high var

(e) $[-0.15, 0]$ low var (f) $[-0.15, 0]$ high var (g) $[0, 0.15]$ low var (h) $[0, 0.15]$ high var

(i) [0.15, 0.4] low var (j) [0.15, 0.4] high var (k) [0.4, ∞] low var (l) [0.4, ∞] high var

Figure 9. Comparison of output probability between calibrated and uncalibrated $\hat{\pi_{\mathcal{D}}}$. The plot shows a histogram of output probability and the number of counts in the dataset with a logarithm scale on the y-axis of the 12 stratified patient groups, where 'low var' and 'high var' mean low and high NEWS2 score variations, respectively.

estimates for the naive baselines, we use the criterion 'higher than $G_{\mathcal{D}}$' as a sanity check: **If a naive baseline can surpass the performance of clinicians according to the OPE results, it suggests that the OPE method may not be reliable.** This is because we expect clinicians who have extensive domain knowledge to generally outperform naive baselines. The table results show that some naive baselines can still surpass clinical experts, regardless of the reward design. This result again supports our position of reevaluating DTR and indicates that model calibration may not be universally helpful in DTR.

Policy Name	WIS	WIS_b	WIS_t	WIS_{bt}	DR
alt	**84.29**	**79.54**	**84.29**	**72.97**	-0.68
max	-90.47	-74.67	-90.47	-77.09	-0.38
min	**84.29**	**84.29**	**84.29**	**84.24**	-0.91
random	**86.0**	**85.48**	**84.3**	**82.87**	-0.58
weight	**86.87**	**63.98**	**86.87**	57.87	-0.5
$G_{\mathcal{D}}$			61.54		

Table 7. Outcome reward

WIS	WIS_b	WIS_t	WIS_{bt}	DR
13.34	**12.28**	**13.34**	**11.42**	-0.54
-9.44	-10.03	-9.44	-9.53	-0.3
13.34	**12.86**	**13.34**	**12.93**	-0.73
14.33	**13.35**	**10.26**	**11.86**	-0.41
16.43	**11.24**	**16.43**	**9.96**	-0.38
		9.47		

Table 8. SOFA reward

WIS	WIS_b	WIS_t	WIS_{bt}	DR
-4.57	-5.63	-4.57	-5.43	**-0.56**
-4.6	-4.85	-4.6	-4.63	**-0.21**
-4.57	-5.59	-4.57	-5.56	**-1.49**
-4.47	-4.77	-5.82	-5.09	**-0.39**
-3.78	-4.53	**-3.78**	-4.61	**-0.35**
		-4.39		

Table 9. NEWS2 reward

In conclusion, our analysis demonstrates that temperature scaling may not be sufficient to mitigate the challenges associated with OPE in the presence of small probabilities. Researchers should be aware of the limitations and consider exploring alternative OPE methods or implementing other calibration methods to obtain more reliable policy evaluations.

H. Full Result

H.1. Full Result on Test set

metric	alt	max	min	random	weight	SL	DQN	CQL	IQL	BCQ
RMSE_{IV}	763.89	861.51	645.83	671.39	645.83	556.48 ± 6.23	624.06 ± 17.85	616.71 ± 17.94	578.13 ± 6.14	620.74 ± 2.12
$\text{RMSE}_{\text{vaso}}$	0.67	0.89	0.32	0.5	0.59	0.31	0.32	0.31	0.31	0.31
WIS	84.29	-89.05	84.29	86.8	86.87	86.66 ± 0.2	87.62 ± 3.34	89.37 ± 2.95	84.29	84.29
WIS_b	81.21	-75.94	84.05	84.35	54.6	85.05 ± 0.95	87.52 ± 1.33	88.26 ± 1.82	77.57 ± 6.87	84.52 ± 0.17
WIS_t	84.29	-89.05	84.29	86.64	86.87	86.3 ± 0.52	87.62 ± 3.34	89.37 ± 2.95	84.29	84.29
WIS_{bt}	79.85	-79.75	84.23	80.84	66.22	84.74 ± 0.99	87.6 ± 1.33	88.28 ± 1.74	75.85 ± 4.38	84.51 ± 0.16
DR	-0.66	-0.38	-0.95	-0.58	-0.49	-0.54 ± 0.01	-0.35 ± 0.05	-0.51 ± 0.08	-0.54 ± 0.03	-0.95
P.F1	0.2	0.02	0.2	0.2	0.0	0.31 ± 0.01	0.2	0.24	0.32 ± 0.01	0.24 ± 0.01
S.F1	0.19	0.02	0.19	0.19	0.0	0.31 ± 0.01	0.2	0.24	0.32 ± 0.01	0.23 ± 0.01
$G_{\mathcal{D}}$						61.54				

Table 10. Outcome all

metric	alt	max	min	random	weight	SL	DQN	CQL	IQL	BCQ
RMSE_{IV}	763.89	861.51	645.83	671.39	645.83	536.36 ± 13.56	628.24 ± 16.8	582.33 ± 10.28	578.13 ± 6.14	627.97 ± 1.67
$\text{RMSE}_{\text{vaso}}$	0.67	0.89	0.32	0.5	0.59	0.3	0.32	0.32	0.31	0.31
WIS	13.34	-12.25	13.34	16.26	16.43	15.49 ± 0.91	12.41 ± 0.72	13.84 ± 1.79	10.24	13.34
WIS_b	12.49	-11.3	12.86	14.19	9.52	13.78 ± 0.66	12.48 ± 0.78	11.85 ± 3.78	11.08 ± 0.62	12.97 ± 0.12
WIS_t	13.34	-12.25	13.34	15.86	16.43	14.36 ± 1.57	12.41 ± 0.72	13.84 ± 1.79	10.24	13.34
WIS_{bt}	12.19	-11.25	12.92	13.34	11.49	13.44 ± 0.87	12.27 ± 1.23	11.76 ± 3.49	10.73 ± 0.44	12.91 ± 0.03
DR	-0.5	-0.27	-0.76	-0.38	-0.37	-0.35 ± 0.01	-0.25 ± 0.06	-0.37 ± 0.1	-0.36 ± 0.02	-0.76
P.F1	0.2	0.02	0.2	0.2	0.0	0.34 ± 0.01	0.17 ± 0.01	0.26 ± 0.02	0.32 ± 0.01	0.22
S.F1	0.19	0.02	0.19	0.19	0.0	0.33 ± 0.01	0.17 ± 0.01	0.26 ± 0.02	0.32 ± 0.01	0.21
$G_{\mathcal{D}}$						9.47				

Table 11. SOFA all

H.2. Full Result on Stratified Subsets

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	880.57	929.48	825.16	821.09	828.29	682.74 ± 8.09	802.31 ± 20.11	793.05 ± 13.33	719.46 ± 13.2	795.1 ± 5.0
$\text{RMSE}_{\text{vaso}}$	0.71	0.87	0.2	0.44	0.55	0.19 ± 0.01	0.2	0.2	0.19 ± 0.01	0.18
WIS	93.3	86.87	90.77	86.87	86.87	89.34 ± 3.42	92.01 ± 1.46	92.72 ± 3.19	91.95 ± 1.68	91.74 ± 0.89
WIS_b	92.12	76.88	91.06	62.66	86.34	88.92 ± 2.47	91.35 ± 0.68	91.05 ± 0.68	90.11 ± 1.23	91.57 ± 0.73
WIS_t	93.3	86.87	90.77	86.87	86.87	89.34 ± 3.42	92.01 ± 1.46	92.72 ± 3.19	91.95 ± 1.68	91.74 ± 0.89
WIS_{bt}	92.32	74.96	90.89	72.06	87.75	89.43 ± 2.0	91.56 ± 0.76	90.97 ± 1.06	90.54 ± 0.78	91.59 ± 0.63
DR	-0.22	-0.43	-0.73	-0.22	-0.22	-0.13 ± 0.12	-0.22	-0.3 ± 0.1	-0.07 ± 0.1	-0.55
P.F1	0.23	0.02	0.23	0.23	0.0	0.32 ± 0.02	0.23 ± 0.01	0.25 ± 0.01	0.36 ± 0.02	0.27 ± 0.01
S.F1	0.22	0.02	0.22	0.22	0.0	0.32 ± 0.02	0.22 ± 0.01	0.24 ± 0.01	0.35 ± 0.01	0.26 ± 0.01
$G_{\mathcal{D}}$							68.51			

Table 12. Outcome sub rate $\in [-\infty, -0.4]$ low std

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	788.91	880.8	756.41	774.78	749.58	637.96 ± 8.04	729.93 ± 14.25	700.73 ± 25.47	648.47 ± 8.58	719.53 ± 5.46
$\text{RMSE}_{\text{vaso}}$	0.54	0.85	0.27	0.47	0.56	0.24 ± 0.02	0.27	0.27	0.27 ± 0.01	0.26
WIS	23.66	26.75	-8.29	88.59	-94.02	88.64	87.46 ± 13.66	94.02 ± 1.51	89.36 ± 1.52	29.48 ± 34.13
WIS_b	14.77	24.94	-0.78	44.85	-39.81	90.85 ± 0.16	83.92 ± 14.47	92.61 ± 0.85	90.81 ± 0.78	33.48 ± 22.12
WIS_t	23.66	26.75	-8.29	88.59	-94.02	88.64	87.46 ± 13.66	94.02 ± 1.51	89.36 ± 1.52	29.48 ± 34.13
WIS_{bt}	26.13	23.87	-0.81	50.99	-13.95	90.43 ± 0.42	84.03 ± 14.76	93.13 ± 1.1	90.28 ± 0.75	30.69 ± 20.3
DR	-0.12	0.0	-0.3	-0.27	-0.4	-0.06 ± 0.08	-0.03 ± 0.07	0.0	-0.11 ± 0.02	-0.12
P.F1	0.25	0.02	0.25	0.25	0.0	0.33 ± 0.02	0.24 ± 0.02	0.28 ± 0.01	0.35 ± 0.01	0.27 ± 0.01
S.F1	0.24	0.02	0.24	0.24	0.0	0.31 ± 0.02	0.23 ± 0.02	0.26 ± 0.01	0.33 ± 0.03	0.26 ± 0.01
$G_{\mathcal{D}}$							78.97			

Table 13. Outcome sub rate $\in [-\infty, -0.4]$ high std

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	732.68	849.88	620.73	649.77	620.52	531.04 ± 6.78	599.25 ± 17.6	584.99 ± 13.05	557.24 ± 7.32	599.44 ± 2.38
$\text{RMSE}_{\text{vaso}}$	0.6	0.87	0.2	0.44	0.55	0.19 ± 0.01	0.2	0.19	0.2 ± 0.01	0.19
WIS	84.29	-87.75	84.29	-94.09	-73.02	-83.9 ± 7.91	84.64 ± 0.75	89.57 ± 3.58	61.66 ± 32.53	84.52 ± 0.21
WIS_b	84.51	-54.45	84.36	-33.01	-43.75	-24.56 ± 13.79	84.4 ± 0.08	84.5 ± 0.29	64.9 ± 25.17	84.76 ± 0.45
WIS_t	84.29	-87.75	84.29	-94.09	-73.02	-83.91 ± 7.92	84.64 ± 0.75	89.57 ± 3.58	61.62 ± 32.57	84.52 ± 0.21
WIS_{bt}	84.41	-76.1	84.3	-15.19	-38.84	-20.09 ± 12.61	84.39 ± 0.11	84.44 ± 0.22	63.09 ± 28.61	84.78 ± 0.39
DR	-0.33	-0.04	-0.67	-0.18	-0.13	-0.07 ± 0.02	-0.06 ± 0.04	-0.13 ± 0.07	-0.08 ± 0.02	-0.67
P.F1	0.21	0.01	0.21	0.21	0.0	0.32 ± 0.01	0.21 ± 0.01	0.23 ± 0.01	0.33 ± 0.02	0.25
S.F1	0.21	0.01	0.21	0.21	0.0	0.32 ± 0.01	0.21	0.23 ± 0.01	0.32 ± 0.02	0.25 ± 0.01
$G_{\mathcal{D}}$							67.49			

Table 14. Outcome sub rate $\in [-0.4, -0.15]$ low std

I. Importance Ratio Histogram of Naive Baselines

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	820.25	887.4	770.94	771.37	767.21	661.62 ± 8.98	745.14 ± 22.99	709.98 ± 19.82	675.19 ± 6.55	738.97 ± 2.38
$\text{RMSE}_{\text{vaso}}$	0.82	1.03	0.65	0.74	0.81	0.64 ± 0.01	0.65	0.64	0.64 ± 0.01	0.64
WIS	84.29	84.29	84.29	88.63	85.02	88.62 ± 0.01	89.24 ± 3.67	91.67 ± 0.59	87.94 ± 2.25	84.29
WIS_b	84.04	69.52	84.84	86.01	86.88	88.25 ± 1.19	89.04 ± 3.6	89.94 ± 0.39	88.07 ± 1.34	85.4 ± 0.59
WIS_t	84.29	84.29	84.29	88.63	85.02	88.62 ± 0.01	89.24 ± 3.67	91.67 ± 0.59	87.94 ± 2.25	84.29
WIS_{bt}	83.92	67.52	81.18	88.22	86.39	88.04 ± 0.99	88.85 ± 3.57	89.59 ± 0.37	87.66 ± 1.73	84.86 ± 0.73
DR	-0.55	-0.45	-0.67	-0.28	-0.26	-0.27 ± 0.01	-0.22 ± 0.08	-0.26 ± 0.12	-0.33 ± 0.06	-0.67
P.F1	0.2	0.02	0.2	0.2	0.0	0.29 ± 0.02	0.2 ± 0.02	0.23 ± 0.01	0.3 ± 0.01	0.23
S.F1	0.19	0.03	0.19	0.19	0.0	0.29 ± 0.02	0.18 ± 0.02	0.22 ± 0.01	0.29 ± 0.01	0.21
$G_{\mathcal{D}}$							68.67			

Table 15. Outcome sub rate $\in [-0.4, -0.15]$ high std

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	688.63	833.0	457.4	532.53	486.77	408.54 ± 9.26	443.51 ± 7.56	442.82 ± 2.72	428.3 ± 8.78	438.72 ± 1.4
$\text{RMSE}_{\text{vaso}}$	0.67	0.89	0.26	0.48	0.58	0.25	0.26	0.26	0.26	0.25 ± 0.01
WIS	84.29	-90.44	84.29	-84.32	-84.21	84.3	92.01 ± 1.19	90.24 ± 2.64	84.29 ± 0.05	84.29
WIS_b	34.74	-35.59	71.32	-60.3	-62.77	84.21 ± 0.37	90.89 ± 0.74	86.12 ± 7.18	83.42 ± 2.45	81.81 ± 4.11
WIS_t	84.29	-90.44	84.29	-84.32	-84.21	84.32 ± 0.05	92.01 ± 1.19	90.24 ± 2.64	83.84 ± 1.0	84.29
WIS_{bt}	40.39	-37.48	64.83	-68.6	-62.39	84.22 ± 0.78	91.06 ± 0.79	85.71 ± 8.78	82.2 ± 5.04	81.1 ± 6.22
DR	-0.51	-0.15	-0.97	-0.3	-0.18	-0.28 ± 0.03	-0.23 ± 0.12	-0.36 ± 0.09	-0.21 ± 0.05	-0.97
P.F1	0.22	0.01	0.22	0.22	0.0	0.32	0.22	0.28 ± 0.01	0.34	0.27 ± 0.01
S.F1	0.21	0.01	0.21	0.21	0.0	0.32	0.21 ± 0.01	0.27 ± 0.01	0.33 ± 0.01	0.26 ± 0.01
$G_{\mathcal{D}}$							65.86			

Table 16. Outcome sub rate $\in [-0.15, 0]$ low std

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	765.09	867.74	669.04	696.63	662.95	571.17 ± 6.95	646.22 ± 19.11	636.61 ± 18.79	597.7 ± 6.59	642.96 ± 2.08
$\text{RMSE}_{\text{vaso}}$	0.66	0.87	0.28	0.48	0.57	0.26 ± 0.01	0.28	0.27 ± 0.01	0.27 ± 0.01	0.27
WIS	84.29	-86.87	84.29	84.29	84.2	84.29	85.88 ± 3.52	89.31 ± 3.79	84.29	84.29
WIS_b	51.74	-50.46	83.73	68.07	68.21	66.15 ± 4.7	85.84 ± 0.92	87.91 ± 3.06	48.93 ± 8.66	84.55 ± 0.33
WIS_t	84.29	-86.87	84.29	84.29	84.2	84.29	85.88 ± 3.52	89.31 ± 3.79	84.29	84.29
WIS_{bt}	52.42	-43.31	83.83	63.38	63.12	65.74 ± 8.39	85.6 ± 0.65	86.94 ± 1.69	47.17 ± 9.39	84.48 ± 0.16
DR	-0.7	-0.36	-0.85	-0.46	-0.3	-0.46 ± 0.02	-0.32 ± 0.11	-0.4 ± 0.11	-0.52 ± 0.05	-0.85
P.F1	0.2	0.02	0.2	0.2	0.0	0.31 ± 0.02	0.2 ± 0.01	0.23	0.33 ± 0.01	0.25
S.F1	0.19	0.02	0.19	0.19	0.0	0.31 ± 0.02	0.19 ± 0.01	0.23	0.32 ± 0.01	0.24
$G_{\mathcal{D}}$							64.19			

Table 17. Outcome sub rate $\in [-0.15, 0]$ high std

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	756.44	875.08	576.23	624.74	597.18	541.54 ± 8.98	561.71 ± 10.18	562.25 ± 10.25	555.25 ± 12.44	559.56 ± 1.79
$\text{RMSE}_{\text{vaso}}$	0.6	0.87	0.13	0.44	0.54	0.13	0.13	0.13	0.15 ± 0.01	0.12
WIS	91.35	-68.28	91.35	84.29	84.29	87.58 ± 1.96	91.35	91.35	86.2 ± 4.71	91.35
WIS_b	86.25	-10.25	80.9	75.39	75.64	86.06 ± 1.04	90.0 ± 1.06	89.88 ± 0.77	85.69 ± 0.54	88.8 ± 1.21
WIS_t	91.35	-68.28	91.35	84.29	84.29	87.64 ± 1.95	91.35	91.35	86.07 ± 4.83	91.35
WIS_{bt}	84.24	-11.44	80.24	72.66	74.64	86.5 ± 0.97	90.14 ± 1.0	89.85 ± 0.55	85.86 ± 0.94	88.38 ± 1.6
DR	-0.51	-0.21	-0.99	-0.85	-0.85	-0.31 ± 0.06	-0.41 ± 0.11	-0.7 ± 0.26	-0.31 ± 0.09	-0.87 ± 0.23
P.F1	0.2	0.01	0.2	0.2	0.0	0.29 ± 0.02	0.2 ± 0.01	0.24	0.31 ± 0.02	0.24 ± 0.01
S.F1	0.19	0.01	0.19	0.19	0.0	0.29 ± 0.02	0.2 ± 0.01	0.24	0.31 ± 0.01	0.24 ± 0.01
$G_{\mathcal{D}}$							60.64			

Table 18. Outcome sub rate $\in [0, 0.15]$ low std

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	760.98	834.61	628.74	637.52	628.39	530.49 ± 8.58	605.18 ± 20.02	587.38 ± 22.11	559.66 ± 7.59	601.25 ± 2.88
$\text{RMSE}_{\text{vaso}}$	0.7	0.88	0.3	0.49	0.58	0.29 ± 0.01	0.3	0.3	0.29 ± 0.01	0.29
WIS	84.29	-84.26	84.29	84.29	84.29	84.29 ± 0.01	89.92 ± 3.21	86.19 ± 3.46	83.45 ± 1.89	84.72 ± 0.39
WIS_b	84.6	-47.35	84.58	47.19	79.95	77.89 ± 6.35	89.05 ± 3.03	86.03 ± 1.7	72.98 ± 16.41	84.95 ± 0.36
WIS_t	84.29	-84.26	84.29	84.29	84.29	84.29	89.92 ± 3.21	86.19 ± 3.46	83.07 ± 2.74	84.72 ± 0.39
WIS_{bt}	84.68	-58.99	84.57	52.31	73.77	80.23 ± 2.9	89.87 ± 2.85	85.97 ± 1.38	72.25 ± 17.75	84.93 ± 0.35
DR	-0.71	-0.37	-1.02	-0.67	-0.55	-0.69 ± 0.03	-0.56 ± 0.15	-0.73 ± 0.15	-0.7 ± 0.08	-0.92 ± 0.1
P.F1	0.17	0.02	0.17	0.17	0.0	0.3 ± 0.01	0.17 ± 0.02	0.2 ± 0.01	0.3 ± 0.01	0.2 ± 0.01
S.F1	0.17	0.02	0.17	0.17	0.0	0.3 ± 0.01	0.16 ± 0.02	0.2 ± 0.01	0.3 ± 0.01	0.19
$G_{\mathcal{D}}$							59.33			

Table 19. Outcome sub rate $\in [0, 0.15]$ high std

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	736.69	853.28	581.25	616.67	592.53	497.49 ± 11.07	560.83 ± 15.47	561.57 ± 14.77	519.75 ± 18.49	559.81 ± 2.82
$\text{RMSE}_{\text{vaso}}$	0.6	0.87	0.24	0.46	0.57	0.23 ± 0.01	0.24	0.24	0.23 ± 0.01	0.23
WIS	−92.24	−92.27	87.46	−88.64	−84.29	84.32 ± 0.15	84.32 ± 0.06	88.32 ± 3.11	84.72 ± 0.99	87.48 ± 0.08
WIS_b	−21.8	−28.01	88.38	−79.87	−80.21	84.5 ± 0.38	85.71 ± 0.7	88.67 ± 1.14	83.33 ± 2.44	88.81 ± 0.31
WIS_t	−92.24	−92.27	87.46	−88.64	−84.29	84.32 ± 0.15	84.32 ± 0.06	88.32 ± 3.11	84.72 ± 0.99	87.48 ± 0.08
WIS_{bt}	−14.44	−33.15	88.22	−81.18	−76.28	84.24 ± 0.66	85.53 ± 0.54	88.52 ± 1.53	82.99 ± 2.66	88.79 ± 0.49
DR	−0.85	−0.65	−1.15	−0.99	−0.93	−0.99 ± 0.01	−0.56 ± 0.06	−0.8 ± 0.19	−1.06 ± 0.07	−1.09 ± 0.07
P.F1	0.2	0.01	0.2	0.2	0.0	0.34 ± 0.01	0.2 ± 0.01	0.27 ± 0.01	0.34 ± 0.01	0.24 ± 0.01
S.F1	0.18	0.01	0.18	0.18	0.0	0.33 ± 0.01	0.2 ± 0.02	0.27 ± 0.01	0.33 ± 0.01	0.24
$G_{\mathcal{D}}$							57.54			

Table 20. Outcome sub rate $\in [0.15, 0.4]$ low std

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	782.93	832.02	699.16	694.47	686.84	594.95 ± 8.31	668.62 ± 28.5	653.72 ± 18.89	624.3 ± 16.96	672.05 ± 2.09
$\text{RMSE}_{\text{vaso}}$	0.62	0.85	0.25	0.45	0.54	0.24 ± 0.01	0.25	0.25	0.24 ± 0.01	0.23
WIS	93.07	−84.29	40.12	−36.02	−84.51	84.2 ± 0.09	89.58 ± 3.95	92.2 ± 1.69	82.16 ± 4.71	40.17 ± 0.03
WIS_b	66.26	−32.94	35.96	−22.56	−73.83	79.76 ± 2.53	89.41 ± 2.37	90.54 ± 0.5	74.85 ± 9.66	37.09 ± 5.96
WIS_t	93.07	−84.29	40.12	−36.02	−84.51	84.2 ± 0.09	89.58 ± 3.95	92.2 ± 1.69	82.16 ± 4.71	40.17 ± 0.03
WIS_{bt}	59.8	−29.34	36.47	−24.04	−76.8	79.8 ± 2.69	89.84 ± 2.51	90.52 ± 0.72	74.11 ± 9.32	37.15 ± 2.23
DR	−0.98	−0.46	−1.12	−0.83	−0.8	−0.7 ± 0.05	−0.29 ± 0.07	−0.46 ± 0.14	−0.83 ± 0.09	−0.76 ± 0.13
P.F1	0.17	0.02	0.17	0.17	0.0	0.29 ± 0.02	0.17 ± 0.01	0.2 ± 0.01	0.28 ± 0.02	0.2
S.F1	0.16	0.02	0.16	0.16	0.0	0.28 ± 0.02	0.16 ± 0.01	0.2 ± 0.01	0.28 ± 0.02	0.19
$G_{\mathcal{D}}$							45.82			

Table 21. Outcome sub rate $\in [0.15, 0.4]$ high std

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	828.61	828.61	626.63	642.26	637.93	553.92 ± 13.64	596.82 ± 30.51	595.55 ± 32.69	537.16 ± 6.65	596.26 ± 4.61
$\text{RMSE}_{\text{vaso}}$	0.87	0.87	0.4	0.53	0.62	0.29 ± 0.03	0.38 ± 0.01	0.38 ± 0.01	0.33 ± 0.02	0.36 ± 0.04
WIS	94.8	−89.68	93.72	−92.49	−90.44	−21.63 ± 25.99	94.25 ± 1.16	93.72	−53.63 ± 20.54	93.72
WIS_b	71.67	−90.58	84.53	−91.33	−91.33	−12.25 ± 18.16	71.45 ± 13.1	89.53 ± 2.77	−25.64 ± 7.78	89.86 ± 1.19
WIS_t	94.8	−89.68	93.72	−92.49	−90.44	−21.63 ± 25.99	94.25 ± 1.16	93.72	−53.63 ± 20.54	93.72
WIS_{bt}	71.72	−90.15	90.31	−90.64	−91.15	−10.64 ± 13.05	73.56 ± 16.71	91.28 ± 1.72	−22.53 ± 19.15	90.54 ± 3.49
DR	−1.92	−1.61	−2.05	−2.49	−1.78	−2.28 ± 0.04	−1.22 ± 0.08	−1.53 ± 0.08	−2.21 ± 0.09	−1.74 ± 0.19
P.F1	0.23	0.03	0.23	0.23	0.02	0.32 ± 0.02	0.23 ± 0.01	0.26 ± 0.01	0.34 ± 0.02	0.26
S.F1	0.23	0.03	0.23	0.23	0.02	0.3 ± 0.02	0.23 ± 0.01	0.26 ± 0.01	0.32 ± 0.02	0.26
$G_{\mathcal{D}}$							23.02			

Table 22. Outcome sub rate $\in [0.4, \infty]$ low std

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	1131.35	1108.49	1131.35	1090.9	1100.26	952.83 ± 9.14	1106.01 ± 22.44	1032.13 ± 14.06	950.08 ± 7.98	1085.45 ± 4.58
$\text{RMSE}_{\text{vaso}}$	0.21	0.86	0.21	0.45	0.54	0.21	0.21	0.2 ± 0.01	0.21 ± 0.01	0.18 ± 0.01
WIS	90.0	3.99	89.54	−68.14	−83.06	92.27 ± 0.01	90.65 ± 2.48	95.1	92.31 ± 0.29	90.27 ± 0.91
WIS_b	81.32	−0.46	70.64	−20.11	−50.82	91.54 ± 0.71	80.52 ± 7.96	91.62 ± 2.23	91.83 ± 0.43	88.1 ± 1.72
WIS_t	90.0	3.94	89.54	−85.19	−83.06	92.07 ± 0.41	90.65 ± 2.48	95.1	92.63 ± 1.01	90.27 ± 0.91
WIS_{bt}	72.37	−3.14	83.27	−47.0	−68.58	90.96 ± 1.05	74.47 ± 11.5	92.32 ± 1.75	91.54 ± 0.91	85.09 ± 2.4
DR	−2.66	−2.05	−2.82	−2.05	−1.99	−2.05	−1.93 ± 0.17	−2.25 ± 0.4	−2.25 ± 0.25	−2.65 ± 0.15
P.F1	0.26	0.02	0.26	0.26	0.0	0.35 ± 0.02	0.25 ± 0.02	0.26 ± 0.01	0.4 ± 0.04	0.33 ± 0.01
S.F1	0.25	0.02	0.25	0.25	0.0	0.34 ± 0.02	0.25 ± 0.01	0.26 ± 0.01	0.39 ± 0.03	0.32 ± 0.01
$G_{\mathcal{D}}$							28.41			

Table 23. Outcome sub rate $\in [0.4, \infty]$ high std

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	880.57	929.48	825.16	821.09	828.29	656.94 ± 10.21	804.13 ± 10.32	722.12 ± 41.74	720.14 ± 13.87	803.12 ± 7.26
$\text{RMSE}_{\text{vaso}}$	0.71	0.87	0.2	0.44	0.55	0.18 ± 0.01	0.21	0.2	0.19 ± 0.01	0.19 ± 0.01
WIS	15.08	16.43	14.47	16.43	16.43	16.43	16.43	16.19 ± 0.5	16.43	14.59 ± 0.15
WIS_b	14.88	12.73	14.52	11.31	16.4	15.21 ± 0.24	15.01 ± 0.49	14.87 ± 0.44	14.97 ± 0.53	14.65 ± 0.06
WIS_t	15.08	16.43	14.47	16.43	16.43	16.43	16.43	16.09 ± 0.51	16.43	14.59 ± 0.15
WIS_{bt}	14.82	12.12	14.63	13.15	16.75	15.07 ± 0.17	15.1 ± 0.42	14.91 ± 0.42	14.91 ± 0.14	14.66 ± 0.03
DR	−0.22	−0.22	−0.32	−0.22	−0.22	−0.01 ± 0.02	−0.22	−0.28 ± 0.09	−0.02 ± 0.03	−0.3 ± 0.14
P.F1	0.23	0.02	0.23	0.23	0.0	0.38 ± 0.02	0.16 ± 0.02	0.26 ± 0.01	0.36 ± 0.02	0.25 ± 0.01
S.F1	0.22	0.02	0.22	0.22	0.0	0.37 ± 0.01	0.15 ± 0.02	0.25 ± 0.01	0.35 ± 0.01	0.24 ± 0.01
$G_{\mathcal{D}}$							11.09			

Table 24. SOFA sub rate $\in [-\infty, -0.4]$ low std

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	788.91	880.8	756.41	774.78	749.58	608.58 ± 16.25	734.77 ± 21.49	661.93 ± 14.83	648.47 ± 8.58	729.71 ± 3.49
$\text{RMSE}_{\text{vaso}}$	0.54	0.85	0.27	0.47	0.56	0.24 ± 0.02	0.27 ± 0.01	0.26	0.27 ± 0.01	0.26 ± 0.01
WIS	1.45	4.98	-4.14	14.24	-12.51	14.24	14.59 ± 0.87	14.52 ± 0.81	14.24 ± 0.01	0.69 ± 6.55
WIS_b	0.09	4.73	-2.51	7.71	-4.69	12.38 ± 1.57	13.64 ± 0.28	14.08 ± 0.69	13.4 ± 0.67	2.25 ± 4.11
WIS_t	1.45	4.98	-4.14	14.24	-12.51	14.24	14.59 ± 0.87	14.52 ± 0.81	14.24 ± 0.01	0.69 ± 6.55
WIS_{bt}	2.14	4.55	-2.64	8.67	-0.84	12.62 ± 2.77	13.59 ± 0.31	14.1 ± 0.39	13.67 ± 0.43	2.27 ± 2.96
DR	-0.09	0.0	-0.35	0.0	-0.18	-0.0 ± 0.01	-0.04 ± 0.08	0.0	-0.01 ± 0.02	-0.21 ± 0.08
P.F1	0.25	0.02	0.25	0.25	0.0	0.34 ± 0.02	0.18 ± 0.01	0.27 ± 0.01	0.35 ± 0.01	0.27
S.F1	0.24	0.02	0.24	0.24	0.0	0.33 ± 0.02	0.17 ± 0.01	0.26 ± 0.01	0.33 ± 0.03	0.25
$G_{\mathcal{D}}$							12.68			

Table 25. SOFA sub rate $\in [-\infty, -0.4]$ high std

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	732.68	849.88	620.73	649.77	620.52	513.03 ± 17.23	604.3 ± 14.72	556.43 ± 16.69	556.7 ± 6.58	605.62 ± 1.0
$\text{RMSE}_{\text{vaso}}$	0.6	0.87	0.2	0.44	0.55	0.18 ± 0.01	0.2	0.2	0.2 ± 0.01	0.19
WIS	13.56	-10.69	13.55	-15.24	-12.41	-6.77 ± 7.5	13.1 ± 0.27	13.84 ± 0.06	8.73 ± 5.08	13.63 ± 0.18
WIS_b	13.57	-7.14	13.56	-4.9	-8.29	-2.04 ± 4.06	13.28 ± 0.15	13.87 ± 0.11	9.39 ± 4.08	13.57 ± 0.14
WIS_t	13.56	-10.69	13.55	-15.24	-12.41	-6.77 ± 7.5	13.1 ± 0.27	13.84 ± 0.06	8.73 ± 5.08	13.63 ± 0.18
WIS_{bt}	13.49	-9.88	13.52	-1.86	-7.61	-1.62 ± 3.65	13.24 ± 0.14	13.85 ± 0.09	9.11 ± 4.52	13.55 ± 0.12
DR	-0.27	-0.07	-0.7	-0.07	-0.1	-0.04 ± 0.02	-0.09 ± 0.04	-0.18 ± 0.1	-0.05 ± 0.03	-0.69 ± 0.02
P.F1	0.21	0.01	0.21	0.21	0.0	0.35 ± 0.01	0.17 ± 0.01	0.26 ± 0.03	0.33 ± 0.02	0.23
S.F1	0.21	0.01	0.21	0.21	0.0	0.34 ± 0.01	0.17 ± 0.01	0.26 ± 0.03	0.32 ± 0.02	0.23 ± 0.01
$G_{\mathcal{D}}$							10.65			

Table 26. SOFA sub rate $\in [-0.4, -0.15]$ low std

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	820.25	887.4	770.94	771.37	767.21	638.76 ± 14.54	746.47 ± 26.15	685.1 ± 16.71	675.19 ± 6.55	747.52 ± 2.79
$\text{RMSE}_{\text{vaso}}$	0.82	1.03	0.65	0.74	0.81	0.63	0.64	0.65	0.64 ± 0.01	0.64
WIS	12.0	11.61	12.0	17.13	12.88	17.09 ± 0.05	14.89 ± 1.23	12.46 ± 1.55	14.58 ± 2.22	12.0
WIS_b	11.46	10.32	11.0	15.43	13.01	15.87 ± 0.29	14.64 ± 1.03	12.22 ± 1.99	14.23 ± 1.4	11.36 ± 0.25
WIS_t	12.0	11.61	12.0	17.13	12.88	17.09 ± 0.05	14.89 ± 1.23	12.46 ± 1.55	14.58 ± 2.22	12.0
WIS_{bt}	11.38	9.96	10.53	15.78	12.97	15.74 ± 0.26	14.68 ± 0.9	12.17 ± 1.96	14.15 ± 1.45	11.35 ± 0.28
DR	-0.51	-0.39	-0.68	-0.31	-0.27	-0.3 ± 0.01	-0.17 ± 0.06	-0.21 ± 0.1	-0.3 ± 0.02	-0.68
P.F1	0.2	0.02	0.2	0.2	0.0	0.32 ± 0.01	0.17	0.26 ± 0.02	0.3 ± 0.01	0.21 ± 0.01
S.F1	0.19	0.03	0.19	0.19	0.0	0.31 ± 0.01	0.16	0.25 ± 0.02	0.29 ± 0.01	0.2
$G_{\mathcal{D}}$							10.62			

Table 27. SOFA sub rate $\in [-0.4, -0.15]$ high std

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	688.63	833.0	457.4	532.53	486.77	390.13 ± 15.74	451.62 ± 3.96	416.64 ± 11.45	428.5 ± 8.67	445.05 ± 1.22
$\text{RMSE}_{\text{vaso}}$	0.67	0.89	0.26	0.48	0.58	0.25	0.27	0.26	0.26	0.26
WIS	13.34	-9.01	13.34	-14.1	-14.12	9.74 ± 0.45	12.86 ± 0.97	13.87 ± 1.76	11.65 ± 1.61	13.34
WIS_b	7.17	-2.14	11.83	-9.45	-10.28	10.83 ± 0.49	12.21 ± 0.65	12.72 ± 2.69	11.65 ± 1.21	12.43 ± 1.26
WIS_t	13.34	-9.01	13.34	-14.1	-14.12	11.08 ± 1.27	12.86 ± 0.97	13.87 ± 1.76	11.9 ± 1.79	13.34
WIS_{bt}	7.84	-2.32	10.89	-10.12	-10.21	11.53 ± 0.44	12.24 ± 0.63	12.57 ± 3.02	11.92 ± 1.45	12.5 ± 1.33
DR	-0.46	-0.14	-0.72	-0.17	-0.15	-0.14 ± 0.02	-0.13 ± 0.04	-0.35 ± 0.19	-0.15 ± 0.03	-0.72
P.F1	0.22	0.01	0.22	0.22	0.0	0.34 ± 0.01	0.19 ± 0.01	0.28 ± 0.01	0.34	0.23 ± 0.01
S.F1	0.21	0.01	0.21	0.21	0.0	0.34 ± 0.01	0.19 ± 0.01	0.28 ± 0.01	0.33 ± 0.01	0.22 ± 0.01
$G_{\mathcal{D}}$							10.11			

Table 28. SOFA sub rate $\in [-0.15, 0]$ low std

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	765.09	867.74	669.04	696.63	662.95	551.05 ± 9.71	650.28 ± 19.61	605.84 ± 15.26	597.7 ± 6.59	652.23 ± 1.63
$\text{RMSE}_{\text{vaso}}$	0.66	0.87	0.28	0.48	0.57	0.25	0.28 ± 0.01	0.28	0.27 ± 0.01	0.27
WIS	12.06	-21.32	12.06	10.24	13.4	10.24	12.71 ± 1.65	13.03 ± 0.7	10.24	12.06
WIS_b	5.79	-13.27	11.83	8.56	10.92	7.87 ± 0.48	12.58 ± 0.82	12.66 ± 0.81	3.81 ± 1.28	11.89 ± 0.19
WIS_t	12.06	-21.32	12.06	10.24	13.4	10.24	12.71 ± 1.65	13.03 ± 0.7	10.24	12.06
WIS_{bt}	5.98	-11.51	11.98	8.09	10.1	8.19 ± 1.45	12.63 ± 1.02	12.75 ± 0.95	3.4 ± 1.44	11.99 ± 0.09
DR	-0.4	-0.19	-0.61	-0.33	-0.31	-0.34 ± 0.02	-0.11 ± 0.03	-0.29 ± 0.17	-0.38 ± 0.04	-0.58 ± 0.02
P.F1	0.2	0.02	0.2	0.2	0.0	0.33 ± 0.01	0.17	0.26 ± 0.02	0.33 ± 0.01	0.22
S.F1	0.19	0.02	0.19	0.19	0.0	0.32 ± 0.01	0.17	0.26 ± 0.02	0.32 ± 0.01	0.21
$G_{\mathcal{D}}$							9.84			

Table 29. SOFA sub rate $\in [-0.15, 0]$ high std

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	756.44	875.08	576.23	624.74	597.18	524.0 ± 20.68	567.79 ± 7.77	549.11 ± 17.14	555.3 ± 4.66	563.58 ± 0.81
$\text{RMSE}_{\text{vaso}}$	0.6	0.87	0.13	0.44	0.54	0.13	0.13 ± 0.01	0.13	0.15 ± 0.01	0.12
WIS	12.98	−7.35	12.98	12.14	12.14	12.49 ± 0.47	13.51 ± 1.17	13.61 ± 1.22	12.93 ± 0.84	12.98
WIS_b	12.67	0.77	11.6	11.92	11.57	12.06 ± 0.51	12.29 ± 0.57	12.2 ± 0.25	12.57 ± 0.75	12.09 ± 0.36
WIS_t	12.98	−7.35	12.98	12.14	12.14	12.58 ± 0.51	13.51 ± 1.17	13.61 ± 1.22	12.95 ± 0.81	12.98
WIS_{bt}	12.39	0.55	11.54	11.45	11.51	12.24 ± 0.53	12.29 ± 0.54	12.3 ± 0.43	12.6 ± 0.75	12.11 ± 0.33
DR	−0.47	−0.19	−0.86	−0.64	−0.68	−0.35 ± 0.05	−0.26 ± 0.16	−0.43 ± 0.12	−0.37 ± 0.04	−0.78 ± 0.1
P.F1	0.2	0.01	0.2	0.2	0.0	0.31 ± 0.02	0.17 ± 0.01	0.25 ± 0.01	0.31 ± 0.02	0.22
S.F1	0.19	0.01	0.19	0.19	0.0	0.31 ± 0.02	0.17 ± 0.01	0.24 ± 0.01	0.31 ± 0.01	0.21 ± 0.01
$G_{\mathcal{D}}$							9.18			

Table 30. SOFA sub rate $\in [0, 0.15]$ low std

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	760.98	834.61	628.74	637.52	628.39	512.98 ± 13.95	608.05 ± 19.71	559.49 ± 24.66	559.66 ± 7.59	609.53 ± 2.49
$\text{RMSE}_{\text{vaso}}$	0.7	0.88	0.3	0.49	0.58	0.28 ± 0.01	0.3	0.3	0.29 ± 0.01	0.29
WIS	9.59	−14.02	9.59	11.62	13.11	11.62	11.46 ± 1.78	12.59 ± 0.87	11.6 ± 0.06	10.01 ± 0.93
WIS_b	10.52	−8.8	10.56	6.07	12.14	9.9 ± 0.98	12.07 ± 1.31	12.33 ± 0.64	9.34 ± 0.92	10.58 ± 0.44
WIS_t	9.59	−14.02	9.59	11.62	13.12	11.62	11.46 ± 1.77	12.38 ± 0.52	11.57 ± 0.12	10.01 ± 0.93
WIS_{bt}	10.7	−10.69	10.54	7.03	11.17	10.31 ± 0.33	11.89 ± 1.49	12.13 ± 0.35	9.1 ± 1.35	10.56 ± 0.45
DR	−0.45	−0.33	−0.82	−0.42	−0.42	−0.36 ± 0.03	−0.49 ± 0.2	−0.5 ± 0.24	−0.37 ± 0.07	−0.82
P.F1	0.17	0.02	0.17	0.17	0.0	0.34 ± 0.01	0.15 ± 0.01	0.25 ± 0.02	0.3 ± 0.01	0.18 ± 0.01
S.F1	0.17	0.02	0.17	0.17	0.0	0.34 ± 0.01	0.15 ± 0.01	0.25 ± 0.02	0.3 ± 0.01	0.17
$G_{\mathcal{D}}$							9.04			

Table 31. SOFA sub rate $\in [0, 0.15]$ high std

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	736.69	853.28	581.25	616.67	592.53	480.48 ± 15.3	568.16 ± 11.04	520.1 ± 15.18	519.96 ± 17.85	566.24 ± 1.92
$\text{RMSE}_{\text{vaso}}$	0.6	0.87	0.24	0.46	0.57	0.22 ± 0.01	0.24	0.24	0.23 ± 0.01	0.23
WIS	−16.39	−16.4	14.4	−17.91	−14.87	11.49 ± 0.63	15.15 ± 0.28	15.27 ± 0.02	12.65 ± 0.85	14.41
WIS_b	−5.02	−5.66	14.2	−15.01	−14.46	11.46 ± 0.81	14.68 ± 0.33	14.76 ± 0.04	11.51 ± 1.8	13.99 ± 0.06
WIS_t	−16.39	−16.4	14.4	−17.91	−14.87	11.49 ± 0.63	15.15 ± 0.28	15.27 ± 0.02	12.65 ± 0.85	14.41
WIS_{bt}	−3.9	−6.52	14.14	−15.0	−13.38	11.45 ± 0.83	14.65 ± 0.36	14.76 ± 0.14	11.62 ± 1.7	14.01 ± 0.12
DR	−0.49	−0.32	−0.65	−0.61	−0.46	−0.47 ± 0.01	−0.26 ± 0.02	−0.37 ± 0.15	−0.5 ± 0.04	−0.65
P.F1	0.2	0.01	0.2	0.2	0.0	0.36 ± 0.01	0.18 ± 0.01	0.28 ± 0.02	0.35 ± 0.02	0.21 ± 0.01
S.F1	0.18	0.01	0.18	0.18	0.0	0.35 ± 0.01	0.18 ± 0.01	0.28 ± 0.01	0.33 ± 0.01	0.2 ± 0.01
$G_{\mathcal{D}}$							8.56			

Table 32. SOFA sub rate $\in [0.15, 0.4]$ low std

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	782.93	832.02	699.16	694.47	686.84	573.43 ± 13.11	670.46 ± 30.69	635.01 ± 10.67	623.91 ± 15.84	680.34 ± 2.92
$\text{RMSE}_{\text{vaso}}$	0.62	0.85	0.25	0.45	0.54	0.22 ± 0.01	0.25	0.24	0.24 ± 0.01	0.23
WIS	15.85	−13.08	7.84	−5.98	−13.33	15.43 ± 0.55	13.43 ± 1.56	15.28 ± 1.25	14.15 ± 1.2	7.84
WIS_b	11.61	−6.02	7.13	−4.04	−12.1	13.74 ± 1.13	13.43 ± 0.87	14.53 ± 0.75	11.91 ± 1.81	7.5 ± 0.79
WIS_t	15.85	−13.08	7.84	−5.98	−13.33	15.43 ± 0.55	13.43 ± 1.56	15.28 ± 1.25	14.15 ± 1.2	7.84
WIS_{bt}	10.6	−5.53	7.35	−4.23	−12.44	13.61 ± 1.0	13.5 ± 1.05	14.48 ± 0.78	11.67 ± 1.9	7.42 ± 0.21
DR	−0.94	−0.4	−0.94	−0.53	−0.56	−0.51 ± 0.01	−0.3 ± 0.12	−0.53 ± 0.14	−0.52 ± 0.04	−0.85 ± 0.09
P.F1	0.17	0.02	0.17	0.17	0.0	0.31 ± 0.01	0.14 ± 0.01	0.23 ± 0.02	0.29 ± 0.02	0.18
S.F1	0.16	0.02	0.16	0.16	0.0	0.31 ± 0.01	0.14 ± 0.01	0.23 ± 0.02	0.28 ± 0.02	0.17
$G_{\mathcal{D}}$							7.19			

Table 33. SOFA sub rate $\in [0.15, 0.4]$ high std

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	828.61	828.61	626.63	642.26	637.93	516.24 ± 10.8	601.13 ± 29.95	563.64 ± 20.54	537.16 ± 6.65	605.02 ± 7.26
$\text{RMSE}_{\text{vaso}}$	0.87	0.87	0.4	0.53	0.62	0.27 ± 0.02	0.39 ± 0.02	0.39 ± 0.02	0.34 ± 0.01	0.38 ± 0.01
WIS	14.88	−15.5	14.54	−14.88	−22.55	−6.23 ± 8.93	14.16 ± 0.21	14.62 ± 0.38	−7.97 ± 9.23	14.54
WIS_b	10.74	−15.99	12.7	−14.89	−18.7	−2.08 ± 4.19	9.72 ± 3.4	13.47 ± 0.54	−4.49 ± 1.9	13.54 ± 0.3
WIS_t	14.88	−15.5	14.54	−14.88	−22.55	−6.23 ± 8.93	14.16 ± 0.21	14.62 ± 0.38	−7.97 ± 9.23	14.54
WIS_{bt}	10.76	−15.92	13.69	−15.33	−19.14	−1.93 ± 4.64	9.96 ± 4.61	13.87 ± 0.44	−3.58 ± 3.76	14.0 ± 0.3
DR	−1.1	−0.87	−1.47	−1.45	−0.88	−1.35 ± 0.06	−1.3 ± 0.02	−1.05 ± 0.02	−1.64 ± 0.09	−1.44 ± 0.06
P.F1	0.23	0.03	0.23	0.23	0.02	0.31 ± 0.03	0.19 ± 0.02	0.26 ± 0.01	0.34 ± 0.02	0.24
S.F1	0.23	0.03	0.23	0.23	0.02	0.29 ± 0.01	0.2 ± 0.02	0.26 ± 0.01	0.32 ± 0.02	0.24 ± 0.01
$G_{\mathcal{D}}$							2.48			

Table 34. SOFA sub rate $\in [0.4, \infty]$ low std

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	1131.35	1108.49	1131.35	1090.9	1100.26	928.45 ± 15.0	1087.4 ± 30.34	996.8 ± 34.24	950.08 ± 7.98	1086.09 ± 7.96
$\text{RMSE}_{\text{vaso}}$	0.21	0.86	0.21	0.45	0.54	0.21 ± 0.01	0.21	0.2	0.21 ± 0.01	0.18 ± 0.01
WIS	12.2	0.64	11.9	−9.46	−13.74	13.48 ± 0.01	15.43 ± 0.15	13.78 ± 1.78	13.61 ± 0.29	11.99 ± 0.21
WIS_b	11.49	−0.25	9.89	−2.63	−8.99	13.13 ± 0.24	12.56 ± 3.55	11.1 ± 2.11	13.31 ± 0.4	12.32 ± 0.5
WIS_t	12.2	0.63	11.9	−11.9	−13.74	13.39 ± 0.16	15.43 ± 0.15	13.78 ± 1.78	13.71 ± 0.63	11.99 ± 0.21
WIS_{bt}	9.75	−0.63	11.76	−6.56	−11.61	13.06 ± 0.32	12.09 ± 4.65	10.44 ± 2.47	13.42 ± 0.59	11.83 ± 0.43
DR	−2.05	−1.69	−2.99	−1.65	−1.63	−1.6 ± 0.02	−1.72 ± 0.29	−1.64 ± 0.34	−1.68 ± 0.09	−2.78 ± 0.12
P.F1	0.26	0.02	0.26	0.26	0.0	0.37 ± 0.02	0.19 ± 0.1	0.29 ± 0.05	0.4 ± 0.04	0.29 ± 0.01
S.F1	0.25	0.02	0.25	0.25	0.0	0.36 ± 0.02	0.18 ± 0.1	0.28 ± 0.05	0.39 ± 0.03	0.28 ± 0.01
$G_{\mathcal{D}}$							3.92			

Table 35. SOFA sub rate $\in [0.4, \infty]$ high std

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	880.57	929.48	825.16	821.09	828.29	683.96 ± 3.08	811.24 ± 26.66	653.9 ± 6.87	706.65 ± 13.36	798.7 ± 10.64
$\text{RMSE}_{\text{vaso}}$	0.71	0.87	0.2	0.44	0.55	0.2	0.39 ± 0.09	0.18 ± 0.01	0.19 ± 0.01	0.19 ± 0.01
WIS	−1.78	−3.78	−3.04	−3.78	−3.78	−3.78	−3.32 ± 1.03	−2.12 ± 0.94	−2.8 ± 0.75	−2.92 ± 0.27
WIS_b	−2.25	−3.78	−2.93	−3.23	−3.59	−3.14 ± 0.09	−2.99 ± 0.18	−2.46 ± 0.56	−2.84 ± 0.35	−2.85 ± 0.08
WIS_t	−1.78	−3.78	−3.04	−3.78	−3.78	−3.78	−3.32 ± 1.03	−2.12 ± 0.94	−2.8 ± 0.75	−2.92 ± 0.27
WIS_{bt}	−2.14	−3.87	−2.99	−3.35	−3.61	−3.18 ± 0.08	−3.04 ± 0.56	−2.36 ± 0.66	−2.83 ± 0.37	−2.93 ± 0.27
DR	−0.49	−0.46	−1.08	−0.35	−0.29	−0.37	−0.29 ± 0.02	−0.62 ± 0.12	−0.5 ± 0.04	−0.95 ± 0.07
P.F1	0.23	0.02	0.23	0.23	0.0	0.32 ± 0.02	0.07 ± 0.04	0.37 ± 0.01	0.4 ± 0.03	0.25 ± 0.02
S.F1	0.22	0.02	0.22	0.22	0.0	0.31 ± 0.02	0.07 ± 0.04	0.36 ± 0.01	0.39 ± 0.02	0.24 ± 0.02
$G_{\mathcal{D}}$							-3.23			

Table 36. NEWS2 sub rate $\in [-\infty, -0.4]$ low std

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	732.68	849.88	620.73	649.77	620.52	528.34 ± 16.69	615.53 ± 13.8	511.36 ± 7.79	559.08 ± 5.67	604.78 ± 7.09
$\text{RMSE}_{\text{vaso}}$	0.6	0.87	0.2	0.44	0.55	0.2	0.37 ± 0.09	0.18 ± 0.01	0.21 ± 0.01	0.19
WIS	−5.08	−5.06	−5.12	−4.02	−4.08	−3.92 ± 0.07	−2.11 ± 1.17	−3.96 ± 1.46	−4.12 ± 0.18	−4.28
WIS_b	−4.77	−4.08	−4.81	−4.11	−4.3	−3.7 ± 0.25	−2.42 ± 0.54	−3.56 ± 0.94	−4.09 ± 0.48	−4.37 ± 0.05
WIS_t	−5.08	−5.06	−5.12	−4.02	−4.08	−3.92 ± 0.07	−2.11 ± 1.17	−3.96 ± 1.46	−4.12 ± 0.19	−4.28
WIS_{bt}	−4.77	−4.61	−4.87	−4.02	−4.37	−3.82 ± 0.06	−2.5 ± 0.54	−3.55 ± 1.04	−4.09 ± 0.47	−4.35 ± 0.05
DR	−0.52	−0.05	−1.55	−0.18	−0.15	−0.2	−0.12 ± 0.04	−0.64 ± 0.12	−0.4 ± 0.1	−1.55
P.F1	0.21	0.01	0.21	0.21	0.0	0.33 ± 0.01	0.07 ± 0.02	0.34 ± 0.01	0.35 ± 0.01	0.24 ± 0.01
S.F1	0.21	0.01	0.21	0.21	0.0	0.32 ± 0.01	0.07 ± 0.02	0.34 ± 0.01	0.34 ± 0.01	0.24 ± 0.01
$G_{\mathcal{D}}$							-4.13			

Table 37. NEWS2 sub rate $\in [-0.4, -0.15]$ low std

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	820.25	887.4	770.94	771.37	767.21	667.05 ± 15.28	752.73 ± 4.24	639.54 ± 5.5	668.83 ± 7.44	744.04 ± 13.4
$\text{RMSE}_{\text{vaso}}$	0.82	1.03	0.65	0.74	0.81	0.64	0.71 ± 0.04	0.64	0.63 ± 0.01	0.64
WIS	−8.81	−3.48	−8.81	−5.15	−5.28	−4.27 ± 0.18	−1.49 ± 0.75	−4.43 ± 0.73	−4.06 ± 0.32	−8.81
WIS_b	−7.33	−4.26	−7.24	−4.99	−4.57	−4.21 ± 0.2	−1.63 ± 0.44	−4.47 ± 0.56	−4.05 ± 0.3	−7.2 ± 0.41
WIS_t	−8.81	−3.48	−8.81	−5.15	−5.28	−4.27 ± 0.18	−1.49 ± 0.75	−4.43 ± 0.73	−4.06 ± 0.32	−8.81
WIS_{bt}	−7.12	−4.22	−7.45	−5.01	−4.81	−4.2 ± 0.22	−1.61 ± 0.4	−4.52 ± 0.5	−4.07 ± 0.3	−7.28 ± 0.3
DR	−0.64	−0.38	−1.46	−0.34	−0.19	−0.36	−0.09 ± 0.02	−0.61 ± 0.03	−0.47 ± 0.03	−1.33 ± 0.09
P.F1	0.2	0.02	0.2	0.2	0.0	0.28 ± 0.01	0.05 ± 0.02	0.3 ± 0.01	0.31 ± 0.02	0.22 ± 0.01
S.F1	0.19	0.03	0.19	0.19	0.0	0.28 ± 0.01	0.05 ± 0.02	0.3 ± 0.01	0.31 ± 0.02	0.21 ± 0.01
$G_{\mathcal{D}}$							-4.49			

Table 38. NEWS2 sub rate $\in [-0.4, -0.15]$ high std

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	688.63	833.0	457.4	532.53	486.77	405.15 ± 3.15	463.11 ± 10.85	398.49 ± 8.64	444.65 ± 15.04	443.3 ± 6.13
$\text{RMSE}_{\text{vaso}}$	0.67	0.89	0.26	0.48	0.58	0.26	0.41 ± 0.09	0.25 ± 0.01	0.26	0.25 ± 0.01
WIS	−4.56	−4.6	−4.56	−5.59	−5.59	−5.12 ± 1.03	−2.31 ± 1.41	−3.61 ± 1.39	−4.15 ± 0.58	−4.56
WIS_b	−4.65	−4.28	−4.69	−5.11	−5.31	−4.33 ± 0.81	−2.57 ± 1.1	−3.68 ± 0.8	−4.12 ± 0.57	−4.58 ± 0.08
WIS_t	−4.56	−4.6	−4.56	−5.59	−5.59	−3.65 ± 0.37	−2.31 ± 1.41	−3.61 ± 1.39	−4.16 ± 0.57	−4.56
WIS_{bt}	−4.53	−4.36	−4.68	−5.12	−5.32	−3.56 ± 0.22	−2.55 ± 0.95	−3.62 ± 0.82	−4.13 ± 0.57	−4.57 ± 0.11
DR	−0.31	−0.1	−1.65	−0.13	−0.08	−0.15	−0.11 ± 0.05	−0.77 ± 0.1	−0.33 ± 0.08	−1.61 ± 0.03
P.F1	0.22	0.01	0.22	0.22	0.0	0.32	0.07 ± 0.02	0.35 ± 0.01	0.35 ± 0.02	0.25 ± 0.02
S.F1	0.21	0.01	0.21	0.21	0.0	0.32	0.07 ± 0.02	0.34 ± 0.01	0.34 ± 0.03	0.24 ± 0.02
$G_{\mathcal{D}}$							-4.3			

Table 39. NEWS2 sub rate $\in [-0.15, 0]$ low std

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	765.09	867.74	669.04	696.63	662.95	577.93 ± 4.79	661.56 ± 8.96	555.14 ± 10.07	595.76 ± 7.32	649.09 ± 9.58
$\text{RMSE}_{\text{vaso}}$	0.66	0.87	0.28	0.48	0.57	0.26 ± 0.01	0.4 ± 0.08	0.25	0.27 ± 0.01	0.27 ± 0.01
WIS	-4.15	-8.69	-4.15	-5.83	-3.16	-5.83	-4.94 ± 1.2	-3.13 ± 1.14	-5.83	-4.15
WIS_b	-4.81	-7.2	-4.56	-5.53	-4.13	-5.66 ± 0.43	-4.4 ± 0.17	-3.42 ± 0.85	-6.18 ± 0.32	-4.49 ± 0.09
WIS_t	-4.15	-8.69	-4.15	-5.83	-3.16	-5.83	-4.94 ± 1.2	-3.14 ± 1.17	-5.83	-4.15
WIS_{bt}	-4.68	-6.85	-4.55	-5.39	-4.61	-5.62 ± 0.24	-4.41 ± 0.26	-3.39 ± 0.93	-6.23 ± 0.26	-4.53 ± 0.06
DR	-0.6	-0.24	-1.58	-0.43	-0.37	-0.44	-0.14 ± 0.05	-0.7 ± 0.11	-0.53 ± 0.04	-1.58
P.F1	0.2	0.02	0.2	0.2	0.0	0.31 ± 0.01	0.05 ± 0.02	0.32 ± 0.02	0.33 ± 0.02	0.23 ± 0.01
S.F1	0.19	0.02	0.19	0.19	0.0	0.31 ± 0.01	0.05 ± 0.02	0.32 ± 0.02	0.32 ± 0.02	0.22 ± 0.01
$G_{\mathcal{D}}$							-4.72			

Table 40. NEWS2 sub rate $\in [-0.15, 0]$ high std

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	756.44	875.08	576.23	624.74	597.18	539.36 ± 6.94	580.52 ± 10.06	525.81 ± 17.46	568.77 ± 18.09	562.86 ± 4.54
$\text{RMSE}_{\text{vaso}}$	0.6	0.87	0.13	0.44	0.54	0.13	0.36 ± 0.11	0.12 ± 0.03	0.16 ± 0.01	0.12
WIS	-3.82	-4.69	-3.82	-5.01	-5.01	-3.19 ± 0.24	-2.14 ± 1.5	-2.58 ± 0.89	-2.97 ± 0.4	-3.8 ± 0.03
WIS_b	-4.25	-4.61	-4.32	-4.24	-4.2	-3.19 ± 0.17	-2.49 ± 0.99	-3.54 ± 0.72	-3.1 ± 0.45	-3.99 ± 0.24
WIS_t	-3.82	-4.69	-3.82	-5.01	-5.01	-3.3 ± 0.21	-2.14 ± 1.5	-2.58 ± 0.89	-3.05 ± 0.5	-3.8 ± 0.03
WIS_{bt}	-4.34	-4.29	-4.28	-4.33	-4.15	-3.27 ± 0.14	-2.48 ± 0.92	-3.62 ± 0.63	-3.19 ± 0.34	-4.05 ± 0.26
DR	-0.39	-0.04	-1.6	-0.39	-0.35	-0.4 ± 0.01	-0.11 ± 0.07	-0.69 ± 0.17	-0.51 ± 0.05	-1.6
P.F1	0.2	0.01	0.2	0.2	0.0	0.29 ± 0.01	0.07 ± 0.03	0.31 ± 0.02	0.32 ± 0.01	0.23 ± 0.01
S.F1	0.19	0.01	0.19	0.19	0.0	0.29 ± 0.02	0.07 ± 0.03	0.31 ± 0.01	0.32 ± 0.01	0.23 ± 0.01
$G_{\mathcal{D}}$							-4.41			

Table 41. NEWS2 sub rate $\in [0, 0.15]$ low std

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	760.98	834.61	628.74	637.52	628.39	531.23 ± 12.54	621.36 ± 11.52	524.17 ± 15.77	563.13 ± 7.26	606.27 ± 6.74
$\text{RMSE}_{\text{vaso}}$	0.7	0.88	0.3	0.49	0.58	0.29	0.42 ± 0.07	0.28 ± 0.01	0.3 ± 0.01	0.29 ± 0.01
WIS	-4.93	-3.75	-4.93	-5.36	-4.92	-5.36	-3.69 ± 1.36	-2.92 ± 1.32	-4.9 ± 1.19	-4.93
WIS_b	-5.03	-3.98	-4.75	-5.81	-5.04	-4.77 ± 0.17	-3.35 ± 0.93	-2.96 ± 0.27	-5.24 ± 0.64	-4.8 ± 0.11
WIS_t	-4.93	-3.75	-4.93	-5.36	-4.92	-5.36	-3.52 ± 1.17	-2.92 ± 1.32	-4.9 ± 1.19	-4.93
WIS_{bt}	-5.15	-4.11	-4.84	-5.52	-4.99	-4.79 ± 0.13	-3.3 ± 0.97	-2.94 ± 0.26	-5.16 ± 0.61	-4.82 ± 0.04
DR	-0.63	-0.22	-1.48	-0.55	-0.46	-0.57	-0.14 ± 0.07	-0.81 ± 0.17	-0.62 ± 0.04	-1.37 ± 0.09
P.F1	0.17	0.02	0.17	0.17	0.0	0.31 ± 0.01	0.05 ± 0.01	0.32 ± 0.01	0.34 ± 0.01	0.19 ± 0.01
S.F1	0.17	0.02	0.17	0.17	0.0	0.3 ± 0.01	0.05 ± 0.01	0.32 ± 0.01	0.33 ± 0.01	0.19 ± 0.01
$G_{\mathcal{D}}$							-4.84			

Table 42. NEWS2 sub rate $\in [0, 0.15]$ high std

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	736.69	853.28	581.25	616.67	592.53	497.22 ± 10.48	577.53 ± 9.49	490.24 ± 12.7	528.45 ± 21.15	564.83 ± 8.02
$\text{RMSE}_{\text{vaso}}$	0.6	0.87	0.24	0.46	0.57	0.23	0.4 ± 0.08	0.22	0.23 ± 0.01	0.23
WIS	-4.65	-4.65	-4.4	-6.55	-7.49	-3.98 ± 0.21	-2.61 ± 1.5	-3.43 ± 0.86	-3.99 ± 0.42	-4.4
WIS_b	-4.16	-5.16	-4.14	-6.48	-6.91	-3.88 ± 0.19	-2.71 ± 0.69	-3.41 ± 0.58	-3.93 ± 0.43	-4.0 ± 0.08
WIS_t	-4.65	-4.65	-4.4	-6.55	-7.49	-3.98 ± 0.21	-2.61 ± 1.5	-3.43 ± 0.86	-3.99 ± 0.42	-4.4
WIS_{bt}	-4.09	-5.07	-4.14	-6.38	-6.85	-3.87 ± 0.19	-2.83 ± 0.82	-3.4 ± 0.64	-3.92 ± 0.44	-4.03 ± 0.13
DR	-0.7	-0.21	-1.69	-0.29	-0.39	-0.3 ± 0.01	-0.1 ± 0.07	-0.81 ± 0.13	-0.64 ± 0.08	-1.61 ± 0.06
P.F1	0.2	0.01	0.2	0.2	0.0	0.33 ± 0.02	0.06 ± 0.02	0.36 ± 0.01	0.35 ± 0.01	0.23 ± 0.01
S.F1	0.18	0.01	0.18	0.18	0.0	0.32 ± 0.01	0.06 ± 0.02	0.34 ± 0.01	0.33 ± 0.02	0.22 ± 0.01
$G_{\mathcal{D}}$							-4.6			

Table 43. NEWS2 sub rate $\in [0.15, 0.4]$ low std

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	782.93	832.02	699.16	694.47	686.84	594.05 ± 9.74	685.2 ± 9.89	584.59 ± 7.08	620.18 ± 15.98	675.9 ± 11.12
$\text{RMSE}_{\text{vaso}}$	0.62	0.85	0.25	0.45	0.54	0.24 ± 0.01	0.37 ± 0.07	0.22 ± 0.01	0.25 ± 0.01	0.23
WIS	-1.94	-6.0	-4.4	-6.75	-7.23	-5.26 ± 0.11	-1.58 ± 0.93	-2.69 ± 1.68	-5.13 ± 0.3	-4.4
WIS_b	-3.26	-5.35	-4.73	-6.34	-6.53	-4.68 ± 0.34	-1.84 ± 0.29	-3.09 ± 1.3	-4.75 ± 0.47	-4.52 ± 0.1
WIS_t	-1.94	-6.0	-4.4	-6.75	-7.23	-5.26 ± 0.11	-1.58 ± 0.93	-2.69 ± 1.68	-5.13 ± 0.3	-4.4
WIS_{bt}	-3.65	-5.35	-4.64	-6.34	-6.71	-4.82 ± 0.22	-1.87 ± 0.32	-3.08 ± 1.27	-4.79 ± 0.43	-4.57 ± 0.04
DR	-0.74	-0.22	-1.54	-0.47	-0.51	-0.48 ± 0.01	-0.22 ± 0.12	-0.85 ± 0.11	-0.64 ± 0.08	-1.36 ± 0.05
P.F1	0.17	0.02	0.17	0.17	0.0	0.28 ± 0.01	0.05 ± 0.01	0.3 ± 0.01	0.32 ± 0.01	0.19 ± 0.01
S.F1	0.16	0.02	0.16	0.16	0.0	0.28 ± 0.01	0.05 ± 0.01	0.29 ± 0.01	0.31 ± 0.01	0.18 ± 0.01
$G_{\mathcal{D}}$							-4.7			

Table 44. NEWS2 sub rate $\in [0.15, 0.4]$ high std

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	828.61	828.61	626.63	642.26	637.93	553.35 ± 10.35	602.95 ± 21.97	514.44 ± 7.44	530.38 ± 16.37	603.67 ± 9.18
$\text{RMSE}_{\text{vaso}}$	0.87	0.87	0.4	0.53	0.62	0.28 ± 0.02	0.5 ± 0.07	0.29 ± 0.01	0.35 ± 0.02	0.38 ± 0.02
WIS	−1.5	−4.3	−1.65	−3.82	−6.8	−3.22 ± 0.16	−3.84 ± 0.9	−1.65	−3.2 ± 0.27	−1.65
WIS_b	−2.35	−4.58	−2.69	−4.0	−5.43	−3.03 ± 0.18	−3.89 ± 0.57	−1.82 ± 0.07	−3.0 ± 0.29	−2.13 ± 0.31
WIS_t	−1.5	−4.3	−1.65	−3.82	−6.8	−3.22 ± 0.16	−3.84 ± 0.9	−1.65	−3.2 ± 0.27	−1.65
WIS_{bt}	−2.36	−4.56	−2.7	−4.16	−5.66	−3.02 ± 0.15	−3.9 ± 0.59	−1.79 ± 0.09	−2.99 ± 0.25	−2.06 ± 0.28
DR	−0.54	−0.36	−1.13	−0.45	−0.54	−0.47 ± 0.02	−0.08 ± 0.02	−0.61 ± 0.33	−0.78 ± 0.14	−1.03 ± 0.13
P.F1	0.23	0.03	0.23	0.23	0.02	0.31 ± 0.01	0.04 ± 0.01	0.33 ± 0.04	0.37 ± 0.02	0.26 ± 0.01
S.F1	0.23	0.03	0.23	0.23	0.02	0.3 ± 0.01	0.04 ± 0.02	0.31 ± 0.03	0.34 ± 0.02	0.25 ± 0.01
$G_{\mathcal{D}}$							-4.25			

Table 45. NEWS2 sub rate $\in [0.4, \infty]$ low std

metric	alt	max	min	random	weight	imitation	dqn	cql	iql	bcq
RMSE_{IV}	1131.35	1108.49	1131.35	1090.9	1100.26	959.1 ± 10.09	1095.13 ± 12.04	932.83 ± 10.43	943.02 ± 3.75	1092.54 ± 18.52
$\text{RMSE}_{\text{vaso}}$	0.21	0.86	0.21	0.45	0.54	0.21 ± 0.01	0.38 ± 0.07	0.19 ± 0.02	0.21 ± 0.01	0.19 ± 0.01
WIS	−2.72	−2.56	−2.89	−5.71	−4.33	−1.53	−1.39 ± 0.52	−1.66 ± 0.66	−1.54 ± 0.01	−2.68 ± 0.11
WIS_b	−2.32	−2.74	−2.66	−4.26	−4.27	−1.87 ± 0.06	−1.71 ± 0.33	−1.76 ± 0.14	−1.88 ± 0.1	−2.33 ± 0.09
WIS_t	−2.72	−2.56	−2.89	−6.16	−4.33	−1.61 ± 0.04	−1.54 ± 0.66	−1.67 ± 0.66	−1.84 ± 0.21	−2.68 ± 0.11
WIS_{bt}	−2.64	−2.78	−2.57	−4.81	−4.34	−1.93 ± 0.08	−1.79 ± 0.41	−1.83 ± 0.22	−2.06 ± 0.09	−2.4 ± 0.12
DR	−0.78	−0.2	−1.97	−0.96	−0.47	−0.95 ± 0.01	−0.13 ± 0.09	−0.57 ± 0.18	−0.82 ± 0.04	−1.89 ± 0.07
P.F1	0.26	0.02	0.26	0.26	0.0	0.35 ± 0.02	0.04 ± 0.02	0.38 ± 0.02	0.4 ± 0.04	0.31 ± 0.03
S.F1	0.25	0.02	0.25	0.25	0.0	0.34 ± 0.03	0.04 ± 0.04	0.37 ± 0.02	0.39 ± 0.03	0.3 ± 0.03
$G_{\mathcal{D}}$							-3.79			

Table 46. NEWS2 sub rate $\in [0.4, \infty]$ high std

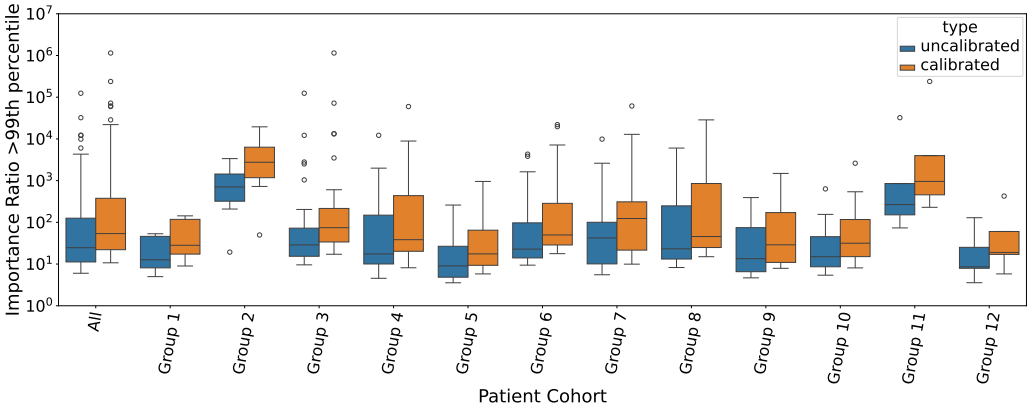


Figure 10. **Importance ratio histogram of random policy > 99th percentile on Outcome reward.**

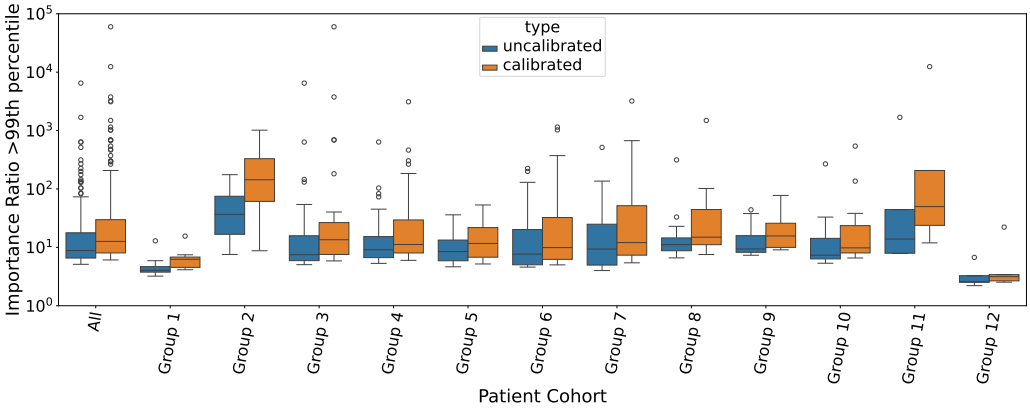


Figure 11. **Importance ratio histogram of min policy > 99th percentile on Outcome reward.**

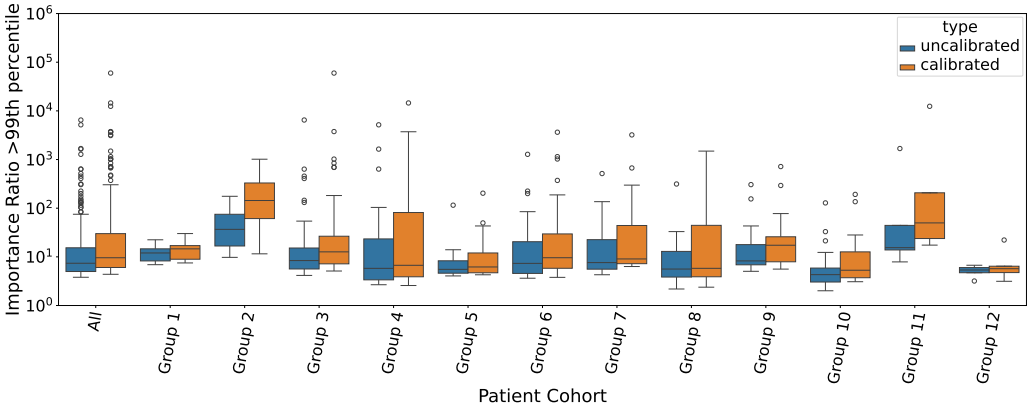


Figure 12. **Importance ratio histogram of max policy > 99th percentile on Outcome reward.**

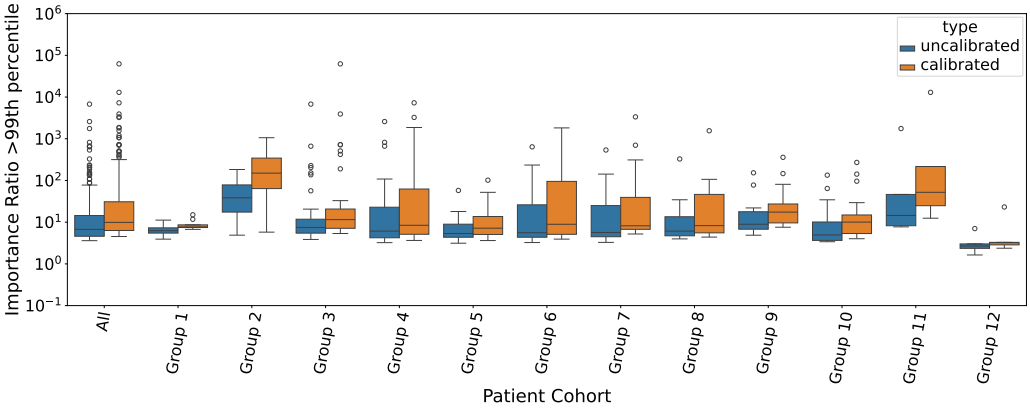


Figure 13. **Importance ratio histogram of alt policy > 99th percentile on Outcome reward.**

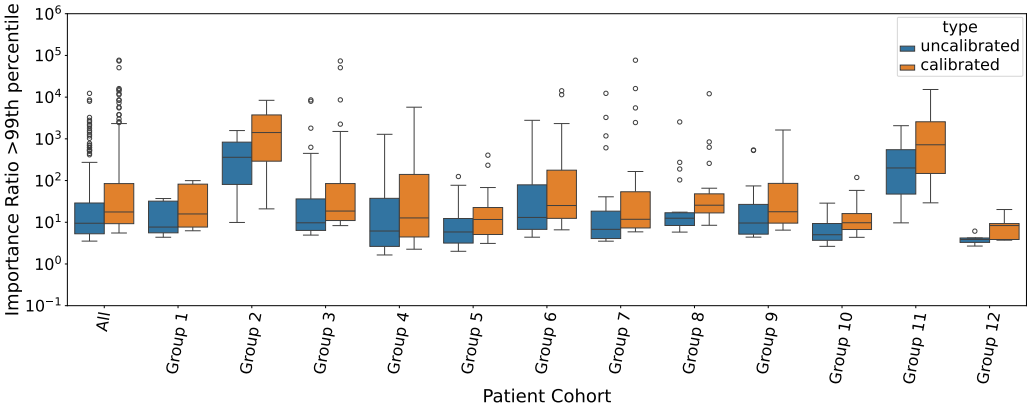


Figure 14. **Importance ratio histogram of weight policy > 99th percentile on Outcome reward.**

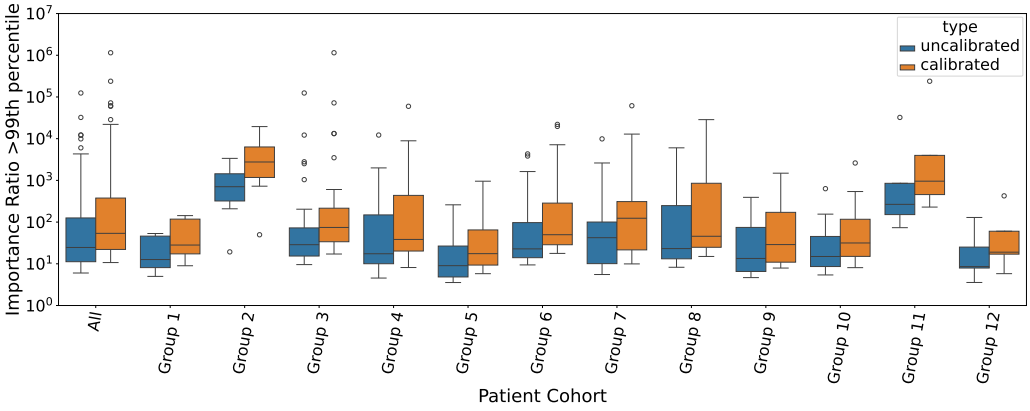


Figure 15. **Importance ratio histogram of random policy > 99th percentile on SOFA reward.**

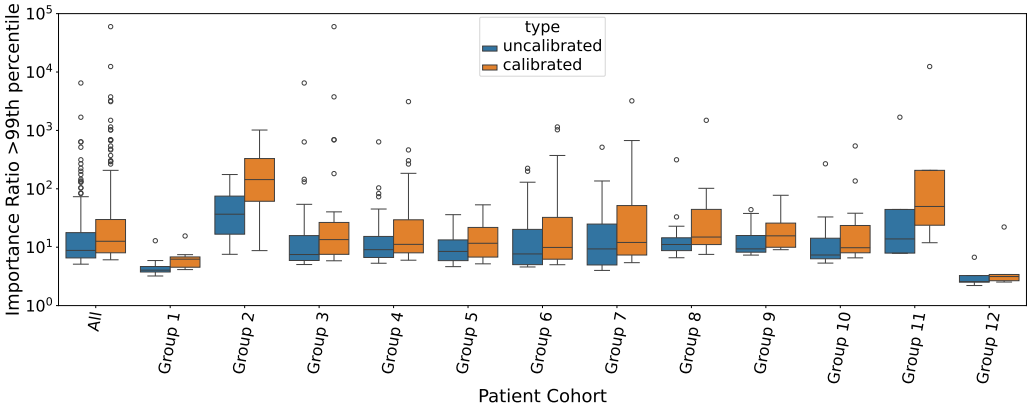


Figure 16. **Importance ratio histogram of min policy > 99th percentile on SOFA reward.**

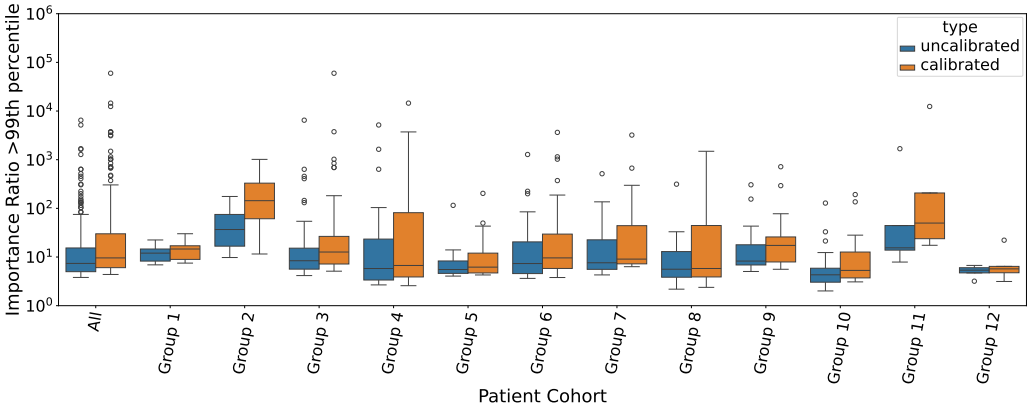


Figure 17. **Importance ratio histogram of max policy > 99th percentile on SOFA reward.**

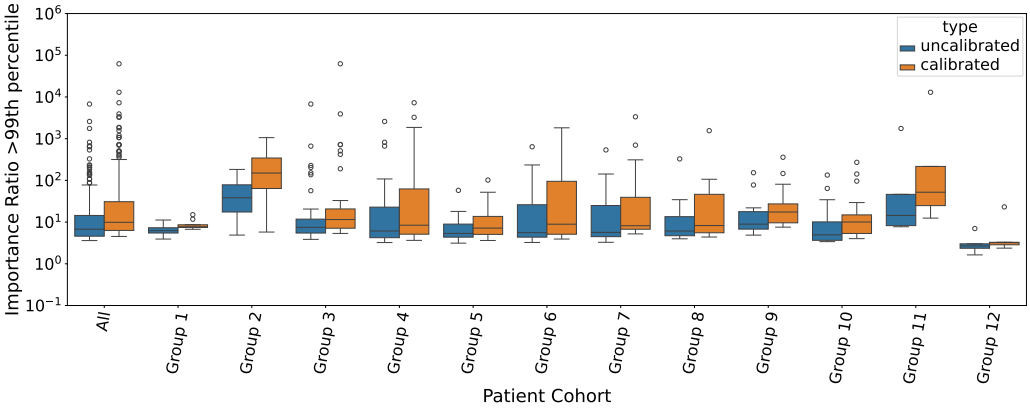


Figure 18. **Importance ratio histogram of alt policy > 99th percentile on SOFA reward.**

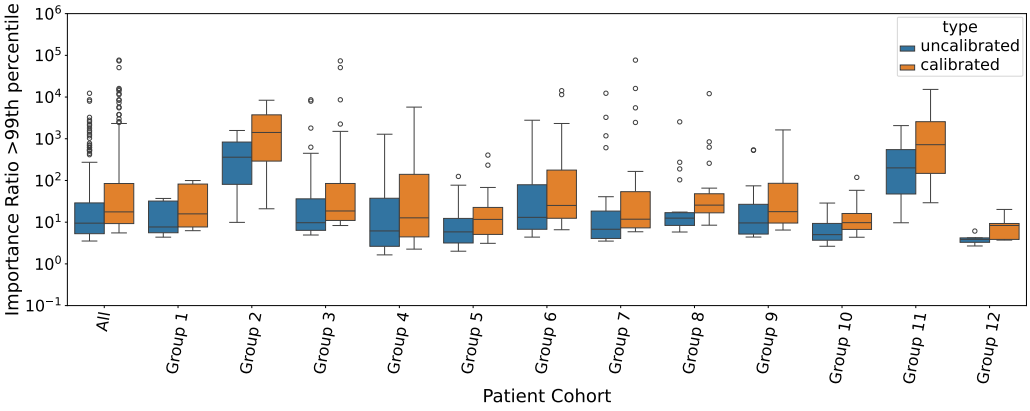


Figure 19. **Importance ratio histogram of weight policy > 99th percentile on SOFA reward.**

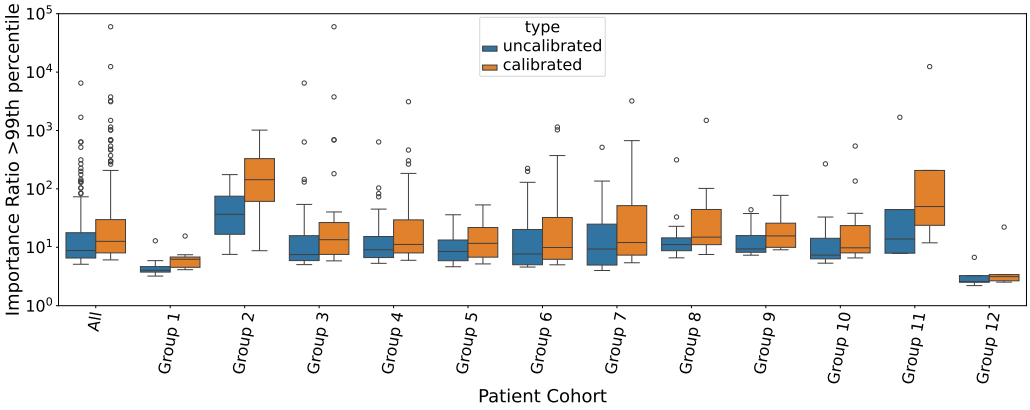


Figure 20. **Importance ratio histogram of min policy > 99th percentile on NEWS2 reward.**

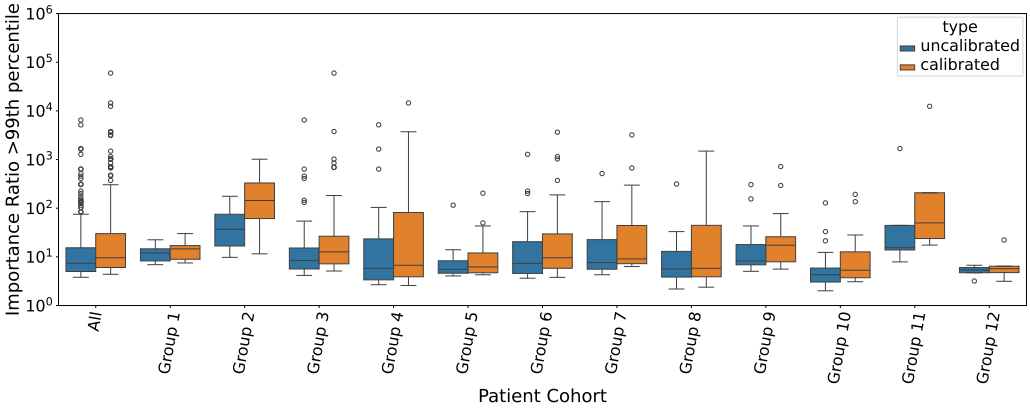


Figure 21. **Importance ratio histogram of max policy > 99th percentile on NEWS2 reward.**

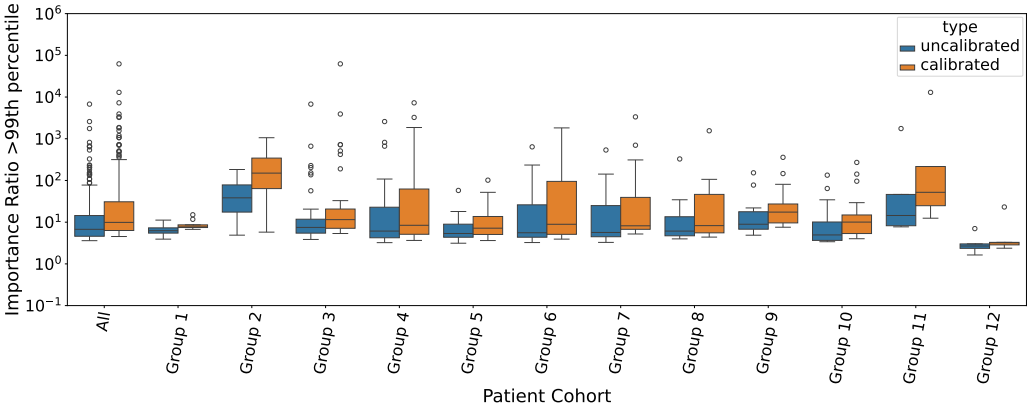


Figure 22. **Importance ratio histogram of alt policy > 99th percentile on NEWS2 reward.**

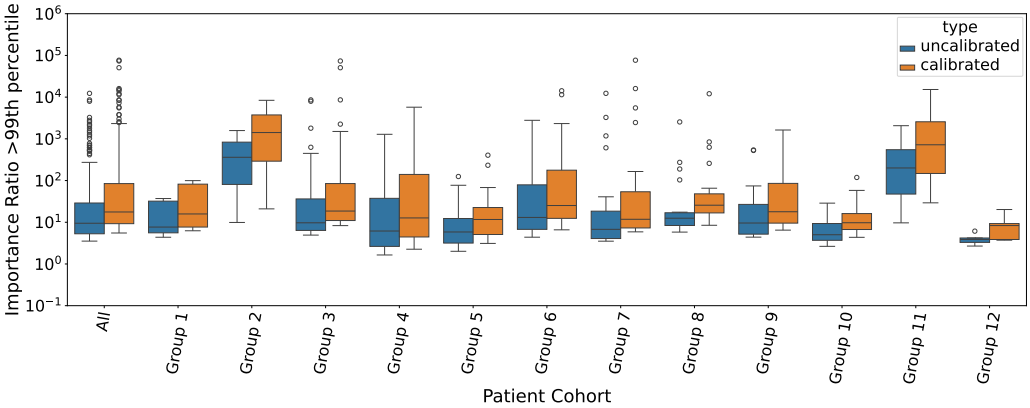


Figure 23. **Importance ratio histogram of weight policy > 99th percentile on NEWS2 reward.**