Title

Evaluating Offline Reinforcement Learning for Vasopressor Management in the ICU

ISBN

9798244888751

Author

Chun, Allen

Publication Date

2026-05-13

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Evaluating Offline Reinforcement Learning for Vasopressor Management in the ICU

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Applied Statistics and Data Science

by

Allen Chun

2026

ABSTRACT OF THE THESIS

Evaluating Offline Reinforcement Learning
for Vasopressor Management in the ICU

by

Allen Chun
Master of Applied Statistics and Data Science
University of California, Los Angeles, 2026
Professor Yuhua Zhu, Chair

This thesis presents an analysis of offline reinforcement learning (RL) methods for vasopressor management in the intensive care unit (ICU) using retrospective clinical data. Vasopressor administration is defined as a sequential decision-making problem, and the problem is formulated as a finite-horizon Markov Decision Process (MDP). The research examines three different approaches: Behavior Cloning (BC), Conservative Q-Learning (CQL), and Batch-Constrained Q-Learning (BCQ).

Policy behavior is evaluated using mean arterial pressure (MAP). All three techniques yield clinically plausible policies that increase the usage of vasopressors in the presence of lower MAP values and reduce their usage in the presence of higher MAP values. BC approximates clinician behavior fairly accurately, CQL offers comparable performance but with more regularization to prevent overestimation, and BCQ yields systematic and conservative divergences. In addition to this, the study explores off-policy evaluation through the application of Weighted Importance Sampling (WIS) and Fitted Q-Evaluation (FQE), which further reveal substantial variability and overlapping confidence intervals across methods. These findings suggest that, while empirical evaluation remains challenging, offline reinforcement learning provides a solid foundation for future exploration in clinical decision support systems.

The thesis of Allen Chun is approved.

Vivian Lew

Rick Schoenberg

Yuhua Zhu, Committee Chair

University of California, Los Angeles

2026

To my parents, who instilled in me the confidence to be great.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

Clinical decision-making processes in the intensive care unit (ICU) involve various treatment decisions made under conditions of uncertainty, time pressure, and incomplete information. Of these, the administration of vasopressors to maintain adequate blood pressure levels is one of the more common and critical decisions. These decisions have to be made while weighing the dangers of hypotension against the potential risks of excessive administration of these drugs.

While vasopressors play a critical role in the ICU, recommendations for dosing decisions are primarily heuristic and context-dependent. Clinical protocols give general instructions, such as maintaining the mean arterial pressure (hereafter MAP) above certain levels, but ultimately, the entire process is influenced by a variety of factors, including the specific characteristics of the patient, clinical protocols, and the judgment of the decision-maker. These processes make vasopressor management highly appropriate for the development of data-driven models, as the decision process is sequential.

1.1 Vasopressor Management as a Sequential Decision Problem

The administration of vasopressors is inherently sequential in nature. This is because decisions made at a given time point directly influence future physiological states, which in turn affect other decisions made at a future time point. This dependency on time is unique in the problem of vasopressor administration and differentiates it from static prediction problems.

Electronically recorded health data collected in the ICU setting contain comprehensive information on the physiology of patients and the decisions made by clinicians. This data

presents an opportunity to examine the process of treatment decision-making and evaluate the potential to derive structured policies based on past practice.

1.2 Reinforcement Learning in Healthcare

Reinforcement learning (RL) is a machine learning method that provides the formal framework for sequential decision-making [1]. It typically involves an "agent" that interacts with an environment through states, actions, and rewards in order to maximize value or develop a best policy. In the past few years, reinforcement learning has seen significant adoption to model a variety of healthcare problems.

However, in the healthcare environment, offline reinforcement learning, a method where policies are learned from static retrospective data, has emerged as a new alternative [2]. In contrast to the standard online RL, offline learning algorithms must address the difficulty that the agent cannot interact to learn new actions or rectify errors. In critical care environments, actions that are not supported may have catastrophic consequences. As a result, it is essential to design algorithms in a conservative manner and interpret outcomes carefully.

1.3 Limitations of Existing Approaches

Conventional supervised learning approaches in healthcare often shape treatment plans as a classification or regression problem in which the algorithm is designed to replicate the clinician's decisions. These algorithms may learn the high-level practice patterns, but the problem is that they do not precisely model the long-term consequences of actions.

On the other hand, unconstrained reinforcement learning algorithms may extrapolate when learning from historic data, assigning high values to actions that are poorly represented or even absent from the original data. In healthcare environments, this can render the models unsuitable.

These limitations highlight the need for further research into conservative offline reinforcement learning algorithms that can differentiate between actions without losing sight of

standard clinical practice.

1.4 Objectives and Research Questions

The goal of this thesis is not to develop or utilize a treatment policy but rather to analyze whether conservative offline reinforcement learning methods can recover clinically credible and interpretable decisions from retrospective ICU data.

This project aims to address the following research questions:

1. Can offline reinforcement learning methods learn vasopressor dosing policies that respond consistently to patient physiology (using MAP)?
2. What makes conservative value-based and support-constrained algorithms different from direct imitation of clinician behavior?
3. How reliable are off-policy evaluation methods in determining estimates of policy value in this setting?

With these questions, the focus is aimed at policy behavior and evaluation accuracy rather than claims of clinical optimality or improved patient outcomes.

1.5 Thesis Organization

The thesis is organized into eight separate chapters. Following the introduction, Chapter 2 looks at prior work done in offline reinforcement learning and healthcare. Chapters 3 and 4 cover the data and problem formulation. The algorithms and evaluation methods are defined in Chapter 5, and the results are explained in Chapter 6. The work wraps up with a discussion of implications in Chapter 7 and final conclusions in Chapter 8.

CHAPTER 2

Background and Related Work

This chapter provides a clinical and methodological background for studying decision-making in vasopressor administration using offline reinforcement learning. We first discuss how vasopressors and MAP are used in critical care settings. Then we will introduce concepts from reinforcement learning applicable to sequential decision making, followed by challenges specific to applying offline reinforcement learning techniques in healthcare settings. Finally, we will review previous work that has applied similar reinforcement learning to problems of ICU treatment.

2.1 Vasopressors and Hemodynamic Management in Critical Care

Vasopressors are a class of medications used to decrease vascular dilation and maintain sufficient blood pressure in patients who have developed hypotension due to conditions such as septic shock, cardiac shock, or postoperative stability. They are commonly used in the ICU to treat hypoxia associated with these conditions. Amongst all available vasopressors, norepinephrine is widely used as a first-line agent because it possesses a favorable balance between its effect on reducing vascular resistance and safety to cardiovascular function [3, 4].

Mean arterial pressure (MAP) is one of the primary goals of vasopressor therapy and reflects the average pressure in the arteries during a cardiac cycle. It serves as an indicator of perfusion of organs. Usually, clinical guidelines recommend keeping MAP above a threshold value (e.g., 65 mmHg) to minimize the risk of organ dysfunction [3, 4]. However, optimal values of MAP may vary across patients based on comorbidities, disease severity, and individual physiological response.

In practice, dosage of vasopressors needs to be adjusted frequently throughout the course of an ICU stay. Clinicians change infusion rate based on current vital signs, lab results, and overall progress of the patient. Clinicians also decide on vasopressor dose based on local protocols and factors unique to each patient [3]. This uniformity of observed treatment algorithm makes management of vasopressor a complex but interesting setting for studying sequential decision making.

2.2 Sequential Decision-Making and Reinforcement Learning

The Reinforcement Learning (RL) framework provides a theoretical basis for developing models to describe the process of making sequential decisions when the outcome of the actions taken will affect what states and rewards occur subsequently. An agent in the RL framework observes a state, selects an action, receives a reward and then interacts with an environment. At some point in time the agent learns how to develop a policy that will maximize the expected total amount of reward received.

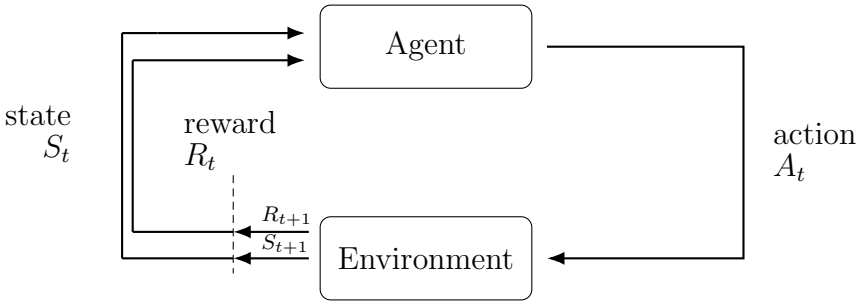

Figure 2.1: Reinforcement learning interaction loop between an agent and its environment.

A typical mathematical model used to represent the environment for RL applications is the Markov Decision Process (MDP) [1]. The MDP defines the environment in terms of states, possible actions, transition probabilities and rewards. The MDP is suitable for modeling clinical decision-making processes when there is a delay in knowing the impact of treatment selections and when those impacts depend on the current state.

RL frameworks differ from supervised learning frameworks because they take into consid-

eration the effect of all of the past actions selected while determining the optimal subsequent action. Thus, unlike supervised learning, RL does not treat each decision independently. This difference is significant when considering health-care environments, where short term decisions can lead to long term patient trajectories.

2.3 Offline Reinforcement Learning and Extrapolation Error

There are several reasons why agents cannot interact directly with their environment in many real world domains, whether it be due to cost or ethical concerns among many others. In offline reinforcement learning, instead of interacting with the environment, agents must learn policies based upon offline data collected during previous interactions with the environment [2].

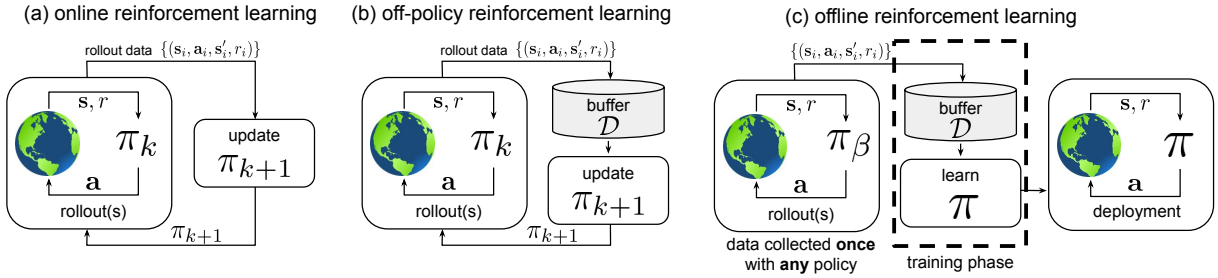

Figure 2.2: Illustration of offline reinforcement learning: data are collected once using a behavior policy, stored in an offline dataset, and used during a training phase to learn a policy for later deployment.

The primary challenge when using offline data to train policies is avoiding extrapolation error; i.e., estimating values for state-action combinations that do not appear frequently in the training data. Most standard RL methods applied to offline data tend to estimate high values for actions that were never observed in the training data because these methods rely on function approximation and bootstrap sampling. When applying these methods to health-care decision-making domains, this type of estimation can potentially identify unsafe or implausible treatments.

To reduce the risk of extrapolation error, researchers have proposed several conservative

methods for offline RL [2, 5, 7]. Conservative methods attempt to limit the scope of the policies learned to be within the bounds of the data collected by imposing penalties on actions that are not supported by observations or by limiting the set of allowable actions. In both cases, conservatism is achieved at the expense of expressiveness.

2.4 Reinforcement Learning for Clinical Decision Support

Reinforcement learning has been successfully applied to numerous health-care related problems including treatment selection, dosing regimens and resource allocation [2, 7]. Prior studies have demonstrated successful application of RL techniques to manage various aspects of critically ill patients including sepsis management, mechanical ventilation therapy, fluid resuscitation, and vasopressor therapy.

While these studies illustrate the capability of RL to model temporally dependent relationships in clinical data, they have also illustrated challenges associated with developing effective RL systems in health-care domains [2, 7]. Challenges include sensitivity to reward definition, instability of off-policy evaluations, and difficulties in interpreting learned policies. Additionally, concerns regarding extrapolation, confounding variables, and reliable off-policy evaluation have led to increased focus on developing conservative learning paradigms along with thorough analysis of learned policies.

Recent work has emphasized understanding policy behavior, interpretability, and robustness of learned policies when limited amounts of data are available for training [6]. This change in perspective is consistent with the emerging acknowledgment that retrospective RL analyses should emphasize safety and transparency over performance claims.

2.5 Positioning of This Work

This study uses offline reinforcement learning to analyze how different types of conservative learning methods would generate clinically plausible and interpretable policies for vasopressor use in ICUs when trained with retrospectively collected data. Unlike prior work that focused

on developing new algorithms or claiming clinical optimality for proposed policies, this study focuses on comparing various conservative learning methods through comparative behavioral analysis and off-policy evaluation. By comparing two conservative learning methods under identical conditions, this study continues an ongoing trend of viewing offline RL as a method for exploratory decision support analysis rather than a means to develop autonomous clinical controls.

CHAPTER 3

Data and Cohort Construction

Table 3.1: Cohort characteristics and dataset summary

Characteristic	**Value**
Number of ICU stays	54,551
Number of patients	54,551
Mean age (years)	64.1 ± 16.5
Female (%)	42.8
ICU length of stay (hours)	57 [37–107]
Total decision points (hourly bins)	5,597,201
ICU stays receiving vasopressors (%)	18.0
Mean arterial pressure (mmHg)	82.4 ± 15.7
In-hospital mortality (%)	10.8

Table 3.1 summarizes key characteristics of the study cohort and the resulting dataset structure. All statistics are reported descriptively and are intended to contextualize the scale and composition of the learning problem rather than to support causal inference.

3.1 Data Source

This study utilizes data from the Medical Information Mart for Intensive Care (MIMIC-IV v3.1) database, a public critical care dataset containing anonymized medical records from patients who have been admitted to the Beth Israel Deaconess Medical Center in Boston, Massachusetts [8]. With data of more than 65,000 patients admitted to the ICU, MIMIC-IV

contains extensive details about the patients' demographics, vital signs, laboratory measurements, procedures, medications, and outcomes related to emergency department and ICU admissions.

3.2 Dataset Construction Overview

All data extraction and processing were done via Google BigQuery, as it allows for scalable SQL-based queries of the MIMIC-IV tables without having to store the entire data locally. This method promotes reproducibility, and also enables explicit expression of cohort construction logic through the use of SQL queries.

In our research, we focus on the ICU cohort, specifically looking at adult patients (18+ years old) during their first ICU admission and staying at the hospital for at least 24 hours. After applying all filters, our final cohort consists of 54,551 unique adult stays. Data for each stay was expanded into one-hour intervals, resulting in roughly 5.6 million hourly observations.

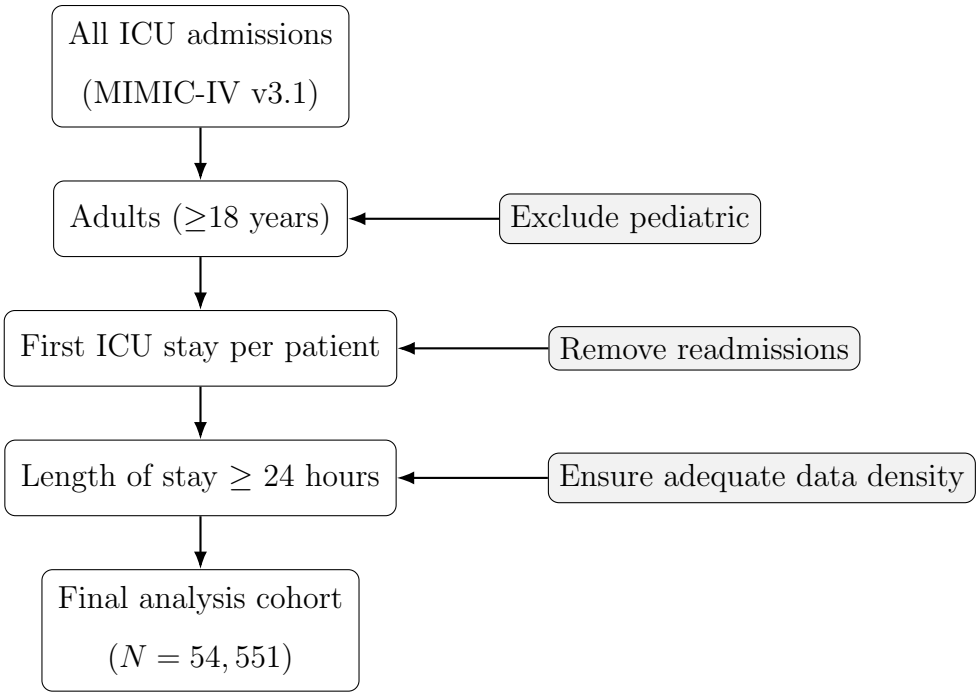

Figure 3.1: Cohort flow diagram showing inclusion/exclusion criteria to form the final ICU analysis cohort from MIMIC-IV v3.1.

The final dataset was created by joining multiple tables:

- `admissions`, `icu`, and `patients` for demographics and outcomes.
- `chartevents` and `labevents` for labs and vitals.
- `inputevents` and `procedureevents` for interventions.

Figure 3.2: Schema of MIMIC-IV tables used and their primary joins. Demographics/outcomes from `patients`, `admissions`; ICU stays from `icu`; time-varying physiology from `chartevents`, `labevents`; interventions from `inputevents`, `procedureevents`.

3.3 Cohort Definition

By limiting the analysis to the first ICU admission for each patient, the dependency between different patient trajectories is reduced and the potential for complications due to repeated hospital admissions with highly correlated treatment histories is eliminated.

In addition to restricting to the first ICU admission for each patient, only ICU admissions where there is adequate duration and data density to enable the creation of hourly discretized states are considered. All ICU admissions with implausible or missing key physiological measurements are also excluded in order to ensure reliable state–action–reward sequence generation.

The cohort was subsequently discretized into hourly time steps to create temporally ordered patient trajectories. A summary description of the study cohort and dataset is provided in Table 1.

3.4 Temporal Discretization

The clinical data were discretized into one-hour time steps, with each one-hour time step being a separate decision point. Although physiological measurements can be made more often than once per hour, finer temporal discretization will increase noise in the data. In contrast, less frequent temporal discretization can obscure clinically relevant changes in the physiological status of the patient. The decision to use a one-hour time step, then, represents a balance between clinical relevance and data availability.

Each one-hour time step has multiple observations summarized using clinically relevant summary statistics. For example, within each one-hour time step, vital signs and laboratory results are summarized to best represent the patient's status at the time of decision making, while treatment variables represent vasopressors administered either during the time step or shortly before the start of the time step. As such, the state-action pair generated for each one-hour time step accurately reflects what information would be available to physicians when deciding how to manage the vasopressor therapy of their patients.

3.5 State Variable Construction

A variety of state variable specifications were created to provide a representative view of a patient's status at each decision point. The state representation includes:

- Demographic characteristics of the patient (e.g., age and sex)
- Vital signs, particularly mean arterial blood pressure
- Laboratory measures providing insight into metabolism and organ function
- Indications of organ dysfunction based upon routine clinical measures
- Recent treatment history, with a focus on vasopressor administration

All state variables were developed solely from data elements that could reasonably be expected to be known to providers at the time treatment decisions were made. None of the state variable specifications include any knowledge about events subsequent to treatment decision making. Therefore, although these representations do not capture all latent physiological processes occurring in critically ill patients, they provide a structured model for understanding the clinical rationale for managing vasopressor therapy.

3.6 Action Definition and Vasopressor Representation

Actions represent vasopressor dosing decisions, but are measured as norepinephrine equivalents. The continuous infusion rate of vasopressors is categorized into a number of distinct dose categories, with each category representing a clinically significant band of vasopressor support.

There are two reasons why actions should be discretized. Clinically, vasopressor titration usually occurs in discrete bands rather than in small increments. Methodologically, discretizing the action space increases the amount of empirical evidence supporting each action. Since we are working in an off-line environment, stable learning and evaluation require sufficient empirical evidence supporting each action. All vasopressor drugs are converted

to norepinephrine equivalents to allow for equivalent representation across drug classes and patients.

3.7 Episode Definition and Termination

Each patient trajectory is considered to be a single episode, starting at ICU admission and ending at ICU discharge or death. Therefore, episodes are always finite and patient-specific in duration. This corresponds to the horizon definition used in our MDP framework.

We also define terminal states to indicate when an episode terminates (i.e., when a patient is discharged or dies). We include terminal states in order to guarantee that learning and evaluation procedures take into consideration episode boundaries, and do not extend value estimates beyond those associated with the observed ICU stay.

3.8 Final Dataset Structure

Our final dataset organization consists of collections of trajectories, where each trajectory is a sequence of hourly state-action-reward tuples. This organization permits us to apply offline RL methodologies while maintaining the temporal ordering and clinical context of vasopressor decision-making.

Summary statistics for both our cohort and dataset size are presented in Table 1. These statistics are descriptive and are intended to provide background information regarding the learning task under consideration rather than to facilitate causal inference.

CHAPTER 4

Problem Formulation

In this chapter we formally establish vasopressor management in the ICU as a sequential decision-making process based on the structure of a finite-horizon Markov Decision Process (MDP). The intent behind establishing this formulation is not to imply that clinical decision-making follows the Markov Property precisely; however, it establishes a structured abstraction that allows us to perform consistent policy-learning and evaluate the performance of these policies using retrospective data.

The MDP structure identifies the basic building blocks necessary for conducting offline reinforcement-learning (states, actions, rewards, transitions, and horizon) and clarifies the constraints and assumptions present in an observational healthcare setting.

4.1 Markov Decision Process Definition

We model the vasopressor decision-making task as an MDP defined by the tuple

$$(\mathcal{S}, \mathcal{A}, P, R, \gamma, T),$$

where $\mathcal{S}$ denotes the state space, $\mathcal{A}$ the action space, P the transition dynamics, R the reward function, γ the discount factor, and T the decision horizon.

An episode in our case represents a single ICU-stay. Episodes begin at ICU-admission and terminate at either discharge or death. We make decisions at an interval of one-hour, as this is the typical period at which clinicians assess patient-condition and modify treatment.

Table 4.1 summarizes the key components of the MDP used in this study.

Table 4.1: Markov Decision Process components for vasopressor decision-making in the ICU

Component	**Definition**
State (s_t)	Patient clinical state at hour t, including demographics, vital signs (with emphasis on MAP), laboratory measurements, indicators of organ function, and recent treatment history derived from electronic health record data
Action (a_t)	Discretized vasopressor dosing decision expressed in norepinephrine-equivalent dose bins
Reward (r_t)	Scalar proxy for short-term physiological stability, constructed to encourage maintenance of adequate mean arterial pressure and penalize extreme instability
Transition	Unknown and stochastic; implicitly defined by observed patient trajectories in retrospective ICU data
Policy (π)	Mapping from patient states to vasopressor dose bins learned from historical clinician behavior
Horizon (T)	Finite; corresponds to the duration of an ICU stay, terminating at discharge or death
Discount (γ)	Scalar in $(0, 1]$ weighting near-term versus future rewards
Learning setting	Offline; policies are learned exclusively from a fixed dataset without environment interaction

4.2 State Space

The state s_t in $\mathcal{S}$ represents the clinical status of a patient at decision-time t. States are generated from data contained within the health record and contain elements that would likely be accessible to clinicians during the decision-making process.

States can include demographic information about the patient, vital-signs, laboratory results, evidence of organ-function and information about recent treatments administered.

Particular attention was paid to MAP, as MAP is often considered the primary physiological signal directing clinicians' use of vasopressors.

To provide additional context regarding recent treatment-decision and physiologic-trends, historical treatment-information were incorporated into states.

Although this state-representation captures no other influences affecting patients' outcomes, it creates a structured collection of observable clinical-data related to the management of vasopressors.

Unmeasured confounding-variables and underlying-latent physiological-processes represent limitations to this representation, as they are further described in Chapter 7.

4.3 Action Space

At each decision-time t, the agent selects an action $a_t \in \mathcal{A}$ corresponding to a decision to administer a dose of vasopressor. Actions are discrete representations of clinically-meaningful ranges of vasopressor-support as norepinephrine equivalent-dose.

Discrete action-spaces serve two purposes. From a clinical perspective, vasopressor dosages are typically modified in coarse increments rather than being continuously-adjusted for optimal precision. Methodologically, discrete action-spaces ensure that each possible-action has been empirically-supported by the dataset, thereby providing a foundation for both reliable learning and evaluation in the offline-setting.

Vasopressor-agents are converted to norepinephrine equivalent-units so that there is a common basis for comparison among different treatments. As such, the resulting action-space is discrete, contains all the supported actions by clinicians, and represents the range of actions that can be taken based upon observed clinician-behavior.

4.4 Transition Dynamics

The transition-function $P(s_{t+1} \mid s_t, a_t)$ represents how patient-states evolve as a result of treatment-decisions. In this study, we do not know the transition-functions and thus they are not modeled explicitly. Rather, the transition-dynamics are defined implicitly via observed patient-trajectories within the dataset.

The implicit nature of our transition-dynamics reflects the complexities and variability inherent in patient physiology. Patient physiology is influenced by many observable and non-observable variables. Therefore, offline reinforcement-learning methods utilize empirical-transition information instead of estimating a parametric-dynamics-model.

4.5 Reward Function

The reward-function $R(s_t, a_t)$ encodes clinicians' immediate goal of achieving stable physiological conditions. Rewards are defined based on proxy-measures derived from patient-state-variables. Emphasis is placed on maintaining stable MAP and minimizing extreme physiological instability.

Our reward-definition reflects an intentional trade-off between maximizing interpretability/alignment with clinicians' decision-making processes versus optimizing long-term outcome metrics such as mortality rates. Although long-term outcome metrics may be critical to clinicians' understanding of overall-patient-benefit, attributing specific-long-term-outcomes to individual-treatment-decisions is difficult in retrospective-data settings. Consequently, rewards within this research should be viewed as measures of relative short-term stability rather than direct measures of patient utility or clinical benefit.

4.6 Horizon and Discounting

The decision horizon T is finite/patient-specific representing the length-of-an ICU-stay. A discount factor $\gamma \in (0, 1]$ is applied to future rewards to encourage policies that value near-

term physiological stability while considering potential downstream effects.

4.7 Offline Learning Setting and Constraints

All policies developed in this research are trained completely in an offline-fashion using a static/dataset-based set of historical-clinician-decisions. Our learning-agent does not interact with its environment nor create any new data. As a result, several challenges emerge due to the offline nature of our learning paradigm, namely distributional-shift and extrapolation-error. These issues arise when learned-policies preferentially select actions that lack sufficient empirical-support in the training-data-set. These issues motivate our utilization of conservative/offline reinforcement-learning methodologies that bound usage of unsupported-actions.

Due to the offline-nature of our learning-paradigm, evaluation is similarly constrained. Since learned-policies cannot be executed in practice, we must estimate their performance utilizing retrospective/off-policy-evaluation techniques that are inherently noisy/approximate. These challenges will be addressed in detail throughout later-chapters using robust uncertainty-reporting/interpretations.

CHAPTER 5

Methods

5.1 Experimental Pipeline and Offline Constraints

This section describes how we learned and tested treatments through offline RL to manage vasopressors. All of our models are trained with retrospective ICU data, so they have never interacted with a patient nor been run in a simulated environment. Consequently, the experimental approach was constrained by the limitations inherent to using observational clinical data along with the need for conservative and interpretable learning processes.

The total experimental process was divided into four phases. First, patient trajectories were generated by combining EHR data into an hourly state-action-reward format, as detailed in Chapter 3. Then, offline RL methods were employed for policy learning with all assumptions about the data remaining constant. After that, we evaluated the quality of the policies via off-policy evaluation techniques to calculate what the expected returns would be without actually running the policies. We further qualitatively assessed policy behavior by evaluating the frequency of different actions relative to clinically important patient states.

An essential aspect of this case is that the clinician's behavior policy, essentially clinician's decision-making history, is the primary data source. There is no way for the learning agent to attempt new options or to correct its own mistakes through interaction with the environment. Consequently, distributional shift challenges policy learning in this case because a policy may assign a high utility to actions that have little evidence in the data. Distributional shift is especially dangerous in clinical settings, where poorly supported actions may represent unsafe or unfeasible interventions [7].

Therefore, this research focuses on methods designed for offline learning. Instead of at-

tempting to maximize performance, these methods emphasize conservatism and the evidence in the data supporting clinician behavior [2, 7]. Behavior Cloning is the baseline model that mimics the clinician's observed behavior. Conservative Q-Learning uses the fact that the value-function should penalize the agent when it takes an action not supported in the data. Batch-Constrained Q-Learning directly restricts the output of the policy to match the actions supported in the data.

Another significant area of interest is policy evaluation. Since the policies learned cannot be directly implemented, their effectiveness must be measured indirectly by estimating performance based upon the data available. Because both the policies and the estimators are learned in an off-policy fashion, the expected return must be approximated taking into consideration the bias and variance of each estimator. Once again, each of the estimators must be compared.

Our basis for selecting methods used in this chapter is to indicate that our objective is not to develop the most effective policies or policies that can be implemented. Our goal is rather to investigate whether offline RL can generate clinically rational and stable policies.

5.2 Behavior Cloning (BC)

Behavior Cloning was chosen as a baseline model to use as a benchmark for developing a policy for making treatment decisions using past clinician behaviors [9]. Behavior Cloning views policy development as a supervised learning problem; therefore, the goal of supervised learning is to develop a predictive model that predicts what action a clinician took in a specific state. With respect to managing vasopressor dosages, this translates into developing a direct mapping between patient physiology and discrete vasopressor dosage bins.

Formally speaking, Behavior Cloning develops a policy $\pi_\theta(a|s)$ by minimizing a supervised loss function over a distribution of state-action pairs (s_t, a_t).

$$\pi_\theta = \arg\min_{\pi} \mathbb{E}_{(s,a)\sim D}[-\log \pi(a \mid s)]$$

In cases where discrete action space exists, this corresponds to minimizing the negative log-likelihood (equivalent to cross-entropy loss) of observed actions taken by the clinician under the learned policy [9]. Behavior Cloning does not require estimation of values nor does it plan over long horizons; therefore, it eliminates much of the instability present in traditional reinforcement learning methods.

Behavior Cloning provides one major benefit in terms of respecting data distribution. By virtue of the fact that Behavior Cloning generates a policy that mirrors what was done by a clinician, it operates strictly within the bounds of the data distribution. This provides Behavior Cloning with considerable appeal in safety-critical domains like healthcare; where generating policies that operate outside of the data distribution can potentially place patients at risk. For these reasons alone, Behavior Cloning can serve as a robust baseline for offline RL.

However, Behavior Cloning has numerous shortcomings. One example of this involves Behavior Cloning's inability to consider outcomes of actions in order to differentiate between actions whose outcome is beneficial versus actions whose outcome is sub-optimal or results in harm to patients. The resulting policy will reflect averages of clinician behaviors present in the training dataset. These include variations attributable to practices among clinicians, institutional variation and other confounding factors. Ultimately, Behavior Cloning will reproduce behaviors that could be acceptable but will not improve them.

Furthermore, Behavior Cloning is very susceptible to distributional differences in training states and testing states. To give you an example, if there is some type of state that a policy needs to make predictions regarding patient states that were not adequately represented in the training data-set, then errors may build up over time until compounding occurs and the model will begin to diverge from typical patient states. This limitation on generalization outside of clinician behaviors is not problematic in this study's completely offline evaluation setting.

Behavior Cloning serves as a reference point representing a pure supervised learning approach to imitation in comparison to more conservative reinforcement learning methods.

Differences between Behavior Cloning and more advanced methods utilizing value functions will demonstrate how having additional knowledge provided by rewards impacts policy behavior.

5.3 Conservative Q-Learning (CQL)

Conservative Q-Learning (CQL) uses an action-value function Q(s,a) to estimate the expected return of taking an action a in state s [10]. An implicit policy is developed by identifying the actions that have the highest estimated value. CQL modifies the traditional Q-learning methodology to minimize the Bellman error; however, it also incorporates a penalty for large Q-values when choosing an action outside of the empirical data distribution. Therefore, CQL will "bias" the estimated value of less-supported actions downwards, and consequently encourage the development of a conservative policy.

5.3.0.1 CQL Objective

Formally, CQL augments the standard Bellman error minimization objective with an additional regularization term that penalizes the expected Q-value under a broad action distribution relative to the Q-values observed in the dataset [10]. A common formulation of the CQL loss is:

$$\mathcal{L}_{\text{CQL}}(Q) = \underbrace{\mathbb{E}_{(s,a,r,s')\sim D}\left[\left(Q(s,a) - \left(r + \gamma\, \mathbb{E}_{a'\sim\pi}[Q(s',a')]\right)\right)^2\right]}_{\text{Bellman error}} + \alpha\, \underbrace{\left(\mathbb{E}_{s\sim D, a\sim\mathcal{A}}[Q(s,a)] - \mathbb{E}_{(s,a)\sim D}[Q(s,a)]\right)}_{\text{conservatism penalty}},$$

where $\mathcal{D}$ denotes the offline dataset, $\alpha > 0$ controls the strength of conservatism, and $\mathcal{A}$ denotes the action space. The second term penalizes large Q-values for actions sampled broadly from the action space relative to those actually observed in the data.

Intuitively, this regularization encourages the learned Q-function to assign lower values to actions that are unlikely under the clinician behavior policy, reducing the risk of extrapolation beyond the dataset support [10].

5.3.0.2 Interpretation in the Clinical Setting

When applied to managing vasopressors, CQL develops a dosing decision policy that recommends dosing decisions both associated with positive short-term stability and historically practiced by clinicians. When evaluating the quality of a dosing decision that is highly ranked by the value function but has insufficient supporting empirical evidence, CQL will significantly lower its weighted value. This type of reasoning parallels clinical thinking where cautious decision making occurs when deviating from typical clinical practices without substantial evidence to justify this departure.

Although CQL does allow for variations from clinician behavior when the variation has some supporting value estimates, it maintains a general conservative bias towards actions that are well-supported by empirical evidence. Thus, CQL exists in the middle of pure imitation and completely unrestricted value estimation.

5.3.0.3 Role of CQL in This Study

This research utilizes CQL to evaluate whether conservatively estimating values creates a policy that differs significantly from pure imitation but is still reasonable and safe. By comparing CQL to Behavior Cloning and Batch-Constrained Q-Learning, we can examine how differing types of conservatism affect the development of a learned policy. Furthermore, since CQL represents a form of conservative offline reinforcement learning, it highlights the trade-offs between expressiveness and safety when developing policies using observational healthcare data.

5.4 Batch-Constrained Q-Learning (BCQ)

Batch-Constrained Q-Learning (BCQ) is another offline reinforcement learning algorithm that induces conservatism through a restriction on the policy's possible actions [11]. In contrast to Conservative Q-Learning, which achieves conservatism through regularizing values, BCQ achieves conservatism by limiting the number of possible actions the policy can take

during decision making. Consequently, BCQ is particularly useful for applications in areas requiring extreme caution, e.g., healthcare.

BCQ consists of three primary components:

1. A generative model that generates an approximate distribution of actions found in the dataset.

2. An action-value function that evaluates proposed actions.

3. An action-selection component that determines among actions that could potentially be taken in a state what should be done.

Together, these three components provide assurance that the learned policy stays within historical clinician behavior and still takes advantage of value estimations to distinguish among feasible actions.

5.4.0.1 Support-Constrained Action Selection

For every state s, BCQ identifies a subset of potential actions for which there is empirical evidence they were taken by clinicians. For discrete action spaces, this can be accomplished by estimating the probability of each action occurring in a given state, and retaining only those actions with probabilities greater than or equal to a predetermined threshold. BCQ then selects the action that maximized its estimated Q-value among those retained [11].

Formally, the BCQ policy can be expressed as:

$$\pi_{\text{BCQ}}(s) = \arg\max_{a \in \mathcal{A}} Q(s, a) \quad \text{s.t.} \quad \hat{p}(a \mid s) \geq \tau,$$

where $\hat{p}(a \mid s)$ is the estimated likelihood of action a under the behavior policy and τ is a threshold controlling how strictly the action space is constrained.

This mechanism ensures that actions rarely or never observed in the dataset are effectively excluded from consideration, regardless of their estimated value [11].

5.4.0.2 Interpretation in the Clinical Context

In the context of administering vasopressors, the support-constraint enforced by BCQ has a straightforward clinical interpretation. The learned policy is able to choose among doses previously used by clinicians in similar physiological conditions; however, it is prohibited from suggesting dose levels that lack empirical evidence. This perspective on decision making is analogous to how clinicians make decisions when they deviate from established clinical practices; namely, they require significant justification before doing so.

Additionally, BCQ separates which actions are permissible from which actions are preferable. Through this separation, BCQ presents a clear mechanism for controlling risk. Value estimations are utilized to rank plausible clinical actions, whereas the support-constraint limits extrapolation into uncharted territory away from empirically-observed practice.

5.4.0.3 Role of BCQ in This Study

BCQ was added to this study as an alternative form of conservatism to CQL. Although both CQL and BCQ seek to limit extrapolation errors in offline RL, they accomplish this through distinct mechanisms: CQL regulates (penalizes) actions outside of clinician behavior through value; whereas BCQ constrains the possible actions through direct regulation.

Thus, comparisons between BCQ and Behavior Cloning and CQL allow researchers to investigate how varying degrees of conservatism shape learned policy behavior. Specifically, BCQ provides insight into whether constraining policies to only include actions supported by clinician behavior results in more stable or understandable decision processes in the vasopressor management domain.

Similar to Behavior Cloning and CQL, BCQ is evaluated using off-policy evaluation and qualitative policy assessment, without claiming to represent an optimal solution or clinical suitability for implementation.

5.5 Off-Policy Evaluation

Since learned policies cannot be deployed in real-world clinical settings, their effectiveness must be evaluated retrospectively using existing data. This study utilizes two complementary off-policy evaluation (OPE) methods, Weighted Importance Sampling (WIS) and Fitted Q-Evaluation (FQE), to estimate the expected utility/return of learned policies using trajectory samples generated by the clinician behavior policy [12, 13]. These estimates are employed to comparatively evaluate policies rather than assert absolute clinical merit.

The use of multiple complementary OPE methods provides robustness checks by utilizing distinct modeling assumptions.

5.5.0.1 Weighted Importance Sampling

Weighted Importance Sampling (WIS) estimates the value of a target policy π by reweighting observed trajectories based on how likely the target policy π would have selected the same actions as the behavior policy μ [12]. For a dataset of N trajectories $\{\tau_i\}_{i=1}^{N}$, the WIS estimator is given by:

$$\hat{V}^{\pi}_{\text{WIS}} = \frac{\sum_{i=1}^{N} \left(\prod_{t=0}^{T_i} \frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)} \right) R(\tau_i)}{\sum_{i=1}^{N} \prod_{t=0}^{T_i} \frac{\pi(a_t^i|s_t^i)}{\mu(a_t^i|s_t^i)}},$$

where $\mathrm{R}(\tau_i)$ denotes the cumulative reward of trajectory τ_i. Normalization reduces variance relative to standard importance sampling, particularly when likelihood ratios are highly variable.

In this study, we treat the clinician policy as the behavior policy μ. We compute likelihood ratios based on learned policy probabilities. WIS provides a model-free direct estimate of policy value but suffers from high variance when there exist large differences between target and behavior policies [13].

5.5.0.2 Fitted Q Evaluation

Fitted Q Evaluation (FQE) is a model-based OPE method that estimates policy value by learning a Q-function under the target policy [14, 15, 16]. Once a target policy π is learned, we train a separate Q-network via Bellman regression using the actions selected by π. The estimated value is then computed as:

$$\hat{V}^{\pi}_{\text{FQE}} = \mathbb{E}_{s\sim\mathcal{D}}\left[Q^{\pi}(s, \pi(s))\right].$$

FQE avoids explicit importance weighting and tends to be more stable than WIS in high-dimensions or long-horizons. However, FQE introduces an approximation bias due to function approximation and relies on accurately modeling value dynamics under the learned policy [14, 15, 16].

5.5.0.3 Interpretation and Limitations

Both WIS and FQE have fundamental limitations inherent in retrospective evaluations [12, 13, 14, 15, 16]. High variance can occur in WIS when considering rare actions, whereas FQE can introduce bias if the learned Q-function extends beyond data support. Confidence intervals are provided using bootstrapping for all OPE estimates.

Importantly, comparative OPE results in this study should be interpreted as reflecting differences in relative policy behaviors under consistent assumptions and not as implying causal effects on patients' outcomes.

5.6 Training Details and Hyperparameters

All models are trained using identical state representations, action space definitions, and dataset splits to enable equitable comparisons among methods. Architectures and optimization parameters were maintained constant wherever appropriate; otherwise, architectural details differed solely as dictated by specific requirements for each method.

Neural networks were trained using d3rlpy, a software package providing standardized implementations for several offline reinforcement learning algorithms [17]. All models were trained using a single fixed dataset and also implemented using the d3rlpy library. In addition, hyperparameters were chosen primarily based on prior literature rather than exhaustive tuning for maximal performance. Off-policy evaluation and qualitative assessments of policies were performed using validation datasets only.

Appendix A provides summary information regarding key hyperparameters for each algorithm. Identical hyperparameters were applied whenever possible across methods to help isolate effects stemming from individual learning algorithms.

CHAPTER 6

Results

In this chapter, we present an empirical evaluation of offline reinforcement learning policies developed from retrospective ICU data. We have separated our findings into two sections. In the first section, we will evaluate policy behaviors based on how they choose to assign actions among clinically relevant patient states. In the second section, we will detail the results from off-policy evaluations (OPE) performed using Weighted Importance Sampling (WIS) and Fitted Q Evaluation (FQE). Throughout this chapter, we describe results descriptively; however, the interpretations of these results are detailed in Chapter 7.

6.1 Dataset and Action Distribution Diagnostics

Before discussing the performance of the developed RL policies, we provide some context about the dataset used for training and the distribution of clinician actions to define the RL problems. The dataset comprises hourly decision points representing adult patients during their first ICU admission. Each decision point represents a specific bin of discretized norepinephrine equivalent vasopressor doses. Clinicians' action frequencies for vasopressors are significantly unbalanced across different bins of discretized norepinephrine equivalent vasopressor doses. Most clinicians' decisions fall within the lower end of the spectrum (lower dose or no vasopressors); higher dose actions were relatively infrequent and often represented by more severe physiological states. As this distribution accurately represents typical clinical practices, it motivated the use of conservative RL methods when developing offline RL algorithms to address unevenly distributed action supports. These distributions provide a baseline against which we can compare the performance of developed RL policies to determine

if the learned policies continue to reflect clinician behaviors or if they begin to extend beyond what was seen in clinician behavior (i.e., extending into areas of poor action support).

6.2 Policy Behavior Across Patient States

Next, we investigate how the developed RL policies behave compared to clinicians' action distributions while considering patient state. We focus on how the policies decide upon actions with respect to MAP, which is a key physiologic parameter influencing vasopressor administration.

For each policy, we compute action frequencies with respect to discrete bins of MAPs so that we may directly compare how policies react to hypotension (low MAP), normotension (normal MAP), and hypertension (high MAP) states. This provides insight into whether the developed RL policies display any clinically meaningful patterns of behavior (e.g., increased vasopressor use as MAP decreases and decreased use as MAP increases).

6.2.0.1 Behavior Cloning vs. Clinician Policy

In Figure 6.1, we illustrate action frequencies for the Behavior Cloning policy and the clinician policy across MAP bins. Because of how the Behavior Cloning policy is constructed, it replicates clinician action distributions nearly perfectly, including both the total frequency of vasopressor administrations and how the frequency changes with decreasing MAP.

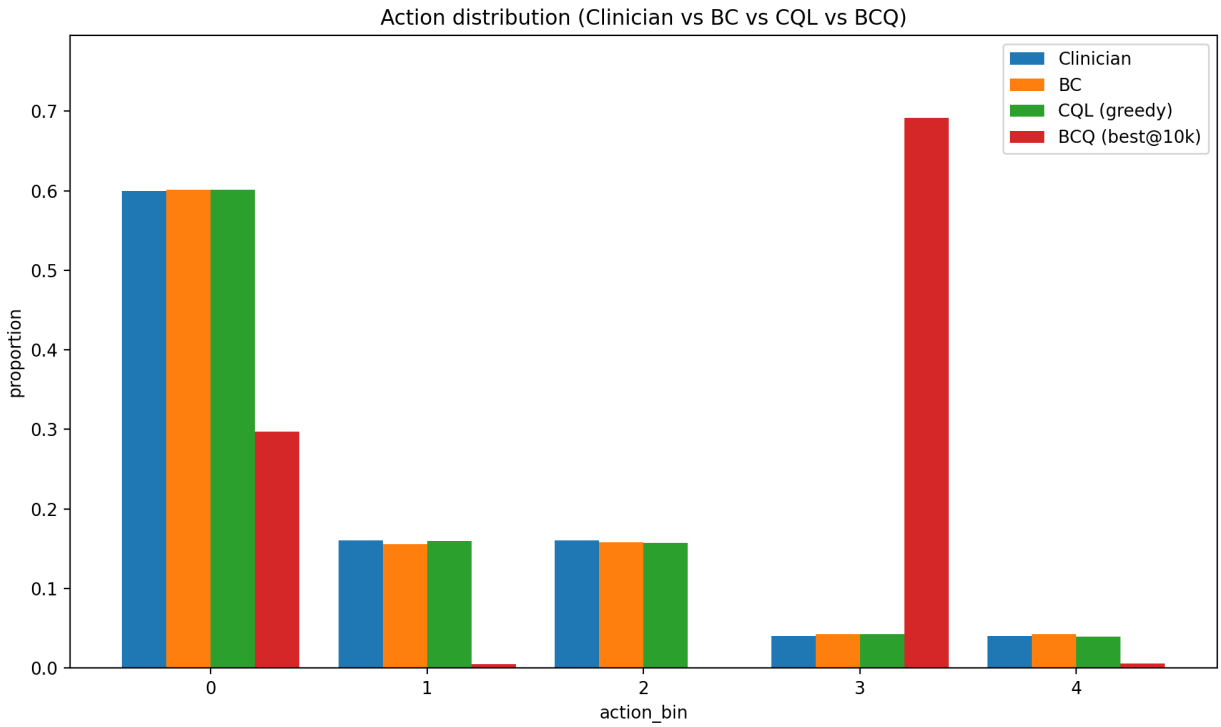

Figure 6.1: Action frequency distributions across discretized mean arterial pressure (MAP) bins for the clinician policy and learned policies. Action frequencies are grouped by vasopressor dose bin and normalized within each MAP bin.

The minor differences between the two distribution profiles stem from model approximation errors and dataset imbalance; however, the same general tendencies have been reproduced. Specifically, both of these policies tend to increase the dose of vasopressors as MAP levels decrease and favor either no or the least amount of vasopressor use when MAP is elevated.

The results confirm that Behavior Cloning provides an accurate imitation of the decision-making process of clinicians. Therefore, Behavior Cloning can serve as a reliable imitation baseline and has successfully replicated the predominant decision-making processes of clinicians without creating many alternative decisions.

6.2.0.2 Batch-Constrained Q-Learning vs. Clinician Policy

Figure 6.2 compares action distributions produced by the Batch-Constrained Q-Learning policy to those of clinicians. While BCQ remains largely within the support of clinician actions, it exhibits systematic differences in how actions are allocated across MAP bins.

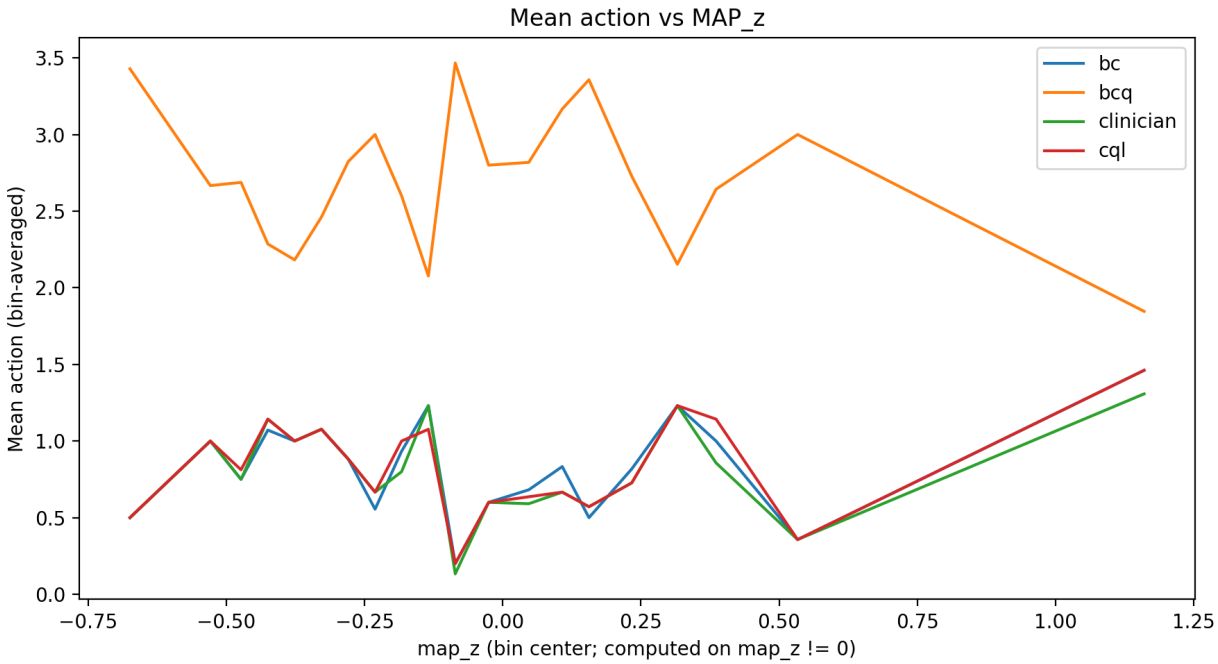

Figure 6.2: Policy action probabilities as a function of standardized mean arterial pressure (MAP). Curves illustrate how learned policies adjust vasopressor dosing across physiological states.

Although BCQ falls well within the bounds of clinician actions, there are some important differences in how BCQ allocates action probability across various MAP levels.

When MAP is low, BCQ tends to place a larger portion of its probability mass on moderate vasopressor doses than the clinician policy and exhibits a preference to avoid very rare high-dosage actions. When MAP is high, BCQ's allocation of action probabilities shifts toward smaller vasopressor doses or no vasopressor use. This demonstrates a heightened awareness of clinical physiology while maintaining conservative constraint boundaries.

Also, as expected, BCQ did not allocate significant amounts of probability to actions that were poorly represented in the dataset. This is due to the explicit support-constrained design of BCQ. All in all, BCQ produces state-dependent action distributions that are significantly different from clinician behavior yet fall within the empirically supported action bounds.

6.2.0.3 Conservative Q-Learning Policy Behavior

To further understand the similarities and differences between the Behavior Cloning policy and Conservative Q-Learning policy, we can look at their respective action distributions. Compared to Behavior Cloning, CQL creates more differentiated action distributions across each level of MAP. The action distributions will be influenced much more sharply by changes in MAP.

Compared to BCQ, CQL assigns non-trivial probabilities to more actions. Even though conservative regularization constrains extreme extrapolation, CQL's value-based mechanism enables broader exploration within the discrete action space.

These patterns illustrate how different forms of value regularization and explicit action constraints influence the learned policy behavior. Ultimately, CQL presents more state-sensitivity relative to imitation and supports a broader range of actions than BCQ.

6.3 Summary of Policy Behavior Results

Table 6.1: Summary statistics of learned policy behavior across the evaluation dataset.

Policy	**Mean Action**	**Action Entropy**	**Support Coverage**
Clinician	0.7605	1.1508	0.7151
Behavior Cloning	0.7689	1.1554	0.7179
Conservative Q-Learning	0.7588	1.1511	0.7152
Batch-Constrained Q-Learning	2.1041	0.6725	0.4179

Across all methods, the developed policies demonstrated a consistent response to MAP. The developed policies increased vasopressor use with decreasing MAP values, whereas the policies decreased vasopressor use with increasing MAP values. Behavior Cloning is able to mirror clinicians' decision making more accurately than the other two methods. Behavior Cloning, BCQ, and CQL each create specific alterations to clinicians' decision making due to the incorporation of reward maximization and conservative constraint methodologies, respectively.

The differences observed among methods were mostly apparent at intermediate MAP levels. At these MAP levels, there are several clinically plausible dosing options available, and the distinction based upon a value metric emerges. Therefore, further analysis was warranted and thus an additional method of comparing methods, which was off-policy evaluation (OPE), is used. The subsequent section presents results derived from OPE.

6.4 Off-Policy Evaluation Results

Using weighted importance sampling (WIS) and Fitted Q evaluation (FQE) we evaluate our learned policies to estimate expected return of each policy under the data collected retrospectively. Due to the fact that our policies are being evaluated without having been deployed or simulated, it is appropriate to consider these estimates as approximate and comparative measures of the quality of our policies.

All of the OPE results are generated by using the clinician policy as the behavior policy, and then computing them against a validation set which has been withheld. In order to provide an indication of the variability inherent within the estimators, we report bootstrapped confidence intervals for all estimates.

Table 6.2: Off-policy evaluation results using weighted importance sampling (WIS) and fitted Q evaluation (FQE). Reported values are bootstrap means. WIS estimates are reported only for learned policies evaluated against the clinician behavior policy. FQE estimates are reported only for learned policies, as value functions for the clinician policy are not defined.

Policy	**WIS Estimate**	**FQE Estimate**
Clinician	—	—
Behavior Cloning	—	-97.9866
Conservative Q-Learning	-0.0084	-97.6813
Batch-Constrained Q-Learning	-0.9373	-107.0433

— indicates that the estimate is not defined or not reported.

6.4.0.1 Weighted Importance Sampling Results

Table 6.2 provides the Weighted Importance Sampling (WIS) estimates for learned policies evaluated relative to the Clinician Behavior Policy. The WIS estimates appear quite variable across all of the methods, and this is reflected in the width of their confidence intervals.

The Behavior Cloning policy produced WIS estimates which were very close to the size of those of the clinician's policy, suggesting that it was designed to imitate the clinician's behavior. While the WIS estimates of CQL and BCQ were slightly different compared to the Behavior Cloning policy, they had considerable overlap among them.

Additionally, there appears to be some instability due to a few samples with large importance weights in several instances. Due to long horizons and discrete action spaces, we would expect this kind of instability. Therefore, the large variance and overlapping confidence intervals suggest that WIS does not provide a clear framework for policy differentiation in this specific setting.

6.4.0.2 Fitted Q Evaluation Results

The estimated values based upon fitted Q evaluation can also be found in Table 6.2. Comparing these with the WIS estimates indicates that the fitted Q evaluation estimates are considerably more stable across bootstrap resamples, and have narrower confidence intervals.

Further, the fitted Q evaluation estimates suggest some systematic differences in terms of estimated values between imitation-based and value-based policies. Specifically, CQL and BCQ have estimated values that are different from Behavior Cloning. Nevertheless, the magnitudes of these differences vary across the runs, and fall well within the range of uncertainty of the estimator.

Since the fitted Q evaluation uses a learned value function, these estimates could include some approximation bias in addition to representing actual differences in the policies. For this reason, the results of the fitted Q evaluation should be used in combination with those of the weighted importance sampling instead of alone. In summary, FQE improves significantly upon the stability of estimates compared to WIS, establishing a more consistent baseline for comparison. Having said that, it should still be kept in mind that these differences between policies are negligible and make conclusive differentiation an ambitious benchmark.

6.4.0.3 Uncertainty and Estimator Instability

Both OPE methods suffer from significant shortcomings when applied to retrospective ICU data. Since WIS can be highly sensitive to action mismatch between a learned policy and clinician behavior, it exhibits significant variance in environments where the action distribution is unbalanced. Since FQE has reduced variance at least partially due to dependency on function approximation and training stability, it suffers less from estimation instability than does WIS. Additionally, since both OPE methods produce overlapping confidence intervals and are sensitive to resampling, the ranking of the policies' values using either method is not robust. This result emphasizes the importance of presenting estimation uncertainty in order to avoid over-interpretation of single-point estimates.

6.5 Results Summary

Overall, our results illustrate that offline RL policies exhibit consistent and clinically sensible behaviors across a variety of patient states. On the other hand, they also show that quantitative evaluations of these behaviors remained uncertain and noisy.

Our behavioral analyses illustrate many important qualitative differences between how the policies select actions given a particular state within the MAP bins. For example, we see greater differences in selecting actions by imitation-based versus conservatively-value-based policies. Our off-line policy evaluation results do not support these qualitative differences. Instead, they reveal considerable overlap in confidence intervals across WIS and FQE.

Therefore, based on our results, we find that analyzing policy behavior is generally more reliable and easier to understand than evaluating the numerical quantities associated with a particular policy. Thus, our results support using quantitative OPE results only as supplementary diagnostic tools rather than definitive proof of whether one policy is better than others.

CHAPTER 7

Discussion

Chapter 7 provides an interpretation of the empirical results presented in Chapter 6 and situates them within the larger framework of offline reinforcement learning in healthcare. Discussion is primarily focused on what can be learned from policy behavior and off-policy evaluation in a retrospective ICU setting; however, methodological and data related limitations are acknowledged throughout the discussion.

7.1 Summary of Key Findings

All methods evaluated in this study resulted in learned policies that exhibited behavior that was clinically coherent with respect to MAP, increasing vasopressor support at low MAP values and decreasing vasopressor support as MAP increases. Behavior Cloning produced learned policies that were close approximations of clinician action distributions, whereas Conservative Q-Learning and Batch-Constrained Q-Learning produced learned policies that had structured deviation from clinician action distributions due to optimized rewards under conservative constraints.

Algorithms differed significantly in terms of how actions were allocated over the range of possible MAP values. Within the intermediate MAP ranges, where multiple dosing decisions are plausible, value-based methods tended to differentiate between actions more strongly than direct imitation methods; yet, both types of methods remained grounded in empirical evidence supporting clinician behavior.

In contrast, quantitative off-policy evaluation results indicated significant variability and overlap in the confidence intervals across the different methods. Notably, fitted Q evalua-

tion provided more stable estimates than weighted importance sampling. However, neither method resulted in clearly defined separation between learned policies.

7.2 Interpretation of Policy Behavior

The results of the behavioral analysis indicate that conservative offline reinforcement learning methods can produce learned policy behavior that is structured and interpretable from retrospective ICU data. Both Conservative Q-Learning (CQL) and Batch-Constrained Q-Learning (BCQ) have produced learned policies that responded meaningfully to physiological signals without producing either extreme or unsupported actions, which is consistent with their respective design objectives.

Behavior Cloning served as a reference point for demonstrating the amount of structure that exists within clinician behavior. Therefore, the deviations introduced by value-based methods can be attributed to the influence of the reward signal and/or conservative constraint(s); thus, there is no indication that these deviations arise from uncontrolled exploration.

Moreover, the differences observed between CQL and BCQ further illustrate that conservatism is not a singular concept. Specifically, value regularization (CQL) and explicit action-space constraints (BCQ) result in distinctly different policy behaviors; despite being trained on the same data. Thus, the need to examine policy behavior directly instead of relying solely on scalar value estimates is reinforced.

7.3 Interpretation of Off-Policy Evaluation Results

As previously stated, OPE results must be interpreted cautiously. Although WIS and FQE provide approximate estimates of policy value, both methods have limitations associated with their use on long-horizon, high-dimensional clinical data.

The high variance observed in WIS estimates arises from its sensitivity to action mismatches and imbalance in action distributions. Conversely, FQE reduces variance but intro-

duces approximation bias through function learning. Furthermore, the lack of consistently clear separation between learned policies across OPE methods indicates that quantitative value estimates alone are insufficient to determine the superior learned policy in this setting.

Therefore, these results support the notion that OPE should be considered as a diagnostic tool, potentially as a means of providing insight into policy behavior rather than replacing it.

7.4 Limitations

There are several limitations to this study. First, this analysis is based upon retrospective observational data; hence, the analysis cannot provide a causal interpretation. Treatment decisions and outcomes are influenced by factors not accounted for within EHRs including unmeasured confounding factors, clinical judgment and hospital-specific protocols.

Secondly, the MDP formulation assumes that partial observability can be resolved through constructing states. Patient condition is determined by unknown variables and time-delayed effects that may not be represented by hourly measurement constructs.

Thirdly, the reward function utilized proxy measures of immediate physiological stability rather than long-term outcome measures (e.g., mortality). This decision enhances interpretability and aligns more closely with clinicians' decision making processes; however, it constrains the conclusions we can draw.

Lastly, OPE methods exhibit large variability especially in environments with highly skewed action distributions and long horizons. As shown by overlapping confidence intervals across methods, the estimates generated by OPE methods should not be taken as exact measures of expected clinical benefit.

7.5 Implications and Scope of Conclusions

The results obtained here do not substantiate claims of clinical optimality or readiness for clinical deployment. Rather, they demonstrate that conservative offline reinforcement learn-

ing methods can recover clinically plausible and interpretable policy behavior from retrospective ICU data.

Thus, these findings suggest that behavioral analysis should serve as a primary component of evaluating learned policies in healthcare; particularly when quantitative value estimates are unreliable or ambiguous. Hence, conservative offline RL methods should be regarded as tools for exploring structured decision support in clinical settings rather than as autonomous decision makers.

7.6 Future Research

A number of areas for future research are suggested by this study. First, richer state representations that incorporate temporal context or latent variable models could potentially mitigate some forms of partial observability inherent in clinical data. Second, expanding the action space to include additional treatments or continuous dosing decisions could allow for capturing more complex clinical decision-making; however, such expansions would likely require new techniques for managing increased uncertainty during evaluation.

Finally, advances in OPE and uncertainty quantification remain necessary for enhancing the reliability of retrospective policy assessments. Additional areas for investigation include potential hybrid approaches that integrate clinician-in-the-loop validation or simulation-based testing prior to any consideration of prospective deployment. Ultimately, careful incorporation of offline reinforcement learning with clinical expertise and rigorous evaluation frameworks will be necessary for converting retrospective insights into safe and effective decision-support tools.

CHAPTER 8

Conclusion

The objective of this dissertation was to evaluate whether conservative offline reinforcement learning methods can produce clinically plausible and interpretable learned policy behavior from retrospective ICU data using electronic health records. To achieve this goal, vasopressor management was framed as a finite horizon sequential decision problem and learned policies were evaluated using offline constraints representative of real-world settings. Specifically, focus was placed on analyzing policy behavior/interpretability rather than determining clinical optimality.

Data used for this analysis consisted of patient trajectories at an hourly resolution created from first ICU admissions in the MIMIC IV dataset. Each trajectory was modeled as a sequence of states-actions-rewards. Three classes of learned policies were evaluated: Behavior Cloning (direct imitation), Conservative Q-Learning (value-based conservatism), and Batch-Constrained Q-Learning (explicit support constraints). Additionally, two types of evaluations were performed: Behavioral Analysis (policy behavior) and Off-Policy Evaluation (quantitative comparison of policy performance).

Results from the behavioral analysis indicated that all three types of learned policies produced clinically coherent behavior with respect to MAP, with all learned policies responding similarly to clinically relevant physiological signals (MAP). All three types of learned policies also resulted in similar responses to variations in MAP with Behavior Cloning resulting in learned policies that were closer approximations of clinician action distributions than Conservative Q-Learning and Batch-Constrained Q-Learning.

Conservative Q-Learning and Batch-Constrained Q-Learning resulted in learned policies that deviated from clinician action distributions; however, these deviations were structured

and remained grounded in empirical evidence supporting clinician behavior. Algorithms also varied in terms of allocating actions across possible MAP values. In intermediate MAP ranges (where multiple dosing decisions are clinically plausible), value-based methods (CQL and BCQ) tended to differentiate more strongly between actions than direct imitation methods (Behavior Cloning), while maintaining grounding in empirical evidence supporting clinician behavior.

Conversely, quantitative comparisons of learned policy performance using off-policy evaluation resulted in significant variability and overlap in estimated confidence intervals across each type of method. Weighted Importance Sampling (WIS) produced less stable estimates than Fitted Q Evaluation (FQE), however both methods failed to clearly separate learned policies.

Overall, the results suggest that conservative offline reinforcement learning can produce clinically plausible and interpretable learned policy behavior from retrospective ICU data; however, quantitative estimates of expected clinical benefits are insufficient to determine superior learned policies. Consequently, rather than supporting claims of clinical optimality or readiness for deployment, this study highlights the role of offline reinforcement learning as a tool for structured analysis of historical clinical decision making under specified safety constraints.

8.1 Future Work

Several areas for future research are suggested by this study. First, utilizing richer state representations that account for temporal context or latent variable models may help reduce aspects of partial observability in clinical data. Second, expanding the action space to include additional treatments or continuous dosing decisions may enable capture of more complex clinical decision-making; although such expansions would likely necessitate development of novel techniques for addressing increased uncertainty during evaluation.

Finally, advancements in off-policy evaluation and uncertainty quantification will be necessary to enhance reliability of retrospective policy assessments. Potential areas for addi-

tional research include investigating hybrid approaches combining clinician-in-the-loop validation and simulation-based testing prior to any consideration of prospective deployment.

APPENDIX A

Hyperparameters

Table A.1: Key hyperparameters for offline reinforcement learning models

Hyperparameter	**BC**	**CQL**	**BCQ**
Seed	42	42	7
Batch size	256	256	256
Discount factor (γ)	–	0.99	0.99
Number of critics	–	2	2
Conservatism coefficient (α)	–	1.0	–
Target smoothing coefficient (τ)	–	0.005	–
Epochs	50	50	5
Steps per epoch	2000	2000	2000
Total gradient steps	100,000	100,000	10,000

A.1 Code and Data Availability

All code, configuration files, and final results used in this study are publicly available at: https://github.com/allenmchun/masds-thesis. The repository includes scripts for data preprocessing, model training, offline policy evaluation, and figure generation, along with configuration files and summarized results necessary to reproduce the analysis.

Due to data use restrictions, raw MIMIC-IV patient data are not included. Users must obtain access to MIMIC-IV independently and follow the provided pipeline to reproduce results.

A.2 Additional Figures

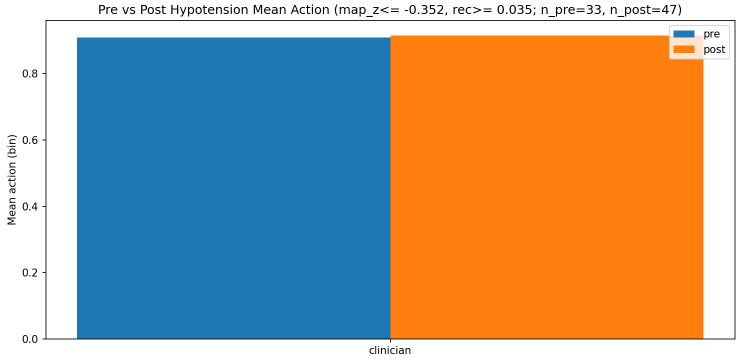

Figure A.1: Mean clinician action before and after hypotensive episodes. Vasopressor use increases only very slightly following hypotension, consistent with expected clinical response.

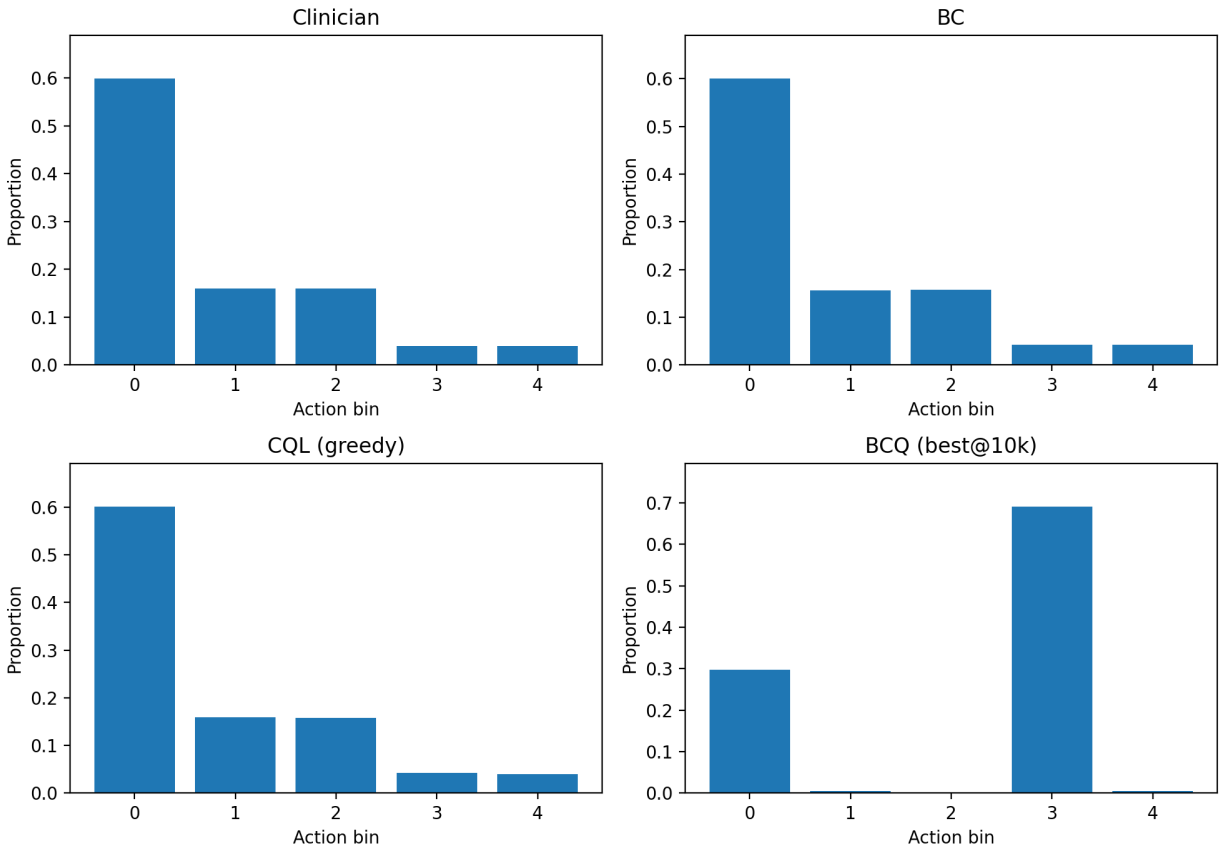

Figure A.2: Action distribution comparisons across clinician policy, Behavior Cloning (BC), Conservative Q-Learning (CQL), and Batch-Constrained Q-Learning (BCQ). Here, BC and CQL closely match clinician behavior, while BCQ exhibits stronger deviations under conservative constraints.

REFERENCES

[1] Richard S. Sutton and Andrew G. Barto. Reinforcement Learning: An Introduction. MIT Press, 2020. http://www.incompleteideas.net/book/RLbook2020.pdf

[2] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline Reinforcement Learning: Tutorial, Review, and Perspectives. arXiv preprint arXiv:2005.01643, 2020. https://arxiv.org/pdf/2005.01643v1

[3] White K. C., Quick L., Durkin Z., McCullough J., Laupland K. B., Blank S., Attokaran A. G., Kumar A., Shekar K., Garrett P., Meyer J., Tabah A., Ramanan M., Luke S., Chaba A., Bellomo R., Lamontagne F., and Young P. J. Mean arterial pressure in critically ill adults receiving vasopressors: A multicentre, observational study. 2025. https://pmc.ncbi.nlm.nih.gov/articles/PMC11938056/

[4] Laura Evans, Andrew Rhodes, Waleed Alhazzani, Massimo Antonelli, Coenraad L. A. Bleck, Mitchell Coopersmith, Craig M. Coopersmith, and others. Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock 2021. Intensive Care Medicine, 2021. https://pmc.ncbi.nlm.nih.gov/articles/PMC8486643/

[5] Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-Based Batch Mode Reinforcement Learning. Journal of Machine Learning Research, 2005. https://www.jmlr.org/papers/volume6/ernst05a/ernst05a.pdf

[6] Sergey Levine. Decisions from Data: How Offline Reinforcement Learning Will Change How We Use Machine Learning. Medium, 2019. https://medium.com/@sergey.levine/decisions-from-data-how-offline-reinforcement-learning-will-change-how-we-use-ml-24d98cb069b0

[7] Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc Bellemare. Safe and Efficient Off-Policy Reinforcement Learning. arXiv preprint arXiv:1906.00949, 2019. https://arxiv.org/pdf/1906.00949

[8] Alistair E. W. Johnson, Lucas Bulgarelli, Tom J. Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data, 2024. https://doi.org/10.13026/kpb9-mt58

[9] Faraz Torabi, Garrett Warnell, and Peter Stone. Behavioral Cloning from Observation. arXiv preprint arXiv:1805.01954, 2018. https://arxiv.org/pdf/1805.01954

[10] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-Learning for Offline Reinforcement Learning. arXiv preprint arXiv:2006.04779, 2020. https://arxiv.org/pdf/2006.04779

[11] Scott Fujimoto, David Meger, and Doina Precup. Off-Policy Deep Reinforcement Learning without Exploration. arXiv preprint arXiv:1812.02900, 2019. https://arxiv.org/pdf/1812.02900

[12] Doina Precup. Eligibility Traces for Off-Policy Policy Evaluation. Technical report, 2000. http://www.incompleteideas.net/papers/PSS-00.pdf

[13] Hongyi Zhou, Josiah P. Hanna, Jin Zhu, Ying Yang, and Chengchun Shi. Demystifying the Paradox of Importance Sampling with an Estimated History-Dependent Behavior Policy in Off-Policy Evaluation. arXiv preprint arXiv:2505.22492, 2025. https://arxiv.org/pdf/2505.22492

[14] Olivier Nachum, Yinlam Chow, Bo Dai, and Lihong Li. DualDICE: Behavior-Agnostic Estimation of Discounted Stationary Distribution Corrections. arXiv preprint arXiv:2005.05951, 2020. https://arxiv.org/pdf/2005.05951

[15] Botao Hao, Xiang Ji, Yaqi Duan, Hao Lu, Csaba Szepesvari, and Mengdi Wang. Bootstrapping Fitted Q-Evaluation for Off-Policy Inference. Proceedings of Machine Learning Research, 2021. https://proceedings.mlr.press/v139/hao21b.html

[16] Ruiqi Zhang, Xuezhou Zhang, Chengzhuo Ni, and Mengdi Wang. Off-Policy Fitted Q-Evaluation with Differentiable Function Approximators: Z-Estimation and Inference Theory. Proceedings of Machine Learning Research, 2022. https://proceedings.mlr.press/v162/zhang22al.html

[17] Takuma Seno and Michita Imai. d3rlpy: An Offline Reinforcement Learning Library. arXiv preprint arXiv:2111.03788, 2021. https://arxiv.org/pdf/2111.03788