

Dynamic Treatment Regime for Electronic Health Record



Zhiyao Luo

Supervised by: Prof. Tingting Zhu

Reuben College

University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy in Engineering Science

Trinity 2025

Dedication

To my beloved parents, Bing Luo and Limin Liu, for their unwavering love and support.

In loving memory of my grandparents, Laifu Luo and Shuqing Zuo, whose guidance and values inspired me to pursue healthcare as my DPhil journey.

To my wife, Wenhui Wu, for her endless encouragement and companionship.

To my supervisor, Prof. Tingting Zhu, for her mentorship and wisdom.

To my colleagues, labmates, and friends in Oxford, for their camaraderie and shared pursuit of knowledge.

And finally, to myself, for the perseverance and dedication that made this work possible.

Acknowledgements

As I write these final words, the bells of St Giles' Church drift faintly across the city, not from my lab in Headington, but from the quiet walks back to Reuben College after long days. Their sound marks the close of four extraordinary years — years lived in code and clinic, shaped by uncertainty, and quietly accompanied by the personal work of growing up.

My deepest gratitude belongs to my parents. Across time zones and borders, your video calls were a tether. Your familiar refrain — “Has AI cured anything yet?” — brought comfort and humour, even when my replies grew weary. In those questions lived something steady and enduring, a reminder that behind the abstractions stood your quiet, unwavering care.

If perseverance and quiet resolve came from my mother, and sincerity and thoughtfulness from my father, then from my grandmother I inherited something gentler: the gift of a light spirit. A teacher of psychology by calling, she lived by quiet truths. Her wisdom — sometimes in words, sometimes in silence — gave me an emotional steadiness no dataset could offer. Since her passing, not a day has gone by without her presence felt.

To you, who stayed up with me through countless nights — reminding me to drink water, or handing me those effervescent vitamin C tablets with a look that said everything — thank you. You turned the longest hours into something warm. I'll carry with me the memory of pigeons scattering in St Mark's Square, the quiet glow of the Chain Bridge in Budapest, the undulating stone of Casa Milà against the Catalan sky, and your soft humming beside me. When the organ sounds in the Sheldonian and I stand beneath the tasselled cap, it's your face I'll look for — the one who could assemble a bridal bouquet with pipette tips.

To my supervisor, thank you for your steady guidance, clear thinking, and belief that ideas can become something more, even when they begin as little more than questions.

These years have left their mark. I'll remember the glow of late-night code, the sting of rejection, the joy of quiet breakthroughs. I'll remember the intellectual debates, the human stories flattened into rows and features, and the discomfort of trying to do justice to lives through data. In all of it, I've come to see that humility isn't just helpful in medical AI — it's essential.

We may never fully grasp the intricacies of the human mind. But with enough care, we can build algorithms that, at the very least, offer understanding. And perhaps, even relief. That, I think, is worth striving for.

Abstract

Dynamic Treatment Regimes (DTRs) aim to personalise treatment through sequential decision policies that adapt to a patient’s evolving state. Reinforcement learning (RL) offers a data-driven approach for learning such policies, but key challenges remain unresolved, including how to reliably evaluate RL methods in clinical settings, how to incorporate meaningful clinical knowledge, and how to ensure generalisability across diverse patient populations.

This thesis addresses these challenges through a structured four-part investigation. First, we introduce DTR-Bench, a modular benchmarking platform that simulates four clinical tasks with noise, pharmacological variation, and missingness. We show that commonly used RL algorithms perform inconsistently under clinical variation.

Second, transitioning from in-silico to real clinical scenarios, we revisit the case of RL for sepsis treatment using retrospective ICU data. Our analysis standardises reward definitions and applies a spectrum of off-policy evaluation across patient subgroups. We advocate for stratification techniques to identify populations where RL may be most beneficial, and we emphasise benchmarking against supervised learning and heuristic baselines.

Third, we focus on intravenous insulin titration and prediction of glycaemia in general ICU patients, including those without diagnosed diabetes. Using a curated MIMIC-III dataset, we develop an ensemble evaluation framework for more robust off-policy evaluation and implement RL algorithms that demonstrate performance comparable to that of clinicians.

Finally, we explore large language models (LLMs) as a minimalist alternative to RL-based treatment recommendations. Within a simulated diabetes-control environment, we find that while RL-based incorporation of clinical priors typically requires extensive engineering, LLMs can effectively absorb structured clinical knowledge through prompting alone, sometimes matching or exceeding RL performance. Nevertheless, we identify clear failure modes in LLM reasoning, such as unit mismatches and unsafe dosing decisions, highlighting important limitations to clinical deployment.

Collectively, these investigations establish reproducible benchmarks, expose failure modes, and provide guidance for safer, more generalisable DTR algorithms. This work lays some essential methodological foundations for the future growth and evaluation of intelligent dynamic treatment algorithms.

Contents

| | |
|---|-------------|
| List of Figures | xiii |
| List of Abbreviations | xvii |
| List of Symbols and Notation | xix |
| 1 Introduction | 1 |
| 1.1 Fundamental Concepts | 2 |
| 1.1.1 Background of DTRs for EHRs: From Supervised Learning to Reinforcement Learning | 2 |
| 1.1.2 Simulation Matters: Building Controlled and Clinically Real- istic Benchmarks | 3 |
| 1.1.3 A Critical Look At Offline RL With Real Data | 4 |
| 1.1.4 Applying RL to a Real Clinical Task | 4 |
| 1.1.5 Towards Minimalist Dynamic Treatment Regimes: Exploring the Potential of Large Language Models | 5 |
| 1.2 Summary | 6 |
| 1.3 Dissertation Structure | 7 |
| 2 Literature Review | 9 |
| 2.1 Introduction | 9 |
| 2.2 Formulating Dynamic Treatment Regime under Markov Decision Process | 11 |
| 2.3 Imitation Learning and DTR Applications | 14 |
| 2.3.1 Imitation-Learning-Based DTR | 16 |
| 2.4 Reinforcement Learning and DTR Application | 17 |
| 2.4.1 Key Concepts and Algorithms in Reinforcement Learning . . | 18 |
| 2.4.2 Evaluation Metrics for Modelling Dynamic Treatment Regimes | 25 |
| 2.4.3 Reinforcement-Learning-Based DTR | 27 |
| 2.5 Open Issues in RL-Driven Dynamic Treatment Regimes | 32 |
| 3 Data | 35 |
| 3.1 MIMIC-III Sepsis Cohort | 36 |
| 3.1.1 Patient Stratification | 36 |
| 3.1.2 Action Discretisation | 38 |
| 3.2 MIMIC-III Insulin Administration Cohort | 40 |

| | | |
|----------|--|-----------|
| 4 | DTR-Bench: An <i>in silico</i> Environment and Benchmark Platform for Reinforcement Learning Based Dynamic Treatment Regime | 55 |
| 4.1 | Introduction | 55 |
| 4.2 | Simulation Models | 59 |
| 4.2.1 | <i>AhnChemoEnv</i> : A Comprehensive Chemotherapy Simulation Model | 60 |
| 4.2.2 | <i>GhaffariCancerEnv</i> : A Mixed Radiotherapy and Chemotherapy Model | 62 |
| 4.2.3 | <i>OberstSepsisEnv</i> : A Sepsis Simulator | 66 |
| 4.2.4 | <i>SimGlucoseEnv</i> : A glucose-insulin simulating environment | 68 |
| 4.3 | Results | 71 |
| 4.3.1 | Overview of Platform Capabilities | 72 |
| 4.3.2 | Benchmark Results | 72 |
| 4.4 | Discussion | 76 |
| 5 | A Critical Re-examination of Reinforcement Learning based Dynamic Treatment Regime | 79 |
| 5.1 | Introduction | 79 |
| 5.2 | Diversity and Inconsistency of Policy Evaluation | 79 |
| 5.3 | Reward Design Choices | 80 |
| 5.4 | Baselines Comparisons | 82 |
| 5.5 | Data and Experiment Setup | 83 |
| 5.5.1 | Implementation | 84 |
| 5.5.2 | Overall Comparison Results | 85 |
| 5.6 | Understanding the Variance of Policy Evaluation | 89 |
| 5.6.1 | Variance from Behavioural Policy | 90 |
| 5.6.2 | Variance from Value Function | 91 |
| 5.6.3 | A Closer Look at the Behavioural Model and Value Estimation Error | 91 |
| 5.6.4 | Investigating Model Calibration for Behavioural Model | 92 |
| 5.7 | Discussion on Considering RL for DTR | 95 |
| 5.8 | Conclusion | 98 |
| 6 | Towards Safer Insulin Dosing in the ICU: Hypoglycaemia Prediction and Offline Short-Acting Insulin Optimisation | 99 |
| 6.1 | Introduction | 99 |
| 6.2 | Methods | 101 |
| 6.2.1 | Data Curation and Preprocessing | 101 |
| 6.2.2 | Clinical Prediction Tasks on Hypoglycaemia and Insulin Dose | 101 |
| 6.2.3 | Behavioural Model Calibration | 103 |

| | | |
|----------|---|------------|
| 6.2.4 | Policy Extraction from Value-Based RL Algorithms | 104 |
| 6.2.5 | Model Selection and Evaluation Metrics | 104 |
| 6.2.6 | Feature Interpretability | 106 |
| 6.3 | Results for Hypoglycaemia Prediction | 106 |
| 6.3.1 | Prediction Performance | 106 |
| 6.3.2 | Sensitivity Analysis | 108 |
| 6.3.3 | Model Interpretation | 109 |
| 6.4 | Results for Insulin Recommendation | 109 |
| 6.4.1 | Behavioural Cloning for Clinician’s Insulin Administration Policy | 110 |
| 6.4.2 | Behavioural Model Calibration | 112 |
| 6.4.3 | Selection of Off-policy Evaluation By Naive Baseline Check | 114 |
| 6.4.4 | Reinforcement Learning for Insulin Administration | 117 |
| 6.5 | Discussion | 118 |
| 7 | Are Large Language Models Dynamic Treatment Planners? An In Silico Study from a Prior Knowledge Injection Angle | 123 |
| 7.1 | Introduction | 123 |
| 7.2 | Problem Formulation and Environment Setting | 124 |
| 7.3 | Reward and Evaluation Protocol | 125 |
| 7.4 | Experimental Setting | 127 |
| 7.4.1 | Small Reinforced Agents in Insulin Dosing | 127 |
| 7.4.2 | LLMs in Insulin Dosing | 129 |
| 7.5 | Results | 132 |
| 7.5.1 | The effectiveness of Prior Knowledge Injection on SRAs | 132 |
| 7.5.2 | Comparison between LLMs and SRAs | 134 |
| 7.5.3 | Analysis of LLM Treatment Behaviour | 135 |
| 7.6 | Failure Modes of Chain-of-Thought Reasoning | 142 |
| 7.6.1 | Case Study 1: Reasoning for Initial Dosage | 142 |
| 7.6.2 | Case Study 2: Reasoning on Glucose Trends and Delayed Effects | 143 |
| 7.6.3 | Summary of Findings on Case Studies | 145 |
| 7.7 | Discussion | 146 |
| 8 | Conclusion | 149 |
| 8.1 | Summary of Results | 149 |
| 8.1.1 | An <i>in silico</i> Environment and Benchmark Platform for Rein- forcement Learning Based Dynamic Treatment Regime | 149 |
| 8.1.2 | A Critical Re-examination of Reinforcement Learning-Based Dynamic Treatment Regime | 150 |

| | | |
|-------------------|--|------------|
| 8.1.3 | Towards Safer Insulin Dosing in the ICU: Hypoglycaemia Prediction and Offline Insulin Optimisation | 151 |
| 8.1.4 | Evaluation of Large Language Models in Dynamic Treatment Regimes | 152 |
| 8.2 | Methodological Contributions | 153 |
| 8.3 | Limitations | 154 |
| 8.4 | Future Work | 155 |
| 8.5 | Closing Remarks | 156 |
| Appendices | | |
| A | Supplementary Results for Chapter 4 | 159 |
| B | Supplementary Results for Chapter 5 | 165 |
| B.1 | Behavioural Policy and Value Visualisation for the training and validation set | 165 |
| B.2 | Full Evaluation Result on Stratified Patient Groups | 166 |
| B.3 | Importance Ratio Histogram of Naive Baselines | 166 |
| C | Supplementary Results for Chapter 6 | 179 |
| C.1 | Reward Definition and Distributional Properties | 179 |
| C.2 | Training Details | 179 |
| C.3 | Integrated Gradients Implementation | 180 |
| C.4 | Per-class Reliability Plot for Behavioural Model Calibration | 181 |
| D | Supplementary Results for Chapter 7 | 185 |
| D.1 | Action Transformation in Continuous Action Space: A Comparative Analysis of clip and tanh | 185 |
| D.1.1 | Theoretical Properties of Tanh Transformation | 185 |
| D.1.2 | Theoretical Limitations of Clip Transformation | 186 |
| D.1.3 | Practical Considerations | 187 |
| D.2 | Heuristic Decision for Exploration Strategy in Discrete Action Space | 187 |
| D.3 | Hyperparameters for Small Reinforced Agent Training | 188 |
| References | | 191 |

List of Figures

| | | |
|-----|--|-----|
| 2.1 | Comparing RL and IL in DTR. | 14 |
| 2.2 | Demonstration of Failure Correction in Behaviour Cloning | 15 |
| 3.1 | Data Stratification using rate of change in NEWS2 score and admission outcome. | 39 |
| 3.2 | Summary of cohort construction and example patient data. | 41 |
| 4.1 | Workflow of the DTR-Bench platform. | 56 |
| 4.2 | A summary of RL algorithms and environments in the DTR-Bench platform. | 58 |
| 5.1 | Number of wins for each policy in the (overall) test set. | 87 |
| 5.2 | A summed number of wins across patient subgroups stratified by mortality risk rate of change. | 88 |
| 5.3 | behavioural and value estimators versus their losses on the testing set. | 90 |
| 5.4 | Loss histograms on different reward settings. | 93 |
| 5.5 | Comparison of output probability between calibrated and uncalibrated $\hat{\pi}_{\mathcal{D}}$ on the training, validation and test set. | 94 |
| 5.6 | Importance ratio histogram of random policy $>$ 99th percentile. | 95 |
| 5.7 | Comparison of output probability between calibrated and uncalibrated $\hat{\pi}_{\mathcal{D}}$ | 96 |
| 6.1 | Performance comparison of nine supervised learning models for hypoglycaemia forecasting. | 107 |
| 6.2 | Sensitivity analyses for hypoglycaemia prediction. | 108 |
| 6.3 | Integrated Gradients attribution heatmap for hypoglycaemia prediction on the test set. | 110 |
| 6.4 | Behavioural cloning performance with and without the previous insulin administration included as an input feature. | 111 |
| 6.5 | Reliability diagrams before and after class-wise temperature scaling. | 113 |
| 6.6 | Evaluation of reinforcement learning policies using ensemble bootstrapped WIS, ensemble bootstrapped WIS lower bound and macro F1 score. | 117 |
| 6.7 | Insulin dosing behaviour of reinforcement learning policies. | 117 |
| 7.1 | Overall normalised return of SRA on the evaluation set for DQN and PPO. | 132 |

| | | |
|------|--|-----|
| 7.2 | LLMs’ normalised return performance compared with the best SRA model | 135 |
| 7.3 | Comparison of insulin dosing performance across LLM families (Qwen2.5 and LLaMA3) for different temperature settings. | 135 |
| 7.4 | Impact of different prompting methods on insulin dosing for Qwen2.5 and LLaMA3 families for prompt settings: base zero-shot prompt (w/o prior), prior knowledge zero-shot prompt (w/ prior) and prior knowledge zero-shot-CoT prompt (w/ prior+CoT). | 136 |
| 7.5 | Detailed analysis of CoT prompting effects on insulin dosing performance within the Qwen2.5 family. | 137 |
| 7.6 | Effects of incorporating meal intake information into CoT prompts on insulin dosing performance. | 138 |
| 7.7 | Insulin administration, glucose and risk behaviours for Qwen2.5-7B model under CoT and zero-shot prompting with prior knowledge system prompt. | 140 |
| 7.8 | Insulin administration as a function of current observed glucose measurement for Qwen2.5-7B model under CoT and zero-shot prompting with prior knowledge system prompt. | 140 |
| 7.9 | Insulin administration, glucose and risk behaviours for Qwen2.5-72B model under CoT and zero-shot prompting with prior knowledge system prompt. | 141 |
| 7.10 | Insulin administration as a function of current observed glucose measurement for Qwen2.5-72B model under CoT and zero-shot prompting with prior knowledge system prompt. | 141 |
| 7.11 | Comparing Qwen2.5 7B and 72B model response on a child patient on step 0 under meal CoT prompt with prior knowledge. | 142 |
| 7.12 | Comparing Qwen2.5 7B and 72B model response on a child patient on step 7 under meal CoT prompt with prior knowledge. | 144 |
| B.1 | Cross entropy and L1 losses vs inference probabilities or Q estimates in the training set. | 165 |
| B.2 | Cross entropy and L1 losses vs inference probabilities or Q estimates in the validation set. | 166 |
| B.3 | Importance ratio histogram of random policy > 99th percentile on Outcome reward. | 173 |
| B.4 | Importance ratio histogram of min policy > 99th percentile on Outcome reward. | 173 |
| B.5 | Importance ratio histogram of max policy > 99th percentile on Outcome reward. | 173 |

| | | |
|------|--|-----|
| B.6 | Importance ratio histogram of alt policy > 99th percentile on Outcome reward. | 174 |
| B.7 | Importance ratio histogram of weight policy > 99th percentile on Outcome reward. | 174 |
| B.8 | Importance ratio histogram of random policy > 99th percentile on SOFA reward. | 174 |
| B.9 | Importance ratio histogram of min policy > 99th percentile on SOFA reward. | 175 |
| B.10 | Importance ratio histogram of max policy > 99th percentile on SOFA reward. | 175 |
| B.11 | Importance ratio histogram of alt policy > 99th percentile on SOFA reward. | 175 |
| B.12 | Importance ratio histogram of weight policy > 99th percentile on SOFA reward. | 176 |
| B.13 | Importance ratio histogram of min policy > 99th percentile on NEWS2 reward. | 176 |
| B.14 | Importance ratio histogram of max policy > 99th percentile on NEWS2 reward. | 176 |
| B.15 | Importance ratio histogram of alt policy > 99th percentile on NEWS2 reward. | 177 |
| B.16 | Importance ratio histogram of weight policy > 99th percentile on NEWS2 reward. | 177 |
| C.1 | Visualisation of the glucose risk index and corresponding reward transformation. | 180 |
| C.2 | Histogram of maximum glucose risk per episode. | 181 |
| C.3 | Distribution of cumulative returns derived from the reward function across all episodes. | 181 |
| C.4 | Per-class reliability diagrams before and after class-wise temperature scaling. | 183 |
| D.1 | Exploration probabilities and expected exploratory insulin dose as a function of discretisation size $ \mathcal{A} $ | 188 |

List of Abbreviations

| | | | |
|----------------------|---|---------------------------------|--|
| BC | Behavioural Cloning | TIR | Time-in-Range |
| BCQ | Batch-Constrained Deep Q-Learning | WIS | Weighted Importance Sampling |
| BEAR | Bootstrapping Error Accumulation Reduction | ALT | Alanine Aminotransferase |
| C51 | Categorical Deep Q-Network | AST | Aspartate Aminotransferase |
| CNN | Convolutional Neural Network | aPTT | Activated Partial Thromboplastin Time |
| CGM | Continuous Glucose Monitoring | BMI | Body Mass Index |
| CQL | Conservative Q-Learning | BUN | Blood Urea Nitrogen |
| DDPG | Deep Deterministic Policy Gradient | CHO | Carbohydrate |
| DDQN | Double Deep Q-Network | CRP | C-Reactive Protein |
| DM | Direct Method | FiO₂ | Fraction of Inspired Oxygen |
| DQN | Deep Q-Network | ICD-9 | International Classification of Diseases, 9th Revision |
| DR | Doubly Robust | ICU | Intensive Care Unit |
| DTR | Dynamic Treatment Regime | IV | Intravenous |
| EHR | Electronic Health Record | NSCLC | Non-Small Cell Lung Cancer |
| IL | Imitation Learning | PaCO₂ | Partial Pressure of Carbon Dioxide (Arterial) |
| IQL | Implicit Q-Learning | PaO₂ | Partial Pressure of Oxygen (Arterial) |
| MAP | Mean Arterial Pressure | PT | Prothrombin Time |
| MDP | Markov Decision Process | SD | Standard Deviation |
| ML | Machine Learning | SIRS | Systemic Inflammatory Response Syndrome |
| OPE | Off-Policy Evaluation | SOFA | Sequential Organ Failure Assessment |
| PID | Proportional–Integral–Derivative | SpO₂ | Peripheral Capillary Oxygen Saturation |
| POMDP | Partially Observable Markov Decision Process | TPN | Total Parenteral Nutrition |
| PPO | Proximal Policy Optimisation | WBC | White Blood Cell (Count) |
| RL | Reinforcement Learning | | |
| SAC | Soft Actor-Critic | | |
| TD3 | Twin Delayed Deep Deterministic Policy Gradient | | |

List of Symbols and Notation

| | |
|----------------------------|---|
| i | Patient (trajectory) index |
| t | Discrete time-step index |
| H | Episode horizon (maximum trajectory length) |
| H_i | Length of the i -th trajectory |
| γ | Discount factor, $\gamma \in [0, 1)$ |
| \mathcal{M} | Markov Decision Process (MDP) $\langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle$ |
| \mathcal{M}_{PO} | Partially Observable MDP (POMDP) $\langle \mathcal{S}, \mathcal{A}, P, R, \gamma, \mathcal{O}, Z \rangle$ |
| \mathcal{S} | Latent state space |
| s | Latent state variable, $s \in \mathcal{S}$ |
| s_t^i | State of patient i at time t |
| s' | Next latent state |
| \mathcal{A} | Action (treatment) space |
| a | Action variable, $a \in \mathcal{A}$ |
| a_t^i | Action for patient i at time t |
| a' | Next-step action used in Bellman targets |
| \mathcal{O} | Observation space |
| o | Observation variable, $o \in \mathcal{O}$ |
| o_t^i | Observation of patient i at time t |
| $P(s' s, a)$ | State-transition probability |
| $R(s, a)$ | Expected immediate reward |
| $Z(o s, a)$ | Observation model in a POMDP |
| $\pi(a s)$ | Evaluation policy |
| π^* | Optimal policy |
| $\beta(a s)$ | Behaviour policy that generated the data |
| $\hat{\beta}(a s)$ | Behaviour-policy estimate from data |
| $\mu(s; \theta^\mu)$ | Deterministic actor in DDPG/TD3 |
| $V^\pi(s)$ | State-value function under π |

| | |
|----------------------------|---|
| $Q^\pi(s, a)$ | Action-value (Q) function under π |
| $V^*(s)$ | Optimal value function |
| $Q^*(s, a)$ | Optimal Q-function |
| \mathcal{D} | Offline dataset of transitions (s, a, r, s') |
| r | Reward realised from the environment |
| r_t^i | Reward at time t for patient i |
| G_i | Discounted return of trajectory i |
| N | Total number of trajectories in \mathcal{D} |
| ρ_i | Cumulative importance ratio of trajectory i |
| η_t | Per-step importance ratio up to time t |
| B | Number of bootstrap resamples |
| N_b | Trajectories in the b -th bootstrap sample |
| \mathcal{B} | Bootstrap sampling distribution |
| τ | Truncation threshold for importance weights (or expectile level in IQL) |
| θ | Generic network parameters (value/Q) |
| θ^- | Target-network parameters (DQN family) |
| θ^Q | Critic parameters in actor–critic methods |
| θ^{Q^-} | Target-critic parameters |
| θ^μ | Actor parameters in DDPG/TD3 |
| θ^{μ^-} | Target-actor parameters |
| ϕ | Parameters of a stochastic policy (e.g. SAC) |
| α | Entropy-temperature parameter (SAC or CQL) |
| $d(\cdot, \cdot)$ | Divergence metric (e.g. MMD, KL) |
| ϵ | Upper bound on $d(\pi, \beta)$ (BEAR constraint) |
| $Z_{s,a}$ | Return-distribution random variable |
| $\{z_i\}_{i=1}^N$ | Support atoms in C51 |
| $\mathbb{I}[\cdot]$ | Indicator function |
| $ \mathcal{A} $ | Cardinality of the discrete action set |
| $\rho_\tau(x)$ | Expectile loss in IQL |
| \mathcal{L} | Supervised loss in imitation learning |
| \mathbb{E} | Expectation operator |
| $\mathcal{P}(\cdot)$ | Probability distribution over \cdot |

Relevant Publications to the Thesis

- **Luo Z**, Zhu M, Liu F, Li J, Pan Y, Zhou J, Zhu T. Dtr-bench: An in silico environment and benchmark platform for reinforcement learning based dynamic treatment regime. arXiv preprint arXiv:2405.18610. 2024 May 28. [Under Review at TPAMI]
- **Luo Z**, Pan Y, Watkinson P, Zhu T. Position: reinforcement learning in dynamic treatment regimes needs critical reexamination. Spotlight poster, forty-first International Conference on Machine Learning
- **Luo Z**, Azam T, Qian S, Zhu T. A Deep Learning Benchmark for Hypoglycaemia Event Prediction in the Intensive Care Unit. [*In submission*]
- **Luo Z**, Qian S, Pan Y, Zhou J, Pan Y, Zhu T. A more Robust Offline Reinforcement Learning System for Short-Acting Insulin Titration in Intensive Care Unit. [*In submission*]
- **Luo Z**, Zhu T. Are Large Language Models Dynamic Treatment Planners? An In-silico Study from A Prior Knowledge Injection Angle. [*In submission*]

Other Publications via Collaborations

- Chou OH, **Luo Z**, Chung CT, Chan J, Li H, Lakhani I, Lee S, Lau DH, Zhang Q, Liu T, Wong WT. Comparison of New-Onset Peripheral Artery Disease in Patients With Type 2 Diabetes Exposed to Sodium–Glucose Cotransporter-2 Inhibitors, Dipeptidyl Peptidase-4 Inhibitors, or Glucagon-Like Peptide-1 Agonists: A Population-Based Cohort Study. Journal of the American Heart Association. 2025 Jun 3;14(11):e034175.
- **Luo Z**, Chou O, Ng Z, To CC, Chan J, Chan R, Lu L, Zhu T, Lee Q, McCEniery C, Wilkinson I. GLP1 receptor agonist with add-on SGLT2 inhibitor therapy is associated with lower risks of major adverse cardiovascular events: a population-based and machine learning causal inference analysis. [2nd revision at NPJ Cardiovascular Health]
- **Luo Z**, Yang J, Zhu T, Wong WC, Zhou J. Language model-based early detection of colorectal cancer occurrence from previous disease diagnosis trajectory data for primary care purpose. The Lancet Regional Health–Western Pacific. 2025 Feb 1;55.

- **Luo Z**, Watkinson P, Zhu T. NurSpecialist: Duel-Agent Reinforcement Learning for Dynamic Hospitalised Intervention Regimes using Electronic Health Records. AAAI 2023-Workshop on Artificial Intelligence for Social Good
- Liu H, **Luo Z**, Zhu T. Best of both worlds: Harmonizing LLM capabilities in decision-making and question-answering for treatment regimes. Advancements In Medical Foundation Models: Explainability, Robustness, Security, and Beyond 2024.

1

Introduction

The modern medical world increasingly recognises that no single treatment is suitable for all. For patients with complex and evolving conditions such as sepsis, diabetes, or cancer, effective care requires a dynamic response that adapts to physiological changes and treatment responses over time. This requirement calls for the concept of *dynamic treatment regimes* (DTRs), which refers to timely decision strategies that personalise treatment based on the current and previous state of the patient.

Dynamic treatment regimes have long been used in clinical practice, often implicitly encoded in the form of rule-based protocols or decision trees derived from expert consensus and population-level evidence [1]. These traditional approaches function as "look-up tables", offering fixed guidance for treatment decisions based on predefined patient characteristics. Although effective in standardised care, such protocols are inherently limited in their ability to adapt to individual patient trajectories or to account for the cumulative impact of sequential decisions.

Recent advances in machine learning have created new opportunities to formalise and extend DTRs beyond rule-based protocols. By leveraging longitudinal data from Electronic Health Records (EHRs), data-driven DTRs can support individualised decision-making that evolves in real time. This shift from rigid, population-based rules toward adaptive, patient-specific strategies marks a significant step forward in the pursuit of personalised medicine, particularly in complex, high-stakes settings such as the intensive care unit, where clinical conditions change rapidly and timely interventions are crucial.

1.1 Fundamental Concepts

1.1.1 Background of DTRs for EHRs: From Supervised Learning to Reinforcement Learning

EHRs have become a cornerstone of modern clinical practice [2], offering longitudinal, patient-specific data that include diagnoses, laboratory results, medications, procedures, and vital signs. These records provide a rich temporal view of the patient’s clinical journey and offer a natural substrate for modelling sequential decision-making in healthcare. The dynamic and often non-linear progression of a patient’s health, coupled with sequential treatment decisions, makes EHRs especially well-suited for learning dynamic treatment regimes. Dynamic treatment regimes aim to personalise treatment strategies over time by adapting to the evolving state of the patient.

With the development of machine learning, most clinical prediction tasks based on EHR data have been addressed using supervised learning (SL) techniques [3]. Although effective for static classification or risk prediction, SL methods face fundamental limitations in the context of sequential treatment planning. One major constraint is the assumption of independent and identically distributed (i.i.d.) data points [4]. In practice, different clinicians may follow distinct heuristics or strategies when treating similar patients, leading to a mixture of behaviour policies. Learning a single SL model over such heterogeneous data is equivalent to approximating a joint distribution over all clinicians’ actions, which is not only difficult but also clinically ambiguous. Moreover, SL models typically imitate clinical behaviours without understanding the ‘goal of action’, which prevents correction from mistakes or adjusting strategies over time.

In contrast, reinforcement learning (RL) is designed to optimise sequential decision-making by directly modelling the interaction between actions and future outcomes. Formulating treatment planning as a Markov Decision Process (MDP), RL enables the learning of policies that maximise cumulative reward over time. However, the theoretical advantages of RL are often difficult to translate into

real-world healthcare applications due to data constraints, ethical concerns related to exploration, and the complexity of partially observed environments.

1.1.2 Simulation Matters: Building Controlled and Clinically Realistic Benchmarks

Given the complexity of real-world clinical data, characterised by irregular sampling, missing values, partial observability, and treatment selection bias, evaluating the performance of RL in healthcare settings presents formidable challenges. Simulation-based evaluation offers a valuable alternative. By constructing synthetic environments where the data-generating process is known and fully controllable, researchers can systematically investigate the behaviour of RL algorithms under various conditions.

Existing simulation-based RL benchmarks in healthcare often fall short in two critical ways. First, they are typically too simplified, relying on toy problems that fail to capture the clinical realism needed for a meaningful evaluation. Second, many simulations lack modularity, making it difficult to introduce real-world complexities incrementally, such as pharmacokinetic variability, physiological noise, data missingness, or heterogeneous subgroups of patients.

This thesis addresses these limitations by introducing a modular simulation framework specifically designed for dynamic treatment planning in healthcare. The framework is designed to balance control and realism, allowing researchers to test RL agents from toy environments to more clinically meaningful scenarios. Each component of the environment (e.g., state dynamics, observation models, and reward design) can be modified independently, enabling fine-grained evaluation of algorithmic robustness, generalisability, and failure modes. We provide simulations covering four treatment scenarios, i.e., chemotherapy, radiotherapy, type-1 diabetic glucose management, and sepsis, collected from clinically validated literature. This benchmark is essential not only for understanding where and why RL methods succeed or fail but also serves as a critical stepping stone toward safe and reliable deployment in real-world settings.

1.1.3 A Critical Look At Offline RL With Real Data

Transitioning from controlled simulations to real-world clinical data introduces new challenges for reinforcement learning. In DTR, exploring new policies directly in practice is often risky for patients and is constrained by ethical barriers. As a result, we must rely on offline RL, where decision policies are learnt from observational EHR data rather than through environmental interactions.

A central challenge in offline settings is policy evaluation. Without the ability to deploy policies online, off-policy evaluation (OPE) techniques are required to estimate performance using offline data. In sparse reward and data-limited settings, typical of EHR data, OPE estimators can be sensitive to variance and bias. Furthermore, the broader literature on RL in healthcare reflects additional challenges. Studies often differ in their definitions of rewards, evaluation metrics, cohort inclusion criteria, and baselines. This methodological heterogeneity makes it difficult to draw generalisable conclusions or compare results across studies.

To address these limitations, the second part of this thesis undertakes a systematic and critical evaluation of offline RL methods using real-world clinical data. Focusing on a large-scale ICU cohort for sepsis management, we compare a range of RL algorithms and OPE estimators under unified preprocessing pipelines and consistent evaluation protocols. Special emphasis is placed on the design of reward functions, estimation of behaviour policies, and the role of subgroup heterogeneity in shaping policy performance. Through this analysis, we develop stronger methodological baselines, analyse the deeper reasons for performance inconsistency, and propose practical guidelines to guide future research.

1.1.4 Applying RL to a Real Clinical Task

Building on the insights gained from simulation studies and critical evaluation of offline RL pipelines, this thesis next focuses on the practical implementation of RL on short-acting insulin titration in the ICU.

Glycaemic control in ICU patients is a high-stakes, time-sensitive task. Insulin dosing must be frequently adjusted in response to fluctuating blood glucose levels,

nutritional intake, and physiological instability. Suboptimal decisions, through either under- or over-dosing, can result in hyperglycaemia, hypoglycaemia, or even an increased risk of mortality. Importantly, this task involves a relatively well-defined action space (i.e., dosing decisions), dense feedback (i.e., glucose measurements), and a clear clinical objective (i.e., avoiding hypoglycaemia while maintaining glucose within the target range), making it a strong scenario for RL-based treatment regimes.

To support this investigation, a high-resolution dataset was curated from a real ICU cohort, capturing insulin administration records and accompanying glucose measurements at fine-grained temporal intervals. To enhance the reliability of policy evaluation, we introduce a bootstrapped ensemble approach to off-policy evaluation, which combines multiple estimators and uncertainty quantification techniques to improve robustness. This ensemble framework enables us to assess policy performance with greater confidence and identify clinically significant differences across models. This work serves as a proof-of-concept for real-world RL-based DTR. More broadly, it highlights how carefully constructed clinical tasks combined with principled RL design can produce actionable insights toward practical utility.

1.1.5 Towards Minimalist Dynamic Treatment Regimes: Exploring the Potential of Large Language Models

The implementation of RL for DTR offers a powerful but complex solution to sequential decision-making. Recent advances in large language models (LLMs) suggest the possibility of a more minimalist approach to dynamic treatment planning. Trained on vast corpora of general text, LLMs have demonstrated impressive capabilities in in-context reasoning, question answering, and even clinical decision support. Crucially, they can perform tasks without task-specific fine-tuning through appropriate prompting. This motivates an exploratory investigation: *Can LLMs serve as zero-shot treatment planners?* More ambitiously, can they operate completely training-free, replacing the need for RL in some scenarios?

To evaluate the potential of LLMs, we propose a comparative investigation between small RL agents and general-purpose LLMs. Specifically, our objective

was to compare generalisation performance across various patient subgroups and to determine whether the integration of prior clinical knowledge leads to a performance gain in both small RL agents and LLMs. The exploration of LLMs as treatment planners challenges us to rethink the boundaries between knowledge-based reasoning and experience-based learning, and to consider how these approaches might be combined to build AI systems that are not only powerful but also safe, transparent, and clinically aligned.

1.2 Summary

This thesis aims to advance the methodological foundations and empirical evaluation of AI-driven dynamic treatment regimes. It is guided by four objectives: **1)** To construct clinically realistic in-silico environments that expose RL algorithms to pharmacological variability, sensor noise, and incompleteness of the data characteristic of real-world healthcare. **2)** To evaluate offline RL methods in clinical datasets with an emphasis on reward specification, evaluation pipelines, and heterogeneity of patient subgroups. **3)** To develop RL-based DTR on a new clinical problem toward more robust model development and evaluation. **4)** To explore whether LLMs, prompted without task-specific fine-tuning, can recommend dynamic treatment regimes and characterise the conditions under which they succeed or fail. To support these objectives, the thesis presents several technical contributions.

- A modular simulation platform is developed to model disease progression across multiple clinical domains, with configurable support for pharmacokinetic/pharmacodynamic (PK/PD) variability, noise, and data missingness.
- A comparative study is conducted in a sepsis ICU cohort, evaluating a variety of RL models and off-policy evaluation methods to identify failure modes in offline learning and evaluation.
- A high-resolution dataset is curated for short-acting insulin titration in the ICU. A bootstrapped ensemble approach to off-policy evaluation is proposed

to improve the robustness of policy assessment. Offline-trained RL agents are shown to outperform historical clinician policies in terms of return.

- Methods for injecting prior clinical knowledge are developed for both LLMs and RL agents. We evaluated these approaches in a simulated insulin management task, comparing the performance of LLMs against RL agents trained offline. LLMs are shown to function as zero-shot treatment planners, outperforming RL agents in certain subgroups of patients. Additionally, we analyse error patterns in LLM reasoning to better understand their limitations.

Overall, these contributions establish a reproducible framework for designing, evaluating, and interpreting AI-based treatment strategies in real clinical settings.

1.3 Dissertation Structure

The remainder of this thesis is organised as follows. Chapter 2 reviews the theoretical foundations of dynamic treatment regimes and their clinical applications. Chapter 3 describes the construction of real-world cohorts for sepsis and glycaemic control using the MIMIC-III database. Chapter 4 presents a new simulation framework for benchmarking RL methods with increasing clinical realism. Chapter 5 critically re-evaluates RL policies for sepsis management, highlighting how off-policy estimators, behaviour policy quality, and reward design can distort apparent gains. Chapter 6 focuses on short-acting insulin dosing, implementing RL algorithms that outperform clinicians' returns under a more robust evaluation metric. Chapter 7 explores the potential of LLMs as zero-shot treatment planners through the injection of prior knowledge. Chapter 8 synthesises the findings, outlines limitations, and proposes future directions toward safe, robust, and clinically aligned DTR systems. Ultimately, this thesis seeks not only to optimise algorithms, but also to ask better questions. Through rigorous simulation, real-world evaluation, and exploratory LLM experimentation, it provides a structured foundation for advancing patient-centred AI in dynamic treatment planning.

2

Literature Review

2.1 Introduction

In recent years, the increasing digitisation of clinical workflows has enabled widespread adoption of Electronic Health Records (EHRs), which systematically capture longitudinal health information, including demographics, laboratory results, vital signs, diagnoses, and prescribed treatments. These data sources have generated a surge of interest in applying Machine Learning (ML) to support clinical decision-making, with demonstrated success in predictive tasks such as in-hospital mortality [5], length of stay [6], and phenotype identification [7]. Beyond risk stratification and diagnostic prediction, a critical and more complex objective in healthcare is to determine how treatment decisions should evolve, and how to personalise care dynamically based on the evolving clinical condition of the patient. This need is particularly acute in hospital settings, where temporally dependent decisions must be made about drug administration, dosage adjustments, or intervention time. The formalism that captures such sequential decision-making processes is known as the Dynamic Treatment Regime (DTR). A DTR comprises a sequence of decision rules that recommend treatments tailored to the patient’s current and past clinical status to optimise long-term outcomes.

Historically, medical treatment strategies followed a “one-size-fits-all” paradigm, where interventions were evaluated under the framework of evidence-based medicine applied to population-level cohorts. The primary objective was to estimate the

average treatment effect across a target population, often overlooking inter-individual variation. However, with the advent of high-throughput data sources, the medical field has experienced a paradigm shift toward leveraging individual-level information to inform treatment decision-making. This shift underpins the broader movement towards precision medicine, of which DTRs are a central component. DTRs have traditionally been formulated using statistical modelling approaches, including regression-based frameworks [8] and marginal structural models [9]. These methods offer interpretable decision rules and have laid a strong theoretical foundation. However, they often struggle to scale with high-dimensional, irregular, and noisy data, which is typical of real-world EHRs, and may not flexibly capture long-term dependencies or complex interactions between variables.

Recent advances in ML, particularly Reinforcement Learning (RL) and Imitation Learning (IL), have introduced new paradigms for learning DTRs from retrospective data. RL models aim to discover treatment policies that maximise cumulative clinical benefit (often framed as a reward function), while IL seeks to replicate clinicians' historical decisions by learning from logged data. These approaches are well-suited to model the sequential and dynamic nature of treatment decisions, and their application to healthcare has generated substantial interest. However, adapting these algorithms to the clinical domain introduces several methodological and practical challenges, such as data sparsity, confounding, partial observability, and the ethical constraints inherent to healthcare.

This review of the literature surveys the state of the art in ML-based approaches to DTRs, with a particular focus on applications to hospital care. The objective of this chapter is to provide a critical synthesis of existing methods, highlight their limitations, and identify open challenges that need to be addressed to enable safe and effective deployment of such models in clinical practice. In Section 2.2, we begin by formally defining the DTR problem as a sequential decision-making task and highlighting its unique challenges in contrast to other ML applications. Section 2.3 and 2.4 introduce the mathematical formulation of DTRs using Markov Decision Processes (MDPs), reviewing two core ML methodologies: imitation learning and

reinforcement learning, with an emphasis on how they have been adapted to the healthcare setting. In Section 2.3.1 and Section 2.4.3, we summarise representative applications of IL and RL in various clinical contexts, such as critical care, oncology, and chronic disease management. Finally, Section 2.5 discusses key open issues, including the challenges of policy evaluation, reward design, and generalisation in healthcare. Through this review, our objective is to clarify the current landscape of DTR research at the intersection of ML and medicine, assess its readiness for real-world deployment, and identify promising directions for future research.

2.2 Formulating Dynamic Treatment Regime under Markov Decision Process

A DTR can be modelled as a Markov Decision Process (MDP) [10] (i.e., Markovian DTR). An MDP is defined by a 5-tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, R, \gamma \rangle$, where:

- \mathcal{S} : the *state space*, representing all possible states $s \in \mathcal{S}$ that describe the agent's condition at a given time t . The state space can be either discrete or continuous.
- \mathcal{A} : the *action space*, representing all possible actions $a \in \mathcal{A}$ the agent may take at each timestep t . Actions influence transitions to future states and determine the rewards received.
- $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$: the *state transition probability function*, where $P(s'|s, a)$ denotes the probability of transitioning to state $s' \in \mathcal{S}$ after taking action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$.
- $R : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$: the *reward function*, where $R(s, a)$ denotes the expected immediate reward received after the action a in state s .
- $\gamma \in [0, 1)$: the *discount factor*, which determines the weight of future rewards relative to immediate rewards. Common values range from 0.95 to 0.99.

We denote the episode horizon by H , where the time step $t \in \{1, 2, 3, \dots, H\}$, and allow trajectory-specific (i.e., patient-specific) lengths H_i when working with variable-length episodes.

The MDP framework assumes the Markov property, which implies that the future state and reward depend only on the current state and action, not on the full history:

$$\begin{aligned} &P(s_{t+1} = s', r_t = r \mid s_t = s, a_t = a) \\ &= P(s_{t+1} = s', r_t = r \mid s_t = s, a_t = a, s_{t-1}, a_{t-1}, r_{t-1}, \dots, s_0, a_0). \end{aligned} \quad (2.1)$$

This assumption simplifies the learning process, as the agent can access to full state information by observing the current state. However, in DTR applications, particularly with observational data such as EHRs, the agent may not have access to the full underlying state s_t . To address this partial observability, the Partially Observable Markov Decision Process (POMDP) is introduced. A POMDP extends the MDP tuple to include an observation mechanism, resulting in a seven-tuple.

$$\mathcal{M}_{PO} = \langle \mathcal{S}, \mathcal{A}, P, R, \gamma, \mathcal{O}, Z \rangle$$

where: \mathcal{O} is the *observation space*, containing all possible observations $o \in \mathcal{O}$ available to the agent at each timestep. $Z : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{P}(\mathcal{O})$ is the *observation model*, a stochastic mapping from state-action pairs to distributions over observations. $Z(o|s, a)$ denotes the probability of observing o after taking action a in the state s .

In healthcare practice, the agent must infer a latent state s_t from an observation o_t , since the true state is not directly observable. Learning an accurate representation of the hidden state space or observation model is critical for modelling treatment effectiveness and decision-making.

Given that agents often operate under partial observability of the environment, it becomes essential not only to model the latent state space effectively but also to determine how the agent should act based on its belief or estimate of this hidden state. This leads to the concept of a *policy*, which governs the agent's decision-making process under uncertainty. A policy π defines the agent's behaviour: It is a mapping $\pi : \mathcal{S} \rightarrow \mathcal{P}(\mathcal{A})$, where $\pi(a|s)$ denotes the probability of selecting the action a in state s . The objective in reinforcement learning is to learn an optimal policy π^* that

maximises the expected cumulative reward. The expected return (i.e., discounted cumulative reward) from state s under policy π is given by the value function:

$$V^\pi(s) = \mathbb{E}_{\pi, P} \left[\sum_{t=0}^{H-1} \gamma^t R(s_t, a_t) \mid s_0 = s \right] \quad (2.2)$$

Similarly, the state-action value function (Q-function) is defined as follows:

$$Q^\pi(s, a) = \mathbb{E}_{\pi, P} \left[\sum_{t=0}^{H-1} \gamma^t R(s_t, a_t) \mid s_0 = s, a_0 = a \right] \quad (2.3)$$

The two are related via the policy:

$$V^\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) Q^\pi(s, a) \quad (2.4)$$

The Bellman equations recursively define the value functions. For a given policy π , we have:

$$\begin{aligned} V^\pi(s) &= \sum_{a \in \mathcal{A}} \pi(a|s) \sum_{s' \in \mathcal{S}} P(s'|s, a) [R(s, a) + \gamma V^\pi(s')] \\ Q^\pi(s, a) &= \sum_{s' \in \mathcal{S}} P(s'|s, a) [R(s, a) + \gamma V^\pi(s')] \end{aligned} \quad (2.5)$$

The objective is to find the optimal policy π^* such that its value function is maximised at every state. The optimal value function V^* and optimal Q-function Q^* satisfy the Bellman optimality equations:

$$\begin{aligned} V^*(s) &= \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} P(s'|s, a) [R(s, a) + \gamma V^*(s')] \\ Q^*(s, a) &= \sum_{s' \in \mathcal{S}} P(s'|s, a) \left[R(s, a) + \gamma \max_{a' \in \mathcal{A}} Q^*(s', a') \right] \end{aligned} \quad (2.6)$$

In practice, computing π^* requires accurate estimation of either Q^* or V^* , which may be challenging under partial observability and sparse observational data, as commonly encountered in clinical environments.

Under the Markov Decision Process (MDP) framework, data-driven approaches to learning a Dynamic Treatment Regime (DTR) can broadly be categorised into *Imitation Learning* and *Reinforcement Learning*, which will be described in the following sections.

2.3 Imitation Learning and DTR Applications

This section focuses on IL-based approaches, which are often easier to implement in healthcare settings but come with specific limitations in terms of adaptability and long-term optimisation. A graphical comparison between IL and RL is shown in Figure 2.1.

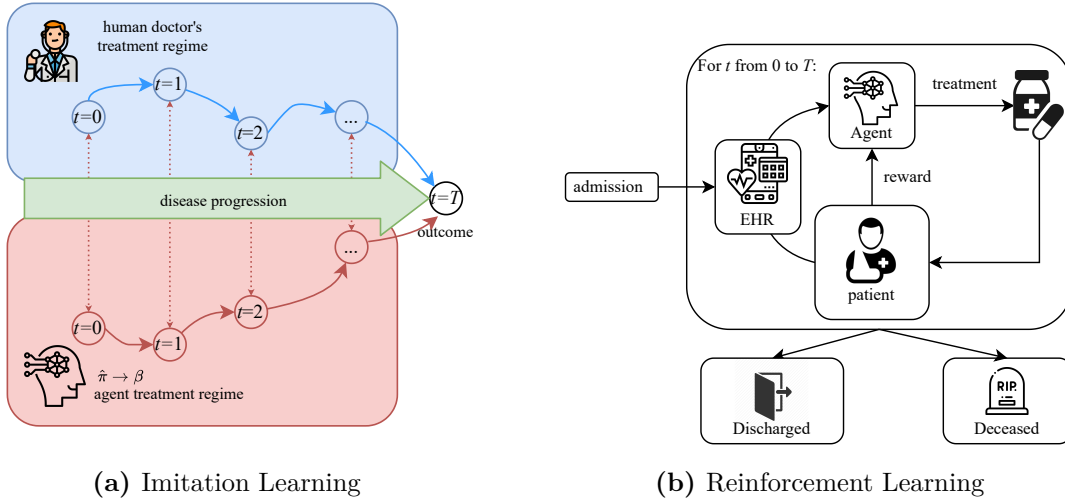


Figure 2.1: Comparing RL and IL in DTR. **a)** An IL agent trying to minimise the distance between its policy $\hat{\pi}$ and human doctors' regimes β for all timesteps. **b)** An RL agent changes its treatment regime at each timestep based on reward.

Imitation Learning (IL) [11] observes and replicates the behaviour demonstrated by an expert. The core idea is that expert behaviour encodes useful prior knowledge about the environment and task, which can be extracted by training the agent to replicate observed state-action pairs. Formally, the goal is to find an imitator policy $\hat{\beta}$ that closely matches the expert policy β by minimising a divergence measure L (e.g., cross-entropy loss or mean square error) between their output distributions over actions:

$$\hat{\beta} = \arg \min_{\pi} \mathbb{E}_{s \sim p} [L(\beta(\cdot|s), \pi(\cdot|s))] \quad (2.7)$$

where p is the state visitation distribution induced by the environment and expert policy, and π is the policy to be learned.

The most straightforward form of IL is *Behavioural Cloning* (BC) [12], which treats the problem as a supervised learning task, predicting the action of the expert

given a state. The agent learns a mapping from state inputs to action outputs by minimising a standard classification or regression loss. BC is attractive for its simplicity and scalability, but has a critical drawback: it assumes that the deployment states will match the training distribution. Unlike other paradigms (such as RL), this is especially problematic for BC because the agent only learns to imitate the expert and lacks mechanisms for on-the-fly correction. Consequently, a small deviation from the expert path can push the agent into unfamiliar states, causing compounding errors, a phenomenon known as covariate shift or distributional drift. Ross and Bagnell [13] demonstrated that if the agent makes a single error, it may enter an unseen state not covered by the training distribution, leading to cascading failures as time progresses. This is illustrated in Figure 2.2, where the agent deviates from the expert trajectory and is unable to recover. To recover from mistakes, Ross and Bagnell [13] proposed the forward training algorithm, which trains a separate policy network $\hat{\pi}_i$ for each time step $t \in \{1, 2, \dots, H\}$. In total, the forward training algorithm learns a non-stationary policy $\hat{\pi} = \{\hat{\pi}_1, \hat{\pi}_2, \dots, \hat{\pi}_T\}$, and executes $\hat{\pi}_t$ accordingly at time t . SMILE, DAgger and AggreVaTe [14] [15] further reduce the time required for expert query by mixing the learnt policy with the expert’s policy and achieving better performance compared to conventional supervised learning models with fewer expert demonstrations on games. However, these algorithms inherently require access to expert demonstrations while interacting with the environment, a setting that is incompatible with DTR.

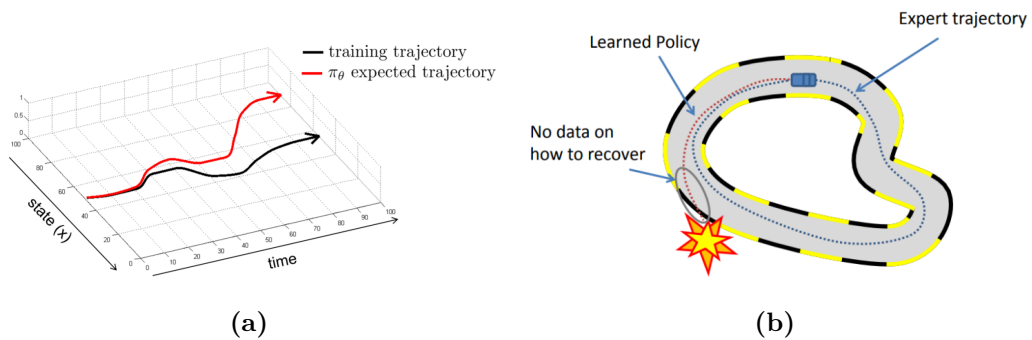


Figure 2.2: Demonstration of Failure Correction in Behaviour Cloning. **a)** Agent trajectory deviates from the expert trajectory as time progresses. **b)** An autopilot example where the Behavioural Cloning agent fails to recover from its initial mistake [16].

In the context of DTR, IL typically corresponds to offline behavioural cloning, as online interaction is rarely feasible due to ethical, safety, and practical constraints in clinical environments. As such, IL approaches rely entirely on retrospective data without querying or exploring alternative actions during learning. While this limits their capacity for treatment optimisation, IL remains attractive for its simplicity and alignment with clinician behaviour.

2.3.1 Imitation-Learning-Based DTR

Imitation learning for DTR focuses on minimising the difference between the learnt policy and the actual treatments prescribed by clinicians. Conceptually, this is closely related to time-series classification or regression, where at each step the model predicts the 'correct' treatment from the features of the patient. Several works demonstrate how standard supervised learning networks can be turned into treatment recommendation systems. For example, Sun et al. [17] leverage patient-similarity scores to recommend cerebral infarction treatments, while Bajor and Lasko [18] employ recurrent neural networks to capture the temporal dynamics of sequential prescriptions. Based on this, DoctorAI uses RNNs trained on large-scale EHR data to predict future diagnoses and medications with strong performance and generalisability [19]. Similarly, the deep diabetologist applies RNNs to predict hypoglycemic medications for diabetic patients, outperforming logistic regression baselines [20]. A broader approach is presented in [21], where deep-sequence models trained on millions of inpatient medication orders achieved high top-k prediction accuracy, making them practical tools for clinical decision support across general inpatient populations. Beyond medication prediction, several works explore drug effectiveness. For example, [22] predicts the efficacy of antihypertensive medications in patients with heart failure using longitudinal blood pressure data, achieving an F1 score close to 0.97. PerDREP [23] models individualised drug responses taking into account patient heterogeneity and leveraging similarity graphs. This not only improves prediction accuracy, but also uncovers clinically meaningful patient subgroups. Addressing the safety of drug combinations, DFNet [24]

integrates knowledge of drug relationships into a dual decision fusion framework that aligns patient and drug representations to recommend safe and effective combinations. [25] presents a graph-integrated model combining EHR data with a Drug-Drug Interaction (DDI) knowledge graph to predict drug combinations and minimise adverse effects, while also forecasting outcomes such as mortality risk for critically ill patients.

A key limitation of IL-based methods is *compounding error*. Any deviation from the expert’s trajectory can accumulate exponentially over many time steps [26]. In a hospital setting, if the sampling rate is hourly over a long stay, even a modest imitation gap at each hour can become substantial. Another limitation involves *class imbalance*: common treatments are over-represented in EHRs, whereas crucial but rarer interventions may be learnt using a behaviour cloning model. Because these rare actions can be essential for certain patients (e.g., life-saving interventions), naively training on unbalanced data often produces suboptimal performance.

2.4 Reinforcement Learning and DTR Application

Compared to IL, RL provides a more powerful and flexible framework for DTRs. RL explicitly optimises long-term outcomes by interacting with the environment or learning from the offline data collected from the environment, rather than simply replicating historical decisions. This makes RL particularly promising for healthcare, where treatment decisions have delayed and compounding effects that are difficult to capture through supervised methods. Formally, the goal of RL is to find a policy π^* that maximises the expected (discounted) cumulative reward:

$$\pi^* = \arg \max_{\pi} \mathbb{E}_{a_t \sim \pi(\cdot|s_t), s_{t+1} \sim P(\cdot|s_t, a_t)} \left[\sum_{t=0}^{H-1} \gamma^t R(s_t, a_t) \right].$$

In offline (batch) RL, we have a fixed dataset \mathcal{D} of transitions $\{(s, a, r, s')\}$ collected under some (possibly unknown) behaviour policy. The objective is similar to

maximise returns, but subject to learning only from \mathcal{D} :

$$\pi^* = \arg \max_{\pi} \mathbb{E}_{(s,a,r,s') \sim \mathcal{D}} \left[\sum_{t=0}^{H-1} \gamma^t R(s_t, a_t) \right] \quad \text{subject to using } \mathcal{D}.$$

Because the agent cannot collect new samples, offline RL algorithms must carefully address the mismatch between the policy learnt (π) and the distribution of states and actions present in \mathcal{D} .

2.4.1 Key Concepts and Algorithms in Reinforcement Learning

To better understand RL methods in the DTR setting, it is important to distinguish three key dichotomies: (1) **Online vs Offline RL:** *Online RL* involves learning through direct interaction with the environment in real time, updating the policy based on feedback from new actions taken. *Offline RL*, by contrast, learns exclusively from previously collected data, often in the form of logged patient trajectories (i.e., observational data). (2) **On-Policy vs Off-Policy RL:** *On-policy* methods learn from data collected using the same policy that is being improved, requiring the agent to interact with the environment using its current policy continually. *Off-policy* methods, however, can learn from data generated by a different policy, such as the historical behaviour of clinicians. (3) **Value-Based vs Policy-Based Methods:** *Value-based* methods (e.g., Q-learning) estimate the expected future reward of actions and derive the policy by maximising these values. *Policy-based* methods (e.g., REINFORCE) directly learn a parameterised policy by optimising the expected return via gradient ascent. Specifically, all offline RL algorithms are value-based, since policy-based optimisation requires an on-policy setting with environmental interactions.

Table 2.1 summarises the key RL approaches applicable to DTR settings, along with their respective training and evaluation characteristics. IL aims to replicate the historical behaviour of the clinician and thus does not involve exploration. Nevertheless, IL models may be prospectively evaluated when deployed, particularly in low-risk scenarios. Both online RL paradigms rely on real-time interaction with

Table 2.1: Taxonomy of RL and IL paradigms for DTRs in healthcare. Env. denotes environment.

| Paradigm | Data source | Exploration during training? | On/Off Policy | Online Evaluation? |
|----------------------|------------------|------------------------------|---------------|--------------------|
| IL | Retrospective | No | On | Yes or No |
| Online On-policy RL | Env. interaction | Yes | On | Yes |
| Online Off-policy RL | Env. interaction | Yes | Off | Yes |
| Standard Offline RL | Retrospective | No | Off | Yes |
| Offline RL for DTR | Retrospective | No | Off | No |

the environment to collect new data, differing in whether they leverage off-policy data during training. Standard offline RL is interaction-free during training, but usually requires online environmental return for policy evaluation. In contrast, Offline RL for DTR is specifically constrained to avoid any patient-facing exploration, focusing solely on retrospective learning. This requires robust off-policy evaluation methods to ensure the reliability of the learnt policies.

Off-policy RL Algorithms for Discrete Action Space

The following describes different RL algorithms for off-policy learning in the discrete action space.

1. **Deep Q-Network (DQN) [27]:** DQN integrates deep neural networks with the value-based Q-learning algorithm. A neural network parameterised by θ approximates the action-value function $Q(s, a; \theta)$. To stabilise training, DQN maintains a separate target network with parameters θ^- . The loss function is:

$$L_{\text{DQN}}(\theta) = \mathbb{E}_{(s,a,r,s') \sim \mathcal{D}} \left[\left(r + \gamma \max_{a'} Q(s', a'; \theta^-) - Q(s, a; \theta) \right)^2 \right], \quad (2.8)$$

where \mathcal{D} is the replay buffer containing transitions (s, a, r, s') . DQN introduced *experience replay* and *fixed Q-targets* to reduce correlation in updates and improve stability.

2. **Double Deep Q-Network (Double DQN) [28]:** Double DQN addresses the overestimation bias present in standard DQN. It uses one set of parameters θ to select actions and another set θ^- (the target network) to evaluate them:

$$L_{\text{DDQN}}(\theta) = \mathbb{E}_{(s,a,r,s') \sim \mathcal{D}} \left[\left(r + \gamma Q\left(s', \arg \max_{a'} Q(s', a'; \theta); \theta^-\right) - Q(s, a; \theta) \right)^2 \right]. \quad (2.9)$$

This decoupling reduces overestimation by ensuring that the selection and evaluation of actions are carried out by different networks.

3. **Dueling Network Architecture [29]:** The dueling network architecture factorises the Q-function into a state-value function $V(s)$ and an advantage function $A(s, a)$. The final Q-function combines them as:

$$Q(s, a) = V(s) + A(s, a) - \frac{1}{|\mathcal{A}|} \sum_{a'} A(s, a'), \quad (2.10)$$

where \mathcal{A} is the action set. This decomposition helps the network to assess how good a state is, separately from how beneficial each action is relative to others.

4. **Categorical DQN (C51) [30]:** C51 enhances DQN by modelling the *distribution* of returns, $Z_{s,a}$, rather than just the expected value. This return distribution is represented on a discrete set of N support points $\{z_1, \dots, z_N\}$. A key step is the distributional Bellman update, which redistributes probability mass from the next state s' and next action a^* onto the supports for (s, a) :

$$\Pr(Z_{s,a} = z_i) = \sum_{j=1}^N \Pr(Z_{s',a^*} = z_j) \Pr(R = r \mid s, a) \mathbb{I}[z_i = r + \gamma z_j], \quad (2.11)$$

where $\mathbb{I}[\cdot]$ is the indicator function. By capturing the entire return distribution, C51 produces richer training signals and often improves performance.

Off-policy RL Algorithms for Continuous Action Space

1. **Deep Deterministic Policy Gradient (DDPG) [31]:** DDPG is a model-free, off-policy actor-critic algorithm for continuous action spaces. It combines

ideas from DQN (target networks and replay buffers) with deterministic policy gradients. Let $Q(s, a; \theta^Q)$ be the critic, and $\mu(s; \theta^\mu)$ the actor. The critic is updated by minimising:

$$L_{\text{DDPG}}(\theta^Q) = \mathbb{E}_{(s,a,r,s') \sim \mathcal{D}} \left[\left(r + \gamma Q(s', \mu(s'; \theta^{\mu^-}); \theta^{Q^-}) - Q(s, a; \theta^Q) \right)^2 \right], \quad (2.12)$$

where θ^{Q^-} and θ^{μ^-} are parameters of the respective target networks. The actor's parameters θ^μ are then updated using the deterministic policy gradient using the current critic.

2. **Twin Delayed DDPG (TD3)** [32]: TD3 addresses the overestimation bias and instability inherent in Deep Deterministic Policy Gradient methods by incorporating several key modifications. First, it employs clipped double-Q learning, wherein two critic networks are used, and the minimum of their predicted Q values is selected as the target, effectively reducing overoptimistic value estimates. Second, the algorithm introduces delayed policy updates, updating the actor network less frequently than the critic networks, which promotes greater stability in the learning process. Third, target policy smoothing is applied by adding noise to the target actions, thereby mitigating the risk of overfitting to sharp and potentially erroneous fluctuations in the estimated Q-function.

3. **Soft Actor-Critic (SAC)** [33]: SAC is an off-policy actor-critic algorithm grounded in the maximum entropy framework, seeking to maximise returns while encouraging exploration via high-entropy policies. Two critics $Q_{\theta_1}, Q_{\theta_2}$ reduce the overestimation. Their parameters θ_i are updated by minimising:

$$L_{\text{SAC}}(\theta_i) = \mathbb{E}_{(s,a,r,s') \sim \mathcal{D}} \left[\left(Q_{\theta_i}(s, a) - \left(r + \gamma \left(\min_{j=1,2} Q_{\theta_j^-}(s', a') - \alpha \log \pi_\phi(a' | s') \right) \right) \right)^2 \right], \quad (2.13)$$

where θ_j^- are parameters of the target critics, π_ϕ is the stochastic actor, α is the temperature parameter, and \mathcal{D} is the replay buffer. The policy π_ϕ is trained to maximise both Q-values and entropy:

$$J_{\text{SAC}}(\phi) = \mathbb{E}_{s \sim \mathcal{D}, a \sim \pi_\phi} \left[\alpha \log(\pi_\phi(a | s)) - \min_{i=1,2} Q_{\theta_i}(s, a) \right]. \quad (2.14)$$

By prioritising entropy, SAC encourages more robust exploration and avoids prematurely converging to suboptimal deterministic actions.

Offline RL Algorithms

Offline RL (also known as *batch* RL) learns from a fixed dataset \mathcal{D} collected by some (possibly unknown) behaviour policy β . No further data collection is possible, so one must avoid relying on out-of-distribution (OOD) state-action pairs. Several offline RL methods address this challenge by constraining Q-value estimates, restricting the policy to the data distribution, or implicitly filtering for high-value actions in the dataset using expectile regression.

1. **Conservative Q-Learning (CQL)** [34]. CQL discourages the policy from selecting actions not supported by the offline dataset by learning a *conservative* Q-function that lower-bounds the true Q-values. One way to write its objective (discrete-action version) is:

$$J_{\text{CQL}}(Q) = \alpha \mathbb{E}_{(s,a) \sim \mathcal{D}} \left[\log \sum_{a' \in \mathcal{A}} \exp(Q(s, a')) - Q(s, a) \right] + L_{\text{Bellman}}(Q), \quad (2.15)$$

where L_{Bellman} represents a standard Bellman backup loss (e.g. MSE between Q and a target). The term multiplied by α penalises the Q-values for actions outside the data distribution, making the learnt Q-function more conservative.

2. **Batch-Constrained Deep Q-Learning (BCQ)** [35]. BCQ limits the agent to actions that resemble those in the offline dataset. An approach is to model the behaviour policy β (or an approximation thereof) with a generative network and constrain the learnt policy π to stay close to β . In discrete settings, for instance:

$$\pi(a | s) \propto \beta(a | s) \mathbf{1}\{Q(s, a) \geq \tau\}, \quad (2.16)$$

where $\mathbf{1}\{\cdot\}$ is the indicator function and τ is a threshold. Hence, BCQ does not deviate too far from the dataset, reducing the selection of OOD actions.

3. **Bootstrapping Error Accumulation Reduction (BEAR)** [36]. While BCQ can be *too* conservative, BEAR relaxes this restriction by allowing small probabilities for actions not seen frequently in \mathcal{D} . The key idea is to constrain the distance between the learned policy π and the behaviour policy β via a suitable divergence (e.g. MMD):

$$\min_{\pi} \mathbb{E}_{s \sim \mathcal{D}} \left[-Q_{\theta}(s, \pi(s)) \right] \quad \text{s.t.} \quad d(\pi(\cdot | s), \beta(\cdot | s)) \leq \epsilon, \quad (2.17)$$

where $d(\cdot, \cdot)$ is a divergence measure. By choosing $\epsilon > 0$ carefully, BEAR allows modest policy improvement beyond the dataset while avoiding drastic OOD actions.

4. **Implicit Q-Learning (IQL)** [37]. IQL aims to learn a policy π that *implicitly* remains close to the data distribution without explicitly restricting actions. The Q-update often uses *expectile regression*:

$$L_{\text{IQL}}(\theta) = \mathbb{E}_{(s,a,r,s') \sim \mathcal{D}} \left[\rho_{\tau} \left(y - Q_{\theta}(s, a) \right) \right], \quad (2.18)$$

where $\rho_{\tau}(x)$ is the expectile loss (a smooth analogue to quantile regression) for some $\tau \in (0, 1)$ and y is a standard Bellman target $(r + \gamma \max_{a'} Q_{\theta}(s', a'))$. Crucially, IQL’s policy extraction step discards Q-values for actions not sufficiently supported by the dataset, preventing exploitation of OOD estimates.

These offline RL approaches are often evaluated on non-healthcare benchmark datasets such as D4RL [38], where data are collected by semi-random or suboptimal policies. However, these datasets may not reflect the sparse and imbalanced nature of real-world clinical scenarios (e.g., dynamic treatment regimes, where some interventions are rarely taken). Evaluating or adapting offline RL algorithms to such complex domains remains an active area of research.

On-policy RL Algorithms

Our study focuses on off-policy and offline reinforcement learning methods due to ethical and safety constraints that prohibit exploratory interactions with real patients. As a result, on-policy approaches are not systematically benchmarked in the subsequent technical chapters. An exception is made in Chapter 7, where on-policy learning is included solely to establish an optimistic performance upper bound for comparison with knowledge-injected Large Language Models (LLMs). For this purpose, we apply Proximal Policy Optimisation (PPO) as a canonical on-policy baseline.

Proximal Policy Optimisation (PPO) [39]. PPO is an actor–critic algorithm that iteratively collects fresh trajectories with the current policy, then updates the policy parameters θ by maximising a *clipped surrogate objective*. Let $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ denote the likelihood ratio between the new and old policies, and \hat{A}_t an estimator of the advantage function. The clipped objective is

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t \left[\min \left(r_t(\theta) \hat{A}_t, \text{clip} \left(r_t(\theta), 1 - \varepsilon, 1 + \varepsilon \right) \hat{A}_t \right) \right], \quad (2.19)$$

where ε (typically 0.1–0.3) bounds how far the policy may move in a single update.

A separate value network $V_\psi(s_t)$ is trained to minimise

$$L^{\text{VF}}(\psi) = \mathbb{E}_t \left[\left(V_\psi(s_t) - \hat{R}_t \right)^2 \right]. \quad (2.20)$$

with \hat{R}_t the empirical return. An entropy bonus $L^{\text{ENT}} = \mathbb{E}_t \left[-\mathcal{H} \left(\pi_\theta(\cdot | s_t) \right) \right]$ encourages exploration. The complete loss combines all three terms, $L_{\text{PPO}}(\theta, \psi) = L^{\text{CLIP}}(\theta) + c_v L^{\text{VF}}(\psi) - c_e L^{\text{ENT}}(\theta)$, with coefficients $c_v, c_e > 0$ controlling the value-function and entropy weightings. By constraining policy updates through clipping rather than trust-region projections, PPO achieves stable learning with modest implementation complexity, making it a popular choice for high-dimensional continuous and discrete tasks alike.

Throughout this thesis, PPO serves solely as a reference point for the best attainable online performance when unlimited interaction is hypothetically permitted. All subsequent algorithmic contributions and empirical studies remain within the stricter off-policy or offline paradigms appropriate for real-world clinical deployment.

2.4.2 Evaluation Metrics for Modelling Dynamic Treatment Regimes

In RL-DTRs, one of the most challenging tasks is evaluating a learnt treatment strategy without deploying it in the real world. This is particularly critical in healthcare, where online evaluation is often infeasible due to ethical and safety constraints. Instead, we rely on *Off-Policy Evaluation*. This framework estimates the value of a target policy π using retrospective data generated by a different behaviour policy β (e.g., the historical clinician policy). OPE is essential to assess whether a learnt treatment policy would outperform existing clinical practices before any real-world implementation. However, accurate estimation is complicated by the distributional mismatch between behaviour and target policies, finite samples, and observational confounding. To address these challenges, we consider multiple estimators, including variants of *Importance Sampling (IS)*, *Direct Method* and *Doubly Robust Estimation*, each with their advantages and trade-offs.

Weighted Importance Sampling with Bootstrapping

The Weighted Importance Sampling (WIS) [40] estimator corrects for the distributional mismatch between π and β by re-weighting observed trajectories based on the likelihood ratio of action probabilities under the two policies. Bootstrapping is incorporated to estimate confidence intervals and reduce variance through resampling. The bootstrapped WIS estimator is defined as:

$$\hat{V}_{\text{WIS}_B}^{\pi} = \frac{1}{B} \sum_{b=1}^B \left(\frac{\sum_{i=1}^{N_b} \rho_i^{(b)} G_i^{(b)}}{\sum_{i=1}^{N_b} \rho_i^{(b)}} \right) \quad (2.21)$$

where B denote the number of bootstrap samples, and N_b represent the number of trajectories in the b -th bootstrap sample. For each trajectory i , the cumulative importance ratio is given by $\rho_i^{(b)} = \prod_{t=0}^{H_i} \frac{\pi(a_t^i | s_t^i)}{\beta(a_t^i | s_t^i)}$, where (s_t^i, a_t^i, r_t^i) denotes the t -th step of the i -th trajectory. The total discounted return for trajectory i is defined as $G_i^{(b)} = \sum_{t=0}^{H_i} \gamma^t r_t^i$, where $\gamma \in [0, 1)$ is the discount factor.

This estimator can be compactly expressed as an expectation:

$$\hat{V}_{\text{WIS}_B}^{\pi} = \mathbb{E}_{b \sim \mathcal{B}} \left[\hat{V}_{\text{WIS}_{(b)}}^{\pi} \right] \quad (2.22)$$

Truncated Weighted Importance Sampling

In practice, the importance weights ρ_i can be extremely large for rare trajectories, leading to high-variance estimates. To mitigate this, we use a *truncation threshold* τ to limit the maximum weight, improving stability at the cost of introducing bias [41]. The Truncated WIS estimator is defined as:

$$\hat{V}_{\text{WIS}_{\text{tr}}}^{\pi} = \frac{\sum_{i=1}^N \min(\rho_i, \tau) G_i}{\sum_{i=1}^N \min(\rho_i, \tau)} \quad (2.23)$$

Combined Bootstrapped and Truncated WIS

To benefit from both variance reduction via bootstrapping and stability via truncation, we define a combined estimator:

$$\hat{V}_{\text{WIS}_{\text{tr},B}}^{\pi} = \mathbb{E}_{b \sim \mathcal{B}} \left[\hat{V}_{\text{WIS}_{\text{tr},(b)}}^{\pi} \right] \quad (2.24)$$

This hybrid estimator allows for more reliable performance estimation of treatment policies by reducing sensitivity to outlier trajectories and finite-sample variability.

Direct Method

The DM estimator for the policy value is formulated as follows:

$$\hat{V}_{DM}^{\pi} = \mathbb{E}^{\pi} \left[\hat{Q}^{\beta}(s, a) \right] \quad (2.25)$$

where $\hat{Q}^{\beta}(s, a)$ is the estimated action-value function under the behaviour policy. Despite its straightforward approach, DM is highly dependent on the accuracy of the model for estimating \hat{Q}^{β} , making it susceptible to model misspecification and bias due to the imbalanced nature of medical data. DM can be implemented with an offline SARSA-based Q-learning approach to estimate the state-action value function $Q(s, a)$. The loss function used during training is the mean squared error of the temporal difference (TD) target:

$$L_{DM} = \mathbb{E}_{(s,a,r,s',a') \sim \mathcal{D}} \left[(r + \gamma Q(s', a') - Q(s, a))^2 \right]$$

where a' is selected according to the behaviour policy β and (s, a, r, s') are transitions observed in the dataset.

Doubly Robust Estimator

The Doubly Robust (DR) estimator [42] combines the strengths of both DM and IS, offering robustness to errors in either component, iteratively calculated per-trajectory estimation as:

$$\hat{V}_{DR}^{(t)} = \hat{V}^\pi(s_t^i) + \eta_t \left(r_t^i + \gamma \hat{V}_{DR}^{(t+1)} - \hat{Q}^\beta(s_t^i, a_t^i) \right) \quad (2.26)$$

with recursion initialised by $\hat{V}_{DR}^{(H_i+1)} := 0$, and $\eta_t = \prod_{k=0}^t \frac{\pi(a_k^i | s_k^i)}{\beta(a_k^i | s_k^i)}$. The overall estimate is then $\hat{V}_{DR}^\pi := \sum_{i=1}^N \hat{V}_{DR}^{H_i}$.

2.4.3 Reinforcement-Learning-Based DTR

Reinforcement learning (RL) offers a promising alternative to imitation learning (IL) by directly optimising a *reward* signal, such as clinical improvement or survival. A growing body of literature has explored RL to construct dynamic treatment regimes (DTRs) for chronic and acute conditions, including cancer, diabetes, sepsis, and anaesthesia [43]. Compared to IL-based approaches, RL has attracted increasing attention due to its potential to uncover treatment policies that outperform historical clinician data [44]. The following sections summarise the application of RL-based DTRs across major clinical domains.

Cancer

In oncology, RL has been applied to personalise chemotherapy, radiotherapy, and combination therapies. For chemotherapy, algorithms such as Q-learning, Monte Carlo tree search, and offline RL have been used to optimise dosing schedules. In particular, Q-learning improved simulated survival rates and reduced grade 4 neutropenia in advanced non-small cell lung cancer (NSCLC) [45, 46]. Conservative Q-learning has also been explored in colon cancer for regimen selection [47], while O. et al. reviewed a broader spectrum of RL techniques across various malignancies. Approaches such as Bayesian data assimilation and multi-objective tree-based RL have further refined dosing personalisation [49–51]. Radiotherapy has seen widespread use of deep RL, particularly Deep Q-Networks (DQN), to optimise dose

distribution and spare surrounding organs. H. et al. applied DQN for adaptive radiotherapy in stage II–III NSCLC. SARSA with convolutional neural networks (CNNs) improved precision in stereotactic radiotherapy for head-and-neck and pancreatic cancers [53, 54]. Further deep RL models have been used for adaptive radiotherapy in NSCLC and cervical cancer [55–58]. Proximal Policy Optimisation (PPO) has been effective in improving proton pencil beam scanning for head-and-neck and liver cancers [59]. Multi-agent RL has coordinated complex planning in intensity-modulated radiotherapy (IMRT) [60–67]. RL has also contributed to adaptive planning under tumour uncertainty [68–80]. In combined modality treatments, RL has supported the scheduling of chemotherapy - immunotherapy (e.g., melanoma, using a deep deterministic policy gradient) [81], intermittent androgen deprivation therapy for prostate cancer [82], and locoregional systemic approaches [74].

Despite promising computational outcomes, the application of cancer using RL in the real-world remains limited. Most models rely on retrospective datasets or simulations, and few have been validated on actual patient data—representing a key translational barrier.

Diabetes

Reinforcement learning (RL) has emerged as a powerful framework for personalising insulin dosing in diabetes care, consistently outperforming conventional algorithms across simulations and select clinical trials. By formulating insulin delivery as a sequential decision-making problem, RL enables agents to maximise time-in-range (TIR) while minimising the risk of hypo- and hyper-glycaemia. Four clinical trials have explored its application across type 1, type 2, and postoperative cohorts. For example, offline model-based RL raised the mean TIR from 61.4% to 85.5% in a retrospective dataset of 12,981 individuals with type 2 diabetes [83]. Similarly, distributional RL improved intravenous insulin dosing accuracy after cardiac surgery,

reducing the mean absolute error (MAE) to 1.1 units across 5,228 cases [84]. Proof-of-concept trials have further demonstrated personalised bolus adjustments and bedside deployment of simulation-trained policies [83, 85].

Across simulation benchmarks, RL algorithms consistently outperform rule-based and proportional–integral–derivative (PID) controllers. Classical Q-learning has achieved TIR levels exceeding 85% while maintaining hypoglycaemia below 2% [86–88]. Contemporary deep RL methods such as Soft Actor–Critic (SAC), Proximal Policy Optimisation (PPO), and shielded SAC variants [89, 90] have demonstrated robust performance across age groups and safety-critical regimes. In particular, SAC-based dual hormone controllers that incorporate glucagon delivery have increased adolescent TIR by approximately 11% [91]. Model-based strategies, including Echo-State planning [92], Dyna-Q [93], and ensemble deep RL [94], further refine performance, achieving glucose prediction errors below 15 mg/dL and less than 0.5% time below range. In parallel, neuro-adaptive fuzzy actors (NAFA) have been developed for subcutaneous insulin delivery in type 2 diabetes [95].

Patient personalisation has also been improved through adaptive RL mechanisms. Transfer-entropy initialisation has allowed TIRs to approach 98% in adult users [96, 97], while meta-RL embeddings can adjust to individual patient dynamics in 24 hours [98]. Hierarchical approaches such as Option-Critic [99] and multi-agent Q-learning [100] have significantly reduced postprandial glucose excursions. Cross-patient generalisation has been further supported by domain randomisation in deep deterministic policy gradients (DDPG) [101], while the temporal difference learning driven by the electronic health record (EHR) has enabled data-efficient training directly from clinical datasets [102]. Meal-agnostic Double DQNs have also been proposed to reduce the reliance on manual meal annotation [103].

Safety and stability remain critical for real-world deployment. Offline algorithms such as Conservative Q-Learning and TD3-BC have demonstrated over 10% TIR improvements compared to behaviour cloning without increasing hypoglycaemic risk [104]. Hybrid training strategies, combining offline pre-training with limited online updates, have yielded an additional 9.7% TIR gain [105]. To maintain

operational safety, Lyapunov-constrained SAC has been introduced, which preserves the behaviour of the bounded system during the execution of the policy [106].

Despite these advances, key limitations persist. Many studies rely heavily on simulators and report heterogeneous metrics, which limits comparability and clinical relevance. In addition, few long-term trials assess generalisability or sustained efficacy in real-world settings. Most current methods assume continuous glucose monitoring (CGM), restricting their use in patients without such infrastructure. In addition, insulin therapy is frequently required in non-diabetic critical care patients, a clinically important subgroup that remains underexplored in the RL literature.

Sepsis

In the management of sepsis, RL has been explored to personalise interventions such as fluid resuscitation, vasopressor dose, and antibiotic timing. Zhang, Zheng, and Liu [107] reported an improvement in average survival (4.1 to 5.7 days) using tailored fluid strategies, while Kim, Ausin, and Chi [108] achieved a reduction in the incidence of septic shocks from 17% to 5.7%. Additional RL approaches that optimised combined fluid-vasopressor regimens led to survival gains of up to 10% [109, 110]. In vasopressor management, RL models have proposed more liberal dosing strategies than traditional protocols [111, 112]. RL-guided recommendations have also supported earlier fluid administration and reduced use of vasopressors [113, 114]. Integrated optimisation of fluid and vasopressor therapy has been shown to further enhance patient outcomes [110, 115–117]. RL has also been applied to antibiotic administration timing and glycaemic control using methods such as DQN, policy iteration, off-policy sampling, temporal difference learning, and partially observable Markov decision processes [118–124]. Reported benefits include mortality reductions of 2.3–22% and survival improvements up to 15.21%. Nevertheless, clinical implementation remains rare. Studies vary widely in reward definitions, evaluation metrics, and optimisation targets, which complicates comparison. These challenges are examined in Section 2.5.

Critical Care Interventions

RL has demonstrated increasing utility in a wide range of critical care applications beyond sepsis, including sedation, mechanical ventilation, vasopressor administration, and anaesthesia control. In sedation management, RL algorithms such as Deep Deterministic Policy Gradient (DDPG) [125], Deep Q-Network (DQN) [126], and Conservative Q-Learning (CQL) [127] have been trained on retrospective datasets like MIMIC-IV to improve dosing precision and reduce the incidence of delirium. In particular, morphine dosing optimisation using Dueling Double Deep Q-Networks trained in MIMIC-III improved pain control and dosing efficiency [128]. RL applications in mechanical ventilation have shown especially promising results. Early work on joint sedation-ventilation policies improved extubation readiness while maintaining patient comfort [129]. More recent studies expanded the RL action space to include fine-grained ventilator settings, surpassing clinician performance in simulated outcomes related to mortality and ventilator-induced lung injury [130]. Techniques such as CQL and Double DQN have improved ventilator parameter selection, leading to improved physiological stability and survival outcomes across datasets such as MIMIC-III and eICU [131–137]. Furthermore, novel algorithms, such as ConformalDQN [138], Dynamic Weighted DQN [137], and deconfounding actor-critic models [139], have addressed uncertainty and confusion in clinical environments. Advanced RL techniques have also informed decision-making in non-invasive ventilation, leveraging Double Dueling DQNs with retrospective MIMIC-III data [140]. RL methods have been applied in both simulation and retrospective settings in anaesthesia control. For example, policy-constrained Q-learning [141] and Watkins Q-learning [142] have guided propofol infusion strategies and optimised anaesthetic depth. Similarly, intravenous heparin dosing has been fine-tuned using DDPG algorithms trained on records from Emory and MIMIC [143], while propofol control strategies have been validated in simulated environments [144]. RL has also proven to be effective in less traditional critical care tasks. In oxygen therapy for COVID-19, DDPG models trained in NYU Langone Health EHR data improved oxygen administration [145, 146]. Additional applications include heart failure

| Evaluation tier | Cancer | Diabetes | Sepsis / Critical care |
|------------------------|---|---|---|
| in-silico only | [49, 53, 54, 56, 57, 59, 61, 63, 68, 78, 79, 151–153] | [86, 88, 89, 92, 94, 95, 101, 103, 104, 106, 154–163] | [108, 111, 113–115, 117–120, 123, 124, 164–172] |
| Retrospective | [47, 51, 74, 173, 174] | [83, 84, 93, 102, 175, 176] | [107, 109, 110, 112, 116, 122, 177–183] |
| Clinical validation | — | [85, 87, 100] | — |
| External-site | [69, 70, 80] | [98, 101] | [112, 168, 178, 179] |

Table 2.2: A selective table on how RL studies evaluate policies. Simulation-only work dominates; genuine prospective validation is rare.

treatment [147], stabilisation of vital signs [148], prevention of delirium during ventilation [149], optimisation of cancer treatment [117], and multiobjective policy learning for complex intervention management [150].

Overall, these applications (though not exhaustive) illustrate how IL and RL have been used to derive DTRs across various clinical contexts. However, substantial barriers remain to the transition from in silico success to clinical integration. The following sections explore outstanding issues, focusing on data curation, reward design, offline policy evaluation, and safety guarantees.

2.5 Open Issues in RL-Driven Dynamic Treatment Regimes

As a cutting-edge interdisciplinary application that bridges healthcare and artificial intelligence, RL-driven DTRs remain at an early stage of development. In the following, we summarise three pressing challenges to address before these methods mature into reliable clinical decision support tools.

Policy Evaluation

Most published studies still assess policies in silico, either in customised simulators or via retrospective EHR replays. As Table 2.2 shows, the simulation-only category dominates across cancer, diabetes, and sepsis. For example, radiotherapy planners are almost exclusively benchmarked in virtual environments [53, 68], while the majority of glucose control articles rely on UVA / Padova or similar simulators [86,

| Reward family | Typical signal | Cancer | Diabetes | Sepsis / ICU |
|---------------------------|-------------------------------------|------------------|------------------------------|-------------------------------|
| Binary outcome | Survival | [49, 151] | — | [44, 108, 109, 117, 119, 183] |
| Dense physiological | Dose, glucose MAP, SOFA | [53, 58, 62, 68] | [86, 88, 104, 106, 156, 161] | [111, 113, 124, 167, 184] |
| Composite multi-objective | TCP/NTCP TIR + hypo MAP + mortality | [51, 59, 79] | [94, 98, 163] | [116, 118, 147] |
| Shaped or hierarchical | Reward shaping safety | [80, 152] | [99, 106, 154] | [169, 181, 185] |

Table 2.3: A selective table showing the diversity of reward definitions. Very few studies compare alternative rewards on the same dataset.

88]. Retrospective EHR studies, for example *AI Clinician* for sepsis [44] or offline RL for type 2 diabetes dosing [83], increase clinical realism, yet remain subject to unmeasured confounding and policy evaluation bias. Only a handful of works have been tested in prospective or proof-of-concept trials, all in diabetes care [85, 87]. Establishing common, sharable simulation benchmarks and reporting standardised policy evaluation metrics will be essential to enable fair comparison of algorithms and build the confidence required for clinical deployment.

Reward Design

The reward definition is highly hand-crafted and varies markedly by clinical domain (Table 2.3). Cancer studies often optimise dose-based surrogates such as tumour control probability or organ at risk preservation [59, 79]; diabetes work typically maximises TIR while penalising hypoglycaemia [104, 106]; sepsis research oscillates between sparse binary survival rewards [109] and dense haemodynamic targets such as mean arterial pressure [111]. Comparison of algorithms across studies is nontrivial because different rewards induce different clinical trade-offs. Exceptionally few papers perform ablation studies to test how alternative reward signals affect learnt policies, and almost none attempt to elicit clinician or patient preferences

explicitly. A systematic framework for reward selection, potentially informed by causal inference and multi-objective optimisation, remains an open research avenue.

Generalisation and Transferability

RL policies trained on a single cohort or simulator frequently overfit to that narrow data distribution. Only isolated studies explore explicit generalisation strategies, such as personalised embeddings for cross-patient adaptation in glucose control [98], domain randomisation in prostate cancer therapy scheduling [80], or external-site validation of haemodynamic policies across European and US ICUs [112]. Off-policy evaluation techniques, such as importance sampling with high-confidence bounds [116] have begun to appear, yet there is no consensus on which estimator is most reliable under clinical covariate shift.

3

Data

This chapter presents the data sources used in this thesis, comprising two real-world datasets derived from the MIMIC-III database and four synthetically generated treatment environments. Real-world datasets provide authentic clinical complexity, while synthetic environments offer controlled experimental settings for rigorous evaluation of RL frameworks applied to dynamic treatment regimes.

The Medical Information Mart for Intensive Care (MIMIC-III) [186] is an openly accessible, de-identified database that includes detailed clinical data on more than 40,000 patients who were admitted to critical care units of the Beth Israel Deaconess Medical Centre in Boston, Massachusetts, between 2001 and 2012. The database captures a wide range of information, including patient demographics, time-stamped vital signs, laboratory results, diagnostic and procedural codes, medication administration, caregiver notes, radiology reports, and mortality outcomes, both in-hospital and post-discharge. The granularity and breadth of MIMIC-III make it highly suitable for modelling DTRs across heterogeneous patient populations and disease presentations within intensive care.

In this thesis, two distinct real-world datasets derived from MIMIC-III are used. Section 3.1 introduces a widely used sepsis cohort curated by [44] to facilitate the learning of treatment regimes involving intravenous fluids and vasopressors. This dataset has been selected due to its frequent adoption in the literature [187–191], serving as a representative benchmark for comparative evaluation. The second dataset reported in Section 3.2 is a novel curated insulin administration cohort

developed as part of this thesis. It enables the evaluation of reinforcement learning-based treatment strategies for glycaemic control in critically ill patients, as well as predictive modelling of hypoglycaemic events.

3.1 MIMIC-III Sepsis Cohort

The Sepsis cohort consists of approximately 10,000 patients and focuses specifically on the administration of intravenous fluids (IV) and vasopressors within the ICU. The dataset targets the critical phase of sepsis management and supports the development of treatment policies using reinforcement learning methodologies. The cohort is extracted according to the sepsis-7 criteria [192], and includes clinical information captured during a 72-hour window surrounding the estimated time of onset of sepsis from 24 hours before to 48 hours after onset. This time frame encapsulates the early resuscitation period, where intervention decisions are the most influential on patient outcomes. In this dataset, IV fluids and vasopressor dosages are discretised into five levels, forming a structured 5×5 action space.

The extracted dataset contains 48 observational variables, including demographics, Elixhauser premorbid status, vital signs, laboratory values, fluids and vasopressors received and fluid balance. Patients' data were coded as multidimensional time series with 4-h time steps. Data variables with multiple measurements within a 4-h bin were averaged (for example, heart rate) or summed (for example, urine output) as appropriate. All features were checked for outliers and errors and imputed following the original paper [44]. Table 3.1 presents the observation space.

3.1.1 Patient Stratification

Patient stratification plays a critical role in modern clinical research and decision-making, particularly given the heterogeneous nature of patient responses and trajectories in intensive care settings [193, 194]. Tailoring interventions to subgroups based on clinically relevant stratification criteria enhances the generalisability and interpretability of computational models, particularly those designed to support high-stakes clinical decisions. Despite its clinical importance, stratification is often

Table 3.1: List of features used in the sepsis cohort. Abbreviations: SOFA, Sequential Organ Failure Assessment; SIRS, Systemic Inflammatory Response Syndrome; SpO₂, peripheral capillary oxygen saturation; BUN, blood urea nitrogen; SGOT, serum glutamic-oxaloacetic transaminase; SGPT, serum glutamic-pyruvic transaminase; PTT, partial thromboplastin time; PT, prothrombin time; INR, international normalised ratio; PaO₂, arterial partial pressure of oxygen; PaCO₂, arterial partial pressure of carbon dioxide; FiO₂, fraction of inspired oxygen.

| Category | Items | Type |
|-------------------------------|--|--------|
| Demographics | Age | Cont. |
| | Gender | Binary |
| | Weight | Cont. |
| | Readmission to intensive care | Binary |
| | Elixhauser score (premorbid status) | Cont. |
| Vital signs | Modified SOFA score | Cont. |
| | SIRS criteria | Cont. |
| | Glasgow Coma Scale | Cont. |
| | Heart rate, blood pressure (systolic, mean, diastolic), shock index | Cont. |
| | Respiratory rate, SpO ₂ | Cont. |
| | Temperature | Cont. |
| Laboratory values | Potassium, sodium, chloride | Cont. |
| | Glucose, BUN, creatinine | Cont. |
| | Magnesium, calcium, ionised calcium, carbon dioxide | Cont. |
| | SGOT, SGPT, bilirubin (total), albumin | Cont. |
| | Haemoglobin | Cont. |
| | White blood cell count, platelets, PTT, PT, INR | Cont. |
| | pH, PaO ₂ , PaCO ₂ , base excess, bicarbonate, lactate, PaO ₂ /FiO ₂ ratio | Cont. |
| | | |
| Ventilation | Mechanical ventilation | Binary |
| | FiO ₂ | Cont. |
| Medications and fluid balance | IV fluid intake (past 4h) | Cont. |
| | Maximum vasopressor dose (past 4h) | Cont. |
| | Urine output (past 4h) | Cont. |
| | Cumulative fluid balance since admission | Cont. |
| Outcomes | Hospital mortality | Binary |
| | 90-day mortality | Binary |

overlooked in the current literature on RL for DTRs, potentially introducing bias and limiting applicability across diverse patient cohorts. To address this limitation, we introduce a stratification approach based on the National Early Warning Score 2 (NEWS2), a validated clinical score to monitor acute physiological deterioration. Our stratification strategy is designed to highlight the differential risk of mortality and stability of the trajectory among patients. Specifically, we define two primary

stratification axes: the rate of change in NEWS2 scores over time, and the variability in these changes, captured by the standard deviation (SD) of the score trajectory.

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. In addition, patients are categorised into two groups based on the standard deviation of their NEWS2 score changes: those with high variance, indicating a more fluctuating state and those with low variance, indicating a more stable deterioration/recovery, split by the SD median in the subgroup. NEWS2 scores are segmented into distinct brackets $[-0.4, -0.15, 0, 0.15, 0.4]$, and the variance is classified as low or high. A comprehensive data distribution plot is shown in Fig 3.1. For the dataset split, we first divided all data into subgroups (2 outcomes, 6 NEWS2 score bins, and 2 SD levels) and ensured that each group was (near) evenly distributed in training, validation, and test sets for fair evaluations.

3.1.2 Action Discretisation

The treatment space in the Sepsis cohort is defined by two continuous clinical interventions, IV fluid administration and vasopressor dosing. To make these suitable for RL-based modelling, both variables are discretised into five ordinal bins, consistent with the scheme proposed by [44]. Each bin represents a clinically relevant range of dosages and is associated with a representative median value, as shown in Table 3.2.

| 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 3.2: Action Discretisation Range and Median

In particular, the median values for each dosage bin differ from those reported in the original supplementary material [44]. These differences arise from sev-

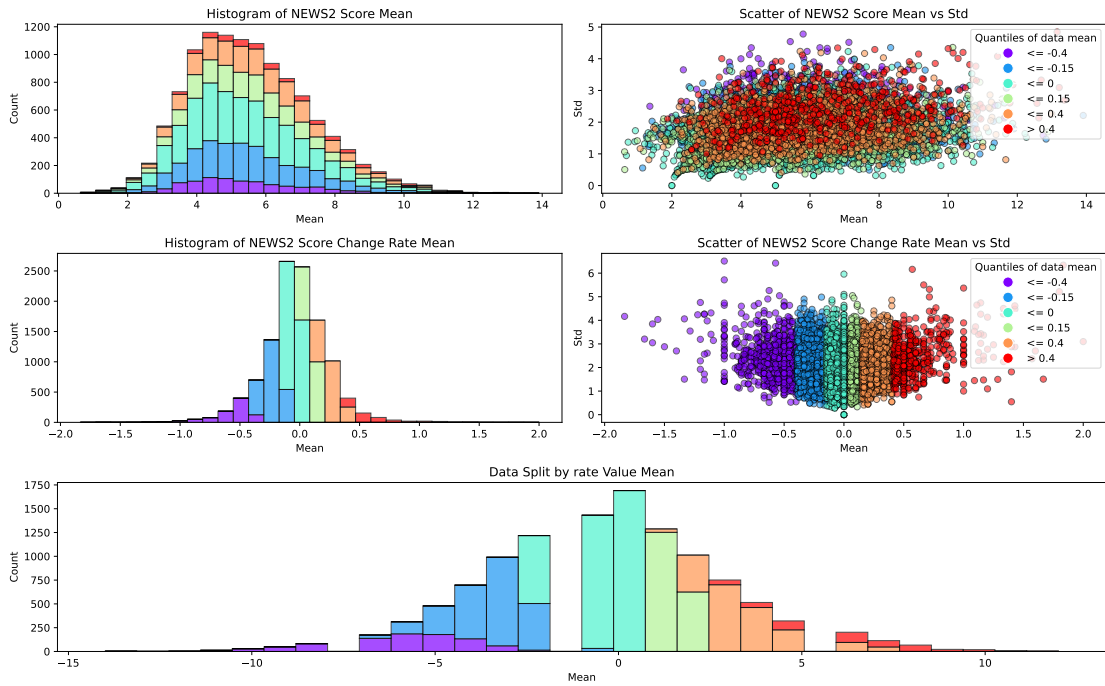


Figure 3.1: Data Stratification using rate of change in NEWS2 score and admission outcome. The plot includes **(top left)** the distribution of mean NEWS2 score, **(top right)** mean NEWS2 against SD, **(middle left)** distribution of NEWS2 rate of change, **(middle right)** NEWS2 rate of change mean 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 patients with different deterioration/recovery speeds are distributed across all risk levels. The plot also shows that the number of patients can significantly differ across risk stratification, supporting the need for RL evaluation on all subsets.

eral deliberate modifications introduced in our work to improve the clinical and methodological reliability. Firstly, we redefine the outcome measure from 90-day to in-hospital mortality, guided by clinical input, to more directly capture the short-term effects of treatment decisions. Secondly, patients with incomplete or inconsistent time-series data are excluded. Specifically, we remove admissions with missing information that spans more than four hours during treatment windows, as these gaps introduce ambiguity in sequential modelling. Unlike the original study, which treated irregularly sampled data as uniformly spaced, our approach adheres

more closely to RL best practices regarding temporal resolution. Furthermore, we omit the ‘input 4-hourly’ feature from the observation space, since this variable directly reflects the administered treatment and therefore constitutes part of the action space. Including it would introduce data leakage and potentially confound model learning and evaluation.

3.2 MIMIC-III Insulin Administration Cohort

This section describes the development of a real-world dataset derived from the MIMIC-III database, designed to characterise glycaemic control patterns among adults admitted to intensive care. The cohort was curated to facilitate detailed modelling of glucose–insulin dynamics across diverse clinical trajectories, with a particular emphasis on supporting the RL decision-making frameworks. An illustrative example of aligned multivariate time series and a schematic summary of the data flow are shown in Fig. 3.2. Dataset construction followed a structured, multi-stage pipeline incorporating clinical expertise and data-driven quality control. Relevant records were extracted from multiple tables within MIMIC-III, including laboratory measurements, physiological observations, medication administrations, and demographic information. In total, 8,472 ICU stays were retained from 7,086 unique patients following the application of predefined exclusion rules.

The curation process begins with the alignment of insulin and glucose events, which were then merged with auxiliary clinical data such as prescription logs, fluid input/output records, and vital signs. Observations were resampled at fixed intervals (1 h or 2 h), enabling a consistent temporal framework across stays. Subsequent processing steps included imputation, unit harmonisation, and normalisation of the measurement scales. Insulin dosages were discretised into clinically informed bins, and each stay was annotated with key stratification attributes, including age, sex, diabetes status, and evidence of hypoglycaemia. The final dataset was partitioned into training, validation, and test sets in a 70%/15%/15% split, stratified to preserve population heterogeneity and prevalence of the outcome. We proposed

two tasks: the first targets insulin dosing recommendation, and the other focuses on hypoglycaemia risk prediction.

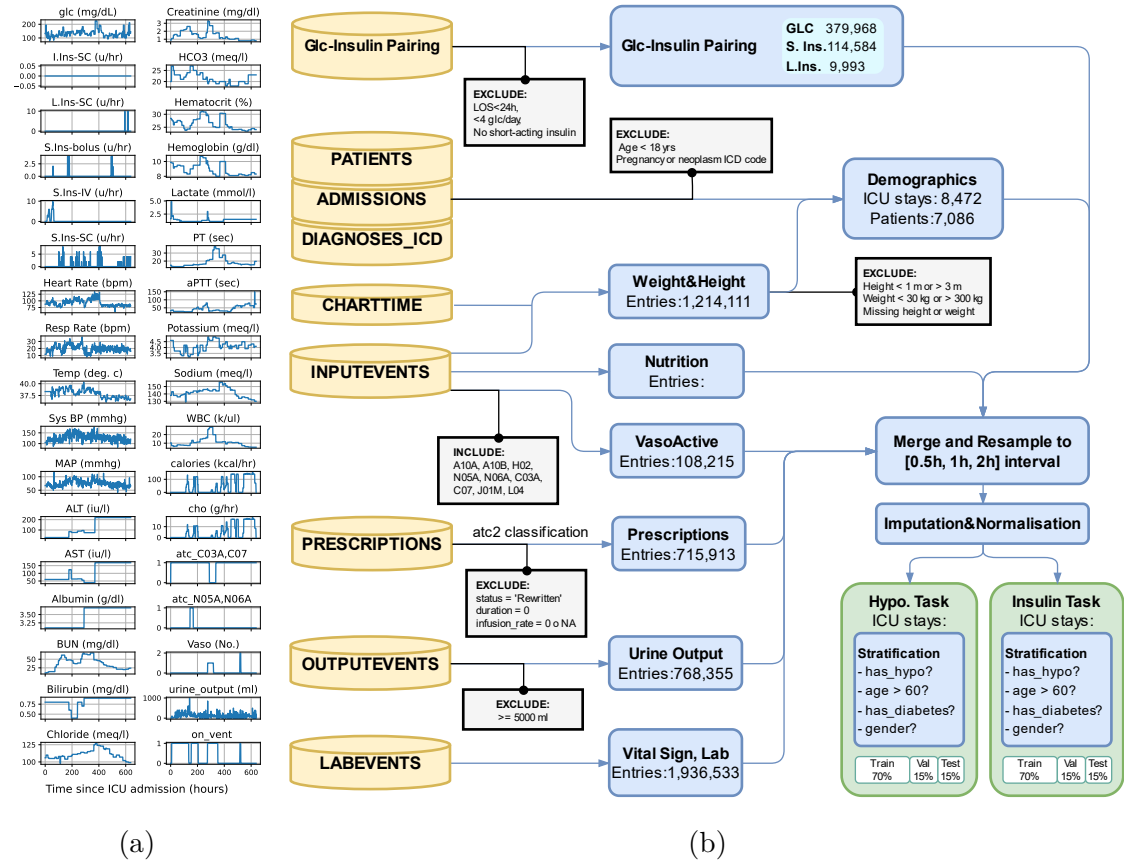


Figure 3.2: Summary of cohort construction and example patient data. (a) Illustrative multivariate time series showing glucose levels, insulin infusions, and selected covariates. For brevity, not all features are shown. The example captures a typical transition from intravenous insulin infusion to subcutaneous injection in an ICU patient, underscoring the complexity and difficulty of the insulin dosing task; (b) Overview of the data curation and preprocessing pipeline for the MIMIC-III insulin cohort. The pipeline incorporates multi-source data integration, exclusion filtering, glucose–insulin pairing, temporal resampling, patient stratification, and task-specific preparation for insulin dosing and hypoglycaemia prediction.

Glucose and Insulin Selection

We begin from a previously curated dataset of glucose-insulin pairs extracted from MIMIC-III based on prior work [195]. Stays in the ICU were included only if they contained at least 24 hours of data, with a minimum of four glucose measurements per day and evidence of short-acting insulin administration (bolus or infusion). To ensure temporal consistency, we truncated each stay in the ICU to the first 30

days after the initial glucose measurement. This glucose-insulin subset includes all admissions to the ICU with at least one insulin administration event, regardless of the type of admission, diagnosis, or demographic of the patient. First and recurrent admissions to the ICU are retained. The cohort includes approximately 9,600 patients for 11,724 ICU stays in the MetaVision system (2008–2012), selected for its improved documentation and greater granularity of insulin events compared to the older CareVue system.

Demographics, Diagnoses, and Anthropometrics

We extract the core demographic and clinical characteristics, including age, sex, height, weight, and diagnosis based on ICD-9 codes. The extraction of weight and height is shown in Table 3.3. Only adult patients (age ≥ 18 years) are retained. Ages above 100 are censored to 90, in line with MIMIC-III de-identification policy. Diagnoses related to neoplasms (ICD-9: 140–239) and pregnancy (ICD-9: 630–679) are excluded to avoid metabolic confounders not related to standard glucose management in the ICU. Height and weight are converted to SI units, and extreme or implausible values (e.g., height < 1 m or > 3 m; weight < 30 kg or > 300 kg) are discarded. ICU stays with missing height or weight are excluded. Diabetes status is determined from ICD-9 codes. Type 1 diabetes is assigned using codes 250.x1 or 250.x3; type 2 diabetes is assigned using 250.x0 or 250.x2. Additional glucose-related disorders are identified using codes 790.21, 790.22, and 790.29. These diagnostic labels are used downstream in both stratification and modelling.

Table 3.3: Height and weight item IDs in MIMIC-III and their unit conversions to SI.

| Measurement | Item IDs (unit) | Conversion to SI |
|-------------|--|---------------------|
| Height (m) | 1394, 226707 (in) | $\times 0.0254$ |
| | 226730 (cm) | $\times 0.01$ |
| Weight (kg) | 580, 581, 763, 224639, 226846, 226512 (kg) | $\times 1$ |
| | 226531 (lbs) | $\times 0.45359237$ |

Prescription Records and Drug Classification

The medication records were extracted from the `PRESCRIPTIONS` table, which includes generic and commercial names, dosage forms, routes of administration, and National Drug Codes (NDC). We standardised drug entries by mapping them to RxNorm Concept Unique Identifiers (RxCUIs), using a hierarchical lookup that prioritises NDCs when available, followed by Generic Sequence Numbers (GSNs), and finally free-text drug names when structured identifiers are missing. To identify medications with potential impact on glucose metabolism, we used the Anatomical Therapeutic Chemical (ATC) classification system maintained by the World Health Organisation. A targeted subset of ATC categories was selected based on their documented metabolic effects and relevance in ICU settings. The selection was based on reviews of the literature and clinical consensus, focusing on drugs known to increase or lower blood glucose or interact with insulin requirements.

Table 3.4: Included ATC medication categories relevant to ICU glucose management, with descriptions, rationale, and occurrence counts in the cohort.

| ATC Code | Description | Reason for Inclusion | Count |
|------------|--|---|--------|
| A10B | Blood Glucose Lowering Drugs, Excluding Insulins | Includes oral agents like metformin, lower glucose. | 1113 |
| C03A, C07 | Thiazide Diuretics; Beta-Blocking Agents | Thiazides raise glucose by reducing insulin secretion [196]. Beta-blockers affect glucose uptake and mask hypoglycaemia symptoms [197]. | 29,382 |
| H02 | Corticosteroids for Systemic Use | Raise glucose via increased gluconeogenesis [198]. | 4,267 |
| L04 | Immunosuppressants | Some, like tacrolimus, cause hyperglycaemia [199]. | 3,280 |
| N05A, N06A | Antipsychotics; Antidepressants | Antipsychotics can lead to weight gain and glucose dysregulation [200]. Antidepressants are associated with hyperglycaemia and diabetes risk [201]. | 10,721 |

Table 3.4 summarises the ATC categories included in our analysis, along with rationale and supporting references. These include insulin and non-insulin glucose-lowering agents, corticosteroids, certain psychotropics, diuretics, beta-blockers,

fluoroquinolones, and immunosuppressants. Each category was chosen due to its known or plausible influence on glycaemic control in critical care. The features derived from these drug classes were incorporated into the modelling pipeline, either as binary indicators or as dose-adjusted exposure metrics. Medications not assigned to the relevant ATC codes or lacking evidence of glycaemic influence were excluded from further analysis. By limiting feature engineering to clinically meaningful drug categories, we reduced the dimensionality and enhanced the interpretability without compromising the coverage of pharmacologic factors influencing glucose dynamics.

Laboratory Measurements

We curated a panel of laboratory measurements relevant to glycaemic control, inflammation, renal and liver function, and the overall severity of critical illnesses. Data were extracted from the `LABEVENTS` table, guided by clinical relevance and item descriptions from `D_LABITEMS`. Only numeric values were retained, and units were standardised where applicable. Non-numeric entries such as textual descriptors (e.g., “trace”, “moderate”) were excluded to ensure consistency in downstream modelling. The final set of laboratory features includes electrolytes, renal biomarkers, liver enzymes, hematologic indices, and markers of inflammation and oxygenation. This includes bicarbonate, sodium, chloride, creatinine, blood urea nitrogen (BUN), haemoglobin, hematocrit, platelet count, white blood cell count (WBC), lactate, C-reactive protein (CRP), and others commonly measured in ICU workflows.

Table 3.5 presents the selected laboratory analytes and their corresponding item IDs. These markers were chosen for their clinical significance in glucose homeostasis and critical illness, and have been frequently included in predictive models involving ICU patients. Where duplicate item IDs captured the same measurement (e.g., across different lab devices), the values were merged and harmonised.

Nutritions, Input and Vitals

To fully capture the factors that influence glucose regulation, we extracted data on nutritional intake, intravenous infusions, vasoactive medications, and routinely charted vital signs. Input records were obtained from both `INPUTEVENTS_CV` and

Table 3.5: Description of selected item IDs used in the dataset, grouped by laboratory, vasopressor, and urine output types.

| Lab (Sec.3.2) | Vasopressor (Sec.3.2) | Urine Output (Sec.3.2) |
|-------------------------|-------------------------|---------------------------------------|
| 50882 – Bicarbonate | 221906 – Norepinephrine | 226559 – Foley |
| 50902 – Chloride | 221289 – Epinephrine | 226560 – Void |
| 50912 – Creatinine | 221749 – Phenylephrine | 226561 – Condom Catheter |
| 50983 – Sodium | 222315 – Vasopressin | 226563 – Suprapubic |
| 51006 – BUN | 221662 – Dopamine | 226564 – Right Nephrostomy |
| 51221 – Haematocrit | 221653 – Dobutamine | 226565 – Left Nephrostomy |
| 51222 – Haemoglobin | 221986 – Milrinone | 226567 – Straight Catheter |
| 51265 – Platelet Count | | 226584 – Ileoconduit |
| 51300 – WBC Count | | |
| 51301 – WBC Count | | 226557 – Right Ureteral Stent |
| 50861 – ALT | | 226558 – Left Ureteral Stent |
| 50862 – Albumin | | 227488 – GU Irrigant Volume In |
| 50878 – AST | | 227489 – GU Irrigant/Urine Volume Out |
| 50885 – Total Bilirubin | | |
| 50813 – Lactate | | |

INPUTEVENTS_MV, encompassing enteral feeds, parenteral nutrition, crystalloid and dextrose infusions, and vasoactive agents. Nutritional formulations such as Ensure, Glucerna, Nepro, and others were mapped to their macronutrient composition using a standardised product dictionary, including caloric density and carbohydrate content per millilitre. Where dilution factors were recorded, nutritional values were adjusted proportionally. Nutrition, dextrose, and lipids were quantified in grams per millilitre, using concentration mappings detailed in Table 3.6 and Table 3.7. Vasoactive agents, including norepinephrine, epinephrine, dopamine, and dobutamine, were identified and harmonised using variants of common names. Dose units were standardised across ICU stays, with normalisation by patient weight where possible. These drugs were selected because of their known physiological impact on glucose metabolism through stress response pathways. Vital signs and respiratory parameters were obtained from CHARTEVENTS. This includes heart rate, systolic and diastolic blood pressure, mean arterial pressure, temperature, respiratory rate, SpO₂, FiO₂, and ventilator settings. All item IDs corresponding to these physiological variables are listed in Table 3.8. Values were converted to consistent units where applicable, and

extreme or non-numeric entries were excluded. These variables were selected for their clinical importance in ICU monitoring and for their potential role as covariates in glucose modelling. Both direct influences (e.g., dextrose administration) and indirect modulators (e.g., vasopressor use or oxygenation status) are represented to enable richer temporal modelling of glycaemic dynamics.

Table 3.6: Ingredients list of Nutritional Product. Each row corresponds to one enteral formula. CHO means carbohydrates.

| Product | Amount (ml) | Calories (kcal/ml) | CHO (g/ml) | Sugars (g/ml) | Fat (g/ml) | Protein (g/ml) |
|-----------------------|-----------------------|------------------------------|----------------------|-------------------------|----------------------|--------------------------|
| Boost Glucose Control | 237 | 0.80 | 0.07 | 0.02 | 0.03 | 0.07 |
| Enlive | 237 | 1.48 | 0.19 | 0.09 | 0.05 | 0.08 |
| Ensure | 237 | 0.93 | 0.14 | 0.06 | 0.03 | 0.04 |
| Ensure Plus | 237 | 1.48 | 0.20 | 0.09 | 0.05 | 0.07 |
| FiberSource HN | 237 | 1.22 | 0.16 | 0.01 | 0.04 | 0.05 |
| Glucerna | 200 | 0.93 | 0.11 | 0.03 | 0.03 | 0.05 |
| Impact | 250 | 1.00 | 0.13 | – | 0.03 | 0.06 |
| Impact with Fiber | 250 | 1.00 | 0.13 | – | 0.03 | 0.06 |
| Isosource 1.5 | 237 | 1.52 | 0.17 | 0.08 | 0.06 | 0.06 |
| Nepro | 220 | 1.82 | 0.15 | 0.03 | 0.10 | 0.08 |
| NovaSource Renal | 237 | 2.00 | 0.18 | 0.08 | 0.10 | 0.09 |
| Nutren 2.0 | 250 | 2.00 | 0.20 | 0.04 | 0.09 | 0.08 |
| Nutren Pulmonary | 250 | 1.50 | 0.10 | 0.05 | 0.09 | 0.07 |
| Nutren Renal | 237 | 2.00 | 0.18 | 0.08 | 0.10 | 0.09 |
| Peptamen 1.5 | 250 | 1.50 | 0.19 | – | 0.06 | 0.07 |
| Peptamen Bariatric | 250 | 1.00 | 0.08 | – | 0.04 | 0.09 |
| Peptamen VHP | 250 | 1.00 | 0.08 | 0.00 | 0.04 | 0.09 |
| Pulmocare | 237 | 1.50 | 0.11 | – | 0.09 | 0.06 |
| Replete | 250 | 1.00 | 0.11 | – | 0.03 | 0.06 |
| Replete with Fiber | 250 | 1.00 | 0.12 | – | 0.03 | 0.06 |
| Two Cal HN | 200 | 2.00 | 0.21 | 0.05 | 0.09 | 0.08 |
| Vivonex | 250 | 1.21 | 0.00 | 0.00 | 0.00 | 0.04 |

Vasopressor Administration Extraction

Vasopressor administration episodes were extracted from the INPUTEVENTS_MV table following the official concept defined in the MIMIC-III code repository¹. Extraction

¹https://github.com/MIT-LCP/mimic-code/blob/main/mimic-iii/concepts_postgres/durations/vasopressor_durations.sql

Table 3.7: Dilution Mapping for Dextrose. TPN means total parenteral nutrition.

| Nutrition Type | Original name | Converted name and factor |
|----------------|--------------------|------------------------------------|
| Dextrose | Dextrose 5% (D5W) | Dextrose, 0.05 g/mL |
| | Dextrose 10% | Dextrose, 0.1 g/mL |
| | Dextrose 20% | Dextrose, 0.2 g/mL |
| | Dextrose 40% | Dextrose, 0.4 g/mL |
| | Dextrose 50% | Dextrose, 0.5 g/mL |
| | Dextrose PN | Dextrose, amount directly in grams |
| Lipids | Lipids 20% | Lipids, 0.2 g/mL |
| | Lipids additive | Lipids, amount directly in grams |
| | TPN with Lipids | Dextrose + Lipids, 0.25 g/mL |
| | TPN without Lipids | Dextrose, 0.25 g/mL |

Table 3.8: ItemIDs and Corresponding Feature Names for Vital Sign Features. Abbreviations: SpO₂, peripheral oxygen saturation; FiO₂, fraction of inspired oxygen; MAP, mean arterial pressure.

| Feature | Item Names | ItemIDs |
|--------------------|---|--|
| Diastolic Pressure | Blood Arterial Blood Pressure diastolic, ART BP Diastolic | 220051, 225310 |
| Systolic Pressure | Blood Arterial Blood Pressure systolic, ART BP Systolic, ABP [Systolic], Arterial BP [Systolic], Manual BP [Systolic], NBP [Systolic] | 220050, 225309, 6, 51, 442, 455 |
| Mean Pressure | Arterial Arterial Blood Pressure mean, ART BP mean, IABP Mean, NBP Mean, Arterial BP Mean | 220052, 225312, 10166, 224322, 456, 52 |
| Heart Rate | Heart Rate | 211, 220045 |
| Temperature | Temperature C, Temperature C (calc), Temperature F, Temperature F (calc), Temperature Fahrenheit, Temperature Celsius | 676, 677, 678, 679, 223761, 223762 |
| SpO ₂ | SpO ₂ , O ₂ saturation pulseoxymetry | 646, 220277 |
| FiO ₂ | FiO ₂ , FIO ₂ | 2981, 3420, 223835 |
| Respiratory Rate | Respiratory Rate | 618, 220210 |

focused on `itemids` representing commonly used vasopressor agents (Table 3.5). Only records with a `statusdescription` not equal to `Rewritten` were retained to ensure inclusion of valid, active medication orders. To derive non-overlapping vasopressor episodes at the patient level, overlapping or consecutive orders were merged using an interval-joining algorithm. Specifically, two administrations were considered part of the same episode if the start of one overlapped with or immediately followed the end of the other. This process yielded a consolidated timeline for vasopressor exposure per stay in the ICU. Episodes with a recorded duration of zero hours (i.e., identical start and end times) were excluded. Furthermore, episodes where the recorded infusion rate was persistently missing or zero throughout the entire administration window were considered clinically implausible and were discarded.

Urine Output

Urine output was extracted following the official fluid balance concept defined in the MIMIC-III code repository². Relevant records were identified from the `OUTPUTEVENTS` table using `itemids` corresponding to documented urine excretion in the MetaVision system. All measurements were standardised to millilitres. Genitourinary irrigant inputs (`itemid` 227488) were treated as negative volumes to account for the net fluid output. Entries with physiologically implausible values ($\geq 5,000$ mL) or missing identifiers for ICU stay were excluded. The remaining observations were aggregated at fixed intervals to produce total urine output, which was used as a proxy for renal function and fluid balance. A summary of the IDS of the urine output item is reported in Table 3.5.

Data Aggregation and Processing

After assembling all constituent data sources, including glucose, insulin, laboratory, vitals, medications, inputs, and outputs, we performed a structured merge using ICU stay identifiers and timestamps. Additional quality filters were applied to ensure internal consistency. Stays with implausible anthropometric values (e.g.,

²https://github.com/MIT-LCP/mimic-code/blob/main/mimic-iii/concepts_postgres/fluid_balance/urine_output.sql

weights below 30 kg or above 300 kg) were removed. Prescription records, which often lacked time-of-day metadata, were assumed to span entire calendar days, from 00:00 to 23:59, unless otherwise indicated by admission timing. This assumption standardises daily drug exposure estimates and facilitates temporal alignment with glucose measurements.

Temporal Resampling and Time Binning

We resampled all dynamic variables into uniform temporal bins to align heterogeneous data sources and enable time-series modelling. Three levels of resolution were implemented: 30-minute, 1-hour, and 2-hour intervals. The binning was anchored at 00:00 each day to preserve the diurnal patterns. This approach was intentionally adopted despite not being standard in ICU data processing, as it supports the modelling of circadian influences on glucose regulation and feeding schedules. A time-of-day feature was also engineered, converting each timestamp into a continuous value between 0 and 1 to represent its relative position within a 24-hour cycle. This variable allows the model to capture circadian variation in physiological states, including glucose metabolism and insulin sensitivity. We applied time-weighted averaging within each bin for continuous infusions, such as dextrose or vasopressors. For intermittent events, such as lab values or insulin boluses, we used forward filling or aggregation (mean, median) depending on the type of feature. The urine outputs were summed within each bin, reflecting the cumulative fluid balance.

Imputation

A structured, rule-based imputation strategy was developed, guided by considerations of clinical plausibility and the preservation of temporal integrity required for sequential modelling. For features derived from input and output events, medication prescriptions, and drug administration records, missing values were assumed to reflect non-occurrence or absence of documentation. In such cases, zero imputation was applied, following established conventions in ICU data analyses and reflecting the clinical assumption that undocumented interventions are unlikely to have been administered. For temporally resolved variables, including physiological

measurements and laboratory results, a hierarchical imputation procedure was used. Within each stay in the ICU, missing values were first addressed using forward and backward propagation to preserve continuity in patient-specific time series. When a variable was entirely unrecorded for a given patient, global median imputation was applied based on the empirical distribution computed across the training cohort. A total of 34 features were fully observed across the dataset. These included key demographic attributes (e.g., age, sex, weight), administrative indicators (e.g., time-step index), categorical comorbidity flags (e.g., diabetes mellitus types 1 and 2), and treatment-related variables (e.g., insulin and vasopressor doses, ventilation status), where missing entries were interpreted as absent interventions or exposures. The proportion of missing data for all partially observed variables is presented in Table 3.9. Features without missingness are omitted for the sake of brevity.

Table 3.9: Percentage of missing observations by feature. Fully observed features are omitted here.

| Feature | Missing Ratio (%) |
|---------------------------------------|-------------------|
| Albumin | 54.06 |
| ALT | 47.52 |
| AST | 47.53 |
| Bilirubin | 47.32 |
| Lactate | 25.11 |
| aPTT | 7.84 |
| PT | 7.59 |
| Temperature | 0.83 |
| Potassium | 0.18 |
| WBC | 0.11 |
| Hemoglobin | 0.11 |
| Hematocrit | 0.10 |
| Chloride | 0.10 |
| BUN | 0.08 |
| Creatinine | 0.08 |
| HCO ₃ | 0.08 |
| Systolic BP | 0.01 |
| <i>Fully-observed features (n=34)</i> | 0.00 |

The Definition of Hypoglycaemia Events Prediction Task

A binary classification task was constructed to identify whether at least one episode of hypoglycaemia (< 70 mg/dL) would occur within a specified future interval. The prediction was performed over fixed temporal horizons of 2, 4, and 6 hours, selected to reflect clinically relevant time frames for anticipatory decision-making in insulin-treated ICU patients. All prediction windows were defined in absolute time units (hours), rather than in terms of time steps, to ensure comparability across datasets with different resampling resolutions. Each sample consisted of an observation window that immediately preceded the prediction window. Observation windows were set to 2, 6, or 12 hours, depending on the modelling configuration, and were designed to capture a temporally coherent representation of the patient’s physiological status and ongoing treatments. A binary label was assigned to each sample based on the presence or absence of any glucose measurement below the defined hypoglycaemia threshold during the prediction window. Samples for which the prediction window extended beyond the available duration of stay in the ICU were excluded to prevent label leakage. Table 3.10 summarises the number of positive and negative labels, along with the number of excluded samples, stratified by prediction horizon and temporal resolution.

Table 3.10: Counts of hypoglycaemia labels at prediction horizons of 2, 4 and 6 hours for each resampling resolution. ‘Excluded’ rows correspond to the final window of each stay that lacks sufficient future data.

| Res. | Horizon | No event | Event | Excluded |
|-------------|----------------|-----------------|--------------|-----------------|
| 0.5 h | 2 h | 2 047 547 | 27 958 | 33 564 |
| | 4 h | 2 002 162 | 39 779 | 67 128 |
| | 6 h | 1 957 972 | 50 405 | 100 692 |
| 1 h | 2 h | 1 026 754 | 13 387 | 16 782 |
| | 4 h | 1 004 006 | 19 353 | 33 564 |
| | 6 h | 981 866 | 24 711 | 50 346 |
| 2 h | 2 h | 516 730 | 6 509 | 8 391 |
| | 4 h | 505 322 | 9 526 | 16 782 |
| | 6 h | 494 230 | 12 227 | 25 173 |

The Definition of Short-acting Insulin Infusion Regimes Task

In this thesis, only short-acting intravenous insulin infusions were considered part of the action space, reflecting their established role in the management of acute dysglycaemia in the intensive care setting. Other forms of insulin administration, including subcutaneous injections, IV boluses, and longer-acting formulations, were excluded from direct modelling and instead treated as covariates. This approach mirrors clinical guidance, which recommends regular insulin administered by continuous variable rate IV infusion for critically ill patients, due to its rapid onset, short half-life, and simple titration [202–204]. We eliminated 5,356 patients whose average frequency of short-acting infusion administration was less than one dose per day, resulting in 3,035 patients remaining as the infusion regime cohort. IV infusion rates were normalised by patient body weight and subsequently discretised into empirically derived bins. The categories were designed to reflect clinically meaningful titration thresholds routinely used in ICU protocols, while also mitigating the sparsity associated with rare dosage levels. Five mutually exclusive bins were defined: no infusion ($0 \text{ units kg}^{-1} \text{ h}^{-1}$), very low ($0\text{--}0.03$), low ($0.03\text{--}0.07$), medium ($0.07\text{--}0.1$), and high (>0.1), as detailed in Table 3.11.

Table 3.11: Frequency of insulin-dose bins by resampling resolution. Infusion rates are body-weight-normalised.

| Res. | Insulin bins (units/kg/h) | | | | |
|-------|---------------------------|----------|-------------|------------|-----------|
| | 0 | [0–0.03) | [0.03–0.07) | [0.07–0.1) | [0.1, +∞) |
| 0.5 h | 551 253 | 74 301 | 75 045 | 18 995 | 16 041 |
| 1 h | 251 011 | 39 759 | 37 440 | 9 284 | 7 802 |
| 2 h | 104 432 | 21 777 | 18 576 | 4 485 | 3 642 |

Data Stratification and Split

To ensure demographic and clinical diversity across the training, validation and test partitions, we applied a stratified splitting procedure based on four binary features: age over 60 years, male sex, body mass index (BMI) above 30 kg/m^2 , and the presence of any glucose-related disorder, including type 1 or type 2 diabetes

or other abnormal glucose diagnoses. These features were combined into a 4-bit stratification code, yielding up to 16 strata of patients. Patients in each stratum were randomly assigned to train, validate and test sets in a 70:15:15 ratio, for both hypoglycaemia prediction and insulin regime tasks. This ensured that all data from a given patient resided entirely within a single partition, thus preserving temporal coherence and preventing data leakage. The numerical features were normalised after splitting, using parameters derived from the training set only. Then these parameters were applied to the validation and test sets to prevent information leakage. Three normalisation strategies were evaluated: Min-Max scaling, z-score normalisation, and robust scaling based on medians and interquartile ranges, with Min-Max used as default.

4

DTR-Bench: An *in silico* Environment and Benchmark Platform for Reinforcement Learning Based Dynamic Treatment Regime

4.1 Introduction

Although integrating reinforcement learning (RL) into dynamic treatment regimes (DTRs) shows promise, it poses significant evaluation challenges. In particular, the ethical and safety concerns of the use of RL require that an algorithm be thoroughly tested *in silico* before any real-world implementation. This highlights the need for a rigorous evaluation of RL-based DTRs in simulated environments first to address the ethical constraints and potential risks associated with directly testing these adaptive algorithms on patients. Thorough *in silico* testing is an essential step to ensure the safety and efficacy of RL-driven personalised medicine approaches. A commonly used approach is off-policy evaluation (OPE) [205, 206], which evaluates the performance of an RL algorithm without new data or experiments. However, OPE’s reliability is compromised by issues such as insufficient data coverage and the limited availability of samples, which are problems exacerbated by the imbalanced nature of medical retrospective data. For example, a study [188] raised concerns about the different evaluation outcomes across OPE methods on the problem of recommending Sepsis treatment, demonstrating that unbalanced data could

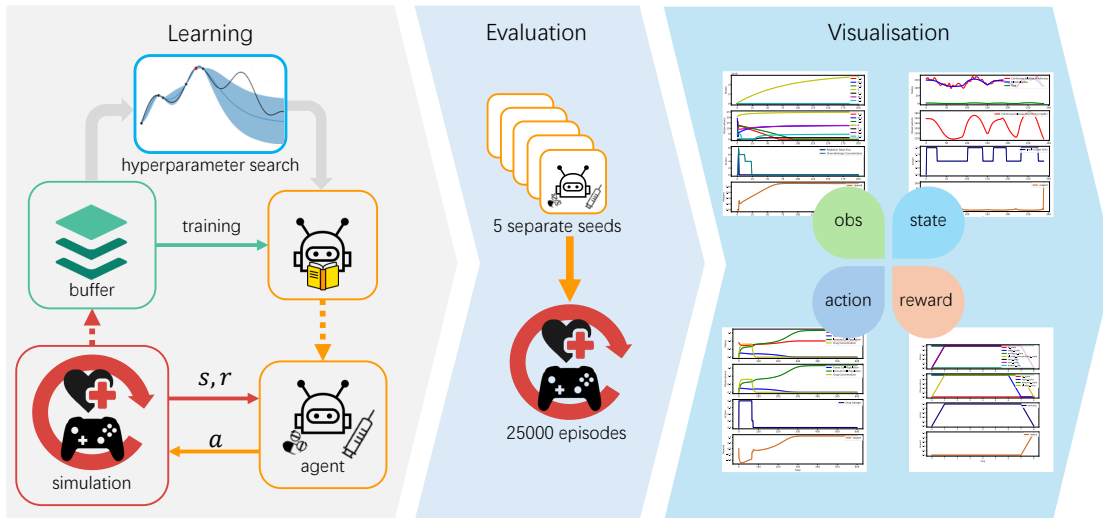


Figure 4.1: Workflow of the DTR-Bench platform. The platform streamlines 3 steps of *DTR-bench*. **Step 1 Learning:** RL algorithms interact with the environment, capturing interaction trajectories in a buffer for efficient off-policy learning. Hyperparameters are tuned using Tree-structured Parzen Estimator (TPE) optimisation [208]. **Step 2: Evaluation:** Models are retrained with the optimised hyperparameters using five distinct seeds, with each model undergoing testing across 5,000 episodes to ensure a fair assessment. **Step 3 Visualisation:** The platform facilitates individual and cohort-averaged trajectory visualisations, supporting intuitive model development and analysis.

increase the variance of the evaluation. The quality of behavioural policy under imbalanced data is important [207], which further calls for a reconsideration of the RL evaluation in DTR problems. Therefore, the complexity of real-world testing and the challenges associated with OPE in healthcare data highlight the broader difficulties of implementing RL in DTR.

Simulation modelling has emerged as an essential tool in response to the difficulties of directly testing RL-based DTRs in clinical settings. By leveraging approaches such as ordinary differential equations (ODEs) and structural causal models (SCMs), simulations can create rich, detailed DTR scenarios based on extensive medical knowledge and patient data. Combined with RL algorithms, these models offer a promising avenue for evaluating the performance of treatment regimes without exposing patients to risk. This methodology has been applied across various medical domains, including cancer chemotherapy [209] and chronic conditions management such as diabetes [210], demonstrating the utility of simulation in developing and refining DTRs. In addition to mathematical modelling, neural

networks (NN), including generative adversarial networks, have also been used to simulate clinical data [211, 212], offering another viable way to create realistic patient datasets for research. Despite their potential, NN approaches can suffer from data coverage, learning bias, and out-of-distribution error under medical data imbalance [213]. This limitation highlights the inherent uncertainty and associated risks when employing NN approaches as environment simulators for RL experiments. Although simulation models are an improvement for evaluating DTRs, the simplified simulations raise valid concerns about whether the model findings will translate accurately to actual clinical settings. To bridge this gap, simulations should more closely mimic the nuances of patient conditions, treatment responses, and outcomes. Mathematical modelling in disease treatment, typically characterised by ODEs and SCMs, faces inherent limitations: (1) **Noise** – Traditional simulation models often overlook noise in observations, a critical aspect of real-world data; (2) **pharmacokinetic/pharmacodynamic (PK/PD) Variance** – Existing models generally use a uniform set of parameters for all patients, neglecting individual differences. This uniformity raises questions about whether RL algorithms genuinely understand treatment principles or simply memorise patterns to achieve high rewards in the simulation environment, as there is no account for individual variability; (3) **Hidden Variables and Missing Values** – Real medical data often includes missing values and inaccessible variables due to clinical decisions or the invasive nature of data collection.

Addressing these issues is crucial for creating simulations that better reflect the complexity of real-life healthcare settings. Historically, many studies have overlooked these important details, resulting in a disconnect between the applicability and reliability of the simulation results.

Simulation-based Dynamic Treatment Regime and Benchmarks

Simulation-based DTRs allow for controlled testing of RL algorithms in a variety of healthcare settings without the ethical concerns of direct patient intervention, such as [214–216]. Simulation-based DTRs benefit from the ability to perform extensive

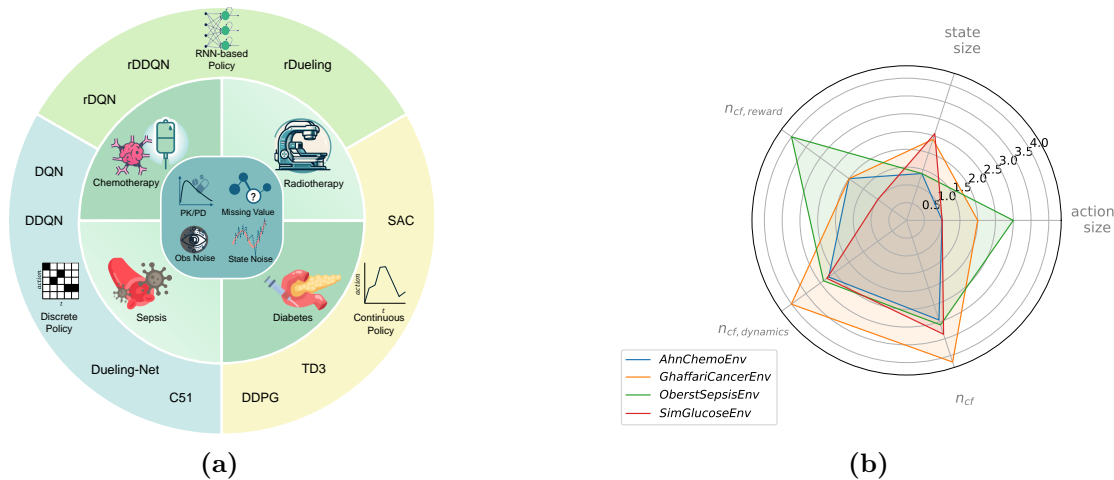


Figure 4.2: A summary of RL algorithms and environments in the DTR-Bench platform. a) DTR-Bench benchmarks discrete-action, continuous-action, and sequential RL algorithms—across four critical healthcare challenges: chemotherapy, radiotherapy, sepsis, and Type-1 diabetes management. The platform evaluates these algorithms rigorously by considering complex factors such as PK/PD modelling, missing values, and noise in observations and states. b) A radar plot showing the environment configurations, where n_{cf} means the counterfactual variables, $n_{cf, reward}$ and $n_{cf, dynamics}$ mean the counterfactual variables affecting the reward function and patient PK/PD dynamics, respectively. n_{cf} , $n_{cf, reward}$ and $n_{cf, dynamics}$ are with Logarithm

testing across a wide range of hypothetical scenarios. For example, an algorithm can perform unlimited trial-and-error iterations on a virtual patient, a practice that is not feasible in real-world scenarios. Despite the extensive benchmarks established for RL in domains such as games [217, 218] and autonomous driving [219], a similar benchmark in DTRs has been conspicuously lacking. This gap highlights a significant need in the field to provide a robust platform that can systematically evaluate and compare the performance of RL algorithms in complex, dynamic healthcare settings. We aim to fill this gap by offering a diverse, standardised set of simulation environments that mimic real-world treatment dynamics and patient responses, providing a critical tool for advancing RL research in healthcare.

Proposed Framework

Here we introduce *DTR-Bench*, a growing in silico RL simulation platform developed specifically to produce more realistic DTR environment simulators. The design of *DTR-Bench* is grounded in a commitment to closely replicate the nuanced intricacies

of healthcare scenarios, providing a robust framework to explore and evaluate RL algorithms. Figure 4.1 shows the overall design workflow. First, we incorporate noise in the observations, the PK/PD variance, hidden variables, and the reality of missing data into our simulation design. This approach aims to enhance the realism of the simulated environments and, by extension, the validity of evaluating RL algorithms in these more nuanced and variable settings. Then, we present a comprehensive benchmarking analysis that leverages the capabilities of *DTR-Bench* across four medical simulation environments, including scheduling chemotherapy, radiotherapy, managing glucose levels in diabetic patients, and formulating sepsis treatment protocols. Our contributions include: (1) developing a comprehensive pipeline that integrates practical clinical considerations into simulation experiment setups; (2) Implementing a collection of mathematical simulation environments for benchmarking RL algorithms, a first in the field of DTR; (3) Standardising RL hyperparameter tuning and evaluation within a treatment regime context, offering a unified and replicable methodology for algorithm evaluation; (4) Establishing the benchmark in DTRs across various clinical research areas, laying the groundwork for broad, interdisciplinary applications as a first in the DTR field.

4.2 Simulation Models

In this section, we introduce the key environmental settings critical to developing more realistic DTR simulations. We aim to bring overly simplified models closer to real-world medical decision-making processes. We experimented with four critical components in designing DTR simulators (i.e., hidden variables, PK/PD, noise, and missing values) and their consequence toward RL performance on various distinct treatment regime scenarios. We detail the fundamental structure of each DTR environment and its problem setup in RL terms, including the observation space, the action space, and the reward design. A summary is shown in Figure 4.2

Our study analyses four simulated medical environments, each targeting a specific disease and type of treatment. These environments are *AhnChemo* [220],

OberstSepsis [221], *GhaffariCancer* [222], and *SimGlucose* [223]¹ They have been chosen for their relevance to different diseases and various treatment approaches, providing a comprehensive range of scenarios to test the RL algorithms. Table 4.1 summarises each environment’s description. Each environment is carefully designed to simulate the complexities and dynamics of real-world medical scenarios. They incorporate varying levels of detail and sophistication in their models, ranging from ordinary differential equations (ODEs) to structural causal models. This variety ensures a robust and comprehensive evaluation of RL algorithms across different healthcare contexts. The details of the design of each simulated environment, including disease dynamics, observation space, action space, and reward functions, are described in the following section.

To enhance clarity, we denote the variables associated with the *AhnChemo*, *GhaffariCancer*, *OberstSepsis*, and *SimGlucose* environments using superscripts $\cdot^{(A)}$, $\cdot^{(G)}$, $\cdot^{(O)}$, and $\cdot^{(S)}$, respectively, where ‘.’ is a placeholder for variables.

Table 4.1: Summary of Simulated Medical Environments in DTRBench. Dynamics abbreviates as Dyn.; Cont./Disc. denotes continuous action spaces that can be discretised.

| Attribute | <i>AhnChemo</i> | <i>GhaffariCancer</i> | <i>OberstSepsis</i> | <i>SimGlucose</i> |
|--------------|-----------------|-----------------------|--|-------------------|
| Disease | Cancer | Cancer | Sepsis | Type-1 Diabetes |
| Treatment | Chemo. | Chemo. Radio. | Antibiotics Ventilation Vasopressors | Insulin Therapy |
| Dynamics | ODE | ODE | SCM | ODE |
| Action Space | Cont./Disc. | Cont./Disc. | Disc. | Cont./Disc. |

4.2.1 *AhnChemoEnv*: A Comprehensive Chemotherapy Simulation Model

The *AhnChemoEnv* is an advanced simulation environment designed to model cellular dynamics under the influence of chemotherapy [220]. Chemotherapy is a way to treat cancer by using medicine to kill cancer cells. This treatment can stop or

¹In the case that the environment is without its original name, we name the environment as the initial of the first author and context.

slow down the growth of tumours, but it can also harm healthy cells because it does not only target cancer cells. The *AhnChemoEnv* incorporates interactions between tumour cells, normal cells, immune cells, and factors in the impact of chemotherapy.

Dynamics Formulation: The dynamics in chemotherapy are modelled by a system of ordinary differential equations (ODEs). The ODEs are based on established mathematical frameworks and biological assumptions reported in [224]. The immune response, competition terms, and control theory for chemotherapy are grounded in known biological interactions, including the immune system's ability to recognise and fight tumour cells, the competition for resources between tumour and normal cells, and the dynamics of drug treatment efficacy and toxicity [225]. The ODEs can be expressed by

$$\begin{cases} \frac{dN^{(A)}}{dt} = r_2^{(A)} N^{(A)} (1 - b_2^{(A)} N^{(A)}) - c_4^{(A)} T^{(A)} N^{(A)} - a_3^{(A)} (q - e^{-B^{(A)}}) N^{(A)}, \\ \frac{dT^{(A)}}{dt} = r_1^{(A)} T^{(A)} (1 - b_1^{(A)} T^{(A)}) - c_2^{(A)} I^{(A)} T^{(A)} - c_3^{(A)} T^{(A)} N^{(A)} - a_2^{(A)} (1 - e^{-B^{(A)}}) T^{(A)}, \\ \frac{dI^{(A)}}{dt} = s^{(A)} + \frac{\rho^{(A)} I^{(A)} T^{(A)}}{\alpha^{(A)} + T^{(A)}} - c_1^{(A)} I^{(A)} T^{(A)} - c_3^{(A)} T^{(A)} N^{(A)} - a_1^{(A)} (1 - e^{-B^{(A)}}) T^{(A)}, \\ \frac{dB^{(A)}}{dt} = -d_2^{(A)} B + u^{(A)}(t) \end{cases} \quad (4.1)$$

The first equation simulates normal cell growth and loss due to tumour competition; the second equation models tumour cell behaviour, including growth and death from interactions with normal and immune cells; the third equation describes changes in the immune cell population, influenced by external sources, tumour presence, competition, and natural death; and the final part focuses on the decay of drug concentration in the blood and the injection of new drugs. The time unit for this ODE system is day, and the default time interval between steps is 6 hours.

Variables Description: The variables within the *AhnChemoEnv* play specific roles in the simulation. A summary of these variables is provided in Table 4.2. The observation space \mathcal{O} comprises the tumour population $T^{(A)}(t)$, the immune population $I^{(A)}(t)$, and the drug concentration $B^{(A)}(t)$. The normal cell population $N^{(A)}(t)$ is hidden as normal cells are not commonly accessible during treatment or diagnosis. Cell counts have been reduced by a factor of 10^{11} so that one unit

represents the carrying capacity of normal cells in the tumour region. It can also prevent overflow problems due to excessively large values. The agent should control $u^{(A)}(t)$, the levels of chemotherapy dosage to eliminate tumour cells in a minimum time with minimal harm to normal cells.

Table 4.2: Variables of the *AhnChemoEnv* ODE system

| Variable | Usage | Description | Unit | Range |
|--------------|---------------|---|--------------------------------|--------|
| $N^{(A)}(t)$ | \mathcal{S} | Normal cell population | No. cells ($\times 10^{11}$) | (0, 2) |
| $T^{(A)}(t)$ | \mathcal{O} | Tumour cell population, representing tumour burden | No. cells ($\times 10^{11}$) | (0, 2) |
| $I^{(A)}(t)$ | \mathcal{O} | Immune cell population, measuring immune response level | No. cells ($\times 10^{11}$) | (0, 2) |
| $B^{(A)}(t)$ | \mathcal{O} | Drug concentration in the bloodstream | unit/L | (0, 1) |
| $u^{(A)}(t)$ | \mathcal{A} | Chemotherapy drug administration rate | unit | (0, 1) |

Parameters Description: The parameter values of the ODEs system are determined based on clinical knowledge [225]. Table 4.2 shows the descriptions and values.

Reward Function: The reward R is defined as

$$r^{(A)}(t) = \frac{N^{(A)}(t)}{N^{(A)}(0)} - \frac{T^{(A)}(t)}{T^{(A)}(0)} + I^{(A)}(t) - u^{(A)}(t) \quad (4.2)$$

It is designed to balance several key factors in cancer treatment: it encourages the increase of normal cells and immune cells, as indicated by the terms $\frac{N(t)}{N(0)}$ and $I(t)$, respectively. Currently, it aims to reduce the number of tumour cells, as shown by the negative term $-\frac{T(t)}{T(0)}$. In addition, the function encourages favourable outcomes with minimal drug dosage, denoted by $-u(t)$. This aspect underscores the importance of reducing drug usage to reduce side effects and costs.

4.2.2 *GhaffariCancerEnv*: A Mixed Radiotherapy and Chemotherapy Model

The *GhaffariCancerEnv* represents a comprehensive simulation environment that models the interactions between normal cells, cancer cells, radiotherapy, and chemotherapy agents at different sites, incorporating the metastatic spread of cancer from a primary to a secondary site [222]. It details how radiotherapy, a treatment

that uses high doses of radiation to eliminate cancer cells and reduce tumour size, and chemotherapy, which uses drugs to destroy cancer cells by inhibiting their growth and division, affect the cellular landscape. Additionally, the model accounts for the time-delayed migration of cancer cells from the primary to the secondary site, adding a crucial dimension to understanding cancer's progression and treatment impacts.

Dynamics Formulation: The dynamics of tumour cell growth and the effect of chemotherapy are determined in studies conducted with mice, in addition to established clinical insights into immune cell reactions to cancer [226, 227]. The influence of radiotherapy and the behaviour of cancer metastasis is determined by mouse data in [222]. The dynamics are formulated in ODEs, which can be expressed by

$$\left\{ \begin{array}{l} \frac{dT_p^{(G)}}{dt} = a_1^{(G)}T_p^{(G)}(1 - b_1^{(G)}T_p^{(G)}) - c_1^{(G)}N_p^{(G)}T_p^{(G)} - D_p^{(G)}T_p^{(G)} - D^{(G)}T_p^{(G)} \\ \quad - \alpha_1^{(G)}T_p^{(G)} + \gamma^{(G)}u^{(G)} - K_{T1}^{(G)} \frac{T_p^{(G)}M^{(G)}}{W_{T1}^{(G)} + T_p^{(G)}} \\ \frac{dN_p^{(G)}}{dt} = e_1^{(G)}C^{(G)} - p_1^{(G)}N_p^{(G)}T_p^{(G)} - f_1^{(G)}N_p^{(G)} - \epsilon^{(G)}D^{(G)}N_p^{(G)} + \gamma_2^{(G)}v^{(G)} \\ \quad - K_{1N}^{(G)} \left(\frac{N_p^{(G)}M^{(G)}}{W_{1N}^{(G)} + N_p^{(G)}} \right) \\ \frac{dL_p^{(G)}}{dt} = -m^{(G)}L_p^{(G)} + j_1^{(G)} \frac{T_p^{(G)}}{k_1^{(G)} + T_p^{(G)}} - q_1^{(G)}L_p^{(G)}T_p^{(G)} + r_{11}^{(G)}N_p^{(G)}T_p^{(G)} + r_{12}^{(G)}C^{(G)}T_p^{(G)} \\ \quad - u_1^{(G)}N_p^{(G)}L_p^{(G)^2} - \epsilon^{(G)}D^{(G)}L_p^{(G)} + \gamma_3^{(G)}x^{(G)} - K_{1L}^{(G)} \left(\frac{L_p^{(G)}M^{(G)}}{W_{1L}^{(G)} + L_p^{(G)}} \right) \\ \frac{dC^{(G)}}{dt} = \alpha^{(G)} - \beta^{(G)}C^{(G)} - K_{1C}^{(G)} \left(\frac{C^{(G)}M^{(G)}}{W_{1C}^{(G)} + C^{(G)}} \right) \\ \frac{dT_s^{(G)}}{dt} = a_2^{(G)}T_s^{(G)}(1 - b_2^{(G)}T_s^{(G)}) - c_2^{(G)}N_s^{(G)}T_s^{(G)} - D_s^{(G)}T_s^{(G)} + \alpha_2^{(G)}T_p^{(G)}(t - \tau^{(G)}) \\ \quad - K_T^{(G)} \left(\frac{T_s^{(G)}M^{(G)}}{W_{2T}^{(G)} + T_s^{(G)}} \right) \end{array} \right. \quad (4.3)$$

$$\left\{ \begin{array}{l}
\frac{dN_s^{(G)}}{dt} = e_2^{(G)} C^{(G)} - p_2^{(G)} N_s^{(G)} T_s^{(G)} - f_2^{(G)} N_s^{(G)} - K_{2N}^{(G)} \left(\frac{N_s^{(G)} M^{(G)}}{W_{2N}^{(G)} + N_s^{(G)}} \right) \\
\frac{dL_s^{(G)}}{dt} = -m_2^{(G)} L_s^{(G)} + j_2^{(G)} \frac{T_s^{(G)}}{k_2^{(G)} + T_s^{(G)}} - q_2^{(G)} L_s^{(G)} T_s^{(G)} + r_{21}^{(G)} N_s^{(G)} T_s^{(G)} \\
\quad + r_{22}^{(G)} C^{(G)} T_s^{(G)} - u_2^{(G)} N_s^{(G)} L_s^{(G)2} - K_{2L}^{(G)} \left(\frac{L_s^{(G)} M^{(G)}}{W_{2L}^{(G)} + L_s^{(G)}} \right) \\
\frac{dc_1^{(G)}}{dt} = \mu_{c1}^{(G)} v_M^{(G)} \left(1 - \frac{c_1^{(G)}}{k_{c1}^{(G)}} \right) \\
\frac{dc_2^{(G)}}{dt} = \mu_{c2}^{(G)} v_M^{(G)} \left(1 - \frac{c_2^{(G)}}{k_{c2}^{(G)}} \right) \\
\frac{dM^{(G)}}{dt} = -\mu_M^{(G)} + v_M^{(G)} \\
\frac{du^{(G)}}{dt} = D^{(G)} T_p^{(G)} - \gamma_1^{(G)} u^{(G)} - \delta^{(G)} u^{(G)} \\
\frac{dv^{(G)}}{dt} = \epsilon^{(G)} D^{(G)} N_p^{(G)} - \gamma_2^{(G)} v^{(G)} - \delta^{(G)} v^{(G)} \\
\frac{dx^{(G)}}{dt} = \epsilon^{(G)} D^{(G)} L_p^{(G)} - \gamma_3^{(G)} x^{(G)} - \delta^{(G)} x^{(G)}
\end{array} \right. \quad (4.4)$$

where $D_p^{(G)} = d_1^{(G)} \frac{L_p^{(G)l}}{s^{(G)} T_p^{(G)l} + L_p^{(G)l}}$, $D_s^{(G)} = d_2^{(G)} \frac{L_s^{(G)l}}{s^{(G)} T_s^{(G)l} + L_s^{(G)l}}$ and $l = 2.09$ is a scalar. The full parameter value table can be found in A.3, A.4 and A.5 in Appendix A.

The ODE system presented in Equation 4.3 describes the complex interactions within the cancer environment under treatment. Specifically, $\frac{dT_p^{(G)}}{dt}$ represents the dynamics of the tumour cell population at the primary site, accounting for natural growth, inhibition by natural killer (NK) cells, reduction due to radiotherapy and chemotherapy, and the effect of metastasis. $\frac{dN_p^{(G)}}{dt}$ and $\frac{dL_p^{(G)}}{dt}$ model the dynamics of NK cells and CD8+ T cells at the primary site, respectively, including the effects of chemotherapy, radiotherapy, and their interactions with tumour cells. $\frac{dC^{(G)}}{dt}$ describes the dynamics of lymphocytes, excluding NK cells and CD8+ T cells. The equations for $\frac{dT_s^{(G)}}{dt}$, $\frac{dN_s^{(G)}}{dt}$, and $\frac{dL_s^{(G)}}{dt}$ mirror those at the primary site but apply to the secondary site, emphasising the role of metastasis. The dynamics of chemotherapy agent concentration in the blood is modelled by $\frac{dM^{(G)}}{dt}$, while $\frac{du^{(G)}}{dt}$, $\frac{dv^{(G)}}{dt}$, and $\frac{dx^{(G)}}{dt}$ capture the populations of cancer cells, NK cells, and CD8+ T

cells exposed to radiation, respectively. Each equation integrates the influence of treatment modalities, cellular interactions, and the migration of cells between sites to provide a comprehensive view of the cancer treatment landscape.

Variables Description: The variables of the ODEs system are summarised in Table 4.3.

Table 4.3: Variables of the *GhaffariCancerEnv* ODEs system

| Variable | Usage | Description | Unit | Range |
|----------------|---------------|--|-----------|----------------|
| $T_p^{(G)}(t)$ | \mathcal{O} | The total tumour cell population at the primary site | cells | $(0, 10^{11})$ |
| $N_p^{(G)}(t)$ | \mathcal{O} | The concentration of NK cells per litre of blood (cells/L) at the primary site | cells / L | $(0, 10^{10})$ |
| $L_p^{(G)}(t)$ | \mathcal{O} | The concentration of CD8+ T cells per litre of blood (cells/L) at the primary site | cells / L | $(0, 10^{10})$ |
| $C^{(G)}(t)$ | \mathcal{O} | The concentration of lymphocytes per liter of blood (cells/L), not including NK cells and active CD8+T | cells / L | $(0, 10^{11})$ |
| $T_s^{(G)}(t)$ | \mathcal{O} | The total tumour cell population at the secondary site | cells | $(0, 10^{11})$ |
| $N_s^{(G)}(t)$ | \mathcal{O} | The concentration of NK cells per liter of blood (cells/L) at the secondary site | cells / L | $(0, 10^{10})$ |
| $L_s^{(G)}(t)$ | \mathcal{O} | The concentration of CD8+ T cells per litre of blood (cells/L) at the secondary site | cells / L | $(0, 10^{10})$ |
| $M^{(G)}(t)$ | \mathcal{S} | The concentration of chemotherapy agent per litre of blood (mg/L) | mg/L | $(0, 10^{10})$ |
| $u^{(G)}(t)$ | \mathcal{S} | The population of cancer cells that have been exposed to radiation | cells | $(0, 10^{11})$ |
| $v^{(G)}(t)$ | \mathcal{S} | The population of NK cells that have been exposed to radiation | cells | $(0, 10^{11})$ |
| $x^{(G)}(t)$ | \mathcal{S} | The population of CD8+ T cells that have been exposed to radiation | cells | $(0, 10^{11})$ |
| $D^{(G)}$ | \mathcal{A} | The radiation dose administered | Gy | $(0, 10)$ |
| $v_M^{(G)}(t)$ | \mathcal{A} | The per-litre blood dosage of chemotherapy agents | mg/L | $(0, 8)$ |

Parameters Description: The parameter values of the ODEs system are determined using mouse and human parameters [227]. The description and values of the parameters are shown in Table 4.3, and Table A.3, A.4, A.5, respectively.

Reward Function: The reward function designed for this environment comprises a per-step tumour reduction reward $r_{T_{ghaf}}(t)$ and an outcome reward $r_{oghaf}(t)$.

$r_{T_{ghaf}}(t)$ is the reward function that encourages the agent to decrease the number of tumour cells at both the primary and secondary sites. A penalty is imposed if the total tumour population exceeds the initial population. In contrast, a reward is given when the agent successfully decreases the tumour population below its initial level. The reward is given by

$$r_T^{(G)}(t) = 1 - \frac{T_p^{(G)}(t) + T_s^{(G)}(t)}{T_p^{(G)}(0) + T_s^{(G)}(0)} \quad (4.5)$$

$r_o^{(G)}(t)$ is the outcome reward designed to encourage the complete elimination of tumours and discourage early termination caused by the oversized tumour population. In particular, +100 is given if the tumour population is 0 at both the primary and secondary site; -100 is given if any of the tumour population is higher than 10^{11} cells/L; and 0 is given otherwise.

$$r_o^{(G)}(t) = \begin{cases} -100 & \text{if } T_p^{(G)}(t) > 10^{11} \text{ or } T_s^{(G)}(t) > 10^{11}, \\ +100 & \text{if } T_p^{(G)}(t) < 1 \text{ and } T_s^{(G)}(t) < 1, \\ 0 & \text{otherwise} \end{cases} \quad (4.6)$$

The total reward function for this environment is a sum of the previous reward functions, denoted by $r^{(G)}(t) = r_T^{(G)}(t) + r_o^{(G)}(t)$.

4.2.3 *OberstSepsisEnv*: A Sepsis Simulator

OberstSepsisEnv is a synthetic environment designed for the treatment of sepsis, based on the framework of a Markov Decision Process (MDP) [221]. This environment incorporates four vital signs: heart rate ($H^{(O)}$), blood pressure ($B^{(O)}$), oxygen concentration ($O^{(O)}$), and glucose levels ($G^{(O)}$), each categorised into discrete states such as low, normal, and high. It offers three treatment options: antibiotics ($a^{(O)}$), vasopressors ($v^{(O)}$), and mechanical ventilation ($m^{(O)}$), available for application at each decision point. A critical condition of death is modelled to occur when at least three vital signs are simultaneously outside the normal range. In addition, the model includes a binary variable that indicates the presence of diabetes ($D^{(O)}$).

Dynamics Formulation: The dynamics of *OberstSepsisEnv* is determined based on the experience of the author [221]. The transition probability for each variable is shown in Table 4.4. The model intricately details how these treatments, when turned on or off, influence the transition of vital signs from states such as low (L), normal (N), high (H), super low (LL), and super high (HH), with respective probabilities. For instance, antibiotic use impacts heart rate and blood pressure, while vasopressors have a more complex effect on blood pressure and glucose levels, particularly distinguishing between diabetic and non-diabetic scenarios. This model also accounts for natural fluctuations in vital signs, both under the influence of treatment and oxygenation. A patient is discharged only when all vital signs have returned to normal and all treatments have been stopped. Death occurs if 3 or more vitals are abnormal.

Table 4.4: Transition probability of the *OberstSepsisEnv*

| Variable | Cur. | New | Change | Effect |
|-----------|------|-----------|---|---|
| $a^{(O)}$ | - | on | - | $H^{(O)}$ H \rightarrow N w.p. 0.5 |
| | | | | $B^{(O)}$ H \rightarrow N w.p. 0.5 |
| | on | off | withdrawn | $H^{(O)}$ N \rightarrow H w.p. 0.1 |
| | | | | $B^{(O)}$ N \rightarrow H w.p. 0.5 |
| $m^{(O)}$ | - | on | - | $O^{(O)}$ L \rightarrow N w.p. 0.7 |
| | on | off | withdrawn | $O^{(O)}$ N \rightarrow L w.p. 0.1 |
| $v^{(O)}$ | - | on | - | $B^{(O)}$ L \rightarrow N w.p. 0.7 (non-diabetic) |
| | | | | N \rightarrow H w.p. 0.7 (non-diabetic) |
| | | | | L \rightarrow N w.p. 0.5 (diabetic) |
| | | | | L \rightarrow H w.p. 0.4 (diabetic) |
| | | | N \rightarrow H w.p. 0.9 (diabetic) | |
| | | | $G^{(O)}$ LL \rightarrow L, L \rightarrow N, N \rightarrow H, | |
| | | | H \rightarrow HH w.p. 0.5 (diabetic) | |
| | | | N \rightarrow L w.p. 0.1 (non-diabetic) | |
| on | off | withdrawn | $B^{(O)}$ H \rightarrow N w.p. 0.1 (non-diabetic) | |
| | | | N \rightarrow L w.p. 0.05 (diabetic) | |
| | | | H \rightarrow N w.p. 0.05 (diabetic) | |
| $H^{(O)}$ | | | | |
| $B^{(O)}$ | | | | |
| $O^{(O)}$ | | fluct. | | • fluctuates ± 1 w.p. 0.1 |
| $G^{(O)}$ | | | | • fluctuates ± 1 w.p. 0.3 (diabetic) |

Variables Description: The variable descriptions for the *OberstSepsisEnv* are shown in Table 4.5

Table 4.5: Variables of the *OberstSepsisEnv* ODEs

| Variable | Usage | Description | Class |
|-----------|---------------|--------------------------|--|
| $H^{(O)}$ | \mathcal{O} | Heart rate | [Low, Normal, High] |
| $B^{(O)}$ | \mathcal{O} | Blood pressure | [Low, Normal, High] |
| $O^{(O)}$ | \mathcal{O} | Oxygen concentration | [Low, Normal, High] |
| $G^{(O)}$ | \mathcal{O} | Glucose levels | [Super Low, Low, Normal, High, Super High] |
| $D^{(O)}$ | \mathcal{S} | The presence of diabetes | [diabetic, non-diabetic] |
| $a^{(O)}$ | \mathcal{A} | Antibiotics | [On, Off] |
| $v^{(O)}$ | \mathcal{A} | Vasopressors | [On, Off] |
| $m^{(O)}$ | \mathcal{A} | Mechanical ventilation | [On, Off] |

Reward Function: The reward function designed in *OberstSepsisEnv* is shown in Equation 4.7.

$$r^{(O)}(t) = \begin{cases} 1 & \text{if discharged,} \\ -1 & \text{if deceased,} \\ 0 & \text{otherwise.} \end{cases} \quad (4.7)$$

A reward of +1 is granted for patient survival, transitioning the system into an absorbing state that continuously accumulates rewards. In contrast, a penalty of -1 is imposed for patient mortality. For all other outcomes, no reward or penalty is administered to the agent. This reward structure is designed to encourage the agent to give treatment toward a discharged state while avoiding the patient’s death.

4.2.4 *SimGlucoseEnv*: A glucose-insulin simulating environment

The *SimGlucoseEnv* describes the glucose-insulin system considering food intake [223]. The model connects plasma glucose and insulin concentrations with their respective fluxes. It simulates how glucose and insulin levels in the bloodstream are influenced by various processes such as glucose absorption, renal excretion, insulin fluxes, and insulin degradation.

Dynamics Formulation: The dynamics are determined based on real-world data from 300 patients, covering a range of metabolic parameters and demographic characteristics. The dynamics are formulated in ODEs, which are developed based on computational models that simulate interactions between insulin dosing, carbohydrate intake, and glucose metabolism as detailed in [228]. The ODEs

can be expressed by

$$\left\{ \begin{array}{l} \frac{dG_p^{(S)}(t)}{dt} = EGP^{(S)}(t) + Ra^{(S)}(t) - U_{ii}^{(S)} - E^{(S)}(t) - k_1^{(S)}G_p^{(S)}(t) + k_2^{(S)}G_t^{(S)}(t) \\ \frac{dG_t^{(S)}(t)}{dt} = -U_{id}^{(S)}(t) + k_1^{(S)}G_p^{(S)}(t) - k_2^{(S)}G_t^{(S)}(t) \\ \frac{dX^{(S)}(t)}{dt} = -p_{2u}^{(S)} \cdot X^{(S)}(t) + p_{2u}^{(S)} \cdot [I^{(S)}(t) - I_b^{(S)}] \\ \frac{dI^{(S)}(t)}{dt} = -k_i^{(S)} \cdot [I^{(S)}(t) - I^{(S)}(t)] \\ \frac{dX^{(S)L}(t)}{dt} = -k_i^{(S)}[X^{(S)L}(t) - I^{(S)}(t)] \\ \frac{dS_{sto}^{(S)}(t)}{dt} = CHO^{(S)}(t) - k_{sto}^{(S)} \cdot S_{sto}^{(S)}(t) \\ \frac{dQ_{sto}^{(S)}(t)}{dt} = k_{sto}^{(S)} \cdot S_{sto}^{(S)}(t) - k_{gut}^{(S)} \cdot Q_{sto}^{(S)}(t) \\ \frac{dQ_{gut}^{(S)}(t)}{dt} = k_{gut}^{(S)} \cdot Q_{sto}^{(S)}(t) - k_{abs}^{(S)} \cdot Q_{gut}^{(S)}(t) \end{array} \right. \quad (4.8)$$

where

$$\left\{ \begin{array}{l} Ra^{(S)}(t) = \frac{f^{(S)} \cdot k_{abs}^{(S)} \cdot Q_{gut}^{(S)}(t)}{BW^{(S)}} \\ E^{(S)}(t) = k_{e1}^{(S)}[G_p^{(S)}(t) - k_{e2}^{(S)}] \\ U_{id}^{(S)}(t) = \frac{[V_{m0}^{(S)} + V_{mx}^{(S)}X^{(S)}(t)]G_t^{(S)}(t)}{K_{m0}^{(S)} + G_t^{(S)}(t)} \\ EGP^{(S)}(t) = k_{p1}^{(S)} - k_{p2}^{(S)}G_p^{(S)}(t) - k_{p3}^{(S)}X^{(S)L}(t) \end{array} \right. \quad (4.9)$$

This ODE system models glucose absorption ($Ra^{(S)}(t)$) from ingested carbohydrates ($CHO^{(S)}(t)$), the body's glucose production ($EGP^{(S)}(t)$), the dynamics of insulin ($I^{(S)}(t)$), and insulin's impact on glucose utilisation ($X^{(S)}(t)$) and its delayed action in the liver ($X^{(S)L}(t)$). The equations track glucose concentrations in plasma ($G_p^{(S)}(t)$) and tissue ($G_t^{(S)}(t)$), account for renal glucose excretion ($E^{(S)}(t)$), and quantify insulin-dependent glucose use ($U_{id}^{(S)}(t)$). In addition, the model delineates the digestion process, distinguishing between the solid ($S_{sto}^{(S)}(t)$) and liquid ($Q_{sto}^{(S)}(t)$) carbohydrate states in the stomach before their absorption in the gut ($Q_{gut}^{(S)}(t)$). The model directly correlates dietary intake and insulin administration with blood glucose levels through these dynamics, offering a sophisticated tool to simulate glucose–insulin interactions and aiding effective diabetes management strategies. The criteria for terminating or truncating this environment are as follows:

termination occurs if the basal plasma glucose level G_p falls below 10 or exceeds 600. If neither condition is met, the environment progresses in 288 steps, which is equivalent to a 24-hour treatment period. Upon completion of these 288 steps, the environment is considered *truncated*.

Variable Description: The variable descriptions for the *SimGlucoseEnv* are shown in Table.4.6.

Table 4.6: Variables of the *SimGlucoseEnv* ODEs

| Variable | Usage | Description | Unit | Range |
|--------------------|---------------|---|-------|-----------|
| $G_p^{(S)}(t)$ | \mathcal{O} | The amount of glucose in plasma | mg/dL | (10, 600) |
| $G_t^{(S)}(t)$ | \mathcal{S} | The amount of glucose in the tissue | mg/dL | – |
| $I^{(S)}(t)$ | \mathcal{S} | The insulin concentration | U/day | – |
| $X^{(S)}(t)$ | \mathcal{S} | The insulin action on glucose utilisation | – | – |
| $X^{(S)L}(t)$ | \mathcal{S} | The delayed insulin action in the liver | – | – |
| $S_{sto}^{(S)}(t)$ | \mathcal{S} | The amount of solid carbohydrates in stomach | mg | – |
| $Q_{sto}^{(S)}(t)$ | \mathcal{S} | The amount of liquid carbohydrates in stomach | mg | – |
| $Q_{gut}^{(S)}(t)$ | \mathcal{S} | The amount of liquid carbohydrates in gut | mg | – |
| $Ra^{(S)}(t)$ | \mathcal{S} | The rate of glucose absorption in the blood | – | – |
| $E^{(S)}(t)$ | \mathcal{S} | The renal excretion of glucose | mg/dL | – |
| $EGP^{(S)}(t)$ | \mathcal{S} | The endogenous glucose production (EGP) | U/day | – |
| $U_{id}^{(S)}(t)$ | \mathcal{S} | The insulin-dependent utilisation takes place in the remote compartment | – | – |
| $CHO^{(S)}(t)$ | \mathcal{S} | The amount of ingested carbohydrates | g | (0, 200) |
| $a^{(S)}(t)$ | \mathcal{A} | The insulin concentration of the insulin pump | U/h | (0, 30) |

Parameters Description: The ODE system is derived from human physiological data, accommodating the creation of 100 virtual adults, 100 adolescents, and 100 children profiles. Given individual variability among patients, specific default values for these parameters are not assigned. Instead, the parameters vary across the simulated individuals to better reflect the diversity found in real-world populations. A description of the virtual patient variable for adolescents is provided in Table A.6. The environment includes 10 adolescents, 10 adults, and 10 children, each with a unique set of dynamic parameters. Full details can be found at [223].

Reward Function: The reward function calculates a numerical reward based on current and subsequent basal glucose ($G_p^{(S)}(t)$) levels in plasma and the status of

the episode. The reward function is defined as $r^{(S)}(t) = r_{\text{risk}}^{(S)}(t) + r_{\Delta}^{(S)}(t) + r_o^{(S)}(t)$. $r_{\text{risk}}^{(S)}$ is the risk index [229] with a \log_{10} transformation. This reward encourages the agent to take actions that reduce diabetes-related risks.

$$r_{\text{risk}}^{(S)}(t) = -\log_{10} \left(\left[1.509 \left(\ln(G_p^{(S)}(t))^{1.084} - 5.381 \right) \right]^2 \right) \quad (4.10)$$

$r_{\Delta}^{(S)}$ is the fluctuation reward, which penalises abrupt fluctuations in basal glucose levels. This reward discourages actions that lead to significant variations in glucose levels, promoting a more stable glucose management.

$$r_{\Delta}^{(S)}(t) = \begin{cases} 0 & \text{if } \Delta G_p^{(S)}(t) < 30, \\ -\frac{1}{30}(\Delta G_p^{(S)}(t) - 30), & \text{if } 30 \leq \Delta G_p^{(S)}(t) < 60, \\ -1 & \text{if } \Delta G_p^{(S)}(t) \geq 60, \end{cases} \quad (4.11)$$

where $\Delta G_p^{(S)}(t) = G_p^{(S)}(t) - G_p^{(S)}(t - 1)$.

$r_o^{(S)}$ is the reward for the outcome, which is given to encourage the agent to avoid early termination due to hypoglycaemia or extreme hyperglycaemia (i.e., when $G_p^{(S)}(t) < 10$ or $G_p^{(S)}(t) > 600$).

$$r_o^{(S)}(t) = \begin{cases} 100 & \text{if } \forall G_p^{(S)}(1:T) \in [10, 600], \\ -100 & \text{if } \exists G_p^{(S)}(t) < 10 \text{ or } \exists G_p^{(S)}(t) > 600, \\ 0 & \text{otherwise.} \end{cases} \quad (4.12)$$

4.3 Results

This section presents the core attributes of the *DTR-Bench* platform, followed by empirical findings from a series of reinforcement learning (RL) experiments conducted across four simulated clinical environments. The platform has been designed to enhance the fidelity and rigour of in silico evaluations for Dynamic Treatment Regimes (DTRs). Preliminary findings suggest that modifications introduced within *DTR-Bench* may improve the realism and discriminative capacity of the RL evaluation pipelines, thus supporting more clinically plausible assessments of algorithmic performance.

4.3.1 Overview of Platform Capabilities

DTR-Bench constitutes an open-source testbed developed to standardise and streamline the evaluation of RL-based DTRs in healthcare. Several key features are incorporated to address the methodological limitations observed in previous simulation frameworks. First, the platform has a structured evaluation framework specifically tailored for RL algorithms. By explicitly modelling key clinical complexities, such as pharmacokinetic and pharmacodynamic variability, unobserved confounding, data missingness, and stochasticity in treatment responses, *DTR-Bench* seeks to approximate the uncertainty and heterogeneity inherent in real-world care trajectories. Second, the system has been designed for seamless integration with widely adopted RL interfaces, including *Gym* [230], *Gymnasium* [231], and *Tianshou* [232]. This compatibility facilitates adoption by the broader RL research community and enables plug-and-play policy experimentation with minimal friction. Third, the platform accommodates flexible development workflows, allowing users to implement novel RL algorithms or extend existing ones with minimal overhead. In addition, support for automated hyperparameter tuning has been implemented through native integration with *Optuna* [233], enabling efficient model calibration through systematic search strategies. Finally, visualisation utilities are included to support intuitive inspection of agent behaviours and outcome distributions. These tools may assist in the quantitative evaluation of policies and the temporal dynamics within the simulated environments. Taken together, these features position *DTR-Bench* as a comprehensive and extensible resource for reproducible development and evaluation of RL methods tailored to high-stakes dynamic clinical decision-making.

4.3.2 Benchmark Results

We evaluate discrete-action algorithms (DQN, DDQN, DDQN-duel, C51, dSAC), their sequential variants (rDQN, rDDQN, rDDQN-duel, rC51, discrete-rSAC), and continuous-action algorithms (DDPG, TD3, SAC) on *AhnChemoEnv*, *GhaffariCancerEnv*, *OberstSepsisEnv*, and *SimGlucoseEnv* with four settings, shown in Table 4.7, 4.8, 4.9, and, 4.10, respectively. All results are based on optimised

Table 4.7: Reward performance of various RL algorithms on *AhnChemoEnv* under realist DTR enhancements (\hat{p} , \hat{p}_1 , \hat{p}_2 , \hat{p}_3). \hat{p} , \hat{p}_1 , \hat{p}_2 , \hat{p}_3 denote the original, PK/PD variant, 'PK/PD+noise' and 'PK/PD+noise+missing' enhancement, respectively. Higher is better. The best reward in each column is highlighted in red and the second best in blue. π_b is the strongest baseline policy (random, zero-drug, or max-drug). Notation and highlighting in the following tables will be the same.

| Policy | \hat{p} | \hat{p}_1 | \hat{p}_2 | \hat{p}_3 |
|-----------|--------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|
| π_b | 62.29 \pm 12.62 | 64.13 \pm 319.24 | 60.60 \pm 324.14 | 60.60 \pm 324.14 |
| DQN | 214.18 \pm 141.64 | 129.24 \pm 367.54 | 95.76 \pm 389.00 | 126.55 \pm 394.22 |
| DDQN | 370.61 \pm 18.36 | 170.02 \pm 375.16 | 134.36 \pm 376.11 | 131.96 \pm 389.68 |
| DDQN-duel | 248.59 \pm 66.77 | 95.04 \pm 349.67 | 105.22 \pm 344.06 | 87.44 \pm 354.02 |
| C51 | 144.30 \pm 90.10 | 108.05 \pm 344.97 | 122.53 \pm 377.36 | 128.05 \pm 386.88 |
| dSAC | 114.06 \pm 82.68 | 137.88 \pm 404.36 | 124.14 \pm 385.20 | 132.34 \pm 357.61 |
| rDQN | 267.78 \pm 62.65 | 113.90 \pm 352.34 | 120.34 \pm 367.73 | 91.13 \pm 368.63 |
| rDDQN | 317.34 \pm 35.24 | 132.60 \pm 350.05 | 101.11 \pm 372.60 | 88.92 \pm 342.95 |
| rC51 | 242.18 \pm 82.29 | 36.58 \pm 277.86 | 64.17 \pm 331.92 | 73.49 \pm 321.97 |
| drSAC | 341.01 \pm 24.63 | 126.53 \pm 378.35 | 99.46 \pm 391.72 | 102.30 \pm 389.97 |
| DDPG | 113.28 \pm 55.93 | 80.95 \pm 327.33 | 124.67 \pm 376.75 | 102.50 \pm 349.72 |
| TD3 | 90.32 \pm 54.04 | 82.10 \pm 339.29 | 87.38 \pm 344.40 | 76.97 \pm 347.48 |
| SAC | 151.72 \pm 88.35 | 89.09 \pm 338.32 | 115.99 \pm 363.45 | 108.80 \pm 374.35 |

Table 4.8: Reward performance on *GhaffariCancerEnv* under realist DTR enhancements. Notations are the same as Table 4.7.

| Policy | \hat{p} | \hat{p}_1 | \hat{p}_2 | \hat{p}_3 |
|-----------|--------------------------------------|--------------------------------------|--------------------------------------|--------------------------------------|
| π_b | -90.78 \pm 2.96 | -90.79 \pm 2.67 | -90.76 \pm 2.72 | -90.76 \pm 2.72 |
| DQN | 134.24 \pm 26.45 | 128.41 \pm 20.04 | 131.14 \pm 28.43 | 90.41 \pm 54.10 |
| DDQN | 86.94 \pm 89.85 | 132.76 \pm 34.16 | 126.40 \pm 18.41 | 115.26 \pm 32.23 |
| DDQN-duel | 129.26 \pm 69.11 | 130.71 \pm 25.39 | 125.29 \pm 17.78 | 113.70 \pm 26.45 |
| C51 | 160.91 \pm 62.61 | 136.28 \pm 34.87 | 139.89 \pm 34.00 | 121.88 \pm 58.38 |
| dSAC | 93.65 \pm 73.98 | 127.65 \pm 12.29 | 75.23 \pm 81.40 | 96.86 \pm 70.69 |
| rDQN | 63.66 \pm 85.66 | 126.53 \pm 13.36 | 124.89 \pm 16.31 | 104.80 \pm 55.21 |
| rDDQN | 57.40 \pm 87.30 | 125.09 \pm 20.41 | 108.64 \pm 43.99 | 67.04 \pm 72.38 |
| rC51 | -82.88 \pm 16.37 | -74.74 \pm 22.22 | -79.81 \pm 19.60 | -83.36 \pm 16.91 |
| drSAC | 143.45 \pm 18.09 | 134.60 \pm 16.05 | 139.64 \pm 15.34 | 114.72 \pm 36.64 |
| DDPG | -33.52 \pm 84.31 | -70.15 \pm 39.87 | -29.19 \pm 73.69 | -48.61 \pm 76.66 |
| TD3 | 62.81 \pm 76.83 | 78.30 \pm 86.14 | 90.96 \pm 75.44 | 107.61 \pm 36.77 |
| SAC | 127.06 \pm 4.53 | 127.26 \pm 6.79 | 124.99 \pm 8.52 | 124.73 \pm 24.80 |

hyperparameters. The test results are provided with the 'mean \pm std' from 25,000 testing episodes in five seeds.

The benchmark results reveal a consistent attenuation in algorithm performance when introducing realistic DTR complexities, specifically PK/PD variance (\hat{p}_1), observational noise (\hat{p}_2), and missing values (\hat{p}_3). This downward trend underscores

Table 4.9: Reward performance on *OberstSepsisEnv* under realist DTR enhancements. Notations are the same as Table 4.7.

| Policy | \hat{p} | \hat{p}_1 | \hat{p}_2 | \hat{p}_3 |
|-----------|-----------------------------------|-----------------------------------|-----------------------------------|------------------------------------|
| π_b | -0.81 ± 0.58 | -0.82 ± 0.57 | -0.82 ± 0.58 | -0.82 ± 0.58 |
| DQN | 0.15 ± 0.73 | 0.12 ± 0.74 | 0.08 ± 0.71 | -0.08 ± 0.46 |
| DDQN | 0.17 ± 0.73 | 0.15 ± 0.73 | 0.09 ± 0.73 | -0.12 ± 0.57 |
| DDQN-duel | 0.16 ± 0.75 | 0.16 ± 0.72 | 0.09 ± 0.72 | -0.08 ± 0.52 |
| C51 | 0.13 ± 0.72 | 0.08 ± 0.64 | 0.05 ± 0.65 | -0.05 ± 0.27 |
| dSAC | 0.16 ± 0.71 | 0.16 ± 0.72 | 0.06 ± 0.65 | -0.06 ± 0.31 |
| rDQN | -0.11 ± 0.77 | -0.22 ± 0.62 | -0.19 ± 0.70 | -0.27 ± 0.53 |
| rDDQN | -0.17 ± 0.74 | -0.15 ± 0.77 | -0.17 ± 0.74 | -0.29 ± 0.53 |
| rC51 | -0.68 ± 0.51 | -0.35 ± 0.48 | -0.16 ± 0.37 | -0.43 ± 0.53 |
| drSAC | 0.17 ± 0.52 | -0.04 ± 0.26 | -0.08 ± 0.30 | -0.15 ± 0.36 |

Table 4.10: Reward performance on *SimGlucoseEnv* under realist DTR enhancements. Notation and highlighting follow Table 4.7.

| Policy | \hat{p} | \hat{p}_1 | \hat{p}_2 | \hat{p}_3 |
|-----------|--------------------------------------|---------------------------------------|--|--|
| π_b | -187.86 ± 0.00 | -182.86 ± 45.51 | -184.36 ± 50.76 | -184.36 ± 50.76 |
| DQN | 119.08 ± 34.50 | -144.56 ± 96.74 | -160.46 ± 102.33 | -150.80 ± 102.16 |
| DDQN | 105.29 ± 12.21 | -151.88 ± 95.48 | -156.85 ± 102.14 | -152.64 ± 108.70 |
| DDQN-duel | 123.52 ± 55.72 | -157.01 ± 95.46 | -189.87 ± 105.58 | -177.73 ± 107.16 |
| C51 | 109.81 ± 44.57 | -153.01 ± 86.86 | -177.68 ± 86.77 | -171.52 ± 97.87 |
| dSAC | -16.28 ± 107.64 | -201.11 ± 89.80 | -179.21 ± 98.02 | -168.73 ± 102.64 |
| rDQN | -7.26 ± 90.57 | -168.98 ± 91.97 | -167.74 ± 102.03 | -167.60 ± 89.42 |
| rDDQN | 114.78 ± 55.44 | -163.81 ± 89.98 | -156.11 ± 96.02 | -156.61 ± 96.38 |
| rC51 | -149.60 ± 164.17 | -195.49 ± 92.99 | -165.10 ± 98.97 | -167.35 ± 89.13 |
| drSAC | 119.10 ± 60.38 | -147.45 ± 97.24 | -151.61 ± 95.20 | -160.24 ± 96.11 |
| DDPG | -304.35 ± 58.25 | -233.84 ± 77.66 | -248.16 ± 74.88 | -248.52 ± 74.94 |
| TD3 | -244.18 ± 120.52 | -233.58 ± 77.77 | -248.63 ± 75.43 | -234.00 ± 82.20 |
| SAC | 170.60 ± 1.49 | -151.57 ± 94.41 | -155.14 ± 106.24 | -152.98 ± 105.25 |

the necessity of incorporating such factors to rigorously evaluate the robustness and adaptability of RL agents in clinical scenarios. Notably, in the *SimGlucoseEnv*, no algorithm achieved a positive reward under the PK/PD variant setting (\hat{p}_1), highlighting the substantial difficulty imposed by individualised PK/PD models across diverse patient demographics (e.g., elderly, adults, children). This variability emphasizes the acute challenge of generalisation in healthcare RL and the critical need for personalised algorithmic approaches.

Our analysis indicates a heterogeneous performance landscape where no single algorithm dominates across all environments and settings. However, discrete-action algorithms, particularly C51 and dSAC, demonstrate commendable adaptability,

achieving competitive results in the *GhaffariCancerEnv* and *SimGlucoseEnv*. Conversely, the hypothesis that incorporating medical history via Recurrent Neural Networks would universally enhance decision-making is not supported by the data. The recurrent variants (e.g., rDQN, rC51) do not consistently outperform their Multilayer Perceptron counterparts. This is most evident in the *OberstSepsisEnv*, where MLP-based baselines often surpass recurrent models (e.g., DDQN achieving 0.17 vs. rDDQN at -0.17 in \hat{p}), suggesting that the underlying causal structure of certain clinical environments may rely less on long-term temporal dependencies than anticipated.

Generalisation to PK/PD Variance ($\hat{p} \rightarrow \hat{p}_1$): Algorithms such as DDQN-duel and drSAC exhibited notable resilience to patient heterogeneity. In the *GhaffariCancerEnv*, DDQN-duel maintained strong performance when transitioning from the original (\hat{p}) to the PK/PD variant setting (\hat{p}_1). Similarly, drSAC demonstrated stability across multiple environments, suggesting potential utility in managing varied treatment effects. However, significant performance drops were observed in other baselines; for instance, standard DQN and DDQN experienced sharp declines in the *AhnChemoEnv* under \hat{p}_1 , and rC51 performance degraded severely in *SimGlucoseEnv*. These results highlight the difficulty standard RL algorithms face in generalising to unobserved biological parameters without specific architectural adaptations.

Robustness Against Noise (\hat{p}_2): The introduction of observational noise in setting \hat{p}_2 further differentiates algorithmic robustness. C51 demonstrated superior stability, particularly in the *OberstSepsisEnv* and *GhaffariCancerEnv*, where it maintained high rewards despite stochastic interference. For example, in *GhaffariCancerEnv*, C51 achieved the highest performance in \hat{p}_2 (139.89 ± 34.00). In contrast, recurrent baselines such as rDQN and rDDQN appeared more vulnerable to noise in specific contexts, showing high variance and performance deterioration in the *SimGlucoseEnv*. This suggests that while distributional RL (C51) may inherently buffer against noisy reward signals, sequential models require careful tuning to prevent overfitting to noisy input trajectories.

Handling Missing Values (\hat{p}_3): The partial observability introduced by missing values in setting \hat{p}_3 presents the most challenging evaluation criterion. Comparing the transition from the noisy setting (\hat{p}_2) to the missing-value setting (\hat{p}_3) reveals distinct adaptability profiles. While most models experienced performance degradation, C51 and dSAC exhibited relative resilience. In the *OberstSepsisEnv*, C51 remained the top-performing discrete algorithm in \hat{p}_3 , and in *GhaffariCancerEnv*, it secured the second-best result. Conversely, rDDQN and rDQN suffered more pronounced drops in environments like *OberstSepsisEnv* when data completeness was compromised. It is worth noting that some fluctuations, such as the slight reward increase for SAC in *SimGlucoseEnv* between \hat{p}_2 and \hat{p}_3 , are accompanied by high standard deviations, rendering the improvement statistically insignificant. Consequently, the selection of algorithms for clinical deployment must prioritise intrinsic mechanisms capable of handling the sparse and irregular sampling characteristic of real-world medical data.

4.4 Discussion

Our research delves into the performance of RL algorithms within various DTR settings, highlighting a notable performance decline when faced with real-world challenges such as PK/PD variance, noise, and missing data. This highlights the critical need to better incorporate these practical challenges in evaluating RL algorithms to reflect their potential effectiveness in real-world healthcare applications. Through our investigation, we discovered several intriguing findings. Contrary to expectations, RNN-based models did not consistently outperform their MLP counterparts. This outcome challenges the assumption that RNNs, with their capacity to process sequential data, would inherently provide superior performance in healthcare contexts. Furthermore, some algorithms, particularly C51 and drSAC, demonstrated remarkable resilience against volatility introduced by PK/PD variance and noise. These observations suggest that certain algorithmic features may be better suited to managing the complexities inherent in healthcare data. The study also acknowledges several limitations that must be considered when interpreting

the results. As the initial open-source platform and benchmark for RL-based DTR, *DTR-Bench* only collected four environments, potentially limiting the external validity of the diseases we have yet to cover. Furthermore, our focus was limited to off-policy RL algorithms, and we did not investigate the essential aspect of safe exploration within treatment regimes.

5

A Critical Re-examination of Reinforcement Learning based Dynamic Treatment Regime

5.1 Introduction

This chapter presents a critical perspective on the effectiveness and necessity of using reinforcement learning in dynamic treatment regimes. Initially, we highlight the various policy evaluation methods used in different studies and point out that their diversity often leads to significant variability in performance among algorithms, complicating fair comparisons. Furthermore, we compare basic baselines, including random policies and supervised learning methods, that many existing studies overlook. We find that RL algorithms can sometimes underperform these simple baselines, raising concerns about the robustness of evaluation metrics currently in use. Moving forward, we propose a standardised evaluation framework to assess the impact of different reward designs, illustrating how changes in reward structure can significantly alter observed policy performance.

5.2 Diversity and Inconsistency of Policy Evaluation

We now consider the issue of policy evaluation, which forms a central component in assessing the effectiveness of reinforcement learning methods for dynamic treatment

regimes. The setting of offline RL in healthcare presents several challenges that make policy evaluation particularly difficult: (1) The dataset is fixed and observational, which means that RL cannot be evaluated by interacting with the environment; (2) Medical decision-making is complex, as the effects of treatments may not be immediately apparent and can be influenced by many confounding factors; and (3) Patients’ responses to treatments are inherently uncertain and variable, making it difficult to assess the true effectiveness of proposed treatment policies. Compared to traditional RL settings, these challenges make it harder to evaluate offline RL algorithms in the context of DTRs.

The recent literature addresses the evaluation challenges in DTR using a range of policy evaluation techniques. Notable among these are Inverse Probability Weighting [234], Weighted Importance Sampling [235, 236], the Direct Method [237, 238], and Doubly Robust [189, 239, 240] estimators. These methods tried to tackle the confounding variables and create a counterfactual estimation based on historical data. To standardise the evaluation procedure across all algorithms, we implement several off-policy evaluation methods, including the doubly robust estimator, WIS, and several WIS variants. These include WIS with bootstrapping (WIS_b, Eq 2.21), WIS with truncation (WIS_t, Eq 2.23), and a combination of both techniques (WIS_{bt}, Eq 2.24). For the bootstrapped variant, we generate $B = 100$ bootstrap samples, each with a sample size $N_b = N$, where N is the size of the dataset. For truncation, we select a fixed threshold of $\tau = 1$ to reduce the influence of extreme importance weights. The definitions and exact formulations of these estimators are provided in Section 2.4.2.

5.3 Reward Design Choices

This section examines how different reward structures influence the performance and interpretability of reinforcement learning algorithms in dynamic treatment regimes. The reward function plays a central role in shaping agent behaviour, and its formulation must reflect both clinical relevance and computational tractability. However, many studies adopt disparate approaches, which complicates comparison

and generalisation. One of the first and most widely cited reward formulations is presented by Komorowski et al. [44], where the reward is defined in a binary fashion. Intermediate time steps receive a reward of zero, and only the terminal outcome is considered, with +100 for patient survival and −100 for death. Although this reduces the complexity of medical trajectories to a single end outcome, it overlooks important fluctuations in a patient’s physiological state and fails to capture intermediate risks such as gradual organ deterioration or transient instability. It also introduces noise into the learning signal, since many factors beyond the modelled treatments (e.g., comorbidities, infections, or surgical interventions) can influence survival. In this setting, the model considers only intravenous fluids and vasopressors, potentially misattributing the outcomes to actions that may have had limited causal influence. From an RL perspective, this form of sparse and delayed feedback can exacerbate the credit assignment problem, reduce sampling efficiency, and increase learning variance.

To address the aforementioned limitations, more recent studies have proposed using intermediate rewards, which offer denser feedback and allow the agent to adapt continuously over time. One common approach is to use established clinical risk scores as proxies for patient stability. For example, the Sequential Organ Failure Assessment (SOFA) score is often used to quantify the severity of a patient’s condition in intensive care settings. SOFA captures the function of multiple organ systems, and changes in this score over time can serve as an indicator of clinical progression. In some formulations, lactate levels are also incorporated to reflect tissue hypoxia and metabolic imbalance, both of which are important predictors of mortality [241]. Let κ_t^i denote the SOFA score at time t for patient i , and v_t^i the lactate level in mmol/L. A typical reward function that combines these features is defined as follows:

$$r_t^i = c_0 \mathbf{1}_{\kappa_t^i = \kappa_{t+1}^i} \cdot \mathbf{1}_{\kappa_{t+1}^i > 0} + c_1 (\kappa_{t+1}^i - \kappa_t^i) + c_2 \tanh(v_{t+1}^i - v_t^i) + \mathbf{1}_{t=T_i} \cdot r_{\text{outcome}} \quad (5.1)$$

where $c_0 = -0.025$, $c_1 = -0.125$, and $c_2 = -2$. The outcome reward r_{outcome} is set to 15 for survivors and -15 for non-survivors. This reward encourages the

agent to reduce organ failure scores and stabilise lactate dynamics throughout the patient trajectory.

Another approach implements the National Early Warning Score 2 (NEWS2), a tool developed for the early detection of clinical deterioration. Unlike SOFA, which focuses on multi-organ failure in critical care, NEWS2 is used more widely across healthcare settings to track vital signs and stratify risk. Recent studies have highlighted its utility in sepsis management [242, 243]. To construct a reward function based on NEWS2, the score is first normalised to the interval $[0, 1]$, representing an approximate probability of mortality. This normalisation provides a consistent signal across patients. The outcome reward is set to -1 for death and 0 otherwise. The reward is then defined as

$$r_t^i = -r_{\text{NEWS2}} + \mathbf{1}_{t=T_i} \cdot r_{\text{outcome}} \quad (5.2)$$

This formulation eliminates the need for manual weighting between multiple components and centres the agent’s focus on minimising mortality risk at every time step.

5.4 Baselines Comparisons

This section discusses the selection of appropriate baselines for evaluating reinforcement learning methods in dynamic treatment regimes. Upon reviewing the literature (see Chapter 2), we observed several inconsistencies: **a)** the use of different baseline sets across different studies, some lacking state-of-the-art offline RL algorithms; **b)** the absence of naive baselines, such as random policy, for essential sanity checks; **c)** the omission of supervised learning baselines. Consequently, this section outlines a set of baselines that we consider appropriate for comparison.

We begin with supervised learning methods, which provide a reference point for understanding whether the complexity of RL is justified by improved performance. Comparing RL with simpler SL algorithms helps us to understand whether 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 clinicians’ decisions in test data). We employ a

Long-Short-Term Memory (LSTM) [244] network to minimise cross-entropy loss, a standard loss function for classification tasks. We also include a set of naive policies, each constructed using simple heuristics without reference to the patient’s state. We include random π_r , zero-drug π_{\min} , max-drug π_{\max} , and alternating policy π_{alt} . Although often neglected in RL-DTR research, naive baseline comparison is vital for two reasons: **a)** to assess whether naive baselines inadvertently outperform clinicians, indicating potential flaws in evaluation metrics; **b)** to establish a lower performance bound, providing a worst-case scenario benchmark for algorithmic assessment. Each naive policy generates actions using fixed probability distributions. Since some actions may be assigned zero probability, we apply a small correction factor ϵ_1 to ensure non-zero support across all actions. This correction is compensated by subtracting a corresponding value ϵ_2 from the dominant actions, maintaining a valid probability distribution. We assume that the action space contains $|\mathcal{A}|$ discrete actions, with a_0 and $a_{|\mathcal{A}|-1}$ denoting the minimum and maximum dose actions, respectively. The exact definitions of the policy are summarised in Table 5.1. In addition to these baselines, we include a range of well-established offline RL algorithms. These include DQN, CQL, BCQ, and IQL. The theoretical motivations of these algorithms are discussed in Section 2.4.1. As the domain of offline RL rapidly evolves, newer algorithms [245, 246] are emerging as potentially more advanced alternatives, but are not reviewed here.

5.5 Data and Experiment Setup

We model the dynamic treatment regime problem as a finite-horizon Markov decision process, as described in Section 2.2. To examine the impact of policy evaluation and reward design, we conduct experiments using the intravenous fluid and vasopressor dosage task for sepsis treatment in the intensive care unit, originally introduced by Komorowski et al. [44]. The dataset is derived from the MIMIC-III database [186], with details provided in Section 3.1. This section presents empirical results on this sepsis treatment task, highlighting how evaluation algorithms, reward definitions, and baselines influence the assessment of reinforcement learning methods. The

Table 5.1: Naive Baselines Equations for OPE Sanity Check

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

Note: p_j denotes the probability of empirical action estimated from the training data by counting the occurrence of the action.

dataset is divided into training, validation, and test subsets in a 70%–15%–15% split. In addition to the main results presented below, we provide full experimental results in Appendix B.2, including performance tables for both outcome-based and SOFA-based rewards (Tables B.1 and B.2), along with results from 35 additional subset experiments.

5.5.1 Implementation

For supervised learning models, we implement a supervised learning model using an LSTM network to estimate the clinician’s policy. This behavioural cloning model is trained in the training set, and model selection is performed using the validation set. The objective is to minimise the cross-entropy loss between the predicted and observed treatment actions over time, formulated as:

$$L_{\text{CE}}(\theta) = -\mathbb{E}_{(s,a) \sim \mathcal{D}} \log \hat{\pi}(a|s; \theta) \quad (5.3)$$

Here, θ denotes the model parameters, and a is the action taken by the clinician at each decision point. The model with the highest patient-level F1 score in the validation set is selected for evaluation. We prioritise the F1 score over the Area Under the ROC Curve, as the latter may be misleading under significant class imbalance. For the RL models, we employ a simplified architecture with a single

linear output layer. This choice is intended to reduce the influence of model-specific deep learning tricks and isolate the effect of algorithmic differences. The input consists of a flattened three-frame temporal window, representing approximately 12 hours of patient history, allowing the model to incorporate short-term clinical trajectories. A grid search is conducted over a standard set of hyperparameters shared across all RL algorithms. The key hyperparameters and their respective values are listed in Table 5.2. Where applicable, algorithm-specific parameters are also included. The seeds were generated using a standard random number generation process in Python.

| 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 5.2: 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.

5.5.2 Overall Comparison Results

We compare naive baselines, supervised learning, and RL (10 policies in total) on nine metrics with 13 patient groups and three 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 B.2.

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 an algorithm outperforms all others. A win is counted when the algorithm is the best on a metric for a particular group of patients 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 summarise the key observations below.

Effectiveness of Naive Baselines: Surprisingly, naive baselines have wins over RL, SL, and even the doctor returns in the overall test set, as shown in Figure 5.1a and Figure 5.1b. Intuitively, naive baselines should not even win once. Table 5.3 and 5.4 reveal 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 B.1, where RL algorithms generally surpass naive baselines. These findings again support our recommendation for including naive baselines as an easy check on the reliability of OPE methods.

| metric | alt | max | min | random | weight | SL |
|----------------------|--------|--------------|---------------|--------------|---------------|----------------------|
| RMSE _{IV} | 763.89 | 861.51 | 645.83 | 671.39 | 645.83 | 557.81 ± 9.27 |
| RMSE _{vaso} | 0.67 | 0.89 | 0.32 | 0.5 | 0.59 | 0.31 |
| WIS | -4.58 | -4.62 | -4.58 | -3.84 | -3.78 | -4.22 ± 0.41 |
| WIS _b | -5.43 | -4.81 | -5.76 | -4.4 | -4.73 | -4.62 ± 0.17 |
| WIS _t | -4.58 | -4.62 | -4.58 | -3.97 | -3.78 | -4.57 ± 0.62 |
| WIS _{bt} | -5.64 | -4.69 | -5.61 | -4.5 | -4.5 | -4.68 ± 0.3 |
| DR | -0.54 | -0.19 | -1.55 | -0.35 | -0.3 | -0.36 |
| P.F1 | 0.2 | 0.02 | 0.2 | 0.2 | 0.0 | 0.31 ± 0.01 |
| S.F1 | 0.19 | 0.02 | 0.19 | 0.19 | 0.0 | 0.3 ± 0.01 |
| $G_{\mathcal{D}}$ | | | | -4.39 | | |

Table 5.3: Rule-based, random, and supervised learning policies on full test set under NEWS2 reward. Best and second-best values are highlighted in red and blue, respectively. Lower is better for RMSE; higher is better for others. This notation will be consistently applied across this chapter.

RL Performances Across Rewards: We observed significant differences in how the RL algorithms perform with different rewards. For example, DQN

| metric | DQN | CQL | IQL | BCQ |
|----------------------|---------------|---------------|----------------|--------------|
| RMSE _{IV} | 638.51 ± 8.63 | 541.67 ± 5.74 | 578.96 ± 10.06 | 626.2 ± 9.56 |
| RMSE _{vaso} | 0.44 ± 0.07 | 0.3 ± 0.01 | 0.31 ± 0.01 | 0.31 |
| WIS | -3.79 ± 0.01 | -4.1 ± 1.43 | -5.83 | -4.58 |
| WIS _b | -3.88 ± 0.73 | -4.48 ± 0.77 | -5.31 ± 0.06 | -5.41 ± 0.17 |
| WIS _t | -3.84 ± 0.11 | -4.1 ± 1.43 | -5.83 | -4.58 |
| WIS _{bt} | -3.87 ± 0.67 | -4.38 ± 0.98 | -5.27 ± 0.05 | -5.55 ± 0.19 |
| DR | -0.14 ± 0.04 | -0.71 ± 0.05 | -0.51 ± 0.04 | -1.54 ± 0.01 |
| P.F1 | 0.06 ± 0.02 | 0.33 ± 0.01 | 0.34 ± 0.01 | 0.23 ± 0.01 |
| S.F1 | 0.06 ± 0.02 | 0.32 ± 0.01 | 0.33 ± 0.01 | 0.22 ± 0.01 |
| $G_{\mathcal{D}}$ | -4.39 | | | |

Table 5.4: RL policies on full test set under NEWS2 reward.

won the most in the NEWS2 reward, while CQL won the most on the outcome reward (see Figure 5.1b). However, in 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 5.4’s DQN column), while we also observe the opposite cases (see Table 5.4’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.

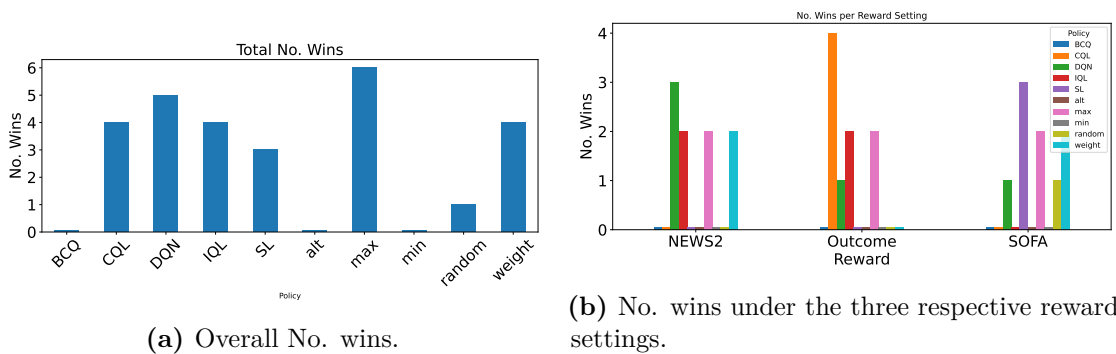
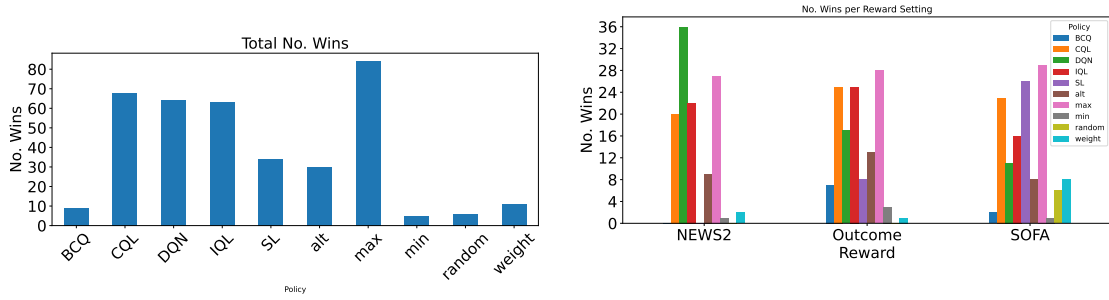


Figure 5.1: Number of wins for each policy in the (overall) test set. Wins are calculated based on the mean performance of 5 random seeds. Alt, min, max, random, and weight policies are naive baselines. This denotation applies to all the following figures.



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

Figure 5.2: 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.

Comparison of DR and IS: Our analysis indicated that DR either overestimates or underestimates IS (see Table 5.4, and Table B.1, B.2 in the appendix, and compare the DR row with other importance sampling rows.), subject to the reward setting. Instead of being doubly robust, our experiments suggest that estimators in DR tend to dual ‘unrobust’ due to behavioural and value approximation error (more details are provided in Section 5.6.) 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 subgroups of patients. A notable illustration can be seen in Table 5.6 and 5.5, 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 individualised treatment solutions.

| metric | alt | max | min | random | weight | SL |
|----------------------|--------|-------|--------|--------|--------|----------------|
| RMSE _{IV} | 788.91 | 880.8 | 756.41 | 774.78 | 749.58 | 637.22 ± 10.04 |
| RMSE _{vaso} | 0.54 | 0.85 | 0.27 | 0.47 | 0.56 | 0.25 |
| WIS | -3.23 | -3.51 | -3.44 | -3.14 | -3.38 | -3.01 ± 0.07 |
| WIS _b | -3.33 | -3.49 | -3.41 | -3.23 | -3.4 | -2.89 ± 0.06 |
| WIS _t | -3.23 | -3.51 | -3.44 | -3.14 | -3.38 | -3.01 ± 0.07 |
| WIS _{bt} | -3.29 | -3.47 | -3.42 | -3.21 | -3.41 | -2.93 ± 0.05 |
| DR | -0.26 | -0.05 | -1.46 | -0.39 | -0.46 | -0.41 ± 0.02 |
| P.F1 | 0.25 | 0.02 | 0.25 | 0.25 | 0.0 | 0.31 ± 0.01 |
| S.F1 | 0.24 | 0.02 | 0.24 | 0.24 | 0.0 | 0.29 ± 0.01 |
| $G_{\mathcal{D}}$ | | | | -3.41 | | |

Table 5.5: Rule-based, random, and SL policy performance on NEWS2 subgroup (rate $[-\infty, -0.4]$ high variance).

| metric | DQN | CQL | IQL | BCQ |
|----------------------|----------------|----------------|--------------|----------------|
| RMSE _{IV} | 734.32 ± 11.14 | 609.62 ± 11.74 | 645.67 ± 6.1 | 719.77 ± 12.57 |
| RMSE _{vaso} | 0.4 ± 0.07 | 0.26 ± 0.01 | 0.25 ± 0.01 | 0.26 |
| WIS | -2.85 ± 0.72 | -1.9 ± 0.36 | -2.98 ± 0.13 | -3.44 |
| WIS _b | -2.85 ± 0.5 | -2.09 ± 0.19 | -2.84 ± 0.02 | -3.34 ± 0.02 |
| WIS _t | -2.85 ± 0.72 | -1.9 ± 0.36 | -2.98 ± 0.13 | -3.44 |
| WIS _{bt} | -2.81 ± 0.49 | -2.1 ± 0.24 | -2.86 ± 0.05 | -3.34 ± 0.05 |
| DR | -0.13 ± 0.09 | -0.48 ± 0.13 | -0.43 ± 0.02 | -1.27 ± 0.03 |
| P.F1 | 0.07 ± 0.02 | 0.33 ± 0.02 | 0.34 ± 0.02 | 0.27 ± 0.01 |
| S.F1 | 0.07 ± 0.02 | 0.31 ± 0.02 | 0.33 ± 0.02 | 0.25 |
| $G_{\mathcal{D}}$ | | | -3.41 | |

Table 5.6: RL policies on NEWS2 subgroup (rate $[-\infty, -0.4]$ high variance).

5.6 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 evaluation of the policy can be unreliable when the behavioural policy or the target value function does not approximate well with the real one. We illustrate this point of view by visualising the inference probability/value of the behavioural ($\hat{\pi}_{\mathcal{D}}$)/value (\hat{Q}^{π}) estimators against their sample losses.

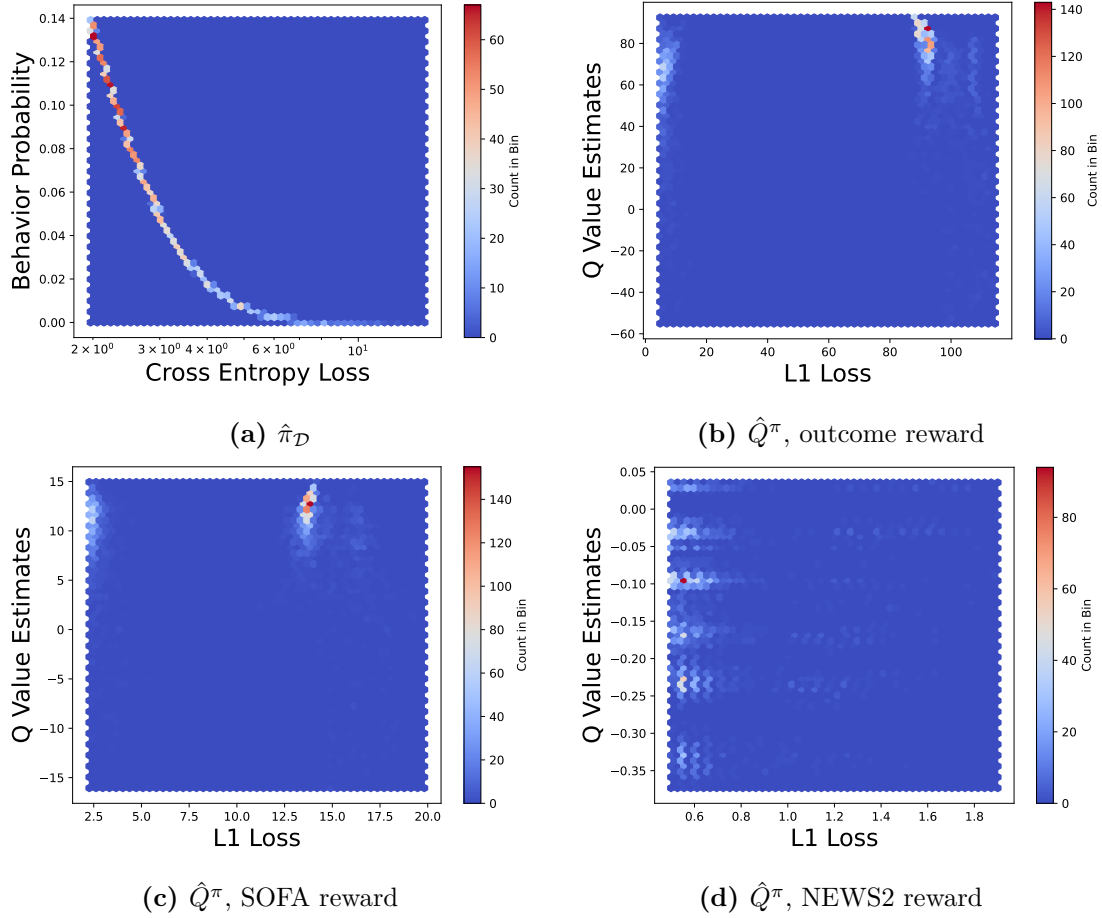


Figure 5.3: behavioural and value estimators versus their losses on the testing set. The count in each bin is indicated by a colour bar, transitioning from blue to red as the number increases. **(a)** depicts the behavioural loss (samples with a cross-entropy loss > 90 th percentile) versus the inference probability. **(b)**, **(c)**, and **(d)** show the direct method estimator loss (samples with L1 loss > 90 th percentile) on the outcome, SOFA, and NEWS2 reward, respectively.

5.6.1 Variance from Behavioural Policy

A common assumption in the OPE literature is that a supervised learning behavioural policy can closely approximate the clinician’s policy, i.e., $\hat{\pi}_{\mathcal{D}} \simeq \pi_{\mathcal{D}}$. This assumption was challenged within the DTR setting by analysing testing samples with substantial loss, as illustrated in Figure 5.3. The variance in $\hat{\pi}_{\mathcal{D}}$ may be interpreted through two critical observations. First, trajectories with small behaviour probabilities tend to dominate importance sampling. Importance sampling assigns greater weight to trajectories with large importance ratios, which typically arise when the behavioural policy assigns very low probability to the

observed actions. As a result, it becomes particularly important for $\hat{\pi}_{\mathcal{D}}$ to accurately model these small probabilities, since errors in this region can disproportionately distort OPE estimates. Second, there appears to be a negative correlation between inference probabilities and prediction errors. Specifically, actions associated with smaller probabilities, those that contribute the most to importance sampling, often correspond to a higher loss, indicating a larger divergence between the predicted and true policies (see Figure 5.3a).

5.6.2 Variance from Value Function

The quality of the estimated value function \hat{Q}^{π} is highly sensitive to the reward specification. This dependency gives rise to inconsistent results in DR-based policy evaluation, as discussed in Section 5.5.2. Figures 5.3b, 5.3c, and 5.3d illustrate this effect across three distinct reward functions. In particular, large estimation errors tend to cluster around high Q values, particularly for the outcome reward ($r_{\text{alive/death}} = \pm 100$) and the SOFA reward ($r_{\text{alive/death}} = \pm 15$), as shown in Figures 5.3(b) and (c), respectively. This suggests that the value estimator may not capture the terminal rewards that fully reflect patient survival. In contrast, the NEWS2 reward does not exhibit this concentration of errors, likely due to its smoother formulation, which ranges from -1 (death) to 0 (otherwise). These observations further support our position that RL-DTR methods warrant careful re-examination. In particular, the findings underscore that reward design is crucial not only for maintaining clinical relevance but also for enabling stable and accurate learning of value functions used in off-policy evaluation.

5.6.3 A Closer Look at the Behavioural Model and Value Estimation Error

To better understand the sources of estimation error in both the behavioural policy and the value function, we conduct a more granular analysis of their respective loss patterns across the training and validation sets. Figures B.1 and B.2 in Appendix B illustrate the correlation between the behavioural cloning loss and the inferred action

probability for $\hat{\pi}_{\mathcal{D}}$, and between the value approximation loss and the estimated state-action value from \hat{Q}^{π} . These patterns complement the findings reported in Section 5.6, and suggest that the observed relationships are not confined to the test set alone, but are also evident in the training and validation sets, and remain consistent across different reward formulations. It is important to note that the observed errors in $\hat{\pi}_{\mathcal{D}}$ and \hat{Q}^{π} are unlikely to be the result of poor model fitting. The behavioural model $\hat{\pi}_{\mathcal{D}}$ achieves a robust F1 score of 0.7 on both the OPE training set and the validation set.¹ Similarly, the selected value function \hat{Q}^{π} is chosen based on its ability to minimise temporal-difference error. Further insight is provided in Figure 5.4, which visualises the loss distributions for both models. The behavioural policy demonstrates generally low loss across most samples, although a minority exhibit substantially higher values. In contrast, the loss distributions for the value functions differ markedly depending on the reward type. For both the Outcome and SOFA rewards, the distributions form prominent Gaussian-shaped peaks centred near 100 and 15, respectively. These values correspond to their respective terminal rewards (± 100 for Outcome; ± 15 for SOFA), suggesting that large prediction errors are concentrated around end-of-trajectory events. This may indicate a mismatch between intermediate value estimates and abrupt terminal feedback. In comparison, the NEWS2 reward exhibits a smoother distribution, likely attributable to its gradual reward structure (ranging from 0 to -1), which appears to mitigate such sharp discontinuities.

5.6.4 Investigating Model Calibration for Behavioural Model

Model calibration refines a model’s predictive probabilities to match the actual likelihoods of outcomes, thereby ensuring that the model’s confidence accurately reflects its empirical accuracy. We investigated the influence of model calibration on behaviour policy and performance change in OPE. The temperature scale [247] was used as the model calibrator. Temperature scaling is a post-processing

¹The term “OPE training set” here refers to the union of the original training and test sets. The validation set retains its conventional definition. This adjustment in terminology aligns with standard RL practice, which decouples evaluation from initial data splits.

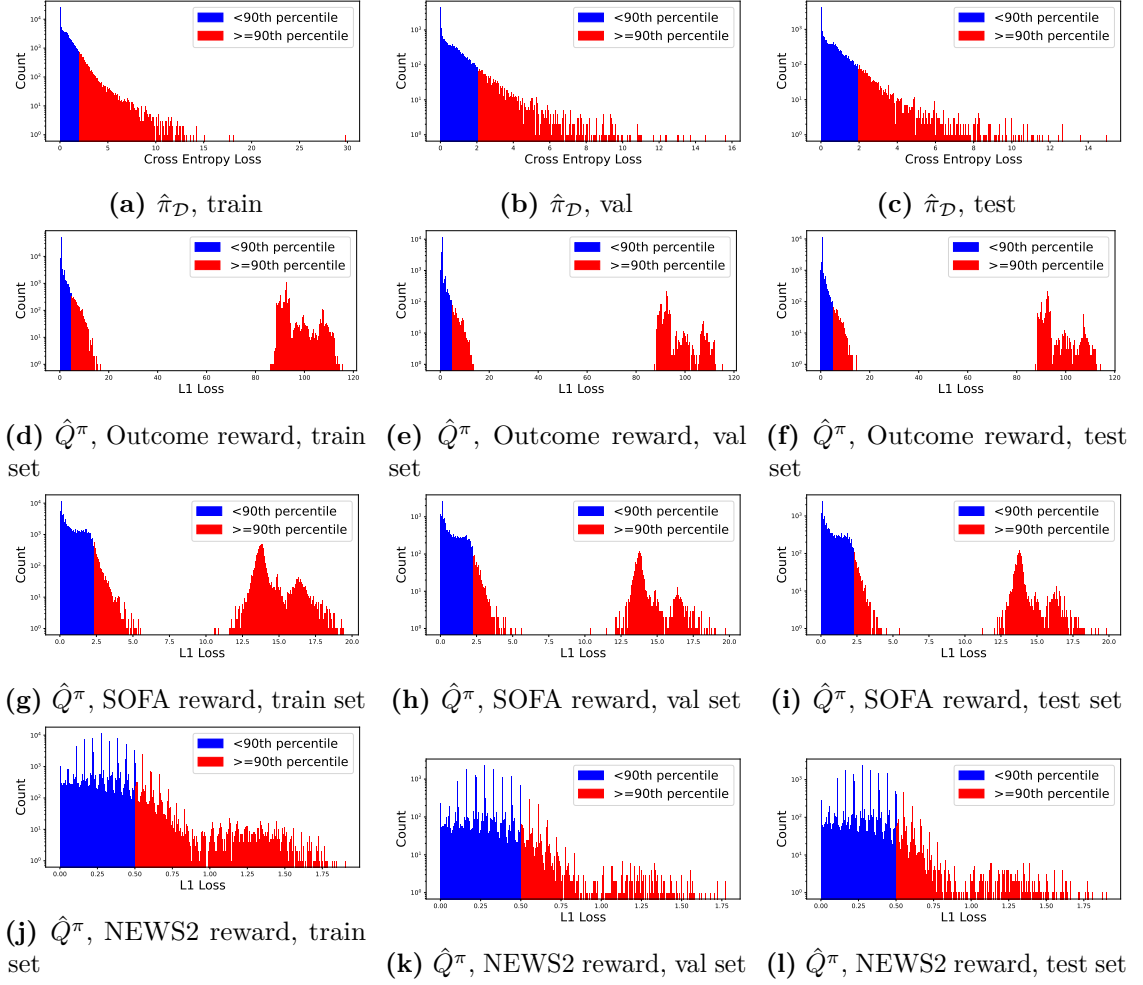


Figure 5.4: Loss histograms on different reward settings. a), b), c) $\hat{\pi}_{\mathcal{D}}$ cross-entropy loss histogram on training, validation and test set. d), e), f) \hat{Q}^{π} L1 loss histogram on Outcome reward, g), h), i), \hat{Q}^{π} L1 loss histogram on SOFA reward, j), k), l) \hat{Q}^{π} L1 loss histogram on NEWS2 reward.

technique that adjusts the model’s predicted probabilities to match the observed probabilities better, thereby (potentially) improving the model’s calibration. We present the implementation details, OPE results, and a visual comparison between the calibrated and uncalibrated models. 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:

$$P_T(y|x) = \text{softmax} \left(\frac{f(x)}{T} \right) \quad (5.4)$$

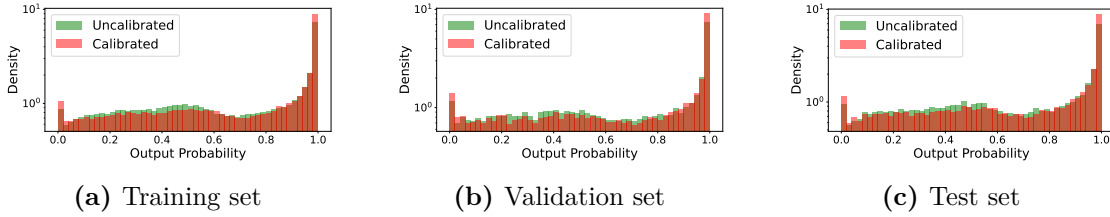


Figure 5.5: Comparison of output probability between calibrated and uncalibrated $\hat{\pi}_{\mathcal{D}}$ on the training, validation and test set. The plot shows histograms of output probability and the number of counts in the dataset with a logarithm scale on the y-axis for training, validation and test set, respectively. It is observed that the frequencies of extreme probabilities (i.e., probabilities near 0 and 1) are higher after calibration.

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. The optimal value of T is typically determined through a calibration process in the validation dataset, to minimise a calibration-specific loss function. Here, we choose the Negative Log Likelihood. We search hyperparameters to determine the best-calibrated model using a learning rate grid $[0.1, 0.05, 0.02, 0.01, 0.001, 0.0005, 0.0001, 0.00005, 0.00001]$ and a batch size of 512.

Figure 5.5 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 5.6.1, importance sampling gives small probabilities more weight, and small probabilities can lead to extremely high losses, indicating that calibration may increase the variance of OPE in this case. To further illustrate this point, we plotted the importance ratio (i.e., $\rho_{1:T_i}^i = \prod_{t=1}^{T_i} \frac{\pi(a_t^i | s_t^i)}{\pi_{\mathcal{D}}(a_t^i | s_t^i)}$) of the random policy π_r in Figure 5.6. The large ratios became even larger after calibration, which implies that model calibration may increase the variance of the OPE and should be applied with caution. An additional 14 figures are reported for the comparison of the baseline ratio in Appendix B.3.

To further investigate the impact of model calibration on OPE, we ran OPE for all naive baselines using the calibrated behaviour policy. The results are presented

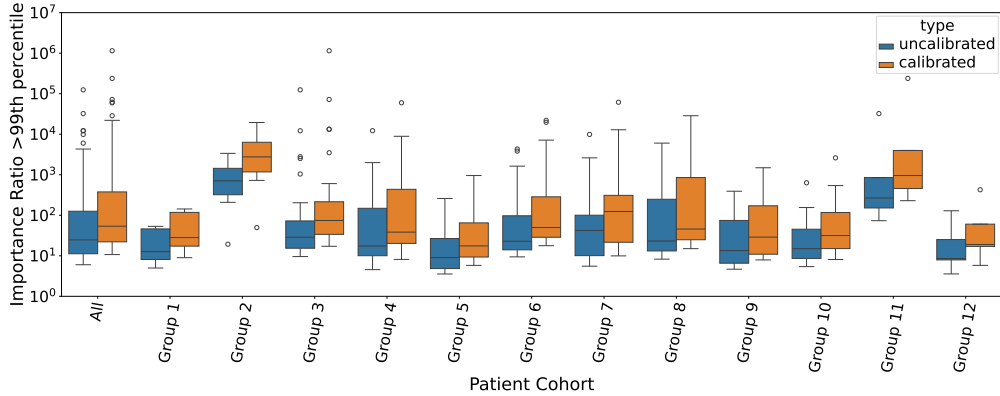


Figure 5.6: Importance ratio histogram of random policy > 99th percentile. The horizontal axis includes different datasets, where 'All' means the test set and the rest are NEWS2 risk-stratified subsets, indexed by the ascending order of NEWS2 change rate. The calibrated model contains more extremely large ratios > 99th percentile. Only ratio outliers (i.e., > 99th percentile) are plotted for visualisation convenience. To view the other 14 ratio plots for 5 baseline policies in 3 different reward settings, please see Appendix B.3.

| Policy | Outcome Reward | | | | | SOFA Reward | | | | | NEWS2 Reward | | | | |
|-------------------|----------------|------------------|------------------|-------------------|--------------|--------------|------------------|------------------|-------------------|--------------|--------------|------------------|------------------|-------------------|--------------|
| | WIS | WIS _b | WIS _t | WIS _{bt} | DR | WIS | WIS _b | WIS _t | WIS _{bt} | DR | WIS | WIS _b | WIS _t | WIS _{bt} | DR |
| alt | 84.29 | 79.54 | 84.29 | 72.97 | -0.68 | 13.34 | 12.28 | 13.34 | 11.42 | -0.54 | -4.57 | -5.63 | -4.57 | -5.43 | -0.56 |
| max | -90.47 | -74.67 | -90.47 | -77.09 | -0.38 | -9.44 | -10.03 | -9.44 | -9.53 | -0.30 | -4.60 | -4.85 | -4.60 | -4.63 | -0.21 |
| min | 84.29 | 84.29 | 84.29 | 84.24 | -0.91 | 13.34 | 12.86 | 13.34 | 12.93 | -0.73 | -4.57 | -5.59 | -4.57 | -5.56 | -1.49 |
| random | 86.00 | 85.48 | 84.30 | 82.87 | -0.58 | 14.33 | 13.35 | 10.26 | 11.86 | -0.41 | -4.47 | -4.77 | -5.82 | -5.09 | -0.39 |
| weight | 86.87 | 63.98 | 86.87 | 57.87 | -0.50 | 16.43 | 11.24 | 16.43 | 9.96 | -0.38 | -3.78 | -4.53 | -3.78 | -4.61 | -0.35 |
| $G_{\mathcal{D}}$ | 61.54 | | | | | 9.47 | | | | | -4.39 | | | | |

Table 5.7: Comparison of evaluation metrics across policies under three reward definitions. Bold indicates the best value within each metric column.

in Tables 5.7. 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 of Table 5.7 show that some naive baselines can still surpass clinical experts, regardless of the reward design.

5.7 Discussion on Considering RL for DTR

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

SL as a baseline: Including SL as a baseline is crucial to evaluate whether RL

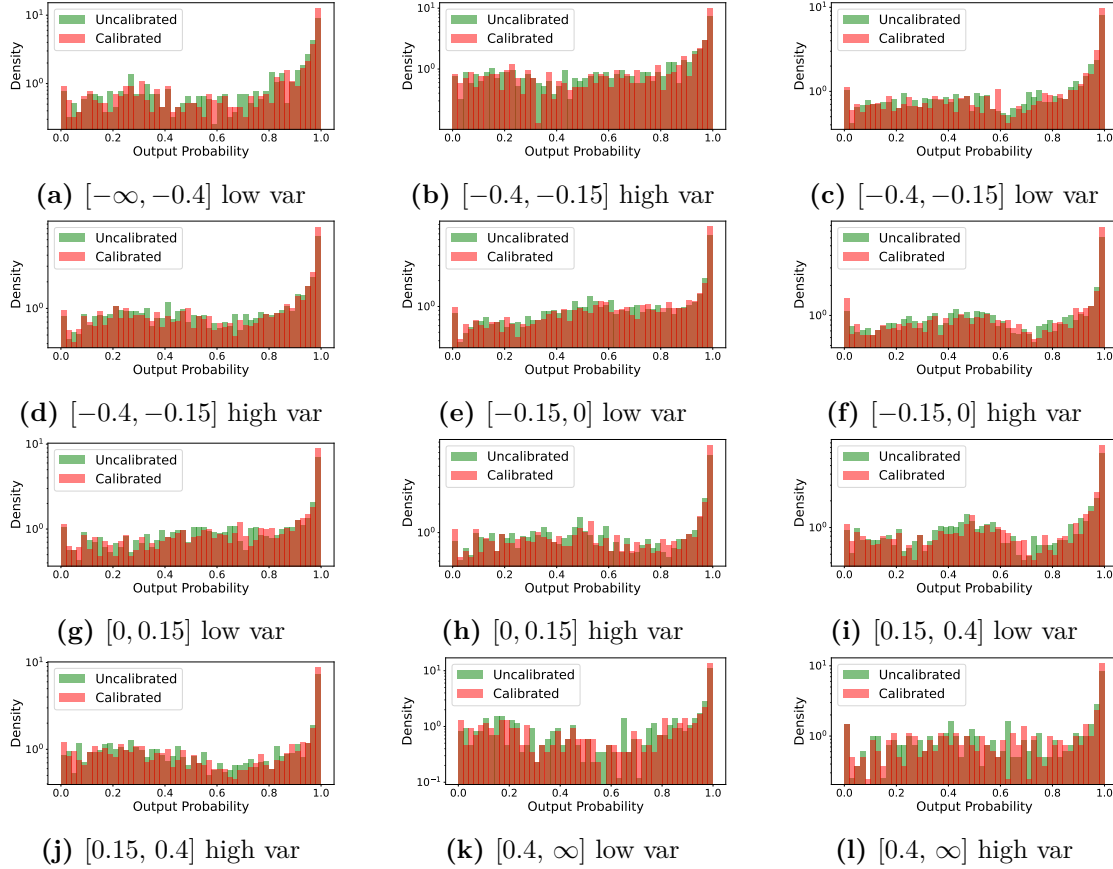


Figure 5.7: Comparison of output probability between calibrated and uncalibrated $\hat{\pi}_{\mathcal{D}}$. The plot displays histograms of output probability and the number of counts in the dataset, with a logarithmic scale on the y-axis for the 12 stratified patient groups. Here, 'low var' and 'high var' refer to low and high variations in NEWS2 scores, respectively.

can outperform a supervised approach using both OPE and supervised learning metrics. Although comparing RL to SL is standard practice in the offline RL community, it may not be common knowledge to a broader audience. Therefore, we emphasise 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 in different subgroups of patients and helps identify

learning bias in treatment, ensuring equitable healthcare outcomes among patient populations. A comparative analysis of Figure 5.1 and Figure 5.2 supports this insight: On the 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 sub-optimal offline data, and its effectiveness may be particularly pronounced in specific patient groups. Stratifying data by groups of patients can be a valuable strategy to quickly pinpoint where RL provides the most benefit. This targeted approach not only facilitates the identification of these advantages, but also allows detailed examination and validation by medical experts. Such a nuanced analysis could lead to more personalised 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 to quantify the dispersion or variance of OPE [248–250] and to avoid overfitting the importance-weighted return as an RL agent [251]. Although these methods provide valuable insights into RL treatment decisions and are encouraged to be used in future RL-DTR works, they do 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, which 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 [252], emphatic TD [253], gradient TD [254], and accelerated TD [255, 256]. These methods present opportunities for further investigation and development.

Behavioural model selection and calibration: Previous work by Raghu et al. [207] suggests that judicious selection and calibration of the behavioural model can help mitigate the variance in OPE. However, our experiments indicated that the effectiveness of the 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.

5.8 Conclusion

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 show that the comparative performance of RL algorithms can vary significantly depending on these three factors through extensive empirical analysis using the medical sepsis dataset. In addition, 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 standardised approach to fully realise the benefits of RL in DTRs. Through this critical analysis, we offer several concrete considerations to enhance the scientific rigour and practical utility of RL in healthcare settings. These include the need to use supervised learning as a baseline, the inclusion of naive policies for sanity checks, the use of stratified evaluations to promote equitable treatment outcomes, and the caution required when calibrating behavioural models. Finally, we call for the development of offline policy evaluation techniques specifically adapted to the context of DTR, with better robustness to small-probability artefacts and improved interpretability for clinical stakeholders.

6

Towards Safer Insulin Dosing in the ICU: Hypoglycaemia Prediction and Offline Short-Acting Insulin Optimisation

6.1 Introduction

Maintaining blood glucose levels within a clinically safe range, typically between 70 and 180 mg/dL, remains a persistent challenge in intensive care units. Dysglycaemia, which encompasses both hyperglycaemia and hypoglycaemia, continues to complicate ICU care, with approximately 40% of critically ill patients experiencing significant hyperglycaemia and roughly 12% experiencing severe hypoglycaemia under insulin treatment [257, 258]. Both extremes are strongly associated with adverse clinical outcomes, including increased mortality, infection rates, renal impairment, and prolonged mechanical ventilation [259]. Existing clinical insulin dosing protocols, such as sliding-scale approaches inspired by the Yale protocol [260], offer simplicity and rapid applicability at the bedside. However, they often fail to respond to the rapidly evolving physiological dynamics common in critically ill patients. Metabolic fluctuations driven by nutritional variability, vasopressor infusions, corticosteroids, and organ dysfunction further complicate the efficacy of static insulin protocols, highlighting the need for dynamic and personalised insulin dosing strategies.

Recent advances in electronic health records and machine learning have created opportunities to develop personalised glycaemic control systems using routinely available clinical data [261, 262]. However, existing predictive and control-focused

studies often have limited generalisability due to their reliance on continuous glucose monitoring devices, limited patient cohorts[263], or narrowly defined input covariates. To date, no publicly available ICU dataset integrates detailed glucose trajectories with concurrent nutrition data, pharmacological interventions, and vital signs at a high temporal resolution, thus limiting the practical application of ML-based glucose management approaches in general ICU practice.

To address the aforementioned limitations, we curate and release an extensive cohort of 8,472 admissions to the adult ICU of the MIMIC-III database. This dataset uniquely integrates glucose measurements at the bedside, minute-resolved short-acting, intermediate, and long-acting insulin infusions, enteral and parenteral nutritional inputs, prescriptions and input in the ICU, and high-frequency vital signs and laboratory results. To our knowledge, this represents the first openly available ICU glycaemic control dataset that fully integrates nutritional, medication, and physiological data at such a granular level. Using this rich clinical dataset, we define two complementary glucose management tasks. The first focuses on hypoglycaemia prediction, and the second addresses short-acting insulin dosing recommendations in the presence of other concurrent insulin therapies. For the prediction task, we benchmark nine machine learning models, including logistic regression, ensemble methods, recurrent neural networks, and transformer architectures. Moving from risk prediction to treatment planning, we frame the second task as insulin titration via offline RL. We have evaluated six RL algorithms, DQN, BCQ, CQL, IQL, QRDQN, and dSAC, using a reward function derived from the Blood Glucose Risk Index. To enable conservative and robust evaluation, we proposed an ensemble bootstrapped weighted importance sampling framework that better captures estimator variability and guards against overestimation. Preliminary results suggest that conservative RL agents, particularly CQL and IQL, can approximate the performance of clinician policies while also showing safer dosing profiles under high-confidence off-policy estimates.

6.2 Methods

6.2.1 Data Curation and Preprocessing

A retrospective cohort was constructed from the MIMIC-III database, comprising 8,472 ICU stays derived from 7,086 adult patients. The inclusion criteria require patients to be 18 years or older, to stay in the ICU for a minimum of 24 hours, to have an average of at least four glucose measurements per day, and to have documented short-acting insulin administration. Records indicating pregnancy, malignancy, implausible anthropometric values (height <1 m or >3 m; weight <30 kg or >300 kg), or missing essential demographic data were excluded. After integrating physiological, laboratory, nutritional, and treatment variables into a uniform temporal resolution, each patient record was represented by a 12-hour sliding window, encapsulating relevant clinical features and associated outcome or treatment labels.

Clinical data were integrated from multiple structured tables, capturing a wide range of physiological, biochemical, and therapeutic variables. In total, 49 covariates were selected, including vital signs, laboratory test results, nutritional input, vasoactive medications, and insulin doses administered through various routes. Where applicable, continuous infusions were time-weighted over their duration, while discrete interventions and categorical indicators were forward-filled within the observation window. Residual missing values were imputed by the last observation carried forward or by median imputation using training set statistics when preceding data were unavailable. The dataset was randomly partitioned into training (70%), validation (15%), and test (15%) sets, stratified by age, sex, obesity, and diagnosis of diabetes, to preserve the heterogeneity of the population across segments. Full curation details can be found in Section 3.2, Chapter 3.

6.2.2 Clinical Prediction Tasks on Hypoglycaemia and Insulin Dose

We investigated the two glucose management tasks using a shared temporal input representation: (1) binary classification of future hypoglycaemia events, defined as any glucose measurement falling below 70 mg/dL within the next six hours;

and **(2)** Modelling of short-acting insulin infusion decisions using both supervised and reinforcement learning approaches. In the following subsections, we provide a detailed description of each task.

Hypoglycaemia Event Prediction For hypoglycaemia prediction, binary labels were assigned based on the forward inspection of glucose trajectories. The prediction timestamp was aligned with the current hour, using a 12-hour look-back window as input. All models were trained using binary cross-entropy loss. We benchmark nine predictive models: logistic regression (LR), random forest (RF), gradient-boosted trees (XGB), multilayer perceptrons (MLP), one-dimensional convolutional networks (CNN1D), gated recurrent units (GRU), long short-term memory networks (LSTM), Transformer architectures [264], and RETAIN [265]. To ensure a fair comparison, all models receive identically preprocessed inputs. Non-sequential models operate on flattened vector representations, while temporal models utilise the full input sequence. More details on architectures and training are provided in Supplementary C.2.

Reinforcement Learning for Short-Acting Insulin Infusion We define insulin titration as a Markov decision process (MDP), with the intention of learning optimal short-acting insulin doses based on retrospective patient data. At each decision point, agents select one of five discrete insulin dosage bands derived from weight-normalised infusion rates: 0, (0, 0.03), [0.03, 0.07), [0.07, 0.1), and ≥ 0.1 U/kg/h. The state space is defined by the same temporal input used in the hypoglycaemia prediction task. The reward function is derived from the Blood Glucose Risk Index (BGRI) proposed by Kovatchev et al. [266], which penalises deviations from normoglycaemia in a clinically meaningful way. Given a glucose reading g_t (in mg/dL), the BGRI is calculated as

$$\text{BGRI}_t = 10 \left[1.509 (\log g_t)^{1.084} - 5.381 \right]^2. \quad (6.1)$$

To better align with clinical safety priorities, a shifted and clipped transformation is applied to yield the learning reward as

$$r_t = 0.01 \cdot (-\min(400, \text{BGRI}_t) + 77). \quad (6.2)$$

This formulation ensured that glucose values within the target range of 70–180 mg/dL yielded positive rewards, and hypoglycemic episodes are heavily penalised. Six offline DTR algorithms are considered: Behavioural Cloning (i.e., supervised learning mimicking doctors’ policy), DQN[267], CQL[34], BCQ[268] and the two algorithms, i.e., dSAC and C51, which achieved the best performance on the simulated insulin administration task in Chapter 4. In this paper, we adopted an improved version of C51, distributional reinforcement learning with quantile regression (QRDQN) [269]. The detailed description of each algorithm can be found in Section 2.4.1 of Chapter 2. Hyperparameters and optimisation details for each RL algorithm are included in the Supplementary C.2.

6.2.3 Behavioural Model Calibration

As discussed in Chapter 5, proper model calibration is critical for reliable off-policy evaluation (OPE), particularly when policy selection relies on importance-weighted estimates sensitive to predicted action probabilities. Miscalibrated behaviour policies, especially those that are overconfident in high-frequency action classes or underconfident in rarely chosen actions, can lead to biased and high-variance OPE estimates. To address these issues, we extend the temperature scaling technique introduced in Section 5.6.4 of Chapter 5 to a class-wise variant, enabling class-specific adjustment of the softmax confidence scores produced by behavioural cloning (BC) policies. Unlike global temperature scaling, which applies a single scalar temperature $T > 0$ uniformly to all logits, class-wise temperature scaling (CTS) introduces a vector of independent temperature parameters $\mathbf{T} = [T_1, \dots, T_{|\mathcal{A}|}]$, one for each of the $|\mathcal{A}|$ discrete insulin dose classes. This allows the model to adjust confidence levels for each action independently, thus improving calibration in long-tailed clinical data, where some classes are underrepresented.

Formally, given a BC model that outputs uncalibrated logits $\mathbf{z} = f_{\boldsymbol{\theta}}(\mathbf{o}) \in \mathbb{R}^{|\mathcal{A}|}$, the CTS-calibrated action probabilities are given by:

$$\hat{p}_k(\mathbf{x}; \mathbf{T}) = \frac{\exp(z_k/T_k)}{\sum_{j=1}^{|\mathcal{A}|} \exp(z_j/T_j)}, \quad \text{for each } k \in \{1, \dots, |\mathcal{A}|\}. \quad (6.3)$$

The temperatures $T_{|A|}$ were optimised on a held-out validation set to minimise the negative log-likelihood of the true labels under the calibrated distribution. This class-wise formulation provides a simple yet effective means of correcting class-specific overconfidence. The calibration result is discussed in Section 6.4.2.

6.2.4 Policy Extraction from Value-Based RL Algorithms

Unlike actor-critic or policy-based reinforcement learning methods that output an explicit action distribution, value-based algorithms (such as DQN, QRDQN, and CQL) learn a state-action value function or state value function, but do not directly produce a stochastic policy. To evaluate and deploy these models in clinical decision-making contexts, we must extract an actionable policy from the learnt value estimates. IQL incorporates a built-in mechanism for this, known as advantage-weighted regression [270], which produces a soft policy by reweighting actions according to their estimated advantage. Similarly, Soft Actor-Critic (SAC) maintains an explicit actor network that outputs a parametrised policy distribution. However, DQN, QRDQN, and CQL do not natively support probabilistic policies. To ensure consistency across all value-based methods, we implemented a standardised softmax-based policy extraction mechanism. Given a learnt Q function $\hat{Q}(s, a)$, we extract a stochastic policy $\pi(a|s)$ using the following temperature-scaled softmax transformation:

$$\pi(a|s) = \text{softmax} \left(\frac{\hat{Q}(s, a) - \max_{a'} \hat{Q}(s, a')}{\tau} \right) \quad (6.4)$$

Here, τ denotes the temperature parameter, which controls the sharpness of the output distribution. Subtracting the maximum Q value before softmax ensures numerical stability, particularly in clinical contexts where extreme Q values may occur due to rare but high-reward outcomes.

6.2.5 Model Selection and Evaluation Metrics

For hypoglycaemia prediction classifiers, the model that produced the lowest cross-entropy loss in validation was selected and evaluated in the test set. Evaluation

metrics for the forecasting task included the area under the receiver operating characteristic curve (AUROC), the area under the precision–recall curve (AUPRC), and the macro-averaged F1 score.

A key lesson from Chapter 5 is that before applying any RL algorithm to clinical decision-making, we must first evaluate the reliability of off-policy evaluation (OPE) estimators using sanity checks. Specifically, naive baseline policies should not appear to outperform clinician-derived policies under any credible metric. This principle is essential for detecting estimator misspecification and high variance. We adopted a rigorous OPE framework designed to provide stable and conservative estimates of expected returns, thereby supporting safe model selection. Before computing OPE estimates, we first min-max normalise the reward by $\frac{r - (-3.23)}{0.77 - (-3.23)}$, where -3.23 and 0.77 are the lower and upper bounds of the per-step reward, respectively. To compute OPE scores, we have reported Importance Sampling (IS) and Weighted Importance Sampling (WIS) following Chapter 5. We also have reported bootstrapped variants [271], which provide empirical confidence intervals around each point estimate. Given the known instability of importance sampling (see Section 5.6 in Chapter 5), we apply a clipping threshold of 1 to the importance weights similar to Chapter 5. This strategy mitigates the influence of extreme values, although it may slightly underestimate performance [272, 273] and violate the performance lower bound. In clinical settings where conservative evaluation is preferred, this trade-off is acceptable.

To improve robustness and address the inconsistency of single-model OPE estimators (as shown in Chapter 5), we proposed an ensemble bootstrapped OPE method, shown in Algorithm 1. Instead of relying on a single BC model to compute importance weights, we train multiple BC estimators and combine their predictions through bootstrapping. This approach reduces estimator variance, captures model uncertainty, and provides more conservative confidence bounds for policy evaluation. Empirically, we set $\delta = 0.05$, corresponding to a 95% confidence interval, and perform $B = 1000$ bootstrap resamples when computing the ensemble

OPE estimates. The comparison between the proposed OPE and the single-model OPE is discussed in Section 6.4.3.

Algorithm 1: Ensemble Bootstrapped Off-Policy Evaluation

Input: Evaluation policy π_e , dataset \mathcal{D} , behaviour models $\{\hat{\beta}_1, \dots, \hat{\beta}_K\}$, number of bootstraps B , confidence level δ

Output: $(1 - \delta)$ confidence interval for ensemble OPE score

Step 1: Compute importance ratios and rewards;

for $i \leftarrow 1$ **to** K **do**

foreach trajectory $t \in \mathcal{D}$ **do**

 Compute ratio $\rho_{i,t}$ and corresponding reward $r_{i,t}$ using $\hat{\beta}_i$;
 Add $(\rho_{i,t}, r_{i,t})$ to global pool \mathcal{P} ;

Step 2: Bootstrap from pooled samples;

for $b \leftarrow 1$ **to** B **do**

 Sample with replacement a bootstrap set $\tilde{\mathcal{P}}_b$ from pool \mathcal{P} ;
 Compute OPE score for bootstrap b :
 $\hat{V}_b \leftarrow \text{OffPolicyEstimateFromPool}(\pi_e, \tilde{\mathcal{P}}_b)$;

Step 3: Aggregate and confidence interval;

$\hat{V}_{\text{ens}} \leftarrow \frac{1}{B} \sum_{b=1}^B \hat{V}_b$;

Sort $\{\hat{V}_b\}_{b=1}^B$ ascending;

$l \leftarrow \lfloor \delta \cdot B \rfloor$;

$\text{CI}_{\text{ens}} \leftarrow [\hat{V}_l, \hat{V}_{B-l}]$;

return \hat{V}_{ens} with CI_{ens} ;

6.2.6 Feature Interpretability

To explore the interpretability of the model and identify key clinical features that influence the predictions, the integrated gradient (IG) method was applied to trained neural networks [274]. This technique attributes changes in model output to individual input features by integrating gradients along a straight path from a reference input to the actual input. Implementation details, including baseline selection and numerical integration, are reported in Supplementary C.3.

6.3 Results for Hypoglycaemia Prediction

6.3.1 Prediction Performance

Performance metrics for hypoglycaemia prediction with a 12-hour observation window and a 6-hour forecast window across nine supervised learning models

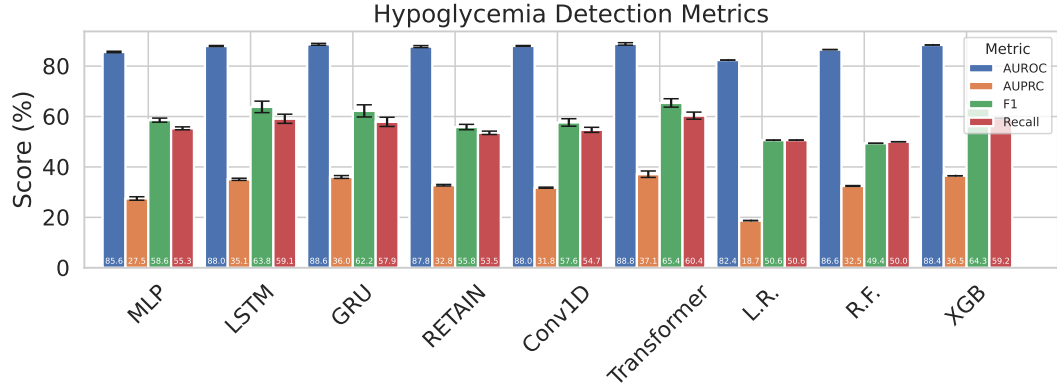


Figure 6.1: Performance comparison of nine supervised learning models for hypoglycaemia forecasting with a 12-hour observation window and a 6-hour forecasting window.

| Pred. | AUROC | AUPRC | Obs. | AUROC | AUPRC | Res. | AUROC | AUPRC |
|-------|------------|------------|------|------------|------------|------|------------|------------|
| 2 | 93.28±0.55 | 42.88±1.90 | 2 | 87.80±0.13 | 35.16±0.15 | 1 | 90.18±0.26 | 38.20±0.66 |
| 4 | 91.35±0.38 | 39.48±0.39 | 6 | 88.83±0.11 | 36.07±0.34 | 2 | 88.34±0.26 | 28.86±0.73 |
| 6 | 88.65±0.15 | 34.37±0.43 | 12 | 90.65±0.33 | 37.15±0.88 | | | |

Table 6.1: Sensitivity analysis of model performance. (Left) Prediction horizon, (Middle) Observation window, (Right) Resolution. ‘Res.’, ‘Obs.’ and ‘Pred.’ denote resolution, observation window and prediction horizon, respectively. The default model configuration uses a prediction horizon of 6 hours, an observation window of 12 hours, and a temporal resolution of 1 hour. Each subtable isolates the effect of varying one parameter while keeping the other two fixed at their default values.

are summarised in Figure 6.1. All models achieve similarly high AUROC scores (above 82%), indicating a strong discriminative capability to classify hypoglycaemic events. However, the corresponding AUPRC values are significantly lower, ranging from 18.7% for logistic regression to 37.1% for the Transformer, highlighting the impact of class imbalance in this task. Sequential deep learning models consistently outperform traditional baselines across all metrics. The Transformer achieves the best overall performance, with the highest AUPRC (37.1%), F1 score (65.4%), and recall (60.4%), demonstrating its ability to capture temporal patterns relevant to impending glucose drops. Interestingly, XGBoost slightly outperforms GRU and LSTM in terms of F1 score and recall, suggesting that informative features can still be extracted without explicit temporal modelling.

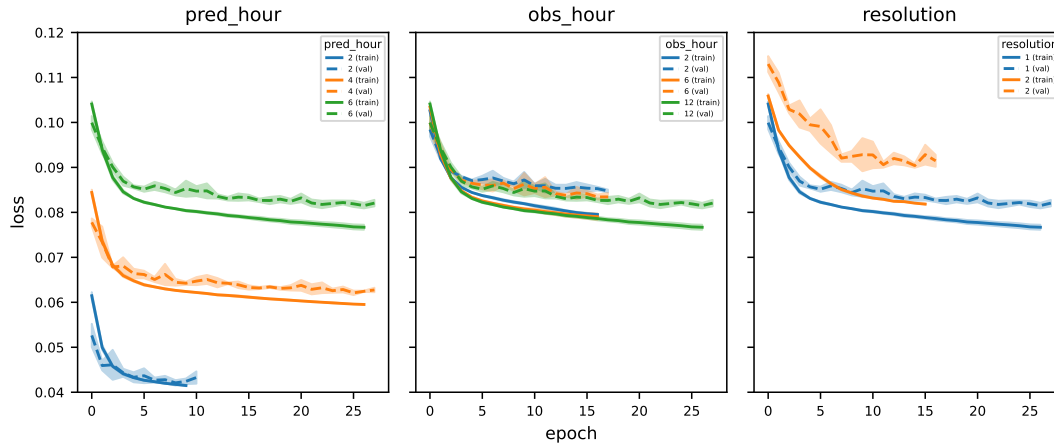


Figure 6.2: Sensitivity analyses for hypoglycaemia prediction, showing the impact of the prediction horizon (*left*), observation window (*centre*), and temporal resolution (*right*) on training and validation cross-entropy loss. The shaded areas represent the standard deviation across four random seeds. Solid lines denote training set loss and dashed lines denote validation set loss.

6.3.2 Sensitivity Analysis

We conducted targeted sensitivity analyses on the best-performing hyperparameter setting of the Transformer model to examine how variations in temporal parameters, namely, prediction horizon, observation window, and resolution, influence hypoglycaemia prediction performance. Figure 6.2 shows the comparison in terms of loss, and Table 6.1 shows the comparison in terms of predictive performance in the test set. The results show that shorter prediction horizons, longer observation windows, and finer temporal resolution consistently improve discrimination and precision. For example, reducing the prediction horizon from 6 to 2 hours increases AUROC from 88.65 to 93.28 and AUPRC from 34.37 to 42.88, while extending the observation window from 2 to 12 hours increases AUROC from 87.80 to 90.65 and AUPRC from 35.16 to 37.15. Similarly, moving from a 2-hour to a 1-hour resolution improves AUROC from 88.34 to 90.18 and AUPRC from 28.86 to 38.20. These findings suggest that capturing richer historical context and maintaining finer temporal granularity enhance predictive accuracy, while longer forecast lead times reduce performance due to physiological uncertainty.

The loss curves in Figure 6.2 further corroborate these findings by demonstrating that longer prediction horizons lead to higher training and validation losses, while extending the observation window reduces validation loss and improves convergence stability. A finer resolution consistently yields lower losses compared with coarser temporal settings. Collectively, the analyses indicate that temporal design choices strongly influence model performance, with shorter horizons, longer windows, and finer resolution producing the most accurate predictions. Nevertheless, we retain a 6-hour prediction horizon as the standard configuration to align with intensive care unit practice, prioritising clinical utility and actionable lead times for insulin therapy adjustments over maximising predictive metrics.

6.3.3 Model Interpretation

Figure 6.3 shows the Integrated Gradients (IG) attribution heatmap averaged over 3,717 positive test samples, identifying the features most influential in the prediction of hypoglycaemia. Glucose dominates the attribution profile across all time steps, with particularly strong contributions in the final 4 hours before prediction. Additional prominent signals arise from weight, potassium, and lactate, reflecting a broader metabolic status. Insulin-related variables (bolus, subcutaneous, infusion) provide moderate but supportive contributions in the last hours before prediction. Nutritional inputs (e.g., Lipids, Dextrose) and selected medications (e.g., ATC A10B, ATC N05A/N06A) contribute modestly, likely due to their sparse occurrence in the dataset. Overall, the attribution patterns support the clinical plausibility of the model, combining proximal glycaemic markers with complementary physiological indicators.

6.4 Results for Insulin Recommendation

This section presents the performance of behaviour cloning models and reinforcement learning RL policies for insulin dosing. We report all RL performance metrics using the format mean(lower bound \sim upper bound), where the mean and 95% confidence intervals are estimated via non-parametric bootstrapping. Specifically, we generate

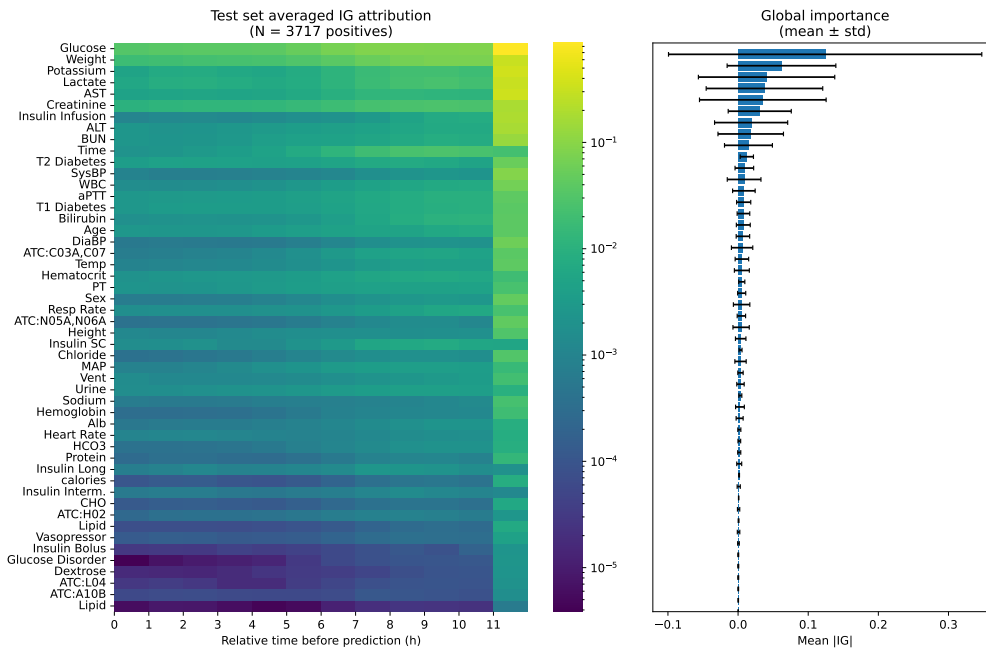


Figure 6.3: Integrated Gradients (IG) attribution heatmap for hypoglycaemia prediction on the test set ($N = 3,717$ positive cases). The heatmap shows the mean absolute IG attribution per feature across 12 temporal steps (hours before prediction). Features are ordered by overall importance (right panel), with error bars denoting standard deviation.

1,000 bootstrap samples, each drawn with replacement and equal in size to the original dataset, to estimate variability. To ensure comparability across metrics, the reward per step was normalised to the interval $[0, 1]$, where clinicians achieve $0.8575(0.8556 \sim 0.8593)$ per step, and the episodic return was scaled to $[0, 111.55]$, where the clinician’s baseline was $95.65(85.97 \sim 105.97)$.

6.4.1 Behavioural Cloning for Clinician’s Insulin Administration Policy

In DTR modelling, including prior treatment decisions in the input state is not a strict requirement from an RL perspective. In typical control tasks or games, where the state of the environment is fully observable or designed to be Markovian, previous actions are often excluded from the input. However, clinical decision-making often depends not only on physiological measurements, but also on recent dosing history, reflecting the clinician’s implicit intent or caution in their reasoning. Therefore, we have conducted behavioural cloning both with and without previous action

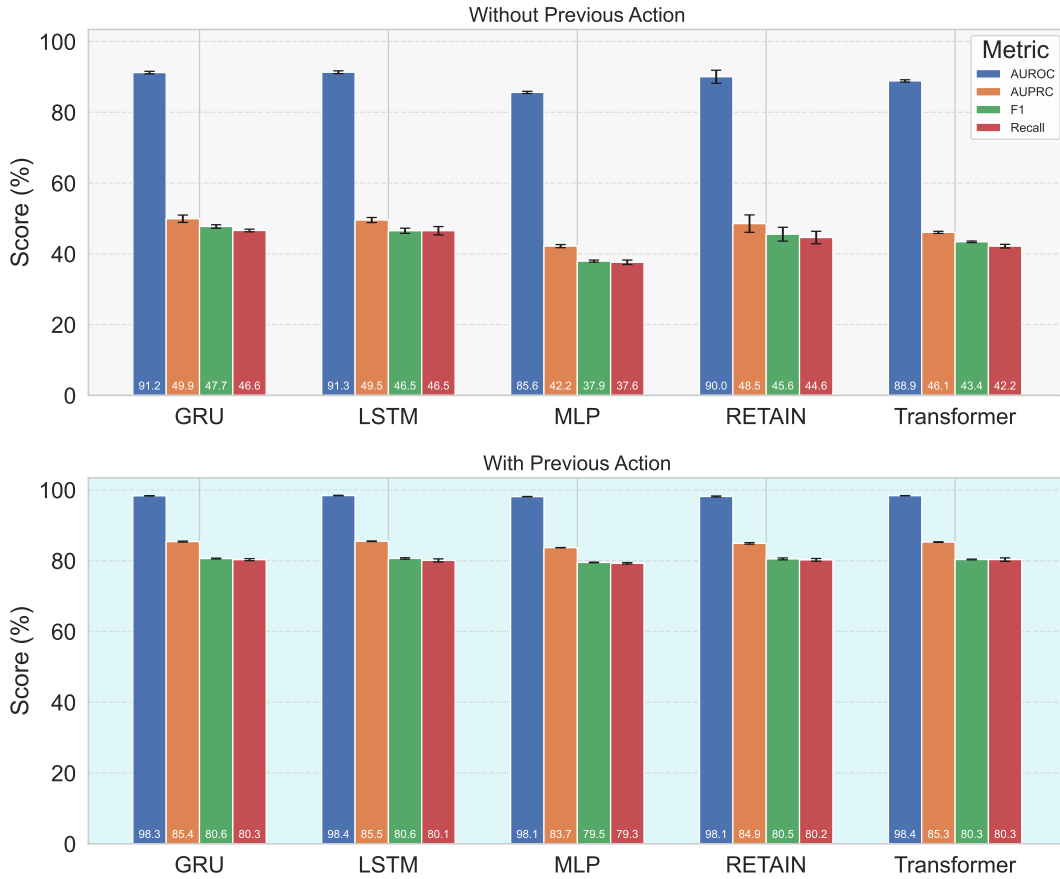


Figure 6.4: Behavioural-cloning performance with (bottom panel) and without (top panel) the previous insulin administration supplied as an input feature. For each encoder (GRU, LSTM, MLP, RETAIN, Transformer), we report mean test performance across four random seeds, where the exact values are printed beneath each bar; error bars denote standard deviations. Metrics are AUROC, AUPRC, macro-F1, and macro-recall.

in the state space. The result is shown in Figure 6.4. Performance comparisons reveal several key patterns. When the previous action is excluded, recurrent models such as GRU, LSTM, RETAIN, and Transformer outperform the MLP by 4 to 6 AUROC points, demonstrating their advantage in capturing temporal dependencies through hidden states. Among these, GRU and LSTM achieve the highest AUROC scores (above 91%), while MLP is slightly behind with an AUROC of 85.6%. Once the previous insulin rate is included as an explicit feature, performance improves substantially across all types of models. The AUROC of all models increases to nearly 98%. AUPRC and macro-F1 scores almost double, with all architectures

increasing to more than 80%. The performance gap between architectures narrows to under 0.3 AUROC points, indicating that the autoregressive signal provided by the last action largely captures the short-term decision context that recurrent models would otherwise infer from sequence history. The sharp performance gains observed with the inclusion of previous dosing suggest that insulin decisions are not fully determined by the current physiological state alone. By explicitly encoding the prior action, the models can more easily replicate these patterns without having to infer them from glucose dynamics alone.

6.4.2 Behavioural Model Calibration

To assess the reliability of predicted action probabilities from behaviour-cloning models, we evaluated the quality of the calibration using reliability diagrams for the four BC models that perform best based on the criteria defined in Section 6.2.5. Figure 6.5 plots the predicted class probabilities against the empirical frequencies observed across binned prediction intervals. We also report the reliability plot per class for each BC model in Figure C.4, Appendix C. Importantly, while all models were trained on the same dataset and achieved near-identical overall metrics, their confidence calibration profiles differ significantly (as evidenced by the variation in the blue lines). This highlights that even strong-performing models can behave differently in terms of uncertainty estimation. Across all four models, we observe that the predicted probabilities near 1.0 are already well calibrated, indicating that high-confidence predictions are generally reliable. However, at the lower end of the confidence spectrum (e.g., probabilities < 0.2), there is a clear mismatch between the predicted and observed frequencies, with models tending to underpredict uncertainty—i.e., assigning overly confident scores to rare or uncertain classes. Even after applying CTS, these discrepancies remain only partially corrected. This residual miscalibration is expected and aligns with our findings in Chapter 5, where we showed that calibration becomes unreliable in extremely low-probability regions due to data sparsity and long-tailed class distributions. The application of CTS improves the calibration, particularly in the rare-class regime, where model

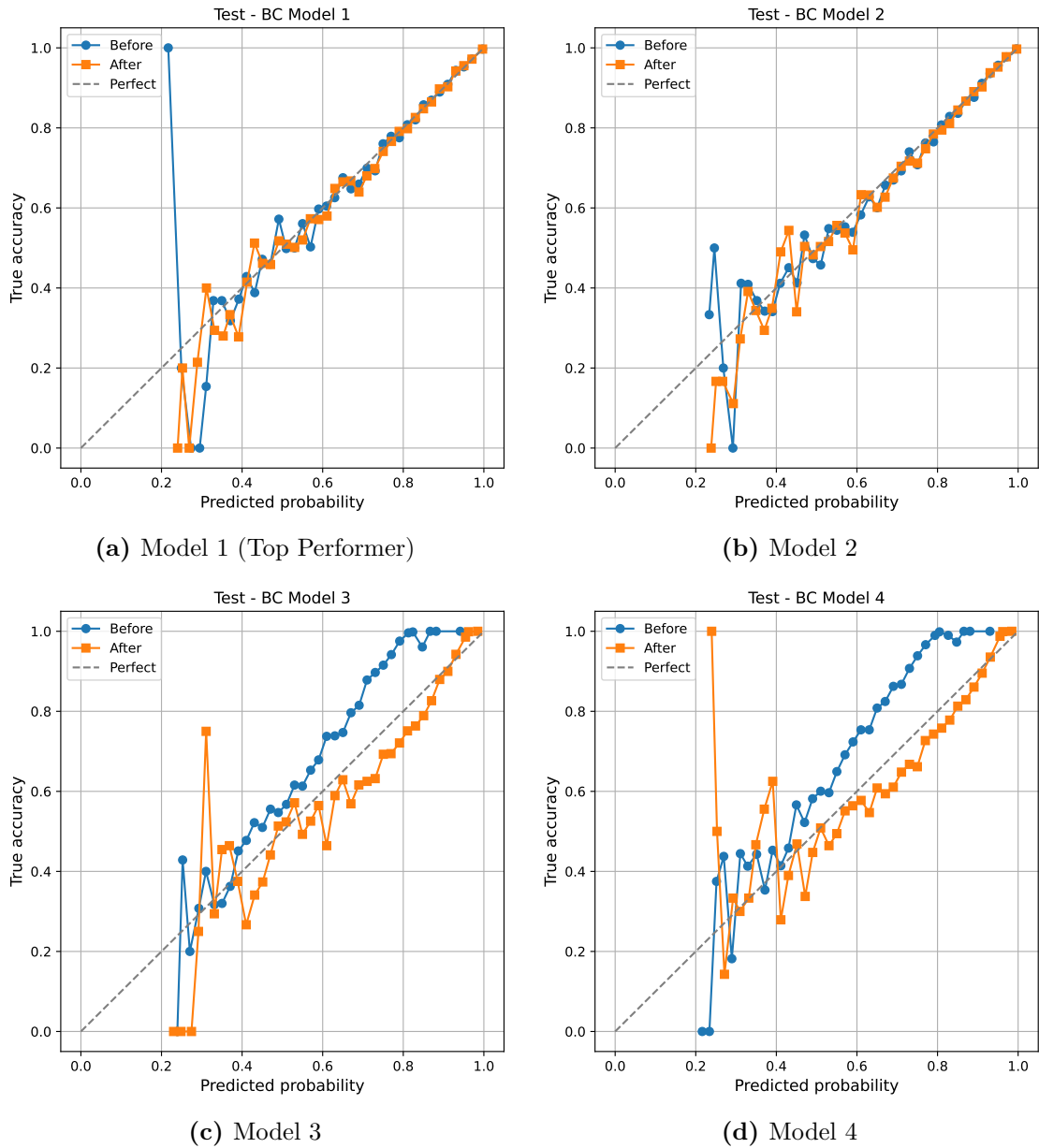


Figure 6.5: Reliability diagrams before and after calibration for the four GRU top-performing behaviour-cloning models used in off-policy evaluation. The blue line shows model predictions before calibration, the orange line shows predictions after class-wise temperature scaling, and the grey diagonal indicates perfect calibration. All models achieve near-perfect AUROC (~ 0.99) and AUPRC (> 0.84), but differ in their probabilistic confidence estimates.

predictions are more prone to overconfidence. Although it cannot completely correct confidence estimates at very small predicted probabilities, it provides a statistically grounded adjustment that enhances the trustworthiness of behavioural policies when used for policy evaluation and comparison.

6.4.3 Selection of Off-policy Evaluation By Naive Baseline Check

To assess the reliability of various OPE metrics, we first compare them with a range of naive baseline policies. This step serves as a sanity check. Since these baseline policies are not learnt from the data and do not incorporate patient-specific information, their performance should not exceed the clinician’s return (see Chapter 5 for rationales). Evaluating these baselines allows us to identify which OPE metrics align with clinical intuition and ground-truth expectations. We consider the following naive policies. **Zero-drug policy** (π_0) does not administer insulin at any time, regardless of the state of the patient. It acts as a conservative baseline and is expected to perform poorly under hyperglycemic conditions. **Max-drug policy** (π_{\max}) always applies the maximum insulin dose available. It represents an aggressive treatment strategy, usually resulting in a high risk of hypoglycaemia. **Uniform random policy** (π_r) selects an action uniformly at random from the insulin classes available in each timestep. This policy does not incorporate any structure and serves as a fully stochastic reference. **Weighted random policy** (π_w) samples actions according to their empirical frequency distribution observed in the training dataset. The **pulse policy**, denoted as $\pi_{P(p,c)}$, selects the insulin class c with probability p and defaults to the zero-drug class otherwise. This design provides a naive and repetitive dosing pattern. Empirically, we define the probability of insulin administration as $p \in \{0.05, 0.2, 0.5\}$, which corresponds to an average of a nonzero dose every 20, 5, or 2 hours, respectively. We select the pulse class c to be either 1 (i.e., 0–0.03 U/kg/hour) or 4 (i.e., > 0.1 U/kg/hour), resulting in six variants of pulse policy $\pi_{P(0.05,1)}, \pi_{P(0.2,1)}, \pi_{P(0.5,1)}, \pi_{P(0.05,4)}, \pi_{P(0.2,4)}, \pi_{P(0.5,4)}$. OPE metrics were computed in the following configurations. Metric_i is the performance metric for the

i -th top-performing BC policy trained using GRU, where $i \in \{1, 2, 3, 4\}$. $\text{Metric}_{b,i}$ is the bootstrapped metric for the BC estimator i , reporting in the form of 'mean (lower bound- upper bound)' of the 95% confidence interval. $\text{Metric}_{\text{ens}}$ is the ensemble metric, computed as the average across all four BC estimators. $\text{Metric}_{\text{ens},b}$ is our proposed ensemble metric with bootstrapped confidence intervals.

Tables 6.2 and 6.3 present detailed comparisons of various off-policy evaluation (OPE) metrics applied to naive baseline policies. The numerical results reveal critical limitations in the use of single-model OPE as estimators. Specifically, IS estimates are frequently near-zero and display minimal variation across policies. Single-model WIS is more stable than IS, but still exhibits considerable sensitivity to the choice of behaviour policy model. For example, under the zero-insulin baseline policy π_0 (first column of Table 6.2), WIS estimates vary widely across the four behaviour models, from 36.96 (WIS_1) to 74.73 (WIS_4). This model-dependent variability introduces uncertainty in policy ranking and challenges reproducibility. To address these issues, we adopt **WIS**_{ens,b}, the ensemble-averaged, bootstrapped variant of WIS, as our primary off-policy evaluation metric. This approach offers several key advantages: (1) reduced bias through weighted importance sampling; (2) improved robustness through ensemble averaging over multiple behaviour models; and (3) uncertainty quantification through bootstrapped confidence intervals. Empirically, $\text{WIS}_{\text{ens},b}$ exhibits consistent scaling across policies and avoids the instability seen in IS or PDIS. An exception is the estimated negative return for the π_{max} policy (-26.31), which appears to violate the theoretical lower bound of zero for the reward function. However, this result is acceptable due to the application of ratio clipping (truncating importance weights to ≤ 1), a standard conservative measure that induces downward bias but improves the stability of the estimator. Although this may slightly underestimate true returns, it aligns with our safety-first philosophy for clinical deployment.

Table 6.2: OPE metrics grouped by behaviour-cloned model, part 1. For OPE with bootstrapping, metrics are reported with the mean and 95% confidence interval. The subscripts 1, 2, 3, and 4 denote the best, second-best, third-best, and fourth-best behaviour policy models, respectively. Values below the reward lower bound are allowed due to the conservative ratio truncation.

| | π_0 | π_{max} | $\pi_{P(0.05,1)}$ | $\pi_{P(0.05,4)}$ | $\pi_{P(0.2,1)}$ |
|----------------------|--------------------------------|-----------------------------------|--------------------------------|--------------------------------|--------------------------------|
| IS _{b,1} | 0.07 _(0.00 0.21) | 0.00 _(0.00 0.00) | 0.00 _(0.00 0.00) | 0.00 _(0.00 0.00) | 0.00 _(0.00 0.00) |
| IS _{b,2} | 0.08 _(0.00 0.24) | 0.00 _(0.00 0.00) | 0.00 _(0.00 0.01) | 0.00 _(0.00 0.01) | 0.00 _(0.00 0.00) |
| IS _{b,3} | 0.30 _(0.08 0.61) | 0.00 _(0.00 0.00) | 0.04 _(0.00 0.12) | 0.04 _(0.00 0.12) | 0.00 _(0.00 0.00) |
| IS _{b,4} | 0.82 _(0.28 1.45) | 0.00 _(0.00 0.00) | 0.13 _(0.00 0.28) | 0.11 _(0.00 0.22) | 0.00 _(0.00 0.00) |
| IS ₁ | 0.07 | 0.00 | 0.00 | 0.00 | 0.00 |
| IS ₂ | 0.08 | 0.00 | 0.00 | 0.00 | 0.00 |
| IS ₃ | 0.31 | 0.00 | 0.04 | 0.04 | 0.00 |
| IS ₄ | 0.82 | 0.00 | 0.13 | 0.11 | 0.00 |
| WIS _{b,1} | 34.29 _(24.64 38.04) | 18.43 _(16.00 22.97) | 33.56 _(24.69 37.09) | 33.05 _(24.54 36.95) | 25.48 _(24.39 27.35) |
| WIS _{b,2} | 32.33 _(12.43 38.12) | 18.64 _(16.00 23.47) | 30.05 _(12.10 36.93) | 29.46 _(10.54 36.93) | 25.24 _(24.20 26.94) |
| WIS _{b,3} | 59.37 _(34.58 96.70) | 18.39 _(16.00 23.30) | 31.58 _(22.63 37.04) | 33.52 _(21.17 38.00) | 25.81 _(21.95 32.04) |
| WIS _{b,4} | 73.40 _(46.10 99.43) | 18.19 _(16.00 23.29) | 38.83 _(33.94 46.23) | 37.72 _(33.28 41.95) | 25.71 _(24.49 29.24) |
| WIS ₁ | 36.96 | 16.00 | 36.41 | 36.49 | 25.45 |
| WIS ₂ | 36.93 | 16.00 | 33.67 | 35.09 | 25.41 |
| WIS ₃ | 59.41 | 16.00 | 34.47 | 36.77 | 24.42 |
| WIS ₄ | 74.73 | 16.00 | 39.12 | 38.35 | 25.32 |
| IS _{ens,b} | 0.09 _(-0.03 0.22) | -0.00 _(-0.00 -0.00) | -0.00 _(-0.03 0.02) | 0.00 _(-0.02 0.03) | 0.00 _(0.00 0.00) |
| WIS _{ens,b} | 21.64 _(3.92 39.74) | -26.31 _(-26.44 -26.44) | 1.90 _(-25.18 16.53) | 8.42 _(-3.93 18.52) | 5.29 _(4.19 6.47) |

Table 6.3: OPE metrics grouped by behaviour-cloned model, part 2. For OPE with bootstrapping, metrics are reported with the mean and 95% confidence interval. The subscripts 1, 2, 3, and 4 denote the best, second-best, third-best, and fourth-best behaviour policy models, respectively. Values below the reward lower bound are allowed due to the conservative ratio truncation.

| | $\pi_{P(0.2,4)}$ | $\pi_{P(0.5,1)}$ | $\pi_{P(0.5,4)}$ | π_r | π_w |
|----------------------|--------------------------------|--------------------------------|--------------------------------|--------------------------------|--------------------------------|
| IS _{b,1} | 0.00 _(0.00 0.00) | 0.00 _(0.00 0.00) | 0.00 _(0.00 0.00) | 0.00 _(0.00 0.00) | 0.00 _(0.00 0.01) |
| IS _{b,2} | 0.00 _(0.00 0.00) | 0.00 _(0.00 0.00) | 0.00 _(0.00 0.00) | 0.00 _(0.00 0.00) | 0.02 _(0.00 0.06) |
| IS _{b,3} | 0.00 _(0.00 0.00) | 0.00 _(0.00 0.00) | 0.00 _(0.00 0.00) | 0.00 _(0.00 0.00) | 0.04 _(0.00 0.11) |
| IS _{b,4} | 0.00 _(0.00 0.00) | 0.00 _(0.00 0.00) | 0.00 _(0.00 0.00) | 0.00 _(0.00 0.00) | 0.04 _(0.00 0.12) |
| IS ₁ | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| IS ₂ | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 |
| IS ₃ | 0.00 | 0.00 | 0.00 | 0.00 | 0.04 |
| IS ₄ | 0.00 | 0.00 | 0.00 | 0.00 | 0.04 |
| WIS _{b,1} | 22.23 _(17.96 26.36) | 25.73 _(24.50 29.44) | 14.53 _(9.78 25.47) | 25.63 _(23.36 26.55) | 30.94 _(24.49 36.90) |
| WIS _{b,2} | 26.13 _(24.67 29.81) | 27.71 _(25.10 30.33) | 28.02 _(23.47 30.41) | 25.64 _(23.36 26.54) | 32.46 _(24.41 36.96) |
| WIS _{b,3} | 18.46 _(9.78 35.45) | 24.91 _(24.52 25.59) | 15.16 _(9.79 25.33) | 25.73 _(24.42 26.54) | 32.07 _(25.41 36.87) |
| WIS _{b,4} | 27.97 _(24.48 31.64) | 25.72 _(25.04 26.28) | 23.79 _(16.83 25.39) | 25.72 _(24.20 26.54) | 32.89 _(25.59 38.26) |
| WIS ₁ | 21.76 | 24.67 | 9.81 | 26.54 | 33.52 |
| WIS ₂ | 25.67 | 28.36 | 30.28 | 26.53 | 36.55 |
| WIS ₃ | 15.69 | 24.74 | 10.49 | 26.45 | 33.60 |
| WIS ₄ | 28.76 | 25.84 | 24.71 | 26.47 | 34.03 |
| IS _{ens,b} | 0.00 _(0.00 0.00) | 0.00 _(0.00 0.00) | 0.00 _(0.00 0.00) | 0.00 _(0.00 0.00) | 0.01 _(0.00 0.02) |
| WIS _{ens,b} | 8.93 _(1.35 15.86) | 15.42 _(10.84 16.97) | 9.34 _(4.10 16.16) | 15.17 _(13.92 15.74) | 18.17 _(17.07 18.60) |

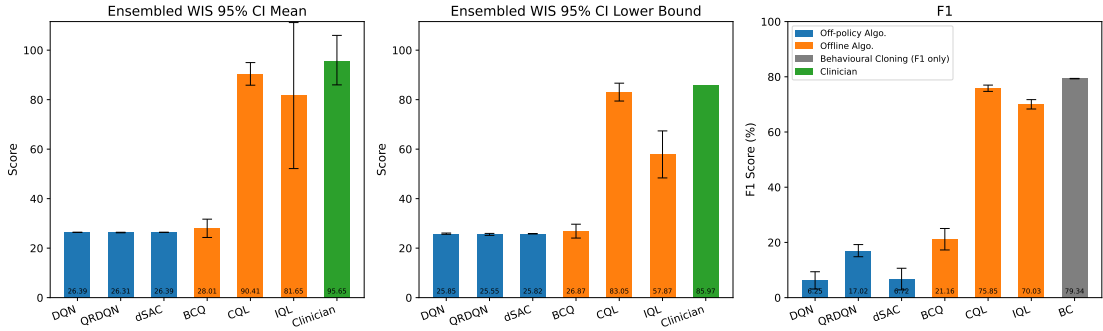


Figure 6.6: Evaluation of reinforcement learning policies using ensemble bootstrapped WIS, ensemble bootstrapped WIS lower bound and macro F1 score. We report the mean performance across four random seeds. Error bars denote one standard deviation. The 95% CI mean and lower bound are shown separately. Algorithms are classified into off-policy (blue), offline (orange) and behavioural cloning (grey). Clinicians’ return is shown in green.

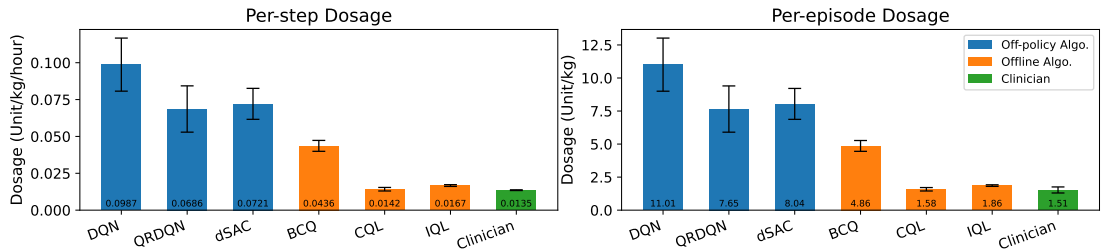


Figure 6.7: Insulin dosing behaviour of reinforcement learning policies. We report the mean value across four random seeds. Error bars denote one standard deviation.

6.4.4 Reinforcement Learning for Insulin Administration

We implemented and evaluated six offline reinforcement learning algorithms, i.e., DQN, BCQ, CQL, IQL, QRDQN and dSAC, trained to recommend weight-adjusted insulin doses using a clinically informed reward function based on the Blood Glucose Risk Index (See Section 6.2.2). All models were trained in the same patient cohort and evaluated using a bootstrapped ensemble WIS estimator to ensure robust off-policy evaluation in high-stakes clinical settings. Figure 6.6 compares the performance of RL-derived policies to clinician dosing behaviour. Offline RL algorithms substantially outperform their off-policy counterparts across all evaluation metrics. Value-based methods such as DQN, QRDQN, and dSAC exhibit tightly clustered WIS mean scores near 26, with lower bounds hovering just above 25. In contrast, offline RL methods, specifically CQL and IQL, achieve markedly higher

WIS means, reflecting their ability to avoid extrapolation errors by conservative learning. CQL achieves a WIS mean of 90.41, and IQL follows with 81.65. In contrast, although BCQ is an offline algorithm, it performs closer to the value-based group with a WIS of 28.01. Among all the policies learnt, CQL emerges as the clear best performer. Its WIS mean of 90.41 approaches that of the clinician benchmark (95.65), and its 95% CI lower bound reaches 83.05, remarkably close to the clinician lower bound of 85.97. This suggests that even with conservative estimates, CQL remains highly competitive. IQL performs well in expectation, but exhibits greater variability across seeds; its lower limit of 57.87, while respectable, falls considerably short of the CQL's, highlighting concerns about stability and worst-case reliability. Interestingly, while reinforcement learning is not explicitly trained to mimic clinician decisions, F1 scores offer a useful sanity check. Higher-performing RL policies tend to show stronger alignment with clinician actions. CQL achieves an F1 of 75.85, and IQL follows at 70.03, approaching the upper bound F1 of the behavioural cloning model of 79.34. In contrast, value-based policies show dramatically lower F1 scores, with DQN at 6.25, QRDQN at 17.02, and dSAC at 6.72. BCQ scores slightly higher at 21.16, but remain distant from the offline leaders. These results reinforce the broader observation from Chapter 5. Policies that deviate significantly from clinician behaviour often perform poorly, suggesting that clinical alignment serves as a meaningful indicator of policy effectiveness rather than a cosmetic feature. Figure 6.7 illustrates the insulin dosing behaviour of different RL models. We observe that DQN, QRDQN, dSAC, and BCQ consistently recommend higher insulin doses compared to clinicians. In contrast, CQL and IQL produce dosing patterns that are more closely aligned with clinical practice. This discrepancy explains the poorer performance of the former group, as overly aggressive dosing strategies often lead to unsafe and suboptimal glucose control.

6.5 Discussion

Our study demonstrates the significant potential of advanced machine learning and reinforcement learning methods to manage glucose instability among critically ill pa-

tients. By systematically benchmarking a wide array of supervised learning models, we establish that sequential neural architectures, particularly Transformer-based models, robustly predict hypoglycaemic events in ICU settings. Furthermore, our reinforcement learning-derived insulin administration policies effectively mirrored and approximated professional clinical protocols, showcasing the practical utility of data-driven approaches for complex treatment decision-making in critical care. These findings support our original hypothesis that the use of multimodal and temporally rich data could enable more accurate risk stratification and personalised insulin titration compared to rule-based protocols currently in use.

A primary contribution of our work is the creation of a thoroughly curated and publicly reproducible dataset derived from the MIMIC-III database, which encompasses a diverse range of clinical variables, including glucose trajectories, insulin administration, nutritional support, vital signs, laboratory measures, and pharmacological interventions. Unlike many previous studies that used narrowly defined patient cohorts or continuous glucose monitoring, our dataset represents a broader adult ICU population. This enhances the ecological validity of our models and improves the generalisability of our findings to typical clinical settings. In particular, our cohort included patients with and without diabetes, with varying comorbidities and demographic backgrounds, increasing applicability across a variety of ICU environments. In doing so, our work responds directly to the gaps identified in the prior literature, where a lack of data diversity has hindered model deployment.

The sensitivity analyses conducted in our study provide clear, clinically relevant insights. The finding that extended observation windows and higher temporal resolutions consistently improve prediction accuracy highlights the importance of capturing detailed physiological dynamics, supporting the clinical necessity of frequent patient monitoring in ICU settings. For example, AUROC increased from 87.80 to 90.65 when the observation windows were expanded from 2 to 12 hours, and prediction accuracy improved substantially when using hourly rather than 2-hourly input resolution. Furthermore, while shorter prediction horizons naturally

yield higher accuracy due to reduced uncertainty, our standard adoption of a 6-hour horizon reflects a clinically meaningful compromise, aligning predictive utility with the actionable timeframe required for long- and intermediate-acting insulin management. This decision reflects real-world trade-offs between early intervention planning and prediction feasibility in fast-changing ICU contexts.

Behavioural cloning analyses further underscored the clinical relevance of incorporating prior insulin administration into predictive models. By explicitly incorporating dosing history, all models achieved markedly superior performance, with GRU's AUROC increasing from 91.2 to 98.3 and AUPRC rising from 49.9 to 85.4. This result reinforces the necessity of capturing autoregressive relationships in treatment decisions and highlights the potential for future model enhancements through more complex contextualisation of therapeutic history. Clinically, it supports the practice of incorporating prior medication events into risk prediction systems, particularly for time-sensitive therapies such as insulin.

Our reinforcement learning analyses demonstrate that while offline RL shows potential for data-driven insulin administration, its capabilities are often overstated in the prior literature. Through sanity checks, we confirmed that single-model OPE produces inconsistent and sometimes misleading estimates, aligned with the findings of Chapter 5. To address this, we introduced an ensemble bootstrapped Weighted Importance Sampling approach, which yields a more stable and conservative policy evaluation. Under this improved evaluation framework, no off-line RL algorithm exceeds the performance of clinicians. However, conservative algorithms, particularly CQL and IQL, are closest to approximating the dosing strategies of clinicians. This is an important and realistic outcome. In high-stakes ICU care, any algorithm that appears to 'easily outperform' well-trained clinicians retrospectively is typically overoptimistic or flawed in evaluation. Our results demystify the perceived superpower of RL and suggest that reliably approximating clinician behaviour is already a meaningful achievement.

At the same time, our findings highlight the inherent limitations of RL in this clinical setting. Insulin dosing in the ICU is a relatively straightforward

control problem. Glucose is a single, well-understood target variable, and the pharmacological effect of insulin is monotonic (higher doses lower glucose). Despite this simplicity, RL agents still struggle to fully avoid unsafe dosing patterns. This shortcoming highlights a deeper challenge that we aim to address in the next chapter: *injecting Prior knowledge into dynamic treatment agent*.

In the future, prospective studies and controlled clinical trials will be the next crucial steps in translating these promising retrospective findings into clinical practice. Future work should also emphasise the integration of real-time clinical feedback mechanisms and the evaluation of clinician acceptance of AI-driven decision support tools. Such prospective validations will not only confirm the robustness and safety of the proposed methods but will also facilitate their seamless integration into routine ICU care. There is also scope for extending the framework to multi-agent settings, incorporating dietary management, glucocorticoid tapering, or other endocrine modulators that affect glucose control. More broadly, our open-source dataset and modelling framework provide a scalable benchmark for future research, enabling global labs to replicate, extend, and validate AI-assisted glycaemic management strategies, ultimately aiming to reduce preventable morbidity and mortality in critical care environments.

7

Are Large Language Models Dynamic Treatment Planners? An In Silico Study from a Prior Knowledge Injection Angle

7.1 Introduction

Chapter 4 demonstrated that RL algorithms could successfully learn effective DTRs for complex clinical tasks, particularly glycaemic control, in carefully controlled settings. However, the practical implementation of RL approaches faces significant challenges, including the need for carefully engineered reward signals, observation space, and action space. To achieve robustness against pharmacokinetic/pharmacodynamic variability, developers must inject sufficient prior knowledge during training, such as policy safety constraints [275], world modelling [276] or direct human advice [277]. However, existing methods are mostly from robotics areas, where adding human demonstrations or applying constraints is at a lower cost. In the field of treatment regimes, there is no silver bullet to inject prior knowledge before or during training.

Recent advances in large language models (LLMs) present an alternative and complementary approach that could mitigate these limitations [278, 279]. LLMs possess distinctive capabilities that make them promising candidates for clinical decision support. First, their proficiency in in-context learning and instruction-following facilitates embedding straightforward prior knowledge, which is typically challenging to encode explicitly in smaller neural network models [280]. Second,

their pre-trained medical knowledge may enhance the quality and safety of clinical recommendations. LLMs also exhibit sophisticated reasoning capabilities, especially when analysing complex historical EHR data [281]. The potential of LLMs for generating treatment recommendations remains largely unexplored, although previous research has extensively explored the reasoning abilities of LLMs across domains such as mathematics, coding, robotics, and medical diagnosis [282]. LLMs can potentially capture the rich temporal dependencies inherent in patient trajectories, provide human-readable rationales, and offer safer, more interpretable treatment policies without explicit reward optimisation.

This chapter investigates whether open-sourced state-of-the-art LLMs can generate clinically meaningful insulin dosing recommendations purely through inference. We directly compare their performance with small neural-network-based RL agents within the same experimental framework. Specifically, our investigation aims at exploring four key research questions (RQs). **(RQ1)** How can expert clinical knowledge be effectively injected into RL, and can LLMs simplify this process? **(RQ2)** Can LLMs outperform small neural-network-based RL policies in zero-shot inference settings? **(RQ3)** Do the parameter scaling law and sampling temperature findings observed in other LLM tasks extend to dynamic treatment regime scenarios? **(RQ4)** Does incorporating chain-of-thought (CoT) prompting improve reasoning quality and clinical performance in dynamic treatment scenarios? To analyse these questions, we use the SimGlucose type-1 diabetes simulator introduced in Chapter 4, a validated in silico environment that rigorously evaluates insulin dosing policies.

7.2 Problem Formulation and Environment Setting

We use the SimGlucose environment to validate the capabilities of LLMs for two critical considerations. First, despite having a seemingly simple observational structure, consisting primarily of blood glucose measurements, the environment exhibits highly complex underlying dynamics. As demonstrated in Chapter 4, maintaining glycaemic control requires precise dosage adjustments, where even

advanced off-policy RL algorithms struggle under realistic noise and inter-patient pharmacokinetic/pharmacodynamic variability. Second, the simulator presents straightforward rules, such as "insulin reduces glucose", which are challenging to embed explicitly into traditional RL policies. In contrast, LLMs can directly incorporate prior knowledge through natural language instructions, making the SimGlucose environment an ideal testbed to assess whether LLMs excel as clinicians who follow instructions.

The patient's state is observed in a discrete timestep of 15 minutes and includes the preceding four-hour glucose trajectory and the insulin infusion history. Similarly to Chapter 4, the agent selects a continuous (for PPO) or 11-bin discrete (for DQN) insulin infusion rate $a_t \in [0, 9]$ U/h, administered as a basal rate over the subsequent 15-minute control interval. The clinical goal is to maintain blood glucose within the target range $[70, 140]$ mg/dL, avoiding severe excursions below 40 mg/dL or above 500 mg/dL. Episodes end after 16 hours (64 time steps) or immediately after glucose excursions above $(40, 500)$ mg / dL.

7.3 Reward and Evaluation Protocol

To rigorously evaluate insulin dosing strategies generated by small reinforced models and LLMs, we adopt three key evaluation metrics:

- **Survival Rate:** Defined as the proportion of episodes that complete the full 16-hour simulation (64 timesteps) without termination due to extreme blood glucose (BG) excursions. Termination occurs if BG drops below 40 mg/dL or exceeds 500 mg/dL.
- **Time-In-Range (TIR):** A standard metric that quantifies the fraction of time steps in which BG lies within the recommended range of 70–180 mg/dL. This assesses the 'control quality' and is computed as the average hit rate across the episode.

- **Normalised Return:** Derived from the total episodic return specified in the reward design (See Eq4.10 in Chapter 4, this metric captures the cumulative efficacy of treatment using the glucose risk index. We min-max scale the return to the $[0, 100]\%$ range using: $\frac{\text{Return} - (-99.7)}{64 - (-99.7)} \times 100$, where -99.7 is the minimum possible return and 64 is the maximum (survival with optimal glycaemic control).

All metrics are computed over three groups of four patients: adults, adolescents, and children. In SimGlucoseEnv, adult physiology is modelled as the least volatile, and paediatric physiology as the most volatile. Consequently, cohorts can be interpreted as representing increasing levels of difficulty: easy (adult), medium (adolescent), and hard (child). Each patient undergoes 20 test episodes (4 seeds \times 5 repeats), delivering sufficient statistical power. We report metrics using bootstrapped confidence intervals. The error bars represent the 95% confidence interval derived by bootstrapping each metric 1,000 times with replacement and selecting the 2.5% and 97.5% percentiles, respectively. The primary reported 'mean' metric is the mean of the bootstrapped result. The notation is used for all the following results, unless otherwise specified.

For small reinforced agents (SRAs), DQN and PPO are trained in adult, adolescent, child and mixed environments, respectively, and the best model is selected by achieving the best overall evaluation return. The best selected model is then evaluated on each stratum in the evaluation environments for generalisation assessment.

For LLMs, the pretrained language models are evaluated directly in the evaluation environments with different prompt methods, which will be introduced in Section 7.4.2.

7.4 Experimental Setting

7.4.1 Small Reinforced Agents in Insulin Dosing

Training Small Neural Networks with Reinforcement Learning

We selected Deep Q-Network (DQN) and Proximal Policy Optimisation (PPO) as representative algorithms for off-policy and on-policy reinforcement learning, respectively. Detailed descriptions of the algorithms are provided in Section 2.4.1, Chapter 2. Although Chapter 4 justified focusing exclusively on off-policy algorithms due to the impracticality of on-policy updates in real clinical settings, we include PPO here to establish an aspirational upper-bound performance benchmark against which we can evaluate LLM policies. Both algorithms were trained separately under four distinct scenarios to account for varying levels of patient variability: adult only (easy), adolescent only (medium), child only (hard) and a mixed group from all three categories of patients. Each training epoch consisted of 480 simulation steps, equivalent to 10 full episodes without early termination. Training was performed over 20 epochs, with model checkpoints saved and evaluated at the end of each epoch in their respective training environments. The best-performing checkpoints were selected based on the highest normalised episodic training return, and were subsequently evaluated on the complete evaluation environment set described in Section 7.2. Detailed hyperparameter configurations and training specifics are provided in Appendix D.3.

Injecting Prior Knowledge in Small Reinforced Agents via Exploration

Integrating clinical prior knowledge directly into RL poses challenges due to the difficulty neural networks have in interpreting natural language or expert intuition directly. Due to the sparse dosage nature of DTR, we choose to incorporate clinical insights into the exploration strategy, thus guiding clinically plausible actions from the early training stages.

For DQN, exploration typically employs an ϵ -greedy strategy, where actions are selected uniformly throughout the action space. In insulin dosing scenarios, this approach often leads to unsafe, clinically implausible actions (e.g., overdosing), as

the optimal insulin dosage is frequently zero. To mitigate this issue, we introduce a domain-informed modification that assigns the zero-dose action a significantly higher exploration probability. Denoting the probability of selecting action 0 (i.e., zero dose) as p_0 and selecting any other action as $p_j, j \in [1, 2, \dots, |A| - 1]$, we design a heuristic

$$\frac{p_0}{p_j} = \frac{|A|(|A| + 1)}{|A| - 1}$$

where $|A|$ is the number of discrete dose levels. Since we use an 11-bin action space for DQN, the probability of selecting zero dosage during exploration is around 0.569. Further justifications of the heuristic design can be found in D.2

For PPO, the integration of prior knowledge differs due to continuous action spaces. We explored two primary action transformations: a smooth hyperbolic tangent (Tanh) function and a hard-clipping function. The clip transformation is commonly used due to simplicity and stability, and is considered the ‘without prior knowledge’ choice. In contrast, the Tanh transformation (considered the ‘with prior knowledge’ choice) provides smooth action saturation, implicitly embedding prior knowledge that favours conservative dosing decisions unless strongly justified by model predictions. Detailed theoretical analyses and justifications comparing these transformations are provided in Appendix D.1.

In addition, we initialise the PPO actor network with near-zero weights and zero biases to ensure that, at the beginning of training, the policy outputs actions close to zero. This encourages safe, low-dose decisions during the early learning phase. To further align initial exploration with prior knowledge, we warm-start the training buffer using a constrained random policy biased toward small doses. For PPO, actions are uniformly sampled from the continuous interval $[0, 2]$ U/h. For DQN, actions are sampled uniformly from the discrete set $\{0, 1, 2\}$, where actions 0,1,2 correspond to dose 0, 0.9 and 1.8 units/hour. This strategy ensures that early-stage trajectories observed by the agent predominantly reflect conservative dosing behaviours, thereby promoting safer exploration from the outset.

7.4.2 LLMs in Insulin Dosing

Basic LLM Policy Construction

To facilitate LLM inference on structured medical time series, we first transformed numerical patient-state representations into descriptive text. Specifically, each observation was formatted to include the patient’s glucose measurements, insulin infusion rates, and corresponding timestamps at 15-minute intervals. This structured format mirrors real-world EHRs and leverages previous research demonstrating improved LLM reasoning when inputs are presented as formatted text [283, 284]. An illustrative example of a four-hour glucose trajectory (covering 16 discrete timesteps) is provided in Prompt 1.

Prompt 1: Observation Example

```
Day 1, Time: 05:00:00 (initial measurement), glucose: 159.27 mg\dl, insulin rate: 0.0000 unit\hour, insulin dose: 0.00 unit.  
Day 1, Time: 05:15:00, glucose: 148.96 mg\dl, insulin rate: 1.0000 unit\hour, insulin dose: 0.25 unit.  
Day 1, Time: 05:30:00, glucose: 149.71 mg\dl, insulin rate: 0.0000 unit\hour, insulin dose: 0.00 unit.  
...
```

Prompt 2: Basic Zero-shot Prompt Template

```
You are a clinical specialist responsible for managing patients with Type-1 Diabetes. Your primary objective is to maintain the patient's blood glucose levels within the safe range of 70-140 mg/dL by administering appropriate insulin doses.  
  
### Monitoring and Decision Frequency  
Blood glucose levels are observed every 15 minutes. Your task is to determine the insulin dose rate in units/hour every 15 minutes based on the latest glucose readings and trends.  
  
### Insulin Administration  
- Insulin lowers blood glucose levels with a delayed effect. Your decisions will define the insulin dose rate to be administered over the next 15 minutes to maintain a normal glucose level in a long run.  
- We do not distinguish between basal and bolus insulin. The dosing decision is based on the latest glucose readings and trends.  
- Rate Range: You MUST provide a dose in the range of [0 to 9] units/hour, inclusive.  
- Administration Interval: The specified dose is distributed evenly over the 15-minute period, i.e., the total dosage is your_action/60*15. You only need to provide the dose rate in units/hour. Do NOT specify the total dose.  
  
###Observations  
<Observation>  
  
###Request  
Determine the optimal insulin rate for the current 15-minute interval to maintain a patient's blood glucose levels within the safe range of 70-140 mg/dL. Choose a dosage value. For example, if you choose 0 units, enter 0. DO NOT say anything else.  
  
###Answer
```

The baseline LLM policy was constructed using a straightforward system prompt to maintain blood glucose within the safe range of 70–140 mg/dL by selecting an appropriate insulin infusion rate every 15 minutes. This prompt explicitly specifies

both the clinical goals and the action constraints, including limiting insulin dosage rates to the clinically acceptable range of 0-9 units/hour. A minimal instruction was added to request a numeric response, forming the zero-shot inference policy prompt (see Prompt 2). This minimalist, zero-shot prompting strategy served as the foundational baseline for all subsequent LLM policy evaluations, based on prior evidence that even minimal natural language supervision can effectively guide LLMs in decision-making scenarios.

Injecting Prior Knowledge in LLMs

Prompt 3: Expert Knowledge System Prompt

```
<Same base system prompt as Prompt 2>

### Hidden Variables
- Food Intake: Food consumption increases blood glucose levels.
- Exercise: Exercise reduces blood glucose levels.
- Estimation: Since food intake and exercise are not directly observable, estimate based on time of day and observed glucose trends using clinical judgment and common sense.

### Penalties and Risks
- Blood Glucose Outside Safe Range (70-140 mg/dL):
  - Above 140 mg/dL: Hyperglycaemia penalties.
  - Below 70 mg/dL: Hypoglycaemia penalties, with increased severity.
  - Above 500 or below 40 mg/dL: EXTREMELY DANGEROUS, your treatment will be considered a failure and the patient will die!
- Insulin Dose Considerations:
  - High Doses: Use cautiously to avoid rapid and excessive lowering of glucose levels.
  - Low Glucose Levels (<70 mg/dL):
    - Action: Immediately cease insulin administration until glucose levels rise above 70 mg/dL.
    - Priority: Prevent hypoglycaemia due to its acute dangers.

### Safety Precautions
- Avoid Overdosing Insulin: Prevent hypoglycaemia by carefully balancing insulin doses.
- Insulin Stacking Awareness: You should consider the accumulated dosage and the delayed effect of insulin on glucose levels carefully.
- Prioritize Patient Safety: Always aim to keep glucose levels within the target range. If uncertainty exists, opt for a lower or zero insulin dose to ensure safety.

### Observations
<Observation>

### Request
Determine the optimal insulin rate for the current 15-minute interval to maintain a patient's blood glucose levels within the safe range of 70-140 mg/dL. Choose a dosage value. For example, if you choose 0 units, enter 0. DO NOT say anything else.

### Answer
```

Prompt 4: Zero-shot CoT Instruction

```
Determine the optimal insulin rate for the current 15-minute interval to maintain a patient's blood glucose levels within the safe range of 70-140 mg/dL. First, analyse the current state step-by-step. Finally, you must choose a dosage value enclosed in answer tags (i.e., <ans> and </ans>]), for example, <ans>0</ans>, without any non-numerical word. Let's think step by step.
```

To add priors to LLMs, we extend the baseline prompt by embedding prior knowledge directly into the system instruction (Prompt 3). This expert-augmented prompt includes additional guidance, such as the delayed effect of insulin, safety

Prompt 5: Zero-shot CoT With Meal Information Instruction

Determine the optimal insulin rate for the current 15-minute interval to maintain a patient's blood glucose levels within the safe range of 70-140 mg/dL. First, analyse the current state step-by-step by estimating the Correction Factor(CF) and Total Daily Insulin(TDI). If the patient has taken any meal, you must consider the meal effect. Finally, you must choose a dosage value enclosed in answer tags (i.e., <ans> and </ans>], for example, <ans>0</ans>, without any non-numerical word. Let's think step by step.

thresholds for hypoglycaemia, and cautionary principles around high-dose administration. It also introduces reasoning about hidden variables, such as estimating food intake or exercise effects from indirect indicators like time of day or rising glucose trends. By explicitly encoding these safety rules and latent clinical patterns, our objective is to improve the plausibility and robustness of the dosing decisions of the model. In recent work, similar approaches have been explored that aim to ground LLM reasoning in clinical heuristics [285]. We then evaluate whether language models can engage in explicit clinical reasoning using the zero-shot CoT prompting strategy (Using the instruction in Prompt 4 and the same system prompt as Prompt 3). Here, the instruction requires the model to analyse the current patient's condition step by step before selecting a dose. The final answer is wrapped in delimiters (<ans>answer</ans>) to isolate the scalar output from the reasoning trace. Lastly, we explore a variant of chain-of-thought prompting that attempts to uncover latent clinical variables, particularly the presence of meals and the patient's insulin sensitivity (Using instruction Prompt 5 and using the same system prompt as Prompt 3). The prompt encourages the model to estimate factors such as total daily insulin (TDI) or correction factor (CF), and to infer hidden states from glucose dynamics when explicit indicators are missing. This is particularly relevant in partially observed settings like SimGlucose, where meal events are recorded as binary tokens, but not described explicitly in natural language. By embedding instructions to reason over such hidden variables, we assess whether LLMs can estimate hidden variables and the environmental PK/PD dynamics. Across all types of prompts, LLMs act solely via inference without training.

7.5 Results

We present an extensive evaluation comparing LLMs with SRAs. Specifically, we examine two representative open-source LLM families, Qwen2.5 [286] and LLaMA3 [287], selected due to their robust general-purpose reasoning capabilities and widespread adoption in the community.

7.5.1 The effectiveness of Prior Knowledge Injection on SRAs

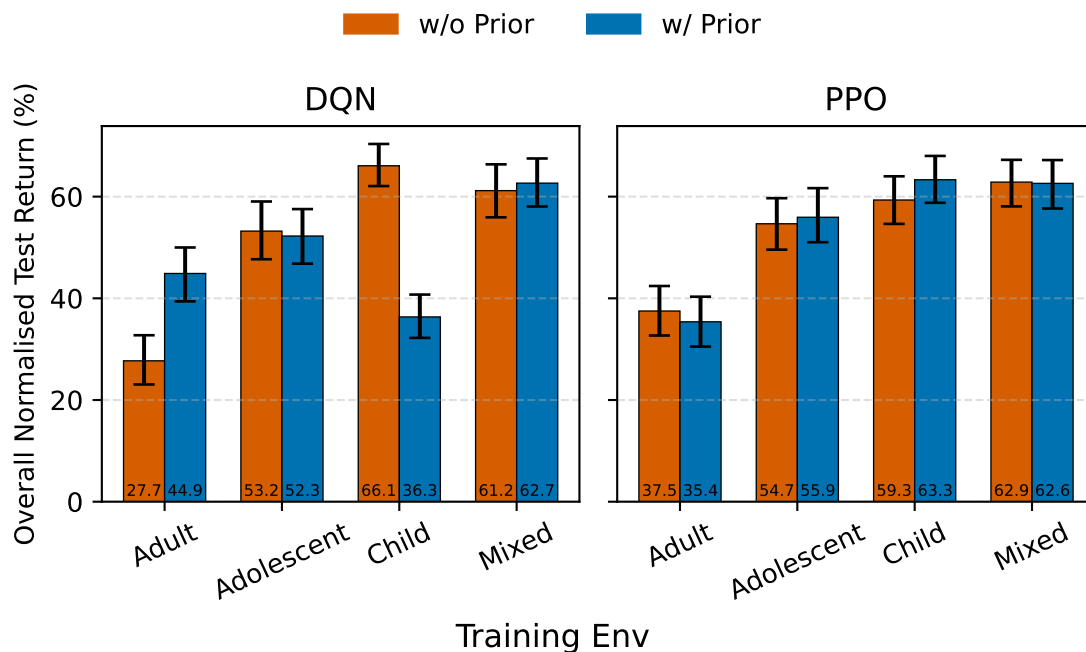


Figure 7.1: Overall normalised return of SRA on the evaluation set for DQN and PPO. Each algorithm is trained on one of four environment types (adult, adolescent, child, or mixed cohort) and evaluated across all environment types. The reported results represent the overall average performance across adult, adolescent and child environments. Evaluation result for each strata is not shown here for clarity, and will be discussed later in Figure 7.2. ‘Prior’ denotes prior knowledge, ‘w/’ means ‘with’ and ‘w/o’ means ‘without’.

Figure 7.1 and Table 7.1 contrast the normalised episodic returns obtained by DQN and PPO trained with and without explicit prior knowledge across four different training environments. The effects of embedding prior knowledge into DQN and PPO appear contradictory. For DQN, the introduction of prior knowledge significantly increases performance in the easiest (adult) environment, improving

| Train Env. | DQN | | PPO | |
|------------|--------------------|--------------|----------------|--------------|
| | w/o → w/ Prior | Δ (%) | w/o → w/ Prior | Δ (%) |
| Adult | 27.7 → 44.9 | 17.2 | 37.5 → 35.4 | -2.1 |
| Adolescent | 53.2 → 52.3 | -0.9 | 54.7 → 55.9 | 1.2 |
| Child | 66.1 → 36.3 | -29.8 | 59.3 → 63.3 | 4.0 |
| Mixed | 61.2 → 62.7 | 1.5 | 62.9 → 62.6 | -0.3 |

Table 7.1: Impact of prior knowledge injection on overall normalised episodic return on the evaluation environments. For clarity, this table only shows the mean value captured from Figure 7.1. Δ represents the performance gain or loss by percentage. ‘Prior’ denotes prior knowledge, ‘w/’ means ‘with’ and ‘w/o’ means ‘without’.

from 27.7% to 44.9%. However, this benefit diminishes sharply with increasing difficulty, becoming negative in the hardest (child) environment, where performance declines markedly from 66.1% to 36.3%. In contrast, PPO demonstrates modest performance improvements upon the introduction of prior knowledge, particularly in harder environments; returns slightly increase from 54.7% to 55.9% in the medium environment and more notably from 59.3% to 63.3% in the child environment. In the mixed environment, neither algorithm shows a statistically significant difference, indicating that embedding explicit knowledge has a negligible marginal impact when training encompasses all levels of difficulty.

Interpreting the aforementioned results, the contradictory behaviour can be explained by the fundamental differences between the learning mechanisms of DQN and PPO. DQN, as an off-policy and value-based algorithm, initially benefits from conservative prior knowledge in easier (such as Adult) environments, since adults’ glucose levels can be well maintained by lower insulin dosage given their stable glucose-insulin dynamics. However, in more challenging environments (such as Children), the same conservative strategy can induce severe hyperglycaemic episodes due to insufficient doses, resulting in substantial penalties. Repetition of these negative experiences rapidly deteriorates the landscape of learnt values, dramatically reducing performance. PPO, as an on-policy actor-critic method, continuously samples trajectories that are aligned with the current policy. Since the conservative prior knowledge injected into PPO is via action scaling, the

prior does not impose strict constraints on policy, but gently biases exploration towards conservative dosing. Consequently, PPO avoids the severe penalties that undermine DQN in more volatile environments. In mixed training environments, prior knowledge injection shows less significant improvement because training already incorporates both slow and fast glucose dynamics, inherently encouraging context-sensitive policy adaptation.

The above findings highlight the algorithm-specific and task-dependent nature of embedding prior knowledge. With extensive experiments on adding or removing prior knowledge, DQN w/o prior training on the child environments stands out as the best SRA due to its superior overall performance on the evaluation environments, outperforming its variants with prior knowledge and all PPO SRAs. Analysis shows that blindly applying prior knowledge to SRAs can lead to counterproductive results and pose challenges when generalised to a wider range of patient groups. These observations naturally lead us to ask: *Can LLMs address these challenges with less manual effort?*

7.5.2 Comparison between LLMs and SRAs

The comparative evaluation between LLMs and the best SRA (i.e., DQN trained on ‘child’ patient environments) reveals several nuanced insights, as shown in Figure 7.2. Without any finetuning, certain LLMs can achieve performance on par with or even surpassing the best-trained SRAs. In terms of overall return, Qwen2.5 models with 7B, 14B, and 32B parameters achieved 62.8%, 62.1%, and 62.3%, respectively, demonstrating the strong zero-shot capabilities of recent LLMs. The advantages of LLMs are primarily driven by superior performance in the adult cohort. In adult evaluation, both the Llama and Qwen2.5 models can overperform the best SRA (78.1%) when the parameter size is greater than 1B. In contrast, across tasks for adolescents and children where physiological dynamics are more volatile, none of the LLM variants outperformed the best-performing SRA.

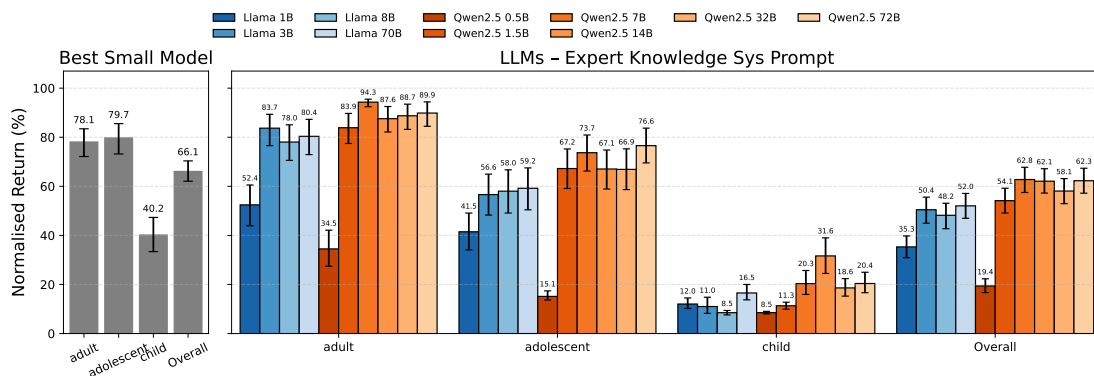


Figure 7.2: LLMs’ normalised return performance compared with the best SRA model (i.e., DQN trained on child environments without prior knowledge) on each sub-evaluation task during testing. The blue colour spectrum represents the Llama3 series and the orange spectrum represents the Qwen2.5 series, respectively. A lighter colour denotes a larger parameter size (in billions).

7.5.3 Analysis of LLM Treatment Behaviour

To better understand the mechanisms behind the clinical decision-making of LLMs, we examine how the model family, the parameter scale, the decoding temperature, and the prompt method influence the insulin dosing behaviour.

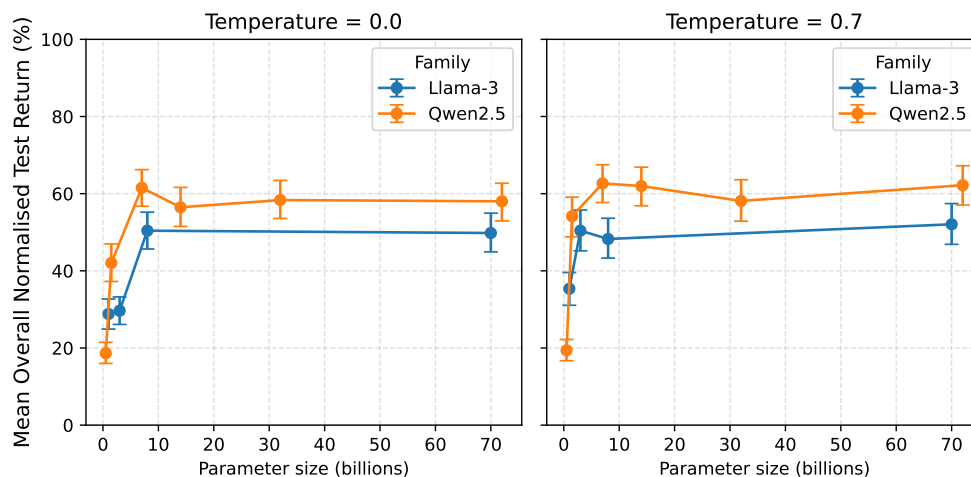


Figure 7.3: Comparison of insulin dosing performance across LLM families (Qwen2.5 and LLaMA3) for different temperature settings.

Impact of Model Family, Sampling Temperature and Parameter Scaling

We evaluated performance variations across different families and sizes of models, presented in Figure 7.3. The Qwen2.5 model family consistently outperforms

the LLaMA3 series across all parameter sizes evaluated, suggesting that intrinsic architectural differences or specific training methodologies significantly influence clinical decision-making capabilities.

We compare deterministic decoding (temperature = 0) with stochastic decoding (temperature = 0.7), where a higher temperature introduces more randomness into the token sampling process. At temperature 0.7, the models consistently produce more diverse responses, yielding superior results compared to the more rigid and repetitive output generated at temperature 0.0. This aligns with observations from other LLM application domains [288].

Clear parameter-scaling trends are observed on both Qwen2.5 and Llama from 1B to 7B, but no further improvement is observed from 7B to 70B. This finding is contradictory to common natural language tasks [289], where the scaling law is typically a log-like curve.

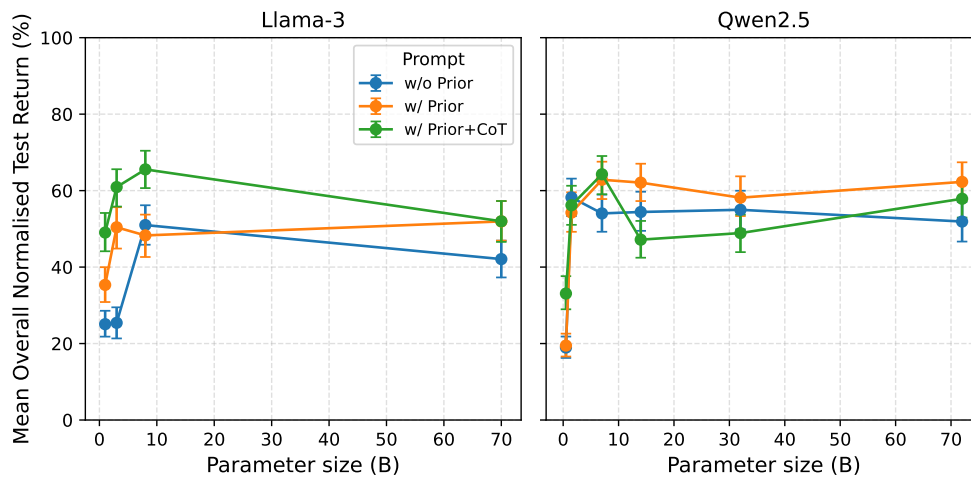


Figure 7.4: Impact of different prompting methods on insulin dosing for Qwen2.5 and LLaMA3 families for prompt settings: base zero-shot prompt (w/o prior), prior knowledge zero-shot prompt (w/ prior) and prior knowledge zero-shot-CoT prompt (w/ prior+CoT).

Influence of Prompting Strategies on Performance To derive a deeper observation of the behaviour of smaller LLMs versus their larger variants, we analyse the scaling trend under various prompt methods, as illustrated in Figure 7.4. Incorporating explicit prior knowledge (i.e., ‘w/ prior’, in orange) into the prompt

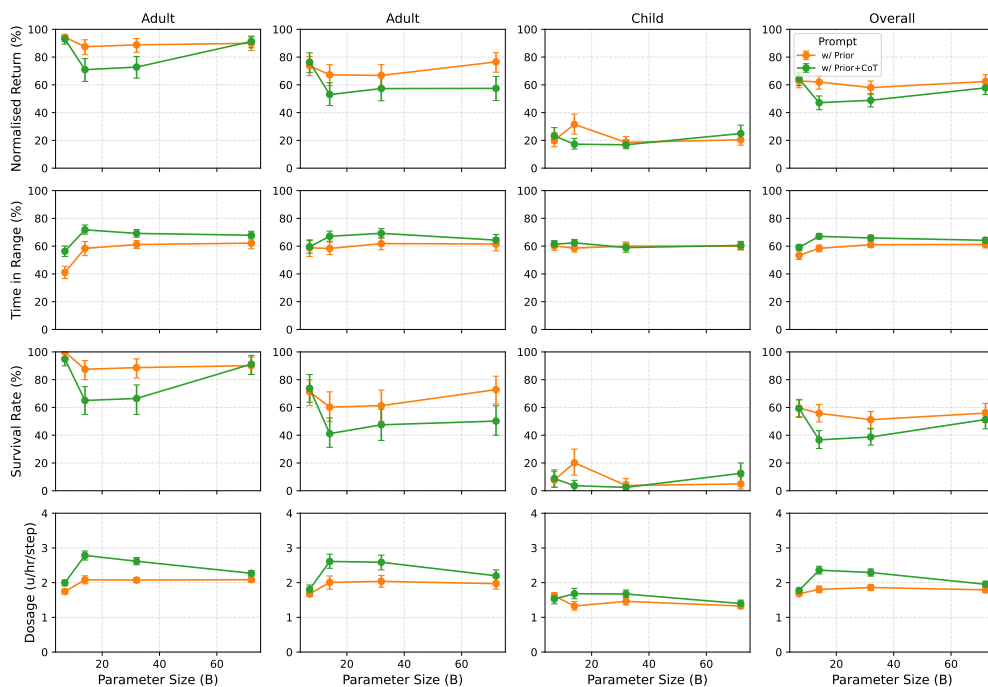


Figure 7.5: Detailed analysis of CoT prompting effects on insulin dosing performance within the Qwen2.5 family. Temperature is set to 0.7. Each column corresponds to the evaluation result for adult (easy), adolescent (medium), child (hard), and overall, respectively. Each row of plots indicates a different evaluation metric, naming normalised return, TIR, survival rate and dosage (See Section 7.3). The x-axis for each subplot indicates the parameter size of the Qwen2.5 model being reported. On the legend, 'w/ prior' means prior knowledge zero-shot prompt and 'w/ prior+CoT' means prior knowledge zero-shot-CoT prompt.

significantly improves the performance of the policy for both model families, with the improvement being particularly significant for Qwen2.5. Interestingly, chain-of-thought (CoT) with prior knowledge (i.e., 'w/ prior+CoT') prompting yields divergent results. For the LLaMA3 family, CoT with prior knowledge generally enhances treatment outcome, while for Qwen2.5, CoT with prior knowledge tends to degrade performance, especially in models exceeding 10B parameters. This unexpected finding challenges conventional wisdom drawn from general-purpose reasoning tasks, where CoT typically improves outcomes.

In-depth Analysis of Chain-of-Thought Prompting Focussing exclusively on the Qwen2.5 family at a fixed decoding temperature of 0.7, we provide a detailed analysis of the nuanced effects of CoT prompting (Figure 7.5). Shown in the last

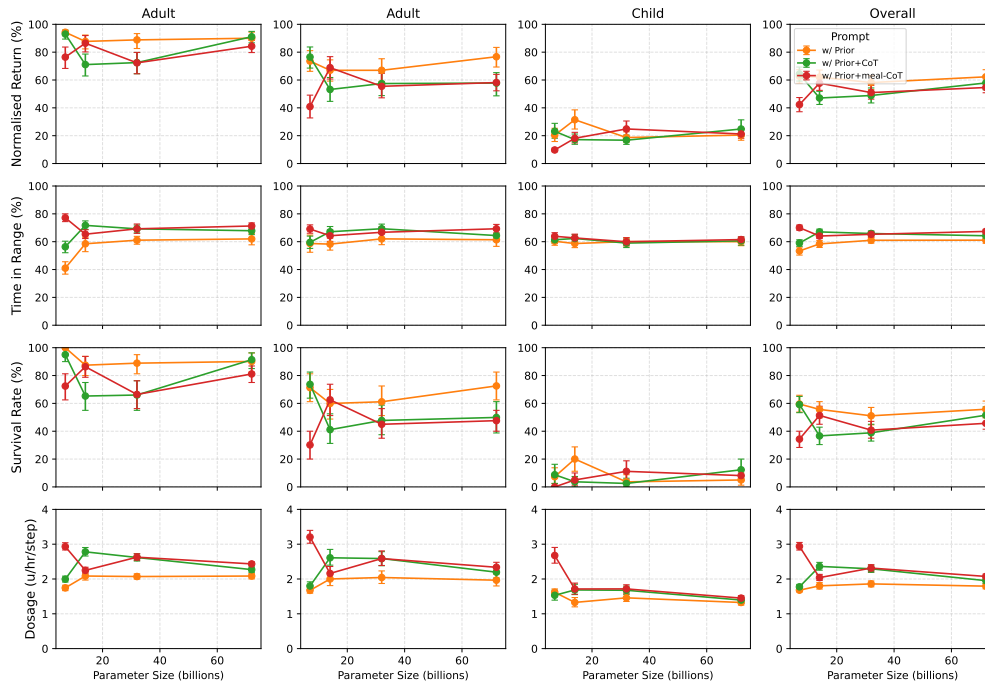


Figure 7.6: Effects of incorporating meal intake information into CoT prompts on insulin dosing performance. Temperature is set to 0.7. Each column corresponds to the evaluation result for adult, adolescent, child, and overall, respectively. Each row of plots indicates a different evaluation metric, naming normalised return, TIR, survival rate and dosage (See Section 7.3). The x-axis for each subplot indicates the parameter size of the Qwen2.5 model being reported. On the legend, 'w/ prior' means prior knowledge zero-shot prompt, 'w/ prior+CoT' means prior knowledge zero-shot-CoT prompt and 'w. Prior+meal-cot' means the zero-shot CoT prompt with prior knowledge and meal information.

row of Figure 7.5, explicit CoT prompts frequently result in overly aggressive insulin dosing strategies. Although this aggressiveness leads to improved time-in-range metrics during periods of stable glucose (second row in Figure 7.5), it concurrently increases severe hypoglycaemic episodes, thus reducing overall survival rates and normalised returns (see the first and third rows in Figure 7.5). Larger models mitigate these adverse effects through improved reasoning capabilities, suggesting that CoT prompts' efficacy critically depends on both task complexity and model scale. When comparing each column, it is shown that LLM tends to recommend less insulin to children than the other two strata.

Evaluating the Impact of Incorporating Meal Information. To explore further improvements in clinical reasoning, we assess the inclusion of meal intake information in CoT prompting (Figure 7.6). First, enriching the prompt with explicit meal-related reasoning (“Prior + meal-CoT”, red curves) consistently harms, rather than helps, the two smaller models (7 B and 13 B). Normalised return, time-in-range, and survival all decline relative to the simpler prior knowledge zero-shot baseline, while insulin dosage increases, indicating that the added reasoning chain is misinterpreted and translated into unnecessarily aggressive interventions. Secondly, the largest model (70 B) is able to absorb most of this complexity penalty. Its red curve returns to parity with the green (“w/ prior+CoT”) and orange baselines, but still fails to produce a systematic gain. In all capacities, there is no metric for which meal-based prompting delivers a statistically significant improvement.

Overall, the analysis indicates that while larger LLMs can ‘undo’ the performance drop caused by the specified meal reasoning prompt, they do not convert the hint of hidden variables into better glycaemic control. Therefore, latent-state inference of meal effects remains an unsolved challenge, and future work will need to look beyond prompt engineering to realise further clinical gains.

Behavioural Analysis of Model Responses We extend our analysis to behavioural patterns resulting from CoT prompting in the 7B (see Figure 7.7 and Figure 7.8) and 72B (see Figure 7.9 and Figure 7.10) parameter models. In the smaller 7B model, CoT prompts produce more reactive dosing strategies that dynamically adjust insulin rates according to glucose levels, resulting in improved short-term glucose control. However, these reactive policies also frequently lead to cumulative insulin overdoses, increasing hypoglycemic risk and reducing overall clinical outcomes. Specifically, the 7B model demonstrates an overly aggressive response to transient hyperglycaemic episodes, resulting in rapid insulin escalation and subsequent hypoglycaemic events. This behaviour indicates a limited ability to accurately estimate long-term insulin effects and highlights sensitivity to the instructional clarity of the prompt structure. In contrast, the larger 72B

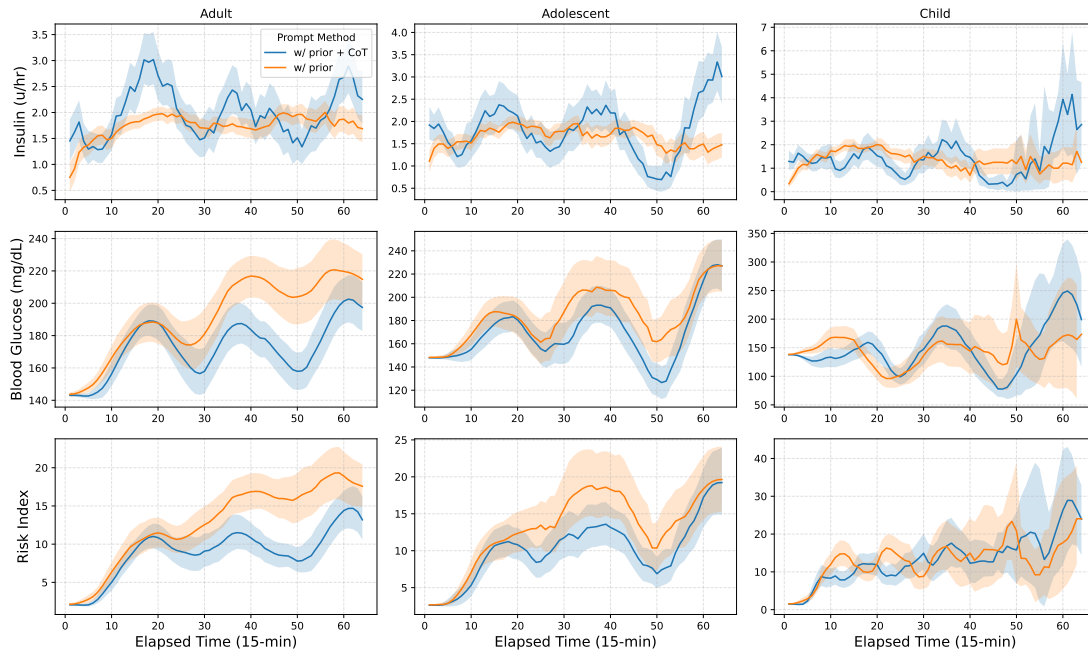


Figure 7.7: Insulin administration, glucose and risk behaviours for Qwen2.5-7B model under CoT and zero-shot prompting with prior knowledge system prompt. The shaded areas mean the 95% confidence interval derived by bootstrapping. The x-axis is the elapsed time step. Each column corresponds to the evaluation result for adult, adolescent and child, respectively. The shaded areas mean the 95% confidence interval derived by bootstrapping.

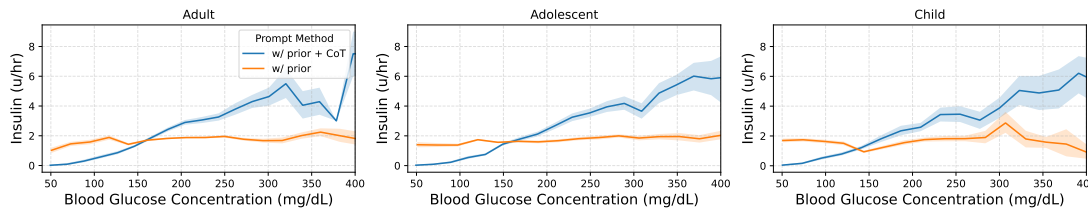


Figure 7.8: Insulin administration as a function of current observed glucose measurement for Qwen2.5-7B model under CoT and zero-shot prompting with prior knowledge system prompt. Notation is the same as Figure 7.7.

model inherently recognises glucose-dose relationships even in zero-shot scenarios, suggesting an intrinsic understanding of physiological insulin dynamics. The introduction of CoT further reinforces this understanding, allowing for a nuanced dosage adjustment more closely aligned with clinical guidelines. Critically, the increased complexity of CoT does not significantly increase hypoglycemic risks, demonstrating a robust enough model capacity to effectively balance aggressive glucose control against safety constraints.

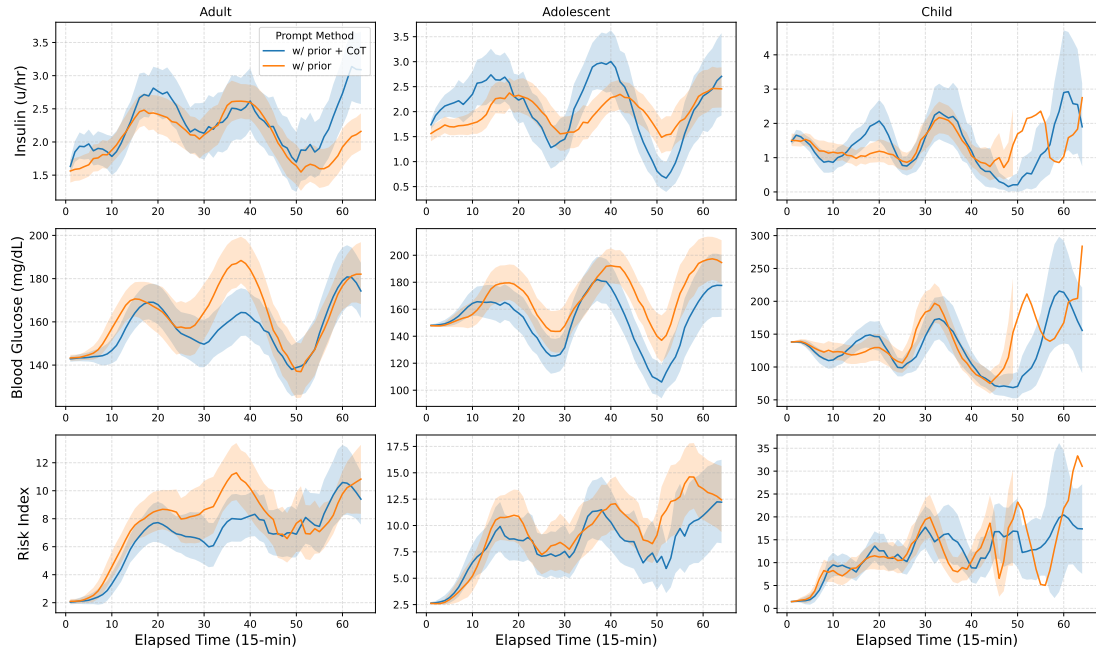


Figure 7.9: Insulin administration, glucose and risk behaviours for Qwen2.5-72B model under CoT and zero-shot prompting with prior knowledge system prompt. The shaded areas mean the 95% confidence interval derived by bootstrapping. The x-axis is the elapsed time step. Each column corresponds to the evaluation result for adult, adolescent and child, respectively. The shaded areas mean the 95% confidence interval derived by bootstrapping. When the shaded area does not show, it means the sample size is too small to present the confidence interval. The mean is used without a confidence interval in such cases.

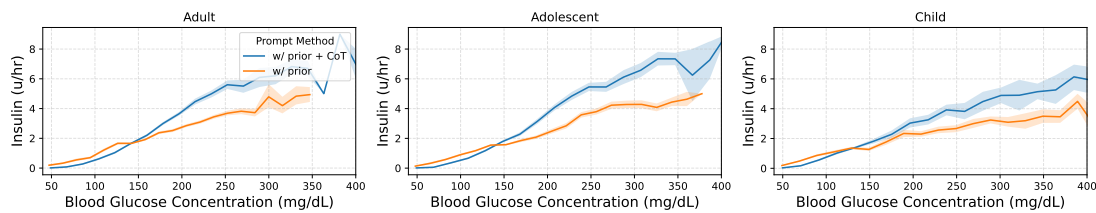


Figure 7.10: Insulin administration as a function of current observed glucose measurement for Qwen2.5-72B model under CoT and zero-shot prompting with prior knowledge system prompt. Notation is the same as Figure 7.7

7.6 Failure Modes of Chain-of-Thought Reasoning

In this section, we analyse common failure modes of CoT reasoning when applied to insulin dosing in paediatric patients with Type 1 Diabetes. We focus on two illustrative case studies and supplement them with broader behavioural trends observed during the evaluation. The outputs of the Qwen2.5-7B and Qwen2.5-72B models are compared with the same prior knowledge and meal-CoT prompt.

7.6.1 Case Study 1: Reasoning for Initial Dosage

| Patient Info | | Prior Knowledge System Prompt |
|---|----------------|---|
| Carbohydrate Factor (CF) : 42.72 | Patient: Child | You are a clinical specialist responsible for managing patients with Type-1 Diabetes. Your primary objective is to maintain the patient's blood glucose levels within the safe range of 70-140 mg/dL by administering appropriate insulin doses. |
| Correction Ratio (CR) : 25 | Age : 9 | |
| Total Daily Insulin (TDI): 17.47 | Step : 0 | |
| Observation | | |
| <p>## Observations You are treating a 9 year old patient. Day 1, Time: 05:00:00 (initial measurement), glucose: 164.02 mg/dL, insulin rate: 0.0000 unit/hour, insulin dose: 0.00 unit.</p> <p>## Request Determine the optimal insulin rate for the current 15-minute interval to maintain a patient's blood glucose levels within the safe range of 70-140 mg/dL. First, analyse the current state step-by-step by estimating Correction Factor(CF) and Total Daily Insulin(TDI). If the patient has taken any meal, you must consider the meal effect. Finally, you must choose a dosage value enclosed in answer tags (i.e., <code><ans></code> and <code></ans></code>), for example, <code><ans>0</ans></code>, without any non-numerical word. Let's think step by step.</p> | | |
| Qwen2.5-7B-Instruct | | Qwen2.5-72B-Instruct |
| <p>## Step 1: Determine the Correction Factor (CF) The Correction Factor (CF) is the amount of insulin required to lower blood glucose levels by 100 mg/dL. It is usually estimated based on the patient's insulin sensitivity. Since the patient's age is 9 years old, we can assume a typical correction factor for this age group. A common assumption is that a child's CF is around 0.5 units per 100 mg/dL.</p> <p>## Step 2: Estimate the Total Daily Insulin (TDI) The Total Daily Insulin (TDI) is the amount of insulin the patient needs to maintain their blood glucose levels within the target range. It is usually estimated based on the patient's age, weight, and activity level. For a 9-year-old child, a common TDI range is around 0.5 to 1.0 unit per kg of body weight per day. Assuming an average weight of 30 kg for a 9-year-old child, the TDI would be around 15 to 30 units per day.</p> <p>## Step 3: Analyze the Current State Given the initial glucose level of 164.02 mg/dL, which is above the target range of 70-140 mg/dL, we need to determine the insulin dose required to bring the glucose level down to within the target range.</p> <p>## Step 4: Calculate the Insulin Dose Using the Correction Factor (CF) and the estimated TDI, we can calculate the insulin dose required to lower the glucose level to within the target range. Insulin Dose = (Glucose level - Target Range) / Correction Factor Since the target range is 70-140 mg/dL, we can choose a target range of 80-120 mg/dL for simplicity. Insulin Dose = (164.02 - 80) / 0.5 Insulin Dose = 164.02 / 0.5 Insulin Dose = 32.804 units</p> <p>## Step 5: Choose the Optimal Insulin Rate Considering the insulin dose calculated above, we need to choose an optimal insulin rate for the current 15-minute interval. Since the patient's insulin rate is currently 0.0000 unit/hour, we can assume that the insulin dose will be administered over the next 15 minutes. To convert the insulin dose to an insulin rate, we can use the following formula: Insulin Rate = Insulin Dose / 15 minutes Insulin Rate = 32.804 units / 15 minutes Insulin Rate = 2.187 units/minute Since the rate range is [0 to 9] units/hour, we need to convert the insulin rate to units/hour: Insulin Rate = 2.187 units/minute * 60 minutes/hour Insulin Rate = 131.22 units/hour However, this value exceeds the upper limit of the rate range. Therefore, we need to choose a lower value within the rate range. Considering the patient's age and the insulin dose calculated above, a reasonable insulin rate for the current 15-minute interval could be: <ans>5</ans></p> | | <p>Day 1, Time: 05:00:00 (initial measurement), glucose: 159.86 mg/dL, insulin rate: 0.0000 unit/hour, insulin dose: 0.00 unit.</p> <p>## Step-By-Step Analysis</p> <ol style="list-style-type: none"> Initial Glucose Level and Time of Day** **Glucose level**: 159.86 mg/dL (slightly above the target range of 70-140 mg/dL). **Time**: 05:00:00 (early morning). Correction Factor (CF)** The Correction Factor (CF) is the amount of glucose (in mg/dL) that 1 unit of insulin will lower the blood glucose level. For a 9-year-old, a common CF is around 50 mg/dL per unit of insulin. **CF**: 50 mg/dL/unit. Total Daily Insulin (TDI)** The Total Daily Insulin (TDI) is the total amount of insulin the patient requires in a day. For a 9-year-old, a typical TDI might be around 0.5 to 1.0 units/kg/day. Assuming the patient weighs around 30 kg, the TDI would be: **TDI**: 0.5 * 30 = 15 units/day to 1.0 * 30 = 30 units/day. Current Glucose Level and Target** **Current Glucose**: 159.86 mg/dL. **Target Range**: 70-140 mg/dL. **Desired Glucose Level**: Midpoint of the target range, approximately 105 mg/dL. **Correction Needed**: 159.86 - 105 = 54.86 mg/dL. **Insulin Dose for Correction**: 54.86 / 50 = 1.0972 units. Insulin Rate for 15 Minutes** The insulin dose needs to be administered over the next 15 minutes. The rate is calculated as: **Rate**: 1.0972 units / (15/60) hours = 4.39 units/hour. Consideration of Time of Day and Potential Food Intake** **Time of Day**: Early morning (05:00:00). The patient is likely not eating at this time, so no meal effect is expected. **Insulin Stacking**: Since the patient has not received any insulin yet, there is no insulin stacking to consider. Final Decision** Given the slight elevation in glucose levels and the lack of immediate food intake, a moderate insulin rate is appropriate to bring the glucose levels back into the target range without causing a rapid drop. **Final Insulin Rate** <ans>4</ans> |

Figure 7.11: Comparing Qwen2.5 7B and 72B model response on a child patient on step 0 under meal CoT prompt with prior knowledge. The system prompt and observation are identical, and responses are shown side by side.

We first examine a controlled single-step dosing scenario to evaluate how language models respond to an initial hyperglycaemic state with minimal context (see Figure 7.11). The virtual patient is a nine-year-old child whose initial glucose level at 05:00 is 164.02 mg/dL. Insulin has not been administered before this point, and no meal intake is recorded. The optimal strategy, derived from the ground truth of the simulator, involves calculating a correction dose based on an estimated correction factor of 25 mg/dL per unit and a total daily insulin of 17.47 units. This corresponds to a correction dose of approximately 2.36 units.

The Qwen2.5-7B model exhibits a clear reasoning failure. It begins by estimating an implausible CF of 0.5 units per 100 mg/dL and proceeds to subtract this from a target glucose level of 80 mg/dL. This leads to a highly exaggerated correction dose that exceeds 30 units. The model then calculates a per-hour dose exceeding 130 units/hour, which it subsequently clips to 5 units/hour in the final answer. Although the final output appears clinically plausible, it is numerically disconnected from the preceding reasoning trace. The model does not validate its recommendation against its own TDI estimate, nor does it apply any internal consistency constraints that would be expected in a clinical dose calculation. This illustrates a form of arithmetic hallucination, in which each individual step may appear locally coherent, but the overall logic is conceptually invalid and yields dangerous recommendations.

In contrast, the Qwen2.5-72B model produces a more restrained response. It estimates a CF of 50 mg/dL per unit, computes a correction dose of approximately 1.1 units, and derives an insulin rate of 4.39 units/hour, which is then rounded to 4.0. This outcome is physiologically reasonable and demonstrates a basic alignment with the underlying insulin-glucose dynamics. However, the model still underdoses the patient relative to the optimal dose derived from the simulator, reflecting a conservative bias when uncertainty exists. Although the internal calculations are more plausible than those of the smaller model, the final recommendation remains somewhat decoupled from the model's own TDI estimate and lacks any mechanism for deferred or conditional action.

7.6.2 Case Study 2: Reasoning on Glucose Trends and Delayed Effects

The second scenario evaluates model behaviour in a temporally extended setting involving multiple insulin interventions, a recorded meal, and a non-linear glucose trend. The virtual patient is again the nine-year-old child reported in Section 7.6.1. A meal event is recorded at 05:30, and insulin doses are administered at varying rates between 0.0 and 9.0 units/hour over the following six time steps.

| Patient Info | | Prior Knowledge System Prompt |
|--|---------------------------------------|--|
| Carbohydrate Factor (CF) : 42.72 Correction Ratio (CR) : 17.45 Total Daily Insulin (TDI) : 25 | Patient: Child Age : 9 Step : 7 | You are a clinical specialist responsible for managing patients with Type-1 Diabetes. Your primary objective is to maintain the patient's blood glucose levels within the safe range of 70-140 mg/dL by administering appropriate insulin doses. |
| <pre> ### Observations You are treating a 9-year-old patient. Day 1, Time: 05:00:00 (initial measurement), glucose: 164.02 mg/dL, insulin rate: 0.0000 unit/hour, insulin dose: 0.00 unit. Day 1, Time: 05:15:00, glucose: 163.42 mg/dL, insulin rate: 0.0000 unit/hour, insulin dose: 0.75 unit. Day 1, Time: 05:30:00, glucose: 161.53 mg/dL, insulin rate: 0.0000 unit/hour, insulin dose: 2.25 unit, the patient eats. Day 1, Time: 05:45:00, glucose: 183.25 mg/dL, insulin rate: 0.0000 unit/hour, insulin dose: 0.25 unit. Day 1, Time: 05:00:00, glucose: 180.17 mg/dL, insulin rate: 0.0000 unit/hour, insulin dose: 0.00 unit. Day 1, Time: 06:15:00, glucose: 157.88 mg/dL, insulin rate: 0.0000 unit/hour, insulin dose: 0.00 unit. Day 1, Time: 06:30:00, glucose: 126.18 mg/dL, insulin rate: 0.3000 unit/hour, insulin dose: 0.00 unit. Day 1, Time: 06:45:00, glucose: 96.84 mg/dL, insulin rate: 0.0000 unit/hour, insulin dose: 0.00 unit. ### Request Determine the optimal insulin rate for the current 15-minute interval to maintain a patient's blood glucose levels within the safe range of 70-140 mg/dL. First, analyse the current state step-by-step by estimating Correction Factor(CF) and Total Daily Insulin(TDI). If the patient has taken any meal, you must consider the meal effect. Finally, you must choose a dosage value enclosed in answer tags (i.e., <ans> and </ans>), for example, <ans>0</ans>, without any non-numerical words, let's think step by step. ### Answer </pre> | | |
| Qwen2.5-7B-Instruct | | Qwen2.5-72B-Instruct |
| <pre> ### Analysis of the Current State #### 1. **Estimate Correction Factor (CF) and Total Daily Insulin (TDI)** - **Correction Factor (CF)**: $ACF = 100 / \text{Glucose level in mg/dL}$ - **Total Daily Insulin (TDI)**: $TDI = CF * \text{Daily insulin requirement}$ #### 2. **Estimate Daily Insulin Requirement** - **Insulin Sensitivity**: $ACF = 100 / \text{Glucose level in mg/dL}$ - **Daily Insulin Requirement**: $\text{Daily insulin requirement} = ACF * \text{Target glucose level}$ - **Target glucose level**: 120 mg/dL #### 3. **Estimate TDI** - **Target glucose level**: 120 mg/dL - **ACF**: $ACF = 100 / 120 = 0.83$ - **Daily insulin requirement**: $\text{Daily insulin requirement} = 0.83 * 120 = 99.6$ units - **TDI**: $TDI = 99.6$ units #### 4. **Estimate CF** - **ACF**: $ACF = 100 / 120 = 0.83$ - **CF**: $CF = 0.83$ #### 5. **Estimate Insulin Rate** - **Insulin rate**: $\text{Insulin rate} = CF * TDI / 24$ - **Insulin rate**: $\text{Insulin rate} = 0.83 * 99.6 / 24 = 3.34$ units/hour #### 6. **Estimate Meal Effect** - **Meal effect**: The patient eats at 05:30:00, so the meal effect is considered. - **Meal effect**: The meal effect is estimated to be 0.5 units/hour. #### 7. **Determine Optimal Insulin Rate** - **Optimal insulin rate**: $\text{Optimal insulin rate} = \text{Insulin rate} - \text{Meal effect}$ - **Optimal insulin rate**: $\text{Optimal insulin rate} = 3.34 - 0.5 = 2.84$ units/hour ### Optimal Insulin Rate The optimal insulin rate for the current 15-minute interval is 2.84 units/hour. </pre> | | <pre> ### Step-by-Step Analysis 1. **Current Glucose Level and Trend** - **Current Glucose**: 124.84 mg/dL - **Previous Glucose**: 112.03 mg/dL (15 minutes ago) - **Trend**: Increase of 12.81 mg/dL over 15 minutes 2. **Insulin Administration** - **Current Insulin Rate**: 2.5 units/hour - **Previous Insulin Rate**: 2.0 units/hour - **Insulin Dose**: 0.42 units over the last 15 minutes 3. **Meal Effect** - The patient ate at 06:15:00, which likely contributed to the rise in glucose levels. 4. **Correction Factor (CF)** - CF is the amount of glucose (in mg/dL) that 1 unit of insulin typically lowers. It can be estimated using the 1800 rule for Type 1 Diabetes: $\text{CF} = \frac{1800}{\text{Total Daily Insulin (TDI)}}$ Since we don't have the TDI, we can estimate it based on the current insulin rate and the observed glucose trend. The patient is currently on 2.5 units/hour, which suggests a moderate insulin requirement. $\text{TDI} = 30 \text{ (units/kg/day)} * \text{times } 0.75 \text{ (units/kg/day)} = 22.5 \text{ (units/day)}$ - A rough estimate for a 9-year-old might be around 0.5 to 1.0 units/kg/day. Assuming an average weight of 30 kg, TDI could be: $\text{TDI} = \frac{1800}{22.5} \approx 80 \text{ (mg/dL per unit)}$ 5. **Total Daily Insulin (TDI)** - **Target Glucose**: 120 mg/dL (midpoint of the safe range) - **Current Glucose**: 124.84 mg/dL - **Difference**: $124.84 - 120 = 4.84$ mg/dL - **Insulin Needed to Correct**: $\text{Insulin Needed} = \frac{\text{Difference} (mg/dL)}{\text{CF} (mg/dL/unit)} \approx 0.86 \text{ (units)}$ - **Adjustment**: Since the glucose is slightly above the target and the trend is increasing, we can slightly increase the insulin rate to counteract the meal effect and bring the glucose back to the target range. 7. **Final Decision** - **New Insulin Rate**: 3.0 units/hour (a moderate increase to counteract the meal effect and bring glucose back to target) ### Answer <ans>3.0</ans> </pre> |

Figure 7.12: Comparing Qwen2.5 7B and 72B model response on a child patient on step 7 under meal CoT prompt with prior knowledge. The system prompt and observation are identical, and responses are shown side by side. Texts highlighted in yellow denote wrong estimations.

The Qwen2.5-7B model fails to produce a coherent or physiologically plausible recommendation. It estimates the TDI using an incorrect and circular formula in which it multiplies a target glucose level by a fraction involving the same target value. This yields an inflated TDI of 99.6 units per day, approximately six times the true requirement for the patient. Based on this estimate, it derives a CF of 0.83 mg/dL per unit and calculates a correction dose of 3.34 units. Despite acknowledging the presence of a residual meal effect, the model merely subtracts a fixed offset of 0.5 units/hour and recommends a final insulin rate of 2.84 units/hour. This recommendation is aggressive and poorly justified, especially since the glucose value is within the safe range and trending downward. The model does not refer to or incorporate prior insulin administration, nor does it reason about possible insulin stacking or delayed pharmacodynamic effects.

The Qwen2.5-72B model performs better in terms of internal numerical consistency, but still fails to apply clinically appropriate decision logic. It estimates a TDI of 22.5 units and a CF of approximately 80 mg/dL per unit. It recognises that the glucose level has decreased over time after a meal, and it acknowledges that the

current glucose level lies within the safe range. However, the model recommends increasing the insulin rate from 2.5 to 3.0 units/hour. This decision is based on a short-term increase in glucose between two adjacent time points, but it does not take into account the broader temporal context. Specifically, the recent increase in glucose is a transient fluctuation following a meal and is likely to self-correct due to existing insulin exposure. The model does not explicitly reason about cumulative insulin effects or the appropriate time lag between administration and glucose response.

7.6.3 Summary of Findings on Case Studies

Our case studies suggest that while CoT enables LLMs to simulate structured clinical reasoning, it also introduces characteristic errors. These errors often emerge not from model incapacity alone, but from the way CoT prompts compel deterministic, single-step answers in a domain that requires cautious, temporally-aware reasoning. To organise these insights, we group the observed failures into the following categories:

- **Arithmetic Hallucination** The model performs invalid or circular arithmetic operations, often producing superficially coherent but numerically incorrect results. These errors typically arise from misapplied formulas or misinterpreted clinical heuristics during chain-of-thought reasoning.
- **Lack of Temporal Abstraction** The model fails to incorporate relevant treatment and meal history, such as insulin stacking or the delayed pharmacodynamic effects of meals and insulin. It tends to overreact to short-term glucose fluctuations while ignoring broader temporal trends, resulting in reactive rather than anticipatory control. This limitation reflects the absence of multi-step planning or memory in the decision-making process.
- **Aggressive Dosing Under Uncertainty** In early or ambiguous states, where glucose trends have not yet stabilised, the model frequently recommends excessive insulin doses. This behaviour increases the risk of hypoglycaemia and contrasts with human clinical intuition, which generally favours cautious observation before intervention under uncertainty.

- **Inconsistent Internal Reasoning** The model’s intermediate reasoning steps often contain internal contradictions or fail to constrain the final output. For example, estimated total daily insulin requirements may not align with the recommended dose, suggesting a lack of internal logical coherence across the reasoning trace.

7.7 Discussion

This chapter provides an exploratory but rigorous evaluation of LLMs as DTR agents for insulin administration in Type 1 diabetes within an *in silico* environment. Our findings provide several insights critical to assessing the potential of LLMs in clinical decision-making. Firstly, our experiments demonstrate that certain smaller-scale LLMs, notably the Qwen2.5-7B model, can achieve performance comparable to extensively trained RL agents under zero-shot inference. This highlights a remarkable efficiency advantage of LLMs, as they do not require interaction-intensive, environment-specific training, positioning them as promising candidates for rapid deployment in clinical decision-support scenarios. Interestingly, our results do not confirm a consistent parameter-scaling law in clinical DTR tasks despite the expectations of the broader literature on LLMs. Larger-scale models, such as the Qwen2.5 7B, did not produce significant performance improvements over their smaller counterparts. Specifically, models from the Qwen2.5 family consistently outperformed those from the LLaMA3 family, indicating that the pre-training quality of foundation models has a substantial influence on clinical decision-making capabilities.

Our investigation of prior knowledge injection strategies revealed contrasting outcomes depending on the underlying RL algorithm. For off-policy algorithms (DQN), embedding prior knowledge significantly improved performance in stable environments but negatively impacted volatile conditions, illustrating the context-sensitive nature of knowledge incorporation. In contrast, PPO, an on-policy algorithm, benefited modestly but consistently across environments from implicit expert-guided exploration. These results highlight the complexity of effectively

embedding clinical heuristics in RL agents, reinforcing the potential advantage of LLMs, which naturally incorporate prior knowledge through intuitive natural language prompting.

A particularly unexpected finding refers to the effectiveness of the Chain-of-Thought prompting method. Unlike general-purpose tasks, where CoT consistently improves reasoning performance, in insulin dosing scenarios, CoT prompts led to mixed outcomes. For smaller models, CoT induced an overly aggressive insulin dosing, which improved short-term glucose control at the expense of increased hypoglycemic episodes, thus reducing overall clinical efficacy. Larger models mitigated but did not eliminate these adverse effects, indicating that CoT prompts, while beneficial for human readability, must be carefully tuned to clinical contexts to avoid unintended risks.

Efforts to explicitly incorporate latent variables such as meal intake into LLM reasoning yielded minimal additional benefit, suggesting a current limitation in LLMs' ability to infer complex, partially observable physiological dynamics purely through text-based prompts. This finding highlights a critical area for future research: the potential integration of explicit physiological models or structured hybrid prompting techniques to further enhance the interpretative capabilities of LLMs.

Finally, our detailed case study of failure modes in CoT reasoning emphasises the risks associated with relying solely on linguistic coherence and structured logic. Despite producing fluent and syntactically well-formed reasoning steps, LLMs often exhibited fundamental clinical errors, such as misinterpreting glucose trends, overestimating insulin requirements, or failing to account for delayed pharmacodynamic effects. The contrast between LLM-generated decisions and standard clinical practice, where deferred intervention is often a safer and more appropriate choice, highlights a critical gap in the temporal and probabilistic reasoning capabilities of current LLMs. This limitation underscores the need for further research into prompt designs that support multi-step planning, uncertainty awareness, and dynamic policy modulation based on evolving patient states.

In summary, our investigation highlights the substantial potential of LLM as interpretable, efficient, and scalable agents for clinical decision support in dynamic treatment scenarios. However, realising this potential requires meticulous validation, context-sensitive prompt engineering, and potentially targeted fine-tuning. These findings advocate for a cautious, yet optimistic outlook on integrating LLMs into clinical workflows, highlighting the need for rigorous empirical evaluations to ensure safety and efficacy in practical healthcare settings.

8

Conclusion

This thesis aimed to answer a central question: Can reinforcement learning and other AI methods be used to develop clinically safe, personalised, and generalisable dynamic treatment regimes? Although the conceptual promise of DTRs is long recognised, their practical implementation remains limited by key barriers, most notably, challenges in evaluation, the fragility of learnt policies in real-world data, and the need for interpretable, scalable, and equitable solutions. To address these issues, this thesis followed a four-part investigation: (1) building clinically grounded simulations for controlled benchmarking; (2) re-evaluating offline RL methods in sepsis care with a focus on reward design and evaluation variance; (3) applying RL to a high-stakes real-world clinical task of insulin dosing; and (4) exploring LLMs as an alternative, minimalist approach for treatment planning. Together, these studies offer methodological insights, practical tools, and a critical framework for advancing the next generation of AI-driven decision support in healthcare.

8.1 Summary of Results

8.1.1 An *in silico* Environment and Benchmark Platform for Reinforcement Learning Based Dynamic Treatment Regime

Motivated by the limitations of real-world clinical data, this thesis began with the development of a simulation benchmark to rigorously test RL algorithms under controlled but clinically meaningful conditions. *DTR-Bench* is introduced as a

modular platform that enables systematic manipulation of key environmental factors, including stochastic dynamics, missingness, and pharmacokinetic/pharmacodynamic variability. The goal is to evaluate whether the RL methods could maintain stable performance across increasingly complex and realistic decision-making scenarios.

The benchmark experiments revealed a critical insight. Algorithm performance rankings are highly sensitive to the underlying structure and variability of the environment. In particular, no single RL method consistently outperformed others across all variants. This result challenges the assumption that models evaluated under idealised or stationary conditions can be reliably translated into clinical settings. One notable finding is that recurrent architectures, which are often preferred for sequential modelling, did not offer consistent advantages, especially in environments where the transition dynamics adhered strictly to a first-order Markov process (i.e., the *OberstSepsis* environment). This highlights the importance of aligning model architecture with environment structure, rather than assuming temporal complexity by default. The aforementioned findings highlight the need to evaluate RL agents not only by their average performance in simulations, but also by their robustness across a range of clinically relevant perturbations. Without such stress-testing, there is a risk of overfitting to oversimplified training environments and drawing misleading conclusions about model reliability. *DTR-Bench* addresses this gap by serving as both a training ground and an adversarial evaluation platform, enabling researchers to better understand the failure modes of RL agents under realistic assumptions. As an open-source and extensible tool, it provides a foundation for reproducible and principled benchmarking in future work on AI-driven treatment strategies.

8.1.2 A Critical Re-examination of Reinforcement Learning-Based Dynamic Treatment Regime

Following the simulation-based analysis, the next phase of this thesis focused on whether offline RL can produce clinically meaningful treatment strategies on real-world data. Although the literature has reported promising results in this area,

closer inspection reveals that many existing studies rely on limited evaluation setups, inconsistent baselines, and unvalidated estimators. These practices raise questions about the robustness, reproducibility, and clinical credibility of offline RL in practice. To address this, a comprehensive re-evaluation is conducted using a unified pipeline on the MIMIC-III sepsis cohort. The study introduced consistent reward definitions, incorporated strong supervised learning baselines, and applied multiple OPE estimators, focusing on variance and calibration. This enabled a more controlled comparison across models, reducing the confusion caused by implementation differences.

The key finding shows that policy rankings are sensitive to the formulation of rewards, behavioural policy estimation, and subgroup stratification. In many settings, supervised learning models, despite lacking return optimisation, matched or outperformed RL agents in some estimation metrics. These results do not invalidate the potential of RL, but reveal how fragile apparent gains can be when the evaluation is noisy or misaligned with clinical priorities. Furthermore, stratified analyses reveal that model performance can vary significantly across patient subgroups, underscoring the need for equity-aware evaluation in AI for healthcare. This study also highlights the importance of methodological discipline in offline RL for dynamic treatment regimes. Without standardised evaluation protocols, calibrated behaviour models, and meaningful baselines, it is difficult to draw trustworthy conclusions. The work concludes with a set of practical recommendations for future studies, including the routine use of stratified baselines, sensitivity checks for reward and estimator design, and transparent reporting of evaluation variance. These practices are essential for transitioning from proof-of-concept RL applications to clinically credible systems.

8.1.3 Towards Safer Insulin Dosing in the ICU: Hypoglycaemia Prediction and Offline Insulin Optimisation

The third study focused on a high-stakes clinical problem, hypoglycaemia prediction and intravenous insulin dosing in the ICU. Managing glucose levels in critically ill patients is complex and time-sensitive, with both under- and over-treatment

posing significant risks. This task provided an ideal testbed to assess whether RL, combined with high-frequency observational data, could support safer and more personalised insulin dosing decisions. To enable this, a new dataset is curated from the MIMIC-III database, capturing dense physiological measurements, insulin administration records, and clinical context for a large and diverse cohort of ICU patients. The study developed supervised and reinforcement learning approaches to predict both hypoglycaemic risk and optimise insulin dosing offline, using a new proposed OPE ensemble.

Key findings suggest that, under well-curated pipelines and conservative evaluation frameworks, offline RL agents can approximate retrospective clinician policies in terms of glycaemia safety. Beyond performance, the study highlights several methodological insights: **(1)** rich temporal modelling improves risk prediction; **(2)** prior treatment history is essential for reliable policy learning; and **(3)** conservative evaluation methods, such as ensemble bootstrapping, are critical for safe model evaluation.

8.1.4 Evaluation of Large Language Models in Dynamic Treatment Regimes

The final study in this thesis revisits a long-standing challenge in dynamic treatment planning: *how to incorporate clinical knowledge into reinforcement learning agents*. Initial attempts to guide small RL models using expert priors showed limited success. Performance gains are highly sensitive to environmental dynamics and task volatility. This motivated us to think whether large language models can serve as treatment planners by leveraging their internalised clinical knowledge through natural language prompts. To evaluate this, we tested several LLMs in a zero-shot insulin control task using the SimGlucose environment for Type 1 diabetes. Surprisingly, the Qwen2.5-7B model performed comparably to the best-trained small RL agent, despite not requiring task-specific fine-tuning. The results also revealed critical limitations in LLMs. Despite producing fluent and well-structured natural language output, LLMs often fail to provide reliable clinical reasoning. We used two case

studies to demonstrate systematic errors in Chain-of-thought, including overdosing, unit confusion, and unsafe timing of insulin delivery. These errors highlight the gap between verbal coherence and clinically sound decision-making.

In summary, this study suggests that LLMs offer a promising new paradigm for dynamic treatment planning. Their zero-shot ability to encode and apply prior knowledge through natural language makes them an attractive option in low-resource or rapidly evolving scenarios. However, the current limitations of LLM in temporal reasoning, uncertainty awareness, and dose safety highlight that reinforcement learning remains irreplaceable for optimising the dynamic treatment regime.

8.2 Methodological Contributions

This thesis contributes a suite of tools, datasets, and methodological insights for the development and evaluation of AI-driven dynamic treatment regimes. These contributions span simulation, evaluation, real-world deployment, and novel paradigms for treatment planning:

(1) *DTR-Bench* (Chapter 4) sets a new standard for simulator realism and modularity. It enables researchers to systematically vary clinical complexity, including stochastic transitions, missing data, and pharmacodynamic variability, and observe how RL agents respond. By enabling controlled adversarial evaluation, it shifts the focus from average-case performance to robustness and failure mode analysis.

(2) The re-examination for offline RL (Chapter 5) provides a unified comparison of offline RL methods in a real-world clinical context. It aligns model implementations, MDP definitions, and evaluation protocols, offering the first variance-aware analysis of RL for sepsis treatment. It exposes how evaluation can fail from misspecified behaviour models and why naive baselines may appear deceptively strong. The chapter concludes with a set of practical guidelines for developing future RL-DTR studies to improve reliability, transparency, and clinical relevance.

(3) The insulin study (Chapter 6) offers a fully curated open-source dataset and benchmark for prediction of hypoglycaemia and insulin optimisation in the ICU. A

novel ensemble-bootstrapped OPE method is proposed to improve evaluation robustness in safety-critical settings. Together, these contributions create a reproducible pipeline for evaluating insulin policies based on RL under real-world constraints.

(4) The prior knowledge study (Chapter 7) explores the possibility of large language models as zero-shot treatment planners. It presents the comparison of LLMs and RL agents in simulated insulin dosing and introduces the concept of prior knowledge injection through prompt engineering.

8.3 Limitations

This thesis advances the use of reinforcement learning and artificial intelligence methods to develop dynamic treatment regimes, but several limitations remain. The first limitation concerns the generalisability of the proposed approaches. The models are trained and evaluated primarily on datasets such as MIMIC-III and SimGlucose, under specific assumptions about patient populations and treatment protocols. As a result, their performance may not translate directly to other clinical environments or diverse groups of patients, limiting immediate deployment without additional validation across multiple settings. A second limitation lies in the reliance on retrospective observational data and offline evaluation. Although rigorous simulation benchmarks and advanced off-policy evaluation techniques are used to mitigate known issues, these methods can only approximate true clinical effectiveness. Offline RL and OPE inherently depend on adequate action coverage under the historical logging policy, which is often violated in intensive care settings where treatments are sparse and highly skewed. This leads to unstable importance weighting and unreliable counterfactual assessments. Ultimately, treatment efficacy cannot be fully established without prospective clinical trials that capture real patient outcomes. A further limitation is the limited interpretability and transparency of many advanced RL approaches. Although language-model-based treatment regimes can generate explicit textual explanations, the quality of the reasoning is limited. Lack of clear explanations for treatment decisions creates barriers to clinician trust and slows adoption in practice. Without interpretable mechanisms, it remains challenging to

integrate AI-driven recommendations into clinical workflows and to refine treatment policies based on clinician feedback.

8.4 Future Work

Future research should first focus on prospective validation through clinical trials. Moving beyond retrospective analysis is essential to determine the safety, effectiveness, and clinical relevance of AI-driven treatment regimes. Adaptive trial designs, such as sequential multiple-assignment randomised trials (SMART), provide a promising pathway to embed RL in real-world clinical studies. Such designs would enable treatment policies to be continuously updated based on real-time patient outcomes and clinician feedback, thereby bridging the gap between offline model development and practical bedside application. A second direction involves improving the generalisability of learnt treatment policies. To ensure robustness across diverse patient populations and clinical settings, future work should incorporate multi-institutional datasets and employ domain adaptation or transfer learning techniques to enhance the generalisability of the findings. By exposing models to a wider range of treatment protocols, physiological responses, and care environments, it becomes possible to develop policies that are less dependent on a single dataset or institution, thus enhancing their external validity. Finally, advancing interpretability and transparency should be a central focus. Future research needs to develop frameworks that not only optimise clinical outcomes, but also provide interpretable rationales for treatment recommendations. To fully release the natural language explainability of LLMs, techniques such as reinforcement learning from human feedback can correct erroneous reasoning chains and generate content that aligns with clinical preferences. Enhancing interpretability would foster clinician trust, facilitate regulatory approval, and support collaborative policy refinement between AI systems and medical professionals.

8.5 Closing Remarks

This thesis is not a celebration of reinforcement learning. Rather, it represents a critical step towards understanding when, where, and how AI can safely and meaningfully support dynamic treatment decisions. By exposing fragilities, introducing practical tools, and sharpening our understanding of the clinical stakes, this thesis contributes to incremental but necessary progress toward a trustworthy AI in healthcare. The end goal remains unchanged: *not just to deliver treatment, but to ensure that the treatment evolves with the patient.*

Appendices

A

Supplementary Results for Chapter 4

Table A.1: Hyperparameter search space. 'Common' hyperparameters are model-agnostic and used identically in all runs. Stack number means the number of stacking frames for RNN policy. In non-RNN case, the stack number is 1.

| Type | Name | Space |
|------------------|------------------------------|--|
| Common | Learning Rate | $[10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}]$ |
| | Batch Size | [128, 256, 512, 1024] |
| | Stack Number | [20, 50, 100] |
| | Batch Normalisation | [True, False] |
| | Dropout | [0.0, 0.25, 0.5] |
| | Target Update Frequency | [1, 1000, 5000] |
| | Update Per Step | [0.1, 0.5] |
| | Update Actor Frequency | [1, 5, 10] |
| | Step Per Collect | [50, 100] |
| | Exploration Noise | [0.1, 0.2, 0.5] |
| | Estimation Step | 1 |
| | γ | 0.95 |
| | τ | 0.001 |
| DQN Series | ϵ_{test} | 0.005 |
| | $\epsilon_{\text{train},0}$ | 1 |
| | $\epsilon_{\text{train},-1}$ | 0.005 |
| C51, rC51 | v_{min} | [-20, -10, -5] |
| | v_{max} | [5, 10, 20] |
| | Number of Atoms | 51 |
| | ϵ_{test} | 0.005 |
| | $\epsilon_{\text{train},0}$ | 1 |
| | $\epsilon_{\text{train},-1}$ | 0.005 |
| SAC, rSAC, drSAC | α | [0.05, 0.1, 0.2] |
| REDQ | α | [0.05, 0.2, 0.4] |
| | Ensemble Size | 10 |
| | Actor Decay | 20 |
| TD3 | Noise Clip | 0.5 |
| | Policy Noise | 0.2 |

Table A.2: Model parameters for the *AhnChemoEnv* ODE system (see Eq. (4.1))

| Param. | Biological meaning | Default value |
|------------------------|--|--|
| $a_i^{(A)}$ | Drug-induced cell-kill coefficients: $a_1^{(A)}$ (immune), $a_2^{(A)}$ (tumour), $a_3^{(A)}$ (normal) | $a_1^{(A)} = 0.2$, $a_2^{(A)} = 0.3$, $a_3^{(A)} = 0.1$ |
| $b_i^{(A)}$ | Inverse carrying capacities (logistic limits): $b_1^{(A)}$ (tumour), $b_2^{(A)}$ (normal) | $b_1^{(A)} = 1.0$, $b_2^{(A)} = 1.0$ |
| $c_i^{(A)}$ | Interaction coefficients: $c_1^{(A)}$ (tumour \rightarrow immune loss), $c_2^{(A)}$ (immune \rightarrow tumour kill), $c_3^{(A)}$ (tumour–normal competition), $c_4^{(A)}$ (normal loss from tumour) | $c_1^{(A)} = 1.0$, $c_2^{(A)} = 0.5$, $c_3^{(A)} = 1.0$, $c_4^{(A)} = 1.0$ |
| $d_1^{(A)}, d_2^{(A)}$ | $d_1^{(A)}$: natural immune-cell death rate; $d_2^{(A)}$: drug elimination rate | $d_1^{(A)} = 0.2$, $d_2^{(A)} = 1.0$ |
| $r_1^{(A)}, r_2^{(A)}$ | Intrinsic growth rates of tumour ($r_1^{(A)}$) and normal cells ($r_2^{(A)}$) | $r_1^{(A)} = 1.5$, $r_2^{(A)} = 1.0$ |
| $s^{(A)}$ | Basal immune-cell influx in the absence of tumour | $s^{(A)} = 0.33$ |
| $\rho^{(A)}$ | Maximum immune proliferation rate stimulated by tumour presence | $\rho^{(A)} = 0.01$ |

Table A.3: Parameters of the *GhaffariCancerEnv* ODE system part I (See Eq 4.3 and Eq 4.4)

| Param. | Units | Description | Default |
|----------------|--------------------------------------|--|------------------------|
| $a_1^{(G)}$ | day ⁻¹ | Primary-tumour logistic growth rate | 4.31×10^{-1} |
| $b_1^{(G)}$ | cells ⁻¹ | $(b_1^{(G)})^{-1}$ = tumour carrying capacity (primary) | 1.02×10^{-9} |
| $c_1^{(G)}$ | cell ⁻¹ day ⁻¹ | NK-mediated tumour-cell kill (primary) | 6.41×10^{-11} |
| $d_1^{(G)}$ | day ⁻¹ | Saturation level in CD8 ⁺ -mediated kill term $D_p^{(G)}$ | 2.34 |
| l | – | Exponent l in fractional CD8 ⁺ -kill term $D_p^{(G)}$ | 2.09 |
| s | – | Steepness coefficient s in $D_p^{(G)}$ | 8.39×10^{-2} |
| $e_1^{(G)}$ | day ⁻¹ | Fraction of circulating lymphocytes becoming NK cells (primary) | 2.08×10^{-1} |
| $f_1^{(G)}$ | day ⁻¹ | NK-cell death rate (primary) | 4.12×10^{-3} |
| $p_1^{(G)}$ | cell ⁻¹ day ⁻¹ | NK inactivation by tumour cells (primary) | 3.42×10^{-4} |
| $m^{(G)}$ | day ⁻¹ | CD8 ⁺ T-cell death rate (primary) | 2.04×10^{-2} |
| $j_1^{(G)}$ | day ⁻¹ | Max. CD8 ⁺ T-cell recruitment rate (primary) | 2.49×10^{-2} |
| $k_1^{(G)}$ | cells ² | Steepness of CD8 ⁺ T-cell recruitment curve (primary) | 3.66×10^7 |
| $q_1^{(G)}$ | cell ⁻¹ day ⁻¹ | CD8 ⁺ T-cell inactivation by tumour (primary) | 1.42×10^{-4} |
| $r_{11}^{(G)}$ | cell ⁻¹ day ⁻¹ | CD8 ⁺ boost from NK-mediated tumour kill (primary) | 1.1×10^{-7} |
| $r_{12}^{(G)}$ | cell ⁻¹ day ⁻¹ | CD8 ⁺ boost from tumour–lymphocyte interaction (primary) | 6.5×10^{-11} |
| $u_1^{(G)}$ | cell ⁻² day ⁻¹ | NK regulation of CD8 ⁺ T-cells (primary) | 3×10^{-10} |

Table A.4: Parameters of the *GhaffariCancerEnv* ODE system part II ((See Eq 4.3 and Eq 4.4))

| Param. | Units | Description | Default |
|--|--------------------------------------|--|------------------------|
| $K_{T1}^{(G)}$ | day ⁻¹ | Chemotherapy-induced tumour-cell kill (primary) | 100 |
| $K_{1L}^{(G)}$ $K_{1N}^{(G)}$ $K_{1C}^{(G)}$ | day ⁻¹ | Chemotherapy-induced immune-cell kill (primary) | 10 |
| $\alpha^{(G)}$ | cell day ⁻¹ | Constant source of circulating lymphocytes | 7.5×10^8 |
| $\beta^{(G)}$ | day ⁻¹ | Natural death/differentiation of circulating lymphocytes | 1.2×10^2 |
| $\mu^{(G)}$ | day ⁻¹ | Chemotherapy-drug decay rate | 9×10^{-1} |
| $a_2^{(G)}$ | day ⁻¹ | Secondary-tumour logistic growth rate | 5 |
| $b_2^{(G)}$ | cells ⁻¹ | $(b_2^{(G)})^{-1}$ = tumour carrying capacity (secondary) | 1×10^{-7} |
| $c_2^{(G)}$ | cell ⁻¹ day ⁻¹ | NK-mediated tumour-cell kill (secondary) | 6.41×10^{-12} |
| $d_2^{(G)}$ | day ⁻¹ | Saturation level in CD8 ⁺ -kill term $D_s^{(G)}$ | 5 |
| $e_2^{(G)}$ | day ⁻¹ | Fraction of lymphocytes becoming NK cells (secondary) | 2.08×10^{-1} |
| $f_2^{(G)}$ | day ⁻¹ | NK-cell death rate (secondary) | 3.5×10^{-2} |
| $p_2^{(G)}$ | cell ⁻¹ day ⁻¹ | NK inactivation by tumour cells (secondary) | 1×10^{-1} |
| $m_2^{(G)}$ | day ⁻¹ | CD8 ⁺ T-cell death rate (secondary) | 1.8×10^{-1} |
| $j_2^{(G)}$ | day ⁻¹ | Max. CD8 ⁺ T-cell recruitment rate (secondary) | 1.6×10^{-2} |
| $k_2^{(G)}$ | cells ² | Steepness of CD8 ⁺ T-cell recruitment curve (secondary) | 3.66×10^7 |

Table A.5: Parameters of the *GhaffariCancerEnv* ODE system part III (See Eq 4.3 and Eq 4.4)

| Param. | Units | Description | Default |
|------------------|--------------------------------------|---|----------------------|
| $q_2^{(G)}$ | cell ⁻¹ day ⁻¹ | CD8 ⁺ T-cell inactivation by tumour (secondary) | 1×10^{-1} |
| $r_{21}^{(G)}$ | cell ⁻¹ day ⁻¹ | CD8 ⁺ boost from NK-mediated tumour kill (secondary) | 2×10^{-1} |
| $r_{22}^{(G)}$ | cell ⁻¹ day ⁻¹ | CD8 ⁺ boost from tumour–lymphocyte interaction (secondary) | 7.5×10^{11} |
| $u_2^{(G)}$ | cell ⁻² day ⁻¹ | NK regulation of CD8 ⁺ T-cells (secondary) | 3×10^{-10} |
| $K_T^{(G)}$ | day ⁻¹ | Chemotherapy-induced tumour-cell kill (secondary) | 100 |
| $K_{2L}^{(G)}$ | day ⁻¹ | Chemotherapy-induced immune-cell kill (secondary) | 10 |
| $K_{2N}^{(G)}$ | | | |
| $K_{2C}^{(G)}$ | day ⁻¹ | Chemotherapy kill of circulating lymphocytes (secondary) | 10 |
| $\gamma_1^{(G)}$ | cell ⁻¹ day ⁻¹ | Recovery rate of damaged cells $u^{(G)}$ | 0.04 |
| $\gamma_2^{(G)}$ | cell ⁻¹ day ⁻¹ | Recovery rate of damaged cells $v^{(G)}$ | 0.10 |
| $\gamma_3^{(G)}$ | cell ⁻¹ day ⁻¹ | Recovery rate of damaged cells $x^{(G)}$ | 0.10 |
| $\epsilon^{(G)}$ | – | Fraction of radiation damaging healthy cells | 0.05 |
| $\alpha_1^{(G)}$ | – | Rate at which cancer leaves the primary site | 0.0001 |
| $\alpha_2^{(G)}$ | – | Arrival/interaction rate at secondary site | 0.00001 |
| $W_{T1}^{(G)}$ | cells ² | Carry-capacity term for tumour (primary) | 0.01 |
| $W_{1N}^{(G)}$ | cells ² | Carry-capacity terms for immune cells (primary) | 1 |
| $W_{1L}^{(G)}$ | | | |
| $W_{2T}^{(G)}$ | cells ² | Carry-capacity term for tumour (secondary) | 1 |
| $W_{2N}^{(G)}$ | cells ² | Carry-capacity terms for immune cells (secondary) | 1 |
| $W_{2L}^{(G)}$ | | | |

Table A.6: Parameters of the *SimGlucoseEnv* ODE system (See Eq 4.8)

| Param. | Unit | Description | Default value |
|-----------------|--------------------------------------|---|-----------------------|
| $k_{p1}^{(S)}$ | mg kg ⁻¹ h ⁻¹ | Rate constant for endogenous glucose production | 11.5048 |
| $k_{p2}^{(S)}$ | (mg dL ⁻¹) ⁻¹ | Sensitivity of glucose production to plasma glucose | 0.0233 |
| $k_{p3}^{(S)}$ | h ⁻¹ | Sensitivity of glucose production to insulin action | 0.0233 |
| $k_{e1}^{(S)}$ | mg kg ⁻¹ h ⁻¹ | Parameter for renal glucose excretion | 0.0005 |
| $k_{e2}^{(S)}$ | mg dL ⁻¹ | Glucose renal-excretion threshold | 339 |
| $V_{m0}^{(S)}$ | mg kg ⁻¹ h ⁻¹ | Maximum glucose utilisation rate | 5.9285 |
| $V_{mx}^{(S)}$ | mg kg ⁻¹ h ⁻¹ | Sensitivity of glucose utilisation to insulin action | 0.0747 |
| $K_{m0}^{(S)}$ | mg dL ⁻¹ | Glucose conc. at half-maximal utilisation | 260.8900 |
| $k_1^{(S)}$ | h ⁻¹ | Transfer rates between plasma and tissue compartments | $k_1^{(S)} = 0.0573,$ |
| $k_2^{(S)}$ | | | $k_2^{(S)} = 0.0677$ |
| $p_{2u}^{(S)}$ | h ⁻¹ | Insulin action time-constant on glucose utilisation | 0.0213 |
| $k_i^{(S)}$ | h ⁻¹ | Plasma insulin kinetics constant | 0.0089 |
| $k_{sto}^{(S)}$ | h ⁻¹ | Gastric emptying rate for solid carbohydrate | 0.0159 |
| $k_{gut}^{(S)}$ | h ⁻¹ | Transfer rate from stomach to gut | 0.0159 |
| $k_{abs}^{(S)}$ | h ⁻¹ | Intestinal absorption rate into plasma | 0.0910 |
| $f^{(S)}$ | – | Fractional bioavailability of ingested carbohydrate | 0.9 |
| $BW^{(S)}$ | kg | Body mass (normalisation factor) | 68.7060 |

B

Supplementary Results for Chapter 5

B.1 Behavioural Policy and Value Visualisation for the training and validation set

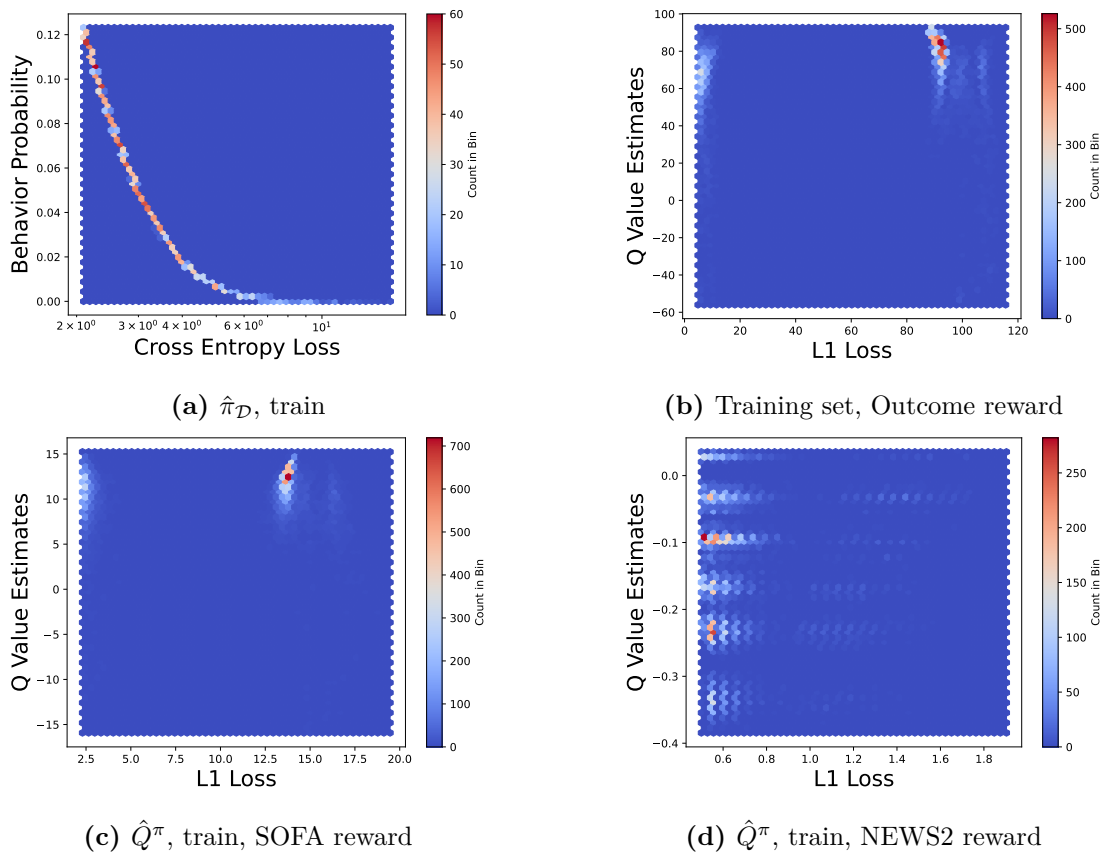


Figure B.1: Cross entropy and L1 losses vs inference probabilities or Q estimates in the training set. (a) Loss of $\hat{\pi}_{\mathcal{D}}$ across predicted action probabilities. (b–d) L1 loss for \hat{Q}^{π} under Outcome, SOFA, and NEWS2 rewards.

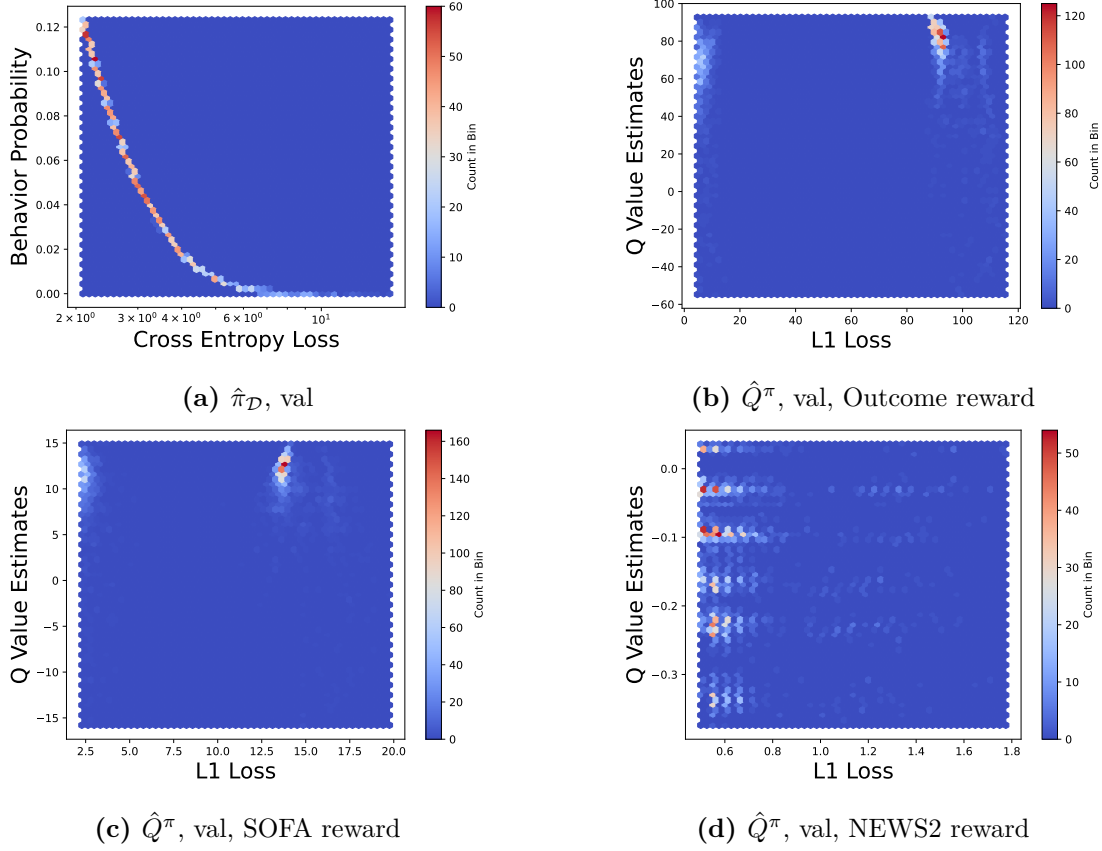


Figure B.2: Cross entropy and L1 losses vs inference probabilities or Q estimates in the validation set. (a) Loss of $\hat{\pi}_{\mathcal{D}}$ across predicted action probabilities. (b–d) L1 loss for \hat{Q}^{π} under Outcome, SOFA, and NEWS2 rewards.

B.2 Full Evaluation Result on Stratified Patient Groups

| 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 |
| RMSE _{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 |
| $\bar{G}_{\mathcal{D}}$ | 61.54 | | | | | | | | | |

Table B.1: Outcome all

B.3 Importance Ratio Histogram of Naive Baselines

| 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 |
| RMSE _{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 |
| $\bar{G}_{\mathcal{D}}$ | | | | | | 9.47 | | | | |

Table B.2: SOFA all

| 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 |
| RMSE _{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} | 93.32 | 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 |
| $\bar{G}_{\mathcal{D}}$ | | | | | | 68.51 | | | | |

Table B.3: 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 |
| RMSE _{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 |
| $\bar{G}_{\mathcal{D}}$ | | | | | | 78.97 | | | | |

Table B.4: 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 |
| RMSE _{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 |
| $\bar{G}_{\mathcal{D}}$ | | | | | | 67.49 | | | | |

Table B.5: Outcome 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 | 661.62 ± 8.98 | 745.14 ± 22.99 | 709.98 ± 19.82 | 675.19 ± 6.55 | 738.97 ± 2.38 |
| RMSE _{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 |
| $\bar{G}_{\mathcal{D}}$ | | | | | | 68.67 | | | | |

Table B.6: Outcome sub rate $\in [-0.4, -0.15]$ high std

| metric | alt | max | min | random | weight | imitation | dqn | cql | iq1 | 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 | |
| RMSE _{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 B.7: Outcome sub rate $\in [-0.15, 0]$ low std

| metric | alt | max | min | random | weight | imitation | dqn | cql | iq1 | 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 | |
| RMSE _{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 B.8: Outcome sub rate $\in [-0.15, 0]$ high std

| metric | alt | max | min | random | weight | imitation | dqn | cql | iq1 | 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 | |
| RMSE _{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 B.9: Outcome sub rate $\in [0.15, 0.4]$ low std

| metric | alt | max | min | random | weight | imitation | dqn | cql | iq1 | 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 | |
| RMSE _{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 B.10: Outcome sub rate $\in [0.15, 0.4]$ high std

| metric | alt | max | min | random | weight | imitation | dqn | cql | iq1 | 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 | |
| RMSE _{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 B.11: 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 | |
| RMSE _{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 | -80.82 | 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 | |
| $\bar{G}_{\mathcal{D}}$ | | | | | | | 28.41 | | | | |

Table B.12: 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 | |
| RMSE _{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 | -1.91 ± 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 | |
| $\bar{G}_{\mathcal{D}}$ | | | | | | | 11.09 | | | | |

Table B.13: 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 | |
| RMSE _{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 | |
| $\bar{G}_{\mathcal{D}}$ | | | | | | | 12.68 | | | | |

Table B.14: 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 | |
| RMSE _{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 | |
| $\bar{G}_{\mathcal{D}}$ | | | | | | | 10.65 | | | | |

Table B.15: 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 | |
| RMSE _{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 | |
| $\bar{G}_{\mathcal{D}}$ | | | | | | | 10.62 | | | | |

Table B.16: 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 | |
| RMSE _{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 B.17: 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 | |
| RMSE _{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 B.18: SOFA sub rate $\in [-0.15, 0]$ 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 | |
| RMSE _{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 B.19: 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 | |
| RMSE _{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 B.20: 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 | |
| RMSE _{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 B.21: SOFA sub rate $\in [0.4, \infty]$ low std

| metric | alt | max | min | random | weight | imitation | dqn | cql | iq1 | 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 | |
| RMSE _{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 B.22: SOFA sub rate $\in [0.4, \infty]$ high std

| metric | alt | max | min | random | weight | imitation | dqn | cql | iq1 | 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 | |
| RMSE _{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 B.23: NEWS2 sub rate $\in [-\infty, -0.4]$ low std

| metric | alt | max | min | random | weight | imitation | dqn | cql | iq1 | 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 | |
| RMSE _{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 B.24: NEWS2 sub rate $\in [-0.4, -0.15]$ low std

| metric | alt | max | min | random | weight | imitation | dqn | cql | iq1 | 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 | |
| RMSE _{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 B.25: NEWS2 sub rate $\in [-0.4, -0.15]$ high std

| metric | alt | max | min | random | weight | imitation | dqn | cql | iq1 | 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 | |
| RMSE _{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 B.26: 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 | 649.09 ± 9.58 |
| RMSE _{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 |
| WIS | -4.15 | -8.69 | -4.15 | -5.83 | -3.16 | -5.83 | -4.94 ± 1.2 | -3.13 ± 1.14 | -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 |
| WIS _t | -4.15 | -8.69 | -4.15 | -5.83 | -3.16 | -5.83 | -4.94 ± 1.2 | -3.14 ± 1.17 | -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 |
| 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 |
| 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 |
| 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 |
| $G_{\mathcal{D}}$ | | | | | | | -4.72 | | |

Table B.27: NEWS2 sub rate $\in [-0.15, 0]$ 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 |
| RMSE _{vaso} | 0.6 | 0.87 | 0.24 | 0.46 | 0.57 | 0.23 | 0.4 ± 0.08 | 0.22 | 0.23 ± 0.01 |
| 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 |
| 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 |
| 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 |
| 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 |
| 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 |
| 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 |
| 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 |
| $G_{\mathcal{D}}$ | | | | | | | -4.6 | | |

Table B.28: 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 |
| RMSE _{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 |
| 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 |
| 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 |
| 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 |
| 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 |
| 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 |
| 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 |
| 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 |
| $G_{\mathcal{D}}$ | | | | | | | -4.7 | | |

Table B.29: 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 |
| RMSE _{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 |
| 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 |
| 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 |
| 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 |
| 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 |
| 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 |
| 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 |
| 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 |
| $G_{\mathcal{D}}$ | | | | | | | -4.25 | | |

Table B.30: 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 |
| RMSE _{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 |
| 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 |
| 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 |
| 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 |
| 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 |
| 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 |
| 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 |
| 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 |
| $G_{\mathcal{D}}$ | | | | | | | -3.79 | | |

Table B.31: NEWS2 sub rate $\in [0.4, \infty]$ high std

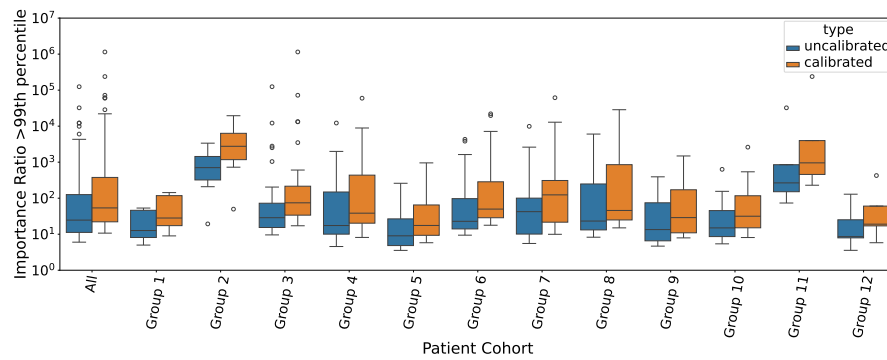


Figure B.3: Importance ratio histogram of random policy $>$ 99th percentile on Outcome reward.

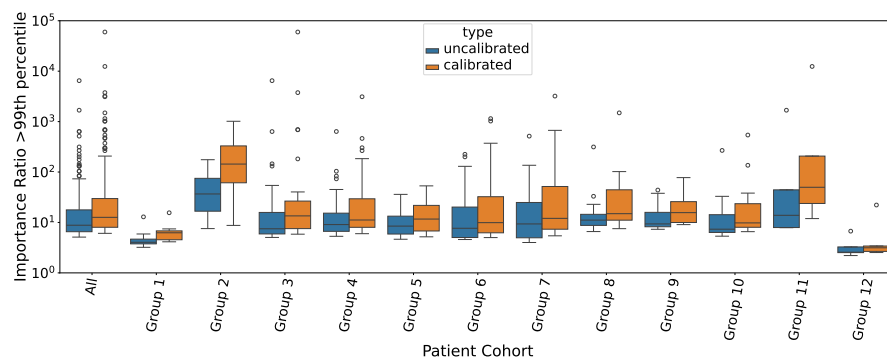


Figure B.4: Importance ratio histogram of min policy $>$ 99th percentile on Outcome reward.

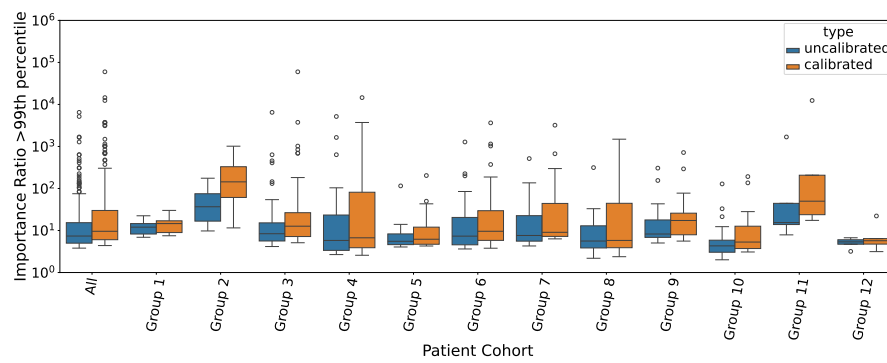


Figure B.5: Importance ratio histogram of max policy $>$ 99th percentile on Outcome reward.

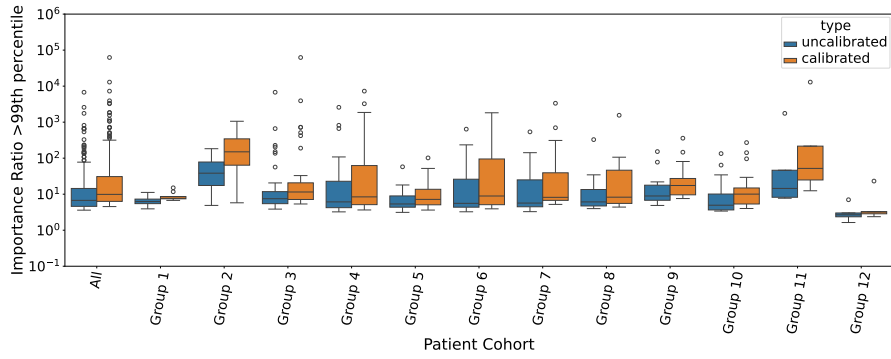


Figure B.6: Importance ratio histogram of alt policy > 99th percentile on Outcome reward.

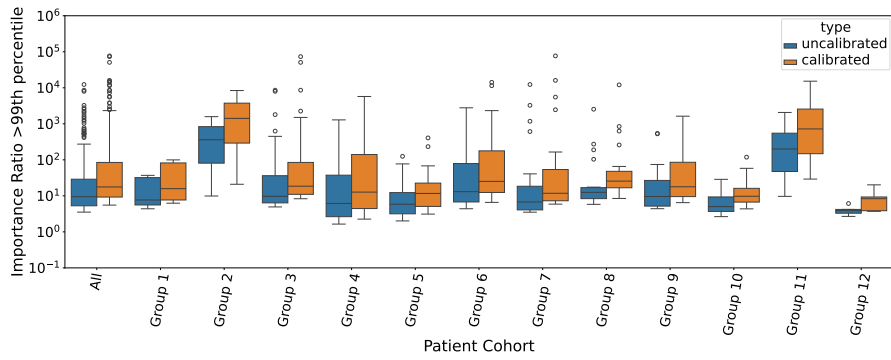


Figure B.7: Importance ratio histogram of weight policy > 99th percentile on Outcome reward.

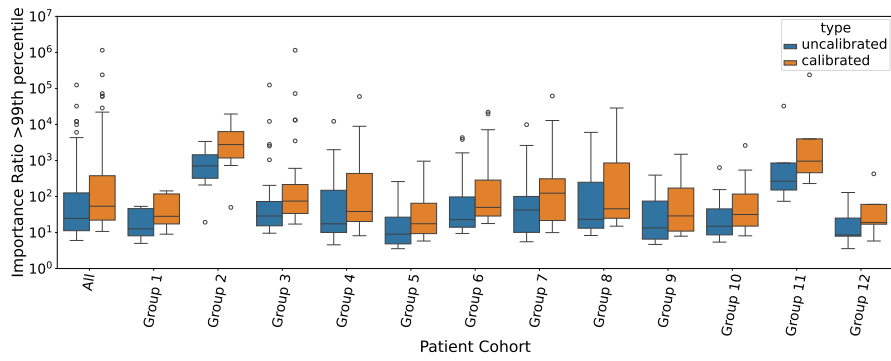


Figure B.8: Importance ratio histogram of random policy > 99th percentile on SOFA reward.

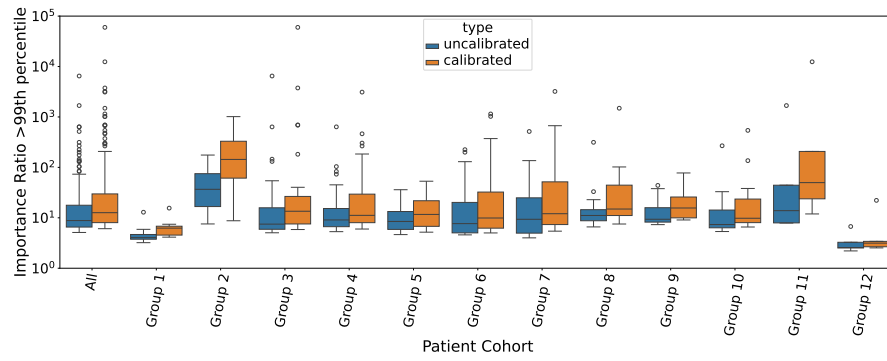


Figure B.9: Importance ratio histogram of min policy $>$ 99th percentile on SOFA reward.

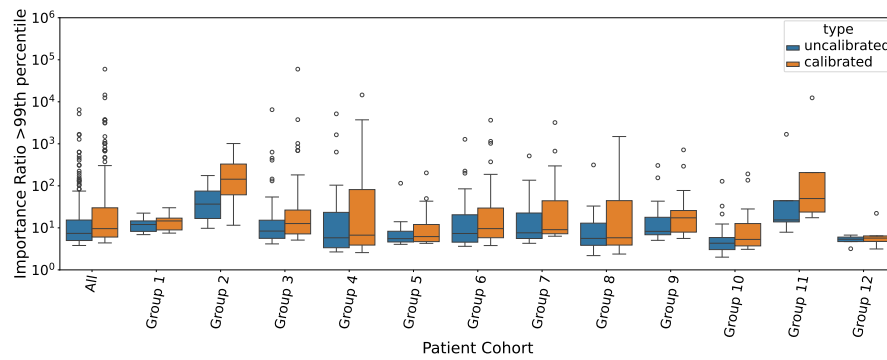


Figure B.10: Importance ratio histogram of max policy $>$ 99th percentile on SOFA reward.

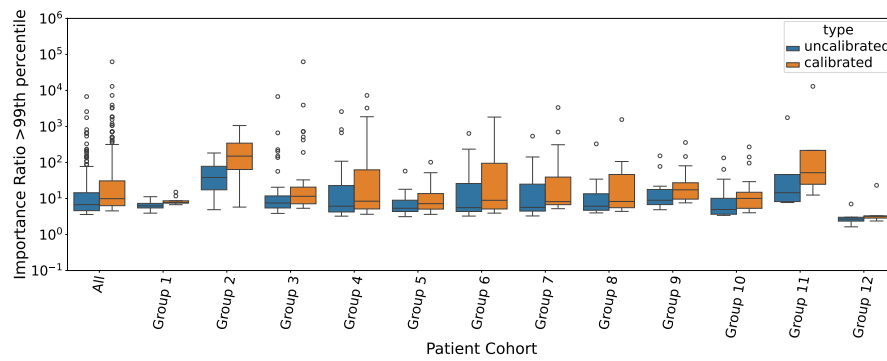


Figure B.11: Importance ratio histogram of alt policy $>$ 99th percentile on SOFA reward.

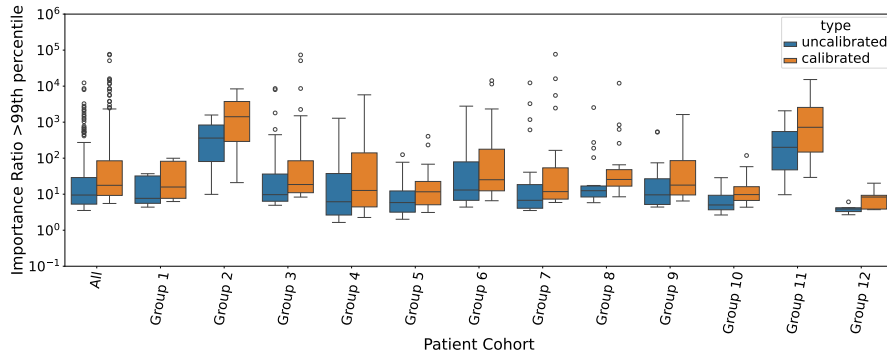


Figure B.12: Importance ratio histogram of weight policy > 99th percentile on SOFA reward.

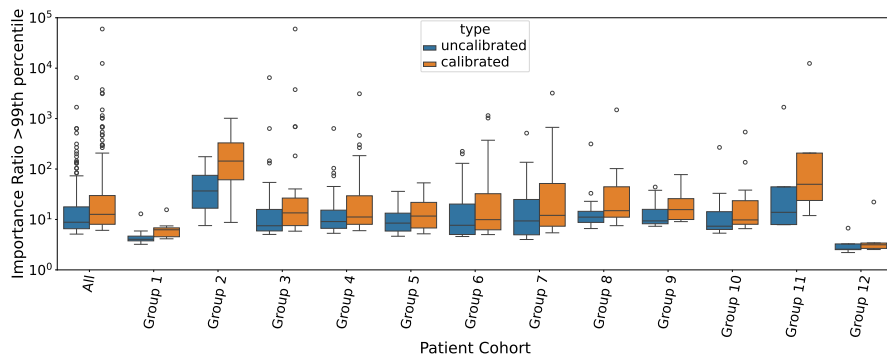


Figure B.13: Importance ratio histogram of min policy > 99th percentile on NEWS2 reward.

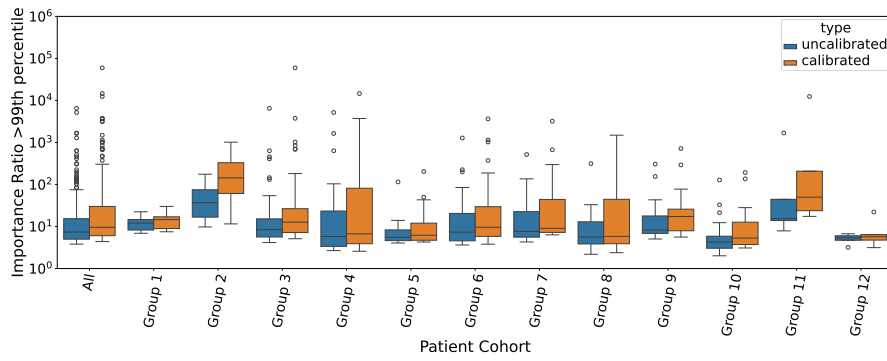


Figure B.14: Importance ratio histogram of max policy > 99th percentile on NEWS2 reward.

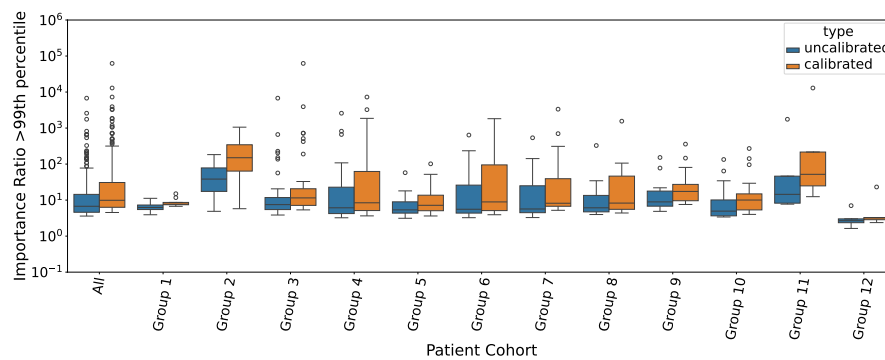


Figure B.15: Importance ratio histogram of alt policy $>$ 99th percentile on NEWS2 reward.

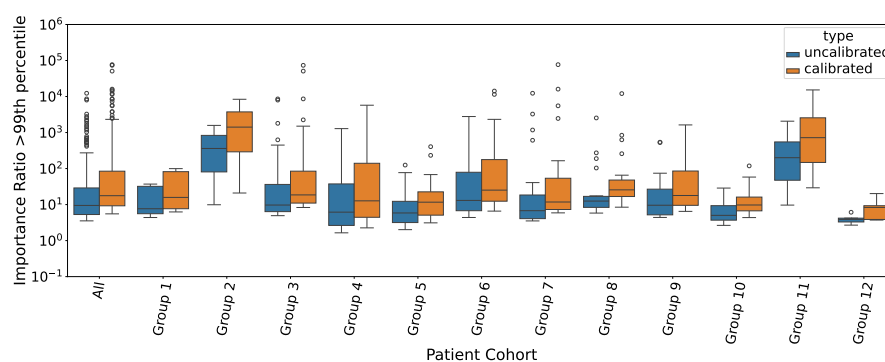


Figure B.16: Importance ratio histogram of weight policy $>$ 99th percentile on NEWS2 reward.

C

Supplementary Results for Chapter 6

C.1 Reward Definition and Distributional Properties

The relationship between glucose risk and the reward signal used in the reinforcement learning framework is illustrated in Figure C.1. A clipped transformation of the Blood Glucose Risk Index (BGRI) was applied to yield bounded and learnable reward values, while preserving clinical asymmetry in penalising hypo- and hyperglycaemic states.

The distribution of maximum glucose-related risk per patient episode is shown in Figure C.2. A right-skewed profile is observed, indicating that a substantial proportion of episodes involve at least one instance of extreme risk.

Figure C.3 presents the distribution of total accumulated returns across the training, validation, and test sets. These distributions were found to be qualitatively similar, consistent with the stratified data splitting approach used during cohort construction.

C.2 Training Details

Grid search is implemented to determine the optimal hyperparameter configurations for both the hypoglycaemia prediction and insulin recommendation tasks. For each candidate configuration, four independent training runs are performed using distinct random seeds. Seeds were generated using `np.random.seed(42)` and

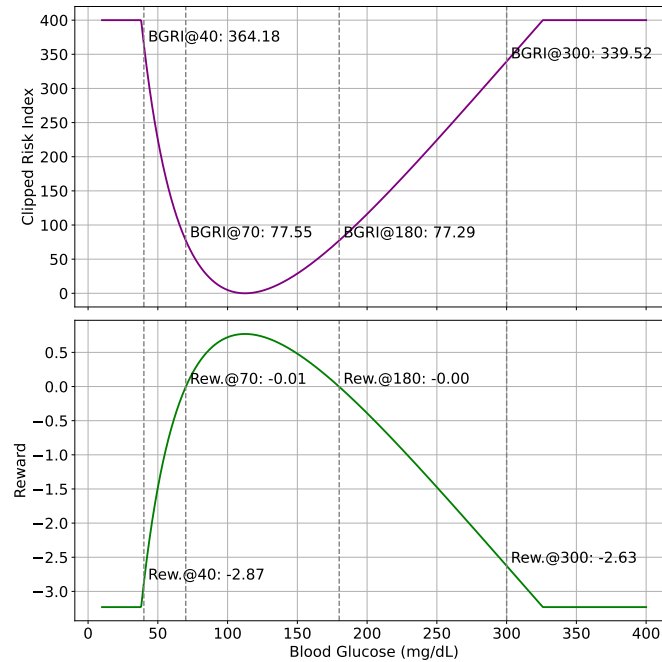


Figure C.1: Visualisation of the glucose risk index and corresponding reward transformation. The risk function exhibits a non-linear penalty that increases with deviation from euglycaemia. The reward function is clipped and rescaled to favour glucose values within the target range.

`np.random.randint(0, 100000, 4)`, resulting in the seeds: 15795, 860, 76820 and 54886.

C.3 Integrated Gradients Implementation

Integrated Gradients (IG) [274] was employed to quantify the contribution of input features to insulin administration predictions made by a GRU-based model. IG computes feature attributions by integrating the model’s gradient with respect to the input along a straight-line path from a baseline input to the actual input.

The baseline inputs were selected to reflect clinically meaningful reference states. Specifically, zero was used for binary indicators (e.g., drug administration), while the mid-point value (0.5) of each normalised channel was used for continuous physiological measurements.

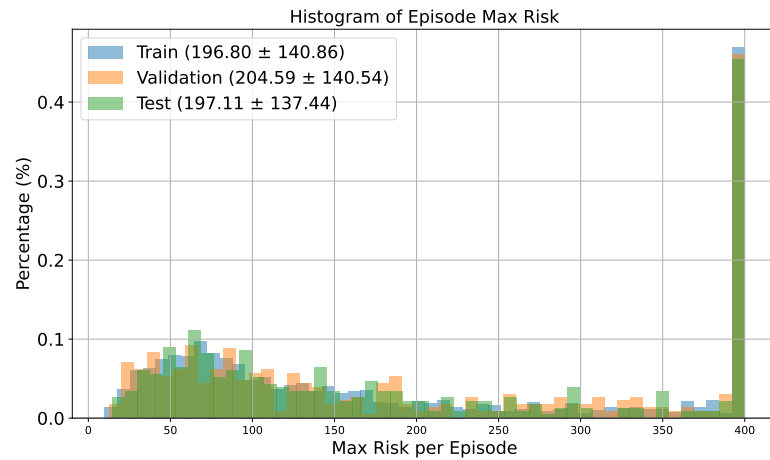


Figure C.2: Histogram of maximum glucose risk per episode. A high frequency of elevated risk values suggests that many patients encountered at least one period of significant glycaemic instability.

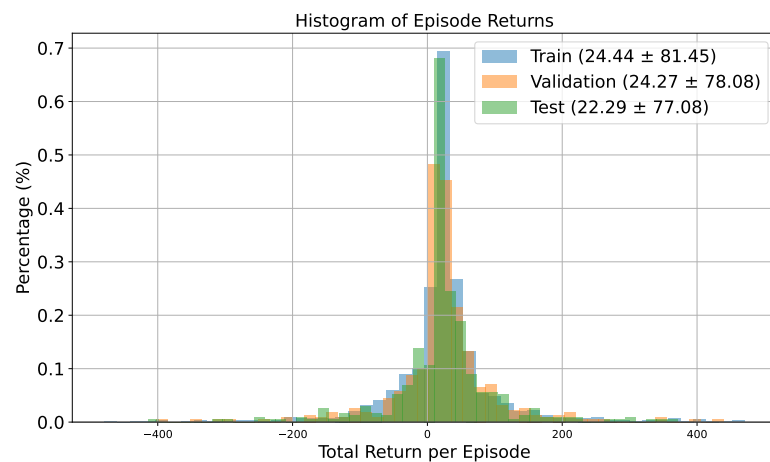


Figure C.3: Distribution of cumulative returns derived from the reward function across all episodes. The alignment of return distributions across subsets reflects stratification by patient phenotype.

C.4 Per-class Reliability Plot for Behavioural Model Calibration

Table C.1: Unified hyperparameter search space. Comma-separated lists denote discrete grid points; single values indicate parameters held constant. Common hyperparameters apply to all neural architectures (MLP, RNN, LSTM, GRU, RETAIN, Transformer), Behaviour Cloning and Reinforcement Learning agents unless stated otherwise.

| Category | Hyperparameter | Values | Applicable Models |
|-------------------|-----------------------------|--|-------------------------|
| Common | Random seed | [15795, 860, 76820, 54886] | All models |
| | Learning rate | $[1 \times 10^{-3}, 1 \times 10^{-4}, 1 \times 10^{-5}]$ | All deep models, BC, RL |
| | Batch size | [128, 1024] | All deep models, BC, RL |
| | Dropout probability | [0, 0.1] | All deep models, BC |
| | Hidden layer width | 128 | All deep models, BC, RL |
| | Number of layers | 4 | All deep models, BC, RL |
| | Training epochs | 50 (patience = 5) | All deep models, BC, RL |
| | Discount factor γ | 0.99 | RL |
| Traditional ML | Target-update interval | [100, 1000, 5000] | RL |
| | Inverse regression strength | [0.01, 0.1, 1, 10] | Logistic regression |
| | Number of trees | [50, 100, 200] | Random forest, XGBoost |
| | Maximum tree depth | [None, 5, 10] | Random forest |
| | Minimum samples split | [2, 5] | Random forest |
| RL-specific | Learning rate | [0.01, 0.1, 0.3] | XGBoost |
| | α | [0.3, 0.5, 0.7] | CQL |
| | Action flexibility | [0.1, 0.3, 0.5] | BCQ |
| | β | [0, 0.3, 0.5] | BCQ |
| | Weight temperature | [1, 10, 20] | IQL |
| | Expectile | 0.7 | IQL |
| Behaviour Cloning | Max weight | [20, 100] | IQL |
| | η (KL weight) | 0, 0.05, 0.1, 0.5 | BC |

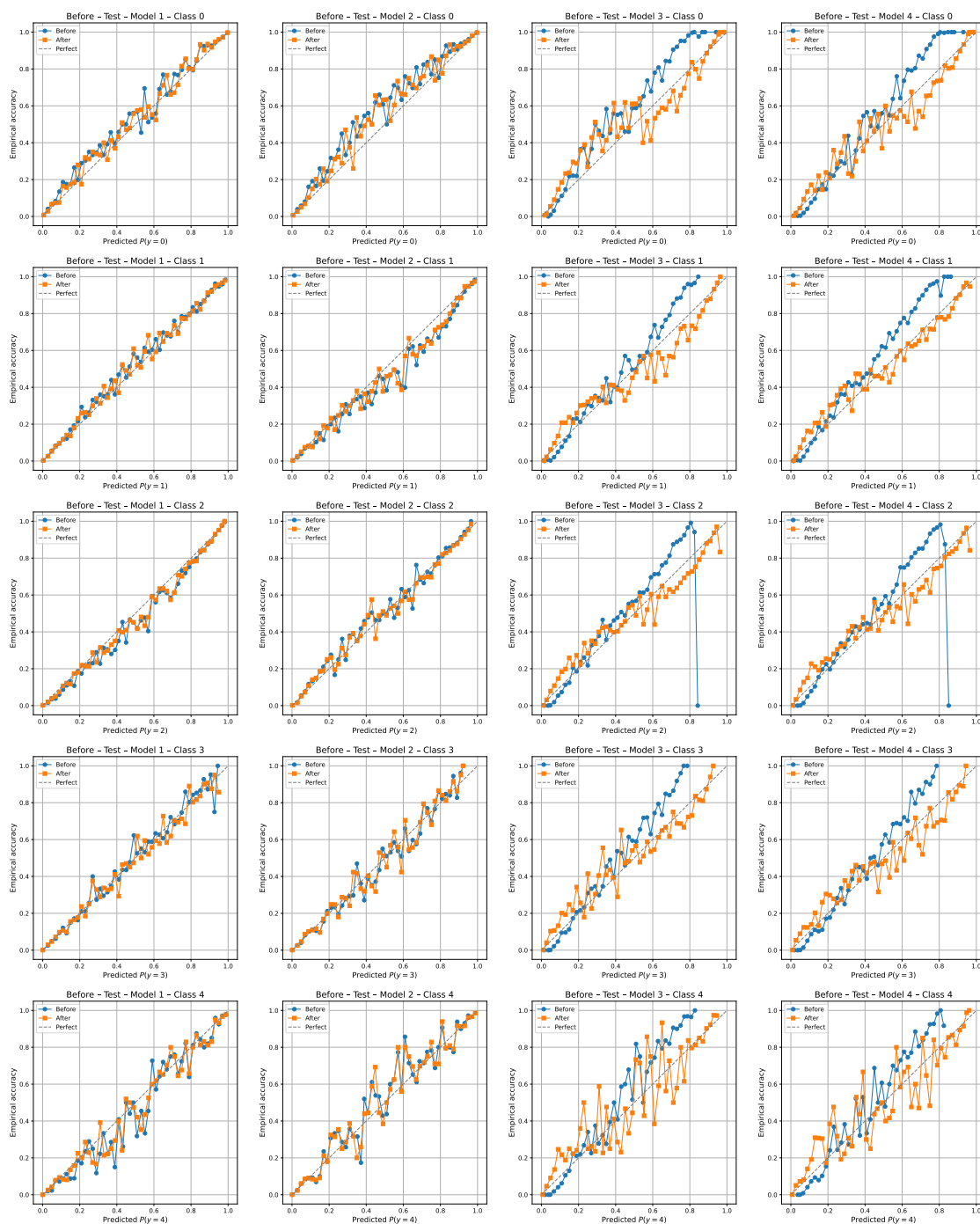


Figure C.4: Per-class reliability diagrams before and after calibration for the four top-performing behaviour-cloning models used in off-policy evaluation. The blue line shows model predictions before calibration, the orange line shows predictions after class-wise temperature scaling, and the grey diagonal indicates perfect calibration.

D

Supplementary Results for Chapter 7

D.1 Action Transformation in Continuous Action Space: A Comparative Analysis of clip and tanh

In continuous control problems with bounded action spaces, reinforcement learning (RL) policy networks must transform unbounded latent logits $z \in \mathbb{R}^d$ into bounded physical actions $a \in [0, D_{\max}]^d$. This is commonly achieved using a two-step transformation:

$$z \sim \mathcal{N}(\mu_\theta(s), \text{diag}(\sigma_\theta^2(s))), \quad a = \frac{D_{\max}}{2} (f(z) + \mathbf{1}),$$

where $f : \mathbb{R}^d \rightarrow [-1, 1]^d$ is an elementwise squashing function. Two common functions f are:

$$f(z) = \begin{cases} \tanh(z), & \text{(smooth squash),} \\ \text{clip}(z, -1, 1), & \text{(hard cut).} \end{cases}$$

D.1.1 Theoretical Properties of Tanh Transformation

The tanh function is differentiable and invertible from $\mathbb{R} \rightarrow (-1, 1)$, allowing precise computation of transformed action probabilities through the change-of-variables formula. Given:

$$a = g(z) = \frac{D_{\max}}{2}(\tanh z + 1), \quad z = g^{-1}(a) = \operatorname{artanh}\left(\frac{2a}{D_{\max}} - 1\right),$$

the log-density of the transformed action is:

$$\log \pi_{\theta}(a | s) = \log \phi(z; \mu_{\theta}(s), \sigma_{\theta}(s)) - \sum_{i=1}^d \log \left(\frac{D_{\max}}{2} (1 - \tanh^2 z_i) \right).$$

This approach maintains consistency in sampling and probability evaluation, preserving the validity of the PPO likelihood ratio and KL divergence computations. Additionally, the tanh Jacobian term penalises extreme logits, implicitly favouring safer, smaller interventions unless explicitly justified by strong model evidence.

D.1.2 Theoretical Limitations of Clip Transformation

In contrast, the clip function $f(z) = \operatorname{clip}(z, -1, 1)$ is neither differentiable nor invertible outside $[-1, 1]^d$. The correct action probability distribution under clipping would need to be represented as a mixture distribution, combining truncated Gaussian densities within bounds and discrete probability mass at boundary points. However, most implementations incorrectly treat clipped actions as directly sampled from a Gaussian distribution without accounting for clipping effects, leading to several critical issues:

- **Incorrect likelihood ratios:** Clipped actions mismatch actual sampling distributions, invalidating the PPO theoretical assumptions.
- **Misleading KL divergence:** The computed KL divergence ignores action saturation, failing to correctly constrain boundary behaviours.
- **Gradient instability:** Non-differentiable clipping results in zero or incorrect gradients, reducing policy training effectiveness.

D.1.3 Practical Considerations

Despite the clip’s theoretical limitations, clipping remains widely used due to its simplicity and numerical stability. While this may be acceptable for general-purpose RL tasks without stringent safety requirements, we argue that the potential clinical risks and theoretical inconsistencies could make clipping unsuitable for medical decision-making tasks. Therefore, we choose to adopt the tanh-based transformation for RL applications in healthcare settings as expert knowledge.

D.2 Heuristic Decision for Exploration Strategy in Discrete Action Space

In reinforcement learning settings involving clinical dosing, naïve uniform exploration often results in implausible or unsafe actions, particularly when the optimal policy frequently prescribes no intervention. To mitigate this, we introduce a domain-informed variant of ϵ -greedy exploration that assigns disproportionately high probability to the zero-dose action, reflecting the real-world sparsity of insulin administration events.

Specifically, given a discretised action space of size $|\mathcal{A}|$, the probability of choosing action 0 (corresponding to a zero-dose) during exploration is defined as

$$p_0 = \frac{|\mathcal{A}|(|\mathcal{A}| + 1)}{2|\mathcal{A}|^2 - |\mathcal{A}| + 1},$$

with the remaining probability mass $(1 - p_0)$ distributed uniformly across the $|\mathcal{A}| - 1$ non-zero actions:

$$p_j = \frac{1 - p_0}{|\mathcal{A}| - 1}, \quad j = 1, \dots, |\mathcal{A}| - 1.$$

This formulation ensures that the ratio between p_0 and any other p_j scales linearly with N , thereby preserving a strong safety bias even as the dose grid becomes more finely discretised.

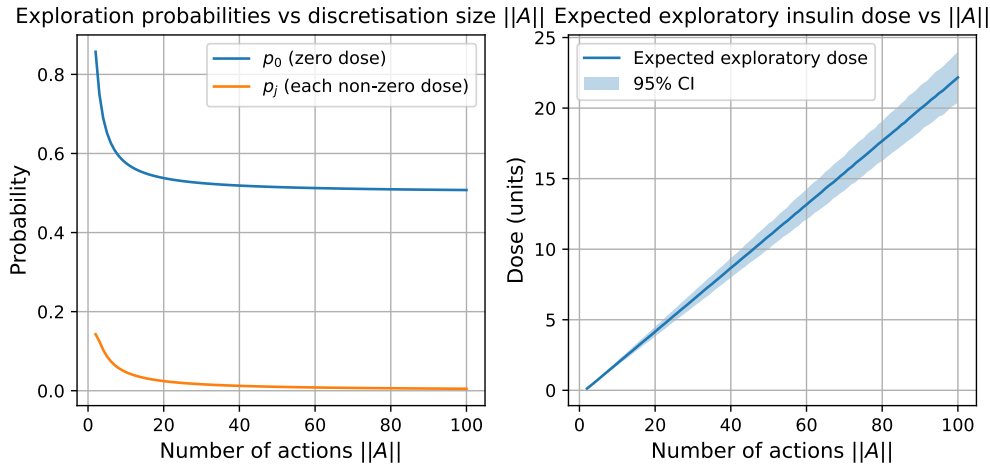


Figure D.1: Exploration probabilities and expected exploratory insulin dose as a function of discretisation size $||\mathcal{A}||$. **Left:** Probability assigned to the zero-dose action and each non-zero action. As $||\mathcal{A}||$ increases, it stabilises above 0.5. **Right:** Expected insulin dose under the exploration distribution, assuming non-zero actions correspond to dose units. The shaded region denotes the 95% bootstrap confidence interval.

Figure D.1 illustrates the evolution of p_0 and p_j as a function of $||\mathcal{A}||$. The probability mass allocated to the zero-dose action quickly stabilises around 0.52 as N increases, while the per-action mass for each non-zero dose diminishes accordingly. This ensures that, despite an expanding action space, the exploratory policy continues to prioritise conservative actions and the expectation of exploratory dose can grow linearly with the number of actions that increase.

D.3 Hyperparameters for Small Reinforced Agent Training

Table D.1: Common hyperparameters across all RL agents

| Parameter | Value(s) |
|---|-----------------------|
| Seed | {1, 100, 1000, 10000} |
| Learning rate (lr) | {1e-3, 1e-4} |
| Batch size | 128 |
| Use expert knowledge | {False, True} |
| Observation mode | Stack(48) |
| Discount factor (γ) | 0.99 |
| Exploration noise | 0.1 |
| Test-time ϵ (ϵ_{test}) | 0.001 |

Table D.2: DQN-specific hyperparameters

| Parameter | Value(s) |
|---------------------------------|------------------------------|
| Target network update frequency | {100, 1000} |
| Update per step | 1 |
| Actor update frequency | 1 |
| Train-time ϵ (initial) | 0.9 |
| Train-time ϵ (final) | 0.1 |
| ϵ schedule | Linear decay from 0.9 to 0.1 |

Table D.3: PPO-specific hyperparameters

| Parameter | Value(s) |
|--|---------------|
| On-policy steps per collect | 192 |
| Repeat per collect | 20 |
| GAE lambda (λ) | 0.95 |
| Conditioned σ | {True, False} |
| Value function coefficient (v_{coef}) | 0.5 |
| Entropy coefficient (ent_{coef}) | 0.001 |
| PPO clip range (ϵ_{clip}) | 0.1 |
| Value function clipping | {True, False} |
| Dual clip | None |
| Advantage normalisation | True |
| Recompute advantage | False |

References

- [1] Bibhas Chakraborty and Erica EM Moodie. *Statistical methods for dynamic treatment regimes*. Vol. 2. Springer, 2013.
- [2] R Scott Evans. “Electronic health records: then, now, and in the future”. In: *Yearbook of medical informatics* 25.S 01 (2016), S48–S61.
- [3] Kyle Swanson et al. “From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment”. In: *Cell* 186.8 (2023), pp. 1772–1791.
- [4] Zachary D Cohen, Jaime Delgado, and Robert J DeRubeis. “Personalized treatment approaches.” In: (2021).
- [5] Philippe Deprez, Pavel V Shevchenko, and Mario V Wüthrich. “Machine learning techniques for mortality modeling”. In: *European Actuarial Journal* 7.2 (2017), pp. 337–352.
- [6] Tahani A Daghistani et al. “Predictors of in-hospital length of stay among cardiac patients: a machine learning approach”. In: *International journal of cardiology* 288 (2019), pp. 140–147.
- [7] Rebecca Howard et al. “Distinguishing asthma phenotypes using machine learning approaches”. In: *Current allergy and asthma reports* 15.7 (2015), pp. 1–10.
- [8] Susan A Murphy. “Optimal dynamic treatment regimes”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65.2 (2003), pp. 331–355.
- [9] Bibhas Chakraborty and Susan A Murphy. “Dynamic treatment regimes”. In: *Annual review of statistics and its application* 1 (2014), pp. 447–464.
- [10] Shao Zhifei and Er Meng Joo. “A review of inverse reinforcement learning theory and recent advances”. In: *2012 IEEE congress on evolutionary computation*. IEEE, 2012, pp. 1–8.
- [11] Takayuki Osa et al. “An algorithmic perspective on imitation learning”. In: *Foundations and Trends® in Robotics* 7.1-2 (2018), pp. 1–179.
- [12] Michael Bain and Claude Sammut. “A Framework for Behavioural Cloning.” In: *Machine Intelligence* 15. 1995, pp. 103–129.
- [13] Stéphane Ross and Drew Bagnell. “Efficient reductions for imitation learning”. In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2010, pp. 661–668.
- [14] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. “A reduction of imitation learning and structured prediction to no-regret online learning”. In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2011, pp. 627–635.
- [15] Wen Sun et al. “Deeply aggravated: Differentiable imitation learning for sequential prediction”. In: *International conference on machine learning*. PMLR, 2017, pp. 3309–3318.
- [16] cse.buffalo.edu. *lecture 14.2 Imitation Learning*. 2019.

- [17] Leilei Sun et al. “Data-driven automatic treatment regimen development and recommendation”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 1865–1874.
- [18] Jacek M Bajor and Thomas A Lasko. “Predicting medications from diagnostic codes with recurrent neural networks”. In: (2016).
- [19] Choi E. et al. “Doctor AI: Predicting Clinical Events via Recurrent Neural Networks”. In: *Machine Learning in Health Care* (2015).
- [20] Mei Jing et al. “Deep Diabetologist: Learning to Prescribe Hypoglycemic Medications with Recurrent Neural Networks”. In: *Medinfo* (2017).
- [21] Rough K. et al. “Predicting Inpatient Medication Orders From Electronic Health Record Data”. In: *Clinical pharmacology and therapy* (2020). URL: <https://europepmc.org/articles/pmc7325318?pdf=render>.
- [22] Shaika Chowdhury Ph.D et al. “Predicting Effectiveness of Antihypertensive Medications for Heart Failure based on Longitudinal Patient Records and Deep Learning”. In: *medRxiv* (2022). URL: <https://www.medrxiv.org/content/medrxiv/early/2022/11/14/2022.11.01.22281804.full.pdf>.
- [23] Dey Sanjoy et al. “PerDREP: Personalized Drug Effectiveness Prediction from Longitudinal Observational Data”. In: *Knowledge Discovery and Data Mining* (2019).
- [24] Wei Jiedong and Zhang Yijia. “DFNet: Dual-Decision Fusion Network for Drug Combination Prediction”. In: *International Conference on Communication, Computing & Security* (2024).
- [25] Panagiotis Symeonidis et al. “Mortality Prediction and Safe Drug Recommendation for Critically-ill Patients”. In: *2022 IEEE 22nd International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE. Nov. 2022, pp. 79–84. URL: <http://dx.doi.org/10.1109/BIBE55377.2022.00025>.
- [26] Seyed Kamyar Seyed Ghasemipour, Richard Zemel, and Shixiang Gu. “A divergence minimization perspective on imitation learning methods”. In: *Conference on Robot Learning*. PMLR. 2020, pp. 1259–1277.
- [27] Volodymyr Mnih et al. “Human-level control through deep reinforcement learning”. In: *nature* 518.7540 (2015), pp. 529–533.
- [28] Hado Van Hasselt, Arthur Guez, and David Silver. “Deep reinforcement learning with double q-learning”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 30. 1. 2016.
- [29] Ziyu Wang et al. “Dueling network architectures for deep reinforcement learning”. In: *International conference on machine learning*. PMLR. 2016, pp. 1995–2003.
- [30] Marc G Bellemare, Will Dabney, and Rémi Munos. “A distributional perspective on reinforcement learning”. In: *International Conference on Machine Learning*. PMLR. 2017, pp. 449–458.
- [31] Timothy P Lillicrap et al. “Continuous control with deep reinforcement learning”. In: *arXiv preprint arXiv:1509.02971* (2015).
- [32] Scott Fujimoto, Herke Hoof, and David Meger. “Addressing function approximation error in actor-critic methods”. In: *International conference on machine learning*. PMLR. 2018, pp. 1587–1596.

- [33] Tuomas Haarnoja et al. “Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor”. In: *International conference on machine learning*. Pmlr. 2018, pp. 1861–1870.
- [34] Aviral Kumar et al. “Conservative q-learning for offline reinforcement learning”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 1179–1191.
- [35] Scott Fujimoto et al. “Benchmarking batch deep reinforcement learning algorithms”. In: *arXiv preprint arXiv:1910.01708* (2019).
- [36] Aviral Kumar et al. “Stabilizing off-policy q-learning via bootstrapping error reduction”. In: *Advances in Neural Information Processing Systems* 32 (2019).
- [37] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. “Offline reinforcement learning with implicit q-learning”. In: *arXiv preprint arXiv:2110.06169* (2021).
- [38] Justin Fu et al. “D4rl: Datasets for deep data-driven reinforcement learning”. In: *arXiv preprint arXiv:2004.07219* (2020).
- [39] John Schulman et al. “Proximal policy optimization algorithms”. In: *arXiv preprint arXiv:1707.06347* (2017).
- [40] A Rupam Mahmood, Hado P Van Hasselt, and Richard S Sutton. “Weighted importance sampling for off-policy learning with linear function approximation”. In: *Advances in Neural Information Processing Systems* 27 (2014).
- [41] Edward L Ionides. “Truncated importance sampling”. In: *Journal of Computational and Graphical Statistics* 17.2 (2008), pp. 295–311.
- [42] Nan Jiang and Lihong Li. “Doubly robust off-policy value evaluation for reinforcement learning”. In: *International Conference on Machine Learning*. PMLR. 2016, pp. 652–661.
- [43] Chao Yu et al. “Reinforcement learning in healthcare: A survey”. In: *ACM Computing Surveys (CSUR)* 55.1 (2021), pp. 1–36.
- [44] Matthieu Komorowski et al. “The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care”. In: *Nature medicine* 24.11 (2018), pp. 1716–1720.
- [45] Zhao Yufan et al. “Reinforcement Learning Strategies for Clinical Trials in Non-small Cell Lung Cancer”. In: (2009).
- [46] Zhao Yufan et al. “Reinforcement Learning Strategies for Clinical Trials in Nonsmall Cell Lung Cancer”. In: *Biometrics* (2011). URL: <https://europepmc.org/articles/pmc3138840?pdf=render>.
- [47] Chamani Shiranthika et al. “Supervised Optimal Chemotherapy Regimen Based on Offline Reinforcement Learning”. In: *IEEE Journal of Biomedical and Health Informatics* 26.9 (Sept. 2022), pp. 4763–4772. URL: <http://dx.doi.org/10.1109/JBHI.2022.3183854>.
- [48] Teplytska O. et al. “Machine Learning Methods for Precision Dosing in Anticancer Drug Therapy: A Scoping Review”. In: *Clinical Pharmacokinetics* (2024). URL: <https://link.springer.com/content/pdf/10.1007/s40262-024-01409-9.pdf>.

- [49] Corinna Maier et al. “Reinforcement learning and Bayesian data assimilation for modelinformed precision dosing in oncology”. In: *CPT: Pharmacometrics & Systems Pharmacology* 10.3 (Mar. 2021), pp. 241–254. URL: <http://dx.doi.org/10.1002/psp4.12588>.
- [50] Corinna Maier. *Bayesian data assimilation and reinforcement learning for model-informed precision dosing in oncology*. 2021. URL: <https://publishup.uni-potsdam.de/51587>.
- [51] Yao Song and Lu Wang. “Multiobjective tree-based reinforcement learning for estimating tolerant dynamic treatment regimes”. In: *Biometrics* 80.1 (Jan. 2024). URL: <http://dx.doi.org/10.1093/biomtc/ujad017>.
- [52] Tseng H. et al. “Deep reinforcement learning for automated radiation adaptation in lung cancer”. In: *Medical Physics (Lancaster)* (2017). URL: <https://aapm.onlinelibrary.wiley.com/doi/pdfdirect/10.1002/mp.12625>.
- [53] X. Dai et al. “Reinforcement Learning Powered Station Parameter Optimized Radiation Therapy (SPORT): A Novel Treatment Planning and Beam Delivery Technique”. In: *International Journal of Radiation Oncology*Biolog*Physics* 117.2 (Oct. 2023), e658. URL: <http://dx.doi.org/10.1016/j.ijrobp.2023.06.2091>.
- [54] J. Zhang et al. “Simulating Planners’ Interactions With the Treatment Planning System: A Reinforcement Learning Study for Pancreas SBRT Planning”. In: *International Journal of Radiation Oncology*Biolog*Physics* 108.3 (Nov. 2020), e255–e256. URL: <http://dx.doi.org/10.1016/j.ijrobp.2020.07.615>.
- [55] Tortora M. et al. “Deep Reinforcement Learning for Fractionated Radiotherapy in Non-Small Cell Lung Carcinoma”. In: *Artif. Intell. Medicine* (2021).
- [56] Chenyang Shen et al. “Intelligent inverse treatment planning via deep reinforcement learning, a proof-of-principle study in high dose-rate brachytherapy for cervical cancer”. In: *Physics in Medicine & Biology* 64.11 (May 2019), p. 115013. URL: <http://dx.doi.org/10.1088/1361-6560/ab18bf>.
- [57] Gang Pu et al. “Deep reinforcement learning for treatment planning in high-dose-rate cervical brachytherapy”. In: *Physica Medica* 94 (Feb. 2022), pp. 1–7. URL: <http://dx.doi.org/10.1016/j.ejmp.2021.12.009>.
- [58] E.B. Wang et al. “Outcome-Oriented Human-Like Automated Brachytherapy Treatment Planning for Cervical Cancer with Deep Reinforcement Learning”. In: *International Journal of Radiation Oncology*Biolog*Physics* 114.3 (Nov. 2022), S162. URL: <http://dx.doi.org/10.1016/j.ijrobp.2022.07.653>.
- [59] Qingqing Wang and Chang Chang. “Automating proton PBS treatment planning for head and neck cancers using policy gradient-based deep reinforcement learning”. In: *arXiv.org* (2024). URL: <https://arxiv.org/abs/2409.11576>.
- [60] Hanlin Wang et al. “An integrated solution of deep reinforcement learning for automatic IMRT treatment planning in non-small-cell lung cancer”. In: *Frontiers in Oncology* 13 (Feb. 2023). URL: <http://dx.doi.org/10.3389/fonc.2023.1124458>.

- [61] Dongrong Yang et al. “Automated treatment planning with deep reinforcement learning for head-and-neck (HN) cancer intensity modulated radiation therapy (IMRT)”. In: *Physics in Medicine & Biology* 70.1 (Dec. 2024), p. 015010. URL: <http://dx.doi.org/10.1088/1361-6560/ad965d>.
- [62] Hunter Stephens et al. “A reinforcement learning agent for head and neck intensity-modulated radiation therapy”. In: *Frontiers in Physics* 12 (Feb. 2024). URL: <http://dx.doi.org/10.3389/fphy.2024.1331849>.
- [63] Damon Sprouts et al. “The development of a deep reinforcement learning network for dose-volume-constrained treatment planning in prostate cancer intensity modulated radiotherapy”. In: *Biomedical Physics & Engineering Express* 8.4 (June 2022), p. 045008. URL: <http://dx.doi.org/10.1088/2057-1976/ac6d82>.
- [64] Chenyang Shen et al. “Improving efficiency of training a virtual treatment planner network via knowledge-guided deep reinforcement learning for intelligent automatic treatment planning of radiotherapy”. In: *Medical Physics* 48.4 (Feb. 2021), pp. 1909–1920. URL: <http://dx.doi.org/10.1002/mp.14712>.
- [65] C. Shen et al. “Automatic Treatment Planning in a Human-like manner: Operating Treatment Planning Systems by a Deep Reinforcement Learning based Virtual Treatment Planner”. In: *International Journal of Radiation Oncology*Biography*Physics* 105.1 (Sept. 2019), S256. URL: <http://dx.doi.org/10.1016/j.ijrobp.2019.06.2538>.
- [66] Bin Liu et al. “Automatic radiotherapy treatment planning with deep functional reinforcement learning”. In: *medRxiv* (June 2024). URL: <http://dx.doi.org/10.1101/2024.06.23.24309060>.
- [67] Yingzi Liu et al. “Automatic inverse treatment planning of Gamma Knife radiosurgery via deep reinforcement learning”. In: *Medical Physics* 49.5 (Mar. 2022), pp. 2877–2889. URL: <http://dx.doi.org/10.1002/mp.15576>.
- [68] Saba Ebrahimi and Gino J. Lim. “A reinforcement learning approach for finding optimal policy of adaptive radiation therapy considering uncertain tumor biological response”. In: *Artificial Intelligence in Medicine* 121 (Nov. 2021), p. 102193. URL: <http://dx.doi.org/10.1016/j.artmed.2021.102193>.
- [69] Dipesh Niraula et al. “A clinical decision support system for AI-assisted decision-making in response-adaptive radiotherapy (ARClIDS)”. In: *Scientific Reports* 13.1 (Mar. 2023). URL: <http://dx.doi.org/10.1038/s41598-023-32032-6>.
- [70] Dipesh Niraula et al. “ARClIDS: A Clinical Decision Support System for AI-assisted Decision-Making in Response-Adaptive Radiotherapy”. In: *medRxiv* (Sept. 2022). URL: <http://dx.doi.org/10.1101/2022.09.23.22280215>.
- [71] D. Niraula et al. “A Decision Support Software for AI-Assisted Decision Making in Response-Adaptive Radiotherapy — An Evaluation Study”. In: *International Journal of Radiation Oncology*Biography*Physics* 114.3 (Nov. 2022), e101–e102. URL: <http://dx.doi.org/10.1016/j.ijrobp.2022.07.895>.
- [72] Niraula D. et al. “Quantum deep reinforcement learning for clinical decision support in oncology: application to adaptive radiotherapy”. In: *Scientific Reports* (2021). URL: <https://www.nature.com/articles/s41598-021-02910-y.pdf>.

- [73] I El Naqa and R Ten. “SUDBRB05: Quantum Learning for KnowledgeBased ResponseAdaptive Radiotherapy”. In: *Medical Physics* 43.6Part4 (June 2016), pp. 3338–3338. URL: <http://dx.doi.org/10.1118/1.4955631>.
- [74] Elisa Tardini et al. “Optimal Treatment Selection in Sequential Systemic and Locoregional Therapy of Oropharyngeal Squamous Carcinomas: Deep Q-Learning With a Patient-Physician Digital Twin Dyad”. In: *Journal of Medical Internet Research* 24.4 (Apr. 2022), e29455. URL: <http://dx.doi.org/10.2196/29455>.
- [75] Elisa Tardini et al. “Optimal policy determination in sequential systemic and locoregional therapy of oropharyngeal squamous carcinomas: A patient-physician digital twin dyad with deep Q-learning for treatment selection”. In: *medRxiv* (Apr. 2021). URL: <http://dx.doi.org/10.1101/2021.04.07.21255092>.
- [76] W.T. Hrinivich, D. Song, and J. Lee. “Machine Parameter Optimization of a Clinical Linear Accelerator Using Deep Reinforcement Learning for Automatic Generation of Deliverable Prostate VMAT Plans”. In: *International Journal of Radiation Oncology*Biography*Physics* 117.2 (Oct. 2023), S56. URL: <http://dx.doi.org/10.1016/j.ijrobp.2023.06.348>.
- [77] William T. Hrinivich and Junghoon Lee. “Artificial intelligence-based radiotherapy planning with reinforcement learning”. In: *Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling*. Ed. by Cristian A. Linte and Jeffrey H. Siewerdsen. SPIE. Feb. 2021, p. 70. URL: <http://dx.doi.org/10.1117/12.2580726>.
- [78] William Thomas Hrinivich and Junghoon Lee. “Artificial intelligencebased radiotherapy machine parameter optimization using reinforcement learning”. In: *Medical Physics* 47.12 (Nov. 2020), pp. 6140–6150. URL: <http://dx.doi.org/10.1002/mp.14544>.
- [79] Azar Sadeghnejad-Barkousaraie et al. “A reinforcement learning application of a guided Monte Carlo Tree Search algorithm for beam orientation selection in radiation therapy”. In: *Machine Learning: Science and Technology* 2.3 (May 2021), p. 035013. URL: <http://dx.doi.org/10.1088/2632-2153/abe528>.
- [80] Kit Gallagher et al. “Learning to Adapt - Deep Reinforcement Learning in Treatment-Resistant Prostate Cancer”. In: *bioRxiv* (Apr. 2023). URL: <http://dx.doi.org/10.1101/2023.04.28.538766>.
- [81] Yao Yao, Frank Youhua Chen, and Qingpeng Zhang. “Optimized patient-specific immune checkpoint inhibitors therapy for cancer treatment based on tumor immune microenvironment modeling”. In: *medRxiv* (Apr. 2024). URL: <http://dx.doi.org/10.1101/2024.04.08.24305526>.
- [82] Lu Yitao et al. “Deep reinforcement learning identifies personalized intermittent androgen deprivation therapy for prostate cancer”. In: *Briefings in Bioinformatics* (2024). URL: <https://academic.oup.com/bib/article-pdf/25/2/bbae071/56995196/bbae071.pdf>.
- [83] Wang Guangyu et al. “Optimized glyceic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial”. In: *Nature Network Boston* (2023). URL: <https://www.nature.com/articles/s41591-023-02552-9.pdf>.

- [84] Jacob M. Desman et al. “GLUCOSE: A Distributional Reinforcement Learning Model for Optimal Glucose Control After Cardiac Surgery”. In: *medRxiv* (Jan. 2025). URL: <http://dx.doi.org/10.1101/2025.01.01.25319851>.
- [85] Adnan Jafar et al. “Personalized insulin dosing using reinforcement learning for high-fat meals and aerobic exercises in type 1 diabetes: a proof-of-concept trial”. In: *Nature Communications* 15.1 (Aug. 2024). URL: <http://dx.doi.org/10.1038/s41467-024-50764-5>.
- [86] Qingnan Sun et al. “A Dual Mode Adaptive Basal-Bolus Advisor Based on Reinforcement Learning”. In: *IEEE Journal of Biomedical and Health Informatics* 23.6 (Nov. 2019), pp. 2633–2641. URL: <http://dx.doi.org/10.1109/JBHI.2018.2887067>.
- [87] Adnan Jafar, Anas El Fathi, and Ahmad Haidar. “Long-term use of the hybrid artificial pancreas by adjusting carbohydrate ratios and programmed basal rate: A reinforcement learning approach”. In: *Computer Methods and Programs in Biomedicine* 200 (Mar. 2021), p. 105936. URL: <http://dx.doi.org/10.1016/j.cmpb.2021.105936>.
- [88] Sayyar Ahmad et al. “Bolus Insulin calculation without meal information. A reinforcement learning approach”. In: *Artificial Intelligence in Medicine* 134 (Dec. 2022), p. 102436. URL: <http://dx.doi.org/10.1016/j.artmed.2022.102436>.
- [89] Jonas Nordhaug Myhre et al. “In-Silico Evaluation of Glucose Regulation Using Policy Gradient Reinforcement Learning for Patients with Type 1 Diabetes Mellitus”. In: *Applied Sciences* 10.18 (Sept. 2020), p. 6350. URL: <http://dx.doi.org/10.3390/app10186350>.
- [90] H. Lim M. et al. “A Blood Glucose Control Framework Based on Reinforcement Learning With Safety and Interpretability: In Silico Validation”. In: *IEEE Access* (2021).
- [91] Taiyu Zhu, Kezhi Li, and Pantelis Georgiou. “Personalized Dual-Hormone Control for Type 1 Diabetes Using Deep Reinforcement Learning”. In: *Studies in Computational Intelligence*. Springer International Publishing, Nov. 2020, pp. 45–53. URL: http://dx.doi.org/10.1007/978-3-030-53352-6_5.
- [92] Yamagata Taku et al. “Model-Based Reinforcement Learning for Type 1 Diabetes Blood Glucose Control”. In: *AAI4H@ECAI* (2020).
- [93] Silvia Del Giorno et al. “A New Glycemic closed-loop control based on Dyna-Q for Type-1-Diabetes”. In: *Biomedical Signal Processing and Control* 81 (Mar. 2023), p. 104492. URL: <http://dx.doi.org/10.1016/j.bspc.2022.104492>.
- [94] Wenzhou Lv et al. “Hybrid Control Policy for Artificial Pancreas via Ensemble Deep Reinforcement Learning”. In: *IEEE Transactions on Biomedical Engineering* 72.1 (Jan. 2025), pp. 309–323. URL: <http://dx.doi.org/10.1109/TBME.2024.3451712>.
- [95] Mohammad Ali Raheb et al. “Subcutaneous insulin administration by deep reinforcement learning for blood glucose level control of type-2 diabetic patients”. In: *Computers in Biology and Medicine* 148 (Sept. 2022), p. 105860. URL: <http://dx.doi.org/10.1016/j.compbimed.2022.105860>.

- [96] Elena Daskalaki, Peter Diem, and Stavroula G. Mougiakakou. “Personalized tuning of a reinforcement learning control algorithm for glucose regulation”. In: *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. July 2013, pp. 3487–3490. URL: <http://dx.doi.org/10.1109/EMBC.2013.6610293>.
- [97] Daskalaki E., Diem P., and Mougiakakou S. “Model-Free Machine Learning in Biomedicine: Feasibility Study in Type 1 Diabetes”. In: *PLoS ONE* (2016). URL: <https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0158722&type=printable>.
- [98] Xuehui Yu et al. “ARLPE: A meta reinforcement learning framework for glucose regulation in type 1 diabetics”. In: *Expert Systems with Applications* 228 (Oct. 2023), p. 120156. URL: <http://dx.doi.org/10.1016/j.eswa.2023.120156>.
- [99] Jenssen C. “Introducing Soft Option-Critic for Blood Glucose Control in Type 1 Diabetes Exploiting Abstraction of Actions for Automated Insulin Administration”. In: (2020).
- [100] Adnan Jafar et al. “Advanced decision support system for individuals with diabetes on multiple daily injections therapy using reinforcement learning and nearest-neighbors: In-silico and clinical results”. In: *Artificial Intelligence in Medicine* 148 (Feb. 2024), p. 102749. URL: <http://dx.doi.org/10.1016/j.artmed.2023.102749>.
- [101] Mohab M. H. Atanasious et al. “Deep Reinforcement Learning Control of Type-1 Diabetes with Cross-Patient Generalization”. In: *2024 32nd Mediterranean Conference on Control and Automation (MED)*. IEEE. June 2024, pp. 221–226. URL: <http://dx.doi.org/10.1109/MED61351.2024.10566186>.
- [102] Li Tianhao et al. “Electronic health records based reinforcement learning for treatment optimizing”. In: *Information Systems* (2022).
- [103] Logé Frédéric, L. Pennec E., and Amadou Boubacar Habiboulaye. “Challenging common bolus advisor for self-monitoring type-I diabetes patients using Reinforcement Learning”. In: *arXiv.org* (2020).
- [104] Harry Emerson, Matthew Guy, and Ryan McConville. “Offline Reinforcement Learning for Safer Blood Glucose Control in People with Type 1 Diabetes”. In: *Journal of Biomedical Informatics* (2022). URL: <https://arxiv.org/abs/2204.03376>.
- [105] Junyoung Yoo et al. “Intelligent Dual Basal–Bolus Calculator for Multiple Daily Insulin Injections via Offline Reinforcement Learning”. In: *IEEE Access* 12 (2024), pp. 192572–192585. URL: <http://dx.doi.org/10.1109/ACCESS.2024.3518832>.
- [106] Thomas Milton and Robert Lieck. “Fully-Automated Patient-Agnostic Diabetes Management with Deep Reinforcement Learning”. In: *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. Dec. 2024, pp. 1085–1091. URL: <http://dx.doi.org/10.1109/BIBM62325.2024.10821736>.
- [107] Zhongheng Zhang, Bin Zheng, and Nan Liu. “Individualized fluid administration for critically ill patients with sepsis with an interpretable dynamic treatment regimen model”. In: *Scientific Reports* 10.1 (Oct. 2020). URL: <http://dx.doi.org/10.1038/s41598-020-74906-z>.

- [108] Yeo Jin Kim, Markel Sanz Ausin, and Min Chi. “Multi-Temporal Abstraction with Time-Aware Deep Q-Learning for Septic Shock Prevention”. In: *2021 IEEE International Conference on Big Data (Big Data)*. IEEE. Dec. 2021, pp. 1657–1663. URL: <http://dx.doi.org/10.1109/BigData52589.2021.9671662>.
- [109] Yunho Choi et al. “Deep reinforcement learning extracts the optimal sepsis treatment policy from treatment records”. In: *Communications Medicine* 4.1 (Nov. 2024). URL: <http://dx.doi.org/10.1038/s43856-024-00665-x>.
- [110] Zeyu Wang et al. “Learning Optimal Treatment Strategies for Sepsis Using Offline Reinforcement Learning in Continuous Space”. In: *International Conference on Health Information Science (2022)*. URL: <https://arxiv.org/abs/2206.11190>.
- [111] Cristian Drudi, Maximiliano Mollura, and Riccardo Barbieri. “Analysis of Reward Formulation Based on Mean Arterial Pressure in Reinforcement Learning for Critically Ill Septic Patient”. In: *Computing in Cardiology Conference (CinC)*. Computing in Cardiology. Nov. 2023. URL: <http://dx.doi.org/10.22489/CinC.2023.383>.
- [112] F Roggeveen Luca et al. “Transatlantic transferability of a new reinforcement learning model for optimizing haemodynamic treatment for critically ill patients with sepsis”. In: *Artif. Intell. Medicine* (2020).
- [113] Jeter R. et al. “Learning to Treat Hypotensive Episodes in Sepsis Patients Using a Counterfactual Reasoning Framework”. In: *medRxiv* (2021). URL: <https://www.medrxiv.org/content/medrxiv/early/2021/03/07/2021.03.03.21252863.full.pdf>.
- [114] Pulakesh Upadhyaya et al. “1478: CHARACTERIZING SEPSIS-INDUCED HYPOTENSION PATIENTS WHO BENEFIT FROM AN EARLY VASOPRESSOR STRATEGY”. In: *Critical Care Medicine* 52.1 (Dec. 2023), S710–S710. URL: <http://dx.doi.org/10.1097/01.ccm.0001004068.30412.e1>.
- [115] Lin Tianlai et al. “A dosing strategy model of deep deterministic policy gradient algorithm for sepsis patients”. In: *BMC Medical Informatics and Decision Making* (2023). URL: <https://bmcmmedinformdecismak.biomedcentral.com/counter/pdf/10.1186/s12911-023-02175-7>.
- [116] Rastogi Anshul and Song Mentor Dongjin. “Importance Sampling to Learn Vasopressor Dosage to Optimize Patient Mortality in an Interpretable Manner”. In: *BigData Congress [Services Society]* (2024).
- [117] XiaoDan Wu et al. “A value-based deep reinforcement learning model with human expertise in optimal treatment of sepsis”. In: *npj Digital Medicine* 6.1 (Feb. 2023). URL: <http://dx.doi.org/10.1038/s41746-023-00755-5>.
- [118] Su L. et al. “Establishment and Implementation of Potential Fluid Therapy Balance Strategies for ICU Sepsis Patients Based on Reinforcement Learning”. In: *Frontiers in Medicine* (2022). URL: <https://www.frontiersin.org/articles/10.3389/fmed.2022.766447/pdf>.
- [119] "Meicheng Yang et al. “Raising High-Risk Awareness in Hemodynamic Treatment with Reinforcement Learning for Septic Shock Patients”. In: *Computing in Cardiology Conference (CinC)*. Computing in Cardiology. Dec. 2022. URL: <http://dx.doi.org/10.22489/CinC.2022.164>.

- [120] JAE-SEUNG YUN et al. “650-P: Artificial Intelligence and Individualized Optimal Glycemic Target in ICU Patients”. In: *Diabetes* 71.Supplement_1 (June 2022). URL: <http://dx.doi.org/10.2337/db22-650-p>.
- [121] Sihan Zhu and Jian Pu. “A self-supervised method for treatment recommendation in sepsis”. In: *Frontiers of Information Technology & Electronic Engineering* 22.7 (July 2021), pp. 926–939. URL: <http://dx.doi.org/10.1631/FITEE.2000127>.
- [122] Quan Zhang et al. “Artificial intelligence can use physiological parameters to optimize treatment strategies and predict clinical deterioration of sepsis in ICU”. In: *Physiological Measurement* 44.1 (Jan. 2023), p. 015003. URL: <http://dx.doi.org/10.1088/1361-6579/acb03b>.
- [123] Saria S. “Individualized sepsis treatment using reinforcement learning”. In: *Nature Network Boston* (2018). URL: <https://www.nature.com/articles/s41591-018-0253-x.pdf>.
- [124] Athanasios Tsoukalas, Timothy Albertson, and Ilias Tagkopoulos. “From Data to Optimal Decision Making: A Data-Driven, Probabilistic Machine Learning Approach to Decision Support for Patients With Sepsis”. In: *JMIR Medical Informatics* 3.1 (Feb. 2015), e11. URL: <http://dx.doi.org/10.2196/medinform.3445>.
- [125] Eghbali Niloufar, Alhanai Tuka, and Ghassemi M. “Patient-Specific Sedation Management via Deep Reinforcement Learning”. In: *Frontiers in Digital Health* (2021). URL: <https://www.frontiersin.org/articles/10.3389/fgdth.2021.608893/pdf>.
- [126] Eghbali Niloufar, Alhanai Tuka, and M. Ghassemi Mohammad. “Reinforcement Learning Approach to Sedation and Delirium Management in the Intensive Care Unit”. In: *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)* (2023).
- [127] Yeul Lee Hong et al. “Reinforcement learning model for optimizing dexmedetomidine dosing to prevent delirium in critically ill patients”. In: *npj Digit. Medicine* (2024). URL: <https://www.nature.com/articles/s41746-024-01335-x.pdf>.
- [128] Martinez D. et al. “Deep Reinforcement Learning for Optimal Critical Care Pain Management with Morphine using Dueling Double-Deep Q Networks”. In: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (2019).
- [129] Niranjani Prasad et al. “A reinforcement learning approach to weaning of mechanical ventilation in intensive care units”. In: *arXiv preprint arXiv:1704.06300* (2017).
- [130] Arne Peine et al. “Development and validation of a reinforcement learning algorithm to dynamically optimize mechanical ventilation in critical care”. In: *NPJ digital medicine* 4.1 (2021), pp. 1–12.
- [131] Peine A. et al. “Development and validation of a reinforcement learning algorithm to dynamically optimize mechanical ventilation in critical care”. In: *npj Digital Medicine* (2021). URL: <https://www.nature.com/articles/s41746-021-00388-6.pdf>.

- [132] Flemming Kondrup et al. “Towards Safe Mechanical Ventilation Treatment Using Deep Offline Reinforcement Learning”. In: *AAAI Conference on Artificial Intelligence* (2022). URL: <https://arxiv.org/abs/2210.02552>.
- [133] Joo Seung Lee, Malini Mahendra, and Anil Aswani. “Methodology for Interpretable Reinforcement Learning for Optimizing Mechanical Ventilation”. In: *arXiv.org* (2024). URL: <https://arxiv.org/abs/2404.03105>.
- [134] F Roggeveen Luca et al. “Reinforcement learning for intensive care medicine: actionable clinical insights from novel approaches to reward shaping and off-policy model evaluation”. In: *Intensive Care Medicine Experimental* (2024). URL: <https://icm-experimental.springeropen.com/counter/pdf/10.1186/s40635-024-00614-x>.
- [135] Yuyu Yuan et al. “Conservative Q-Learning for Mechanical Ventilation Treatment Using Diagnose Transformer-Encoder”. In: *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. Dec. 2023, pp. 2346–2351. URL: <http://dx.doi.org/10.1109/BIBM58861.2023.10385663>.
- [136] Zhang Bo, Qiu Xihe, and Tan Xiaoyu. “Balancing therapeutic effect and safety in ventilator parameter recommendation: An offline reinforcement learning approach”. In: *Engineering applications of artificial intelligence* (2024).
- [137] Zhu Zhichen, Du Jiang, and Chen Yixiang. “AgentMV: A Deep Reinforcement Learning Model For Mechanical Ventilation”. In: *2024 IEEE International Conference on Medical Artificial Intelligence (MedAI)* (2024).
- [138] Eghbali Niloufar, Alhanai Tuka, and M. Ghassemi Mohammad. “Distribution-Free Uncertainty Quantification in Mechanical Ventilation Treatment: A Conformal Deep Q-Learning Framework”. In: (2024).
- [139] Yin Changchang et al. “Deconfounding Actor-Critic Network with Policy Adaptation for Dynamic Treatment Regimes”. In: *Knowledge Discovery and Data Mining* (2022). URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9466407>.
- [140] Feng Xue et al. “Reinforcement Learning Model for Managing Noninvasive Ventilation Switching Policy”. In: *IEEE journal of biomedical and health informatics* (2023).
- [141] Xiuding Cai et al. “Towards Real-World Applications of Personalized Anesthesia Using Policy Constraint Q Learning for Propofol Infusion Control”. In: *arXiv* (2023). URL: <https://arxiv.org/abs/2303.10180>.
- [142] Padmanabhan Regina, Meskin N., and Haddad W. “Closed-loop control of anesthesia and mean arterial pressure using reinforcement learning”. In: *IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning* (2014).
- [143] Lin Rongmei et al. “A Deep Deterministic Policy Gradient Approach to Medication Dosing and Surveillance in the ICU”. In: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (2018).
- [144] Ribba B. et al. “Model enhanced reinforcement learning to enable precision dosing: A theoretical case study with dosing of propofol”. In: *CPT: Pharmacometrics & Systems Pharmacology* (2022). URL: <https://onlinelibrary.wiley.com/doi/pdfdirect/10.1002/psp4.12858>.

- [145] Hua Zheng et al. “Reinforcement learning assisted oxygen therapy for COVID-19 patients under intensive care”. In: *BMC Medical Informatics and Decision Making* 21.1 (Dec. 2021). URL: <http://dx.doi.org/10.1186/s12911-021-01712-6>.
- [146] Zheng Hua et al. “Reinforcement Learning Assisted Oxygen Therapy for COVID-19 Patients Under Intensive Care”. In: *BMC Medical Informatics and Decision Making* (2021). URL: <https://bmcmidinformatik.biomedcentral.com/track/pdf/10.1186/s12911-021-01712-6>.
- [147] Drudi Cristian et al. “A Reinforcement Learning Model for Optimal Treatment Strategies in Intensive Care: Assessment of the Role of Cardiorespiratory Features”. In: *IEEE Open Journal of Engineering in Medicine and Biology* (2024).
- [148] Zhang Yudong et al. “Advanced deep learning methods for biomedical information analysis: An editorial”. In: *Neural Networks* (2020).
- [149] Lee Hyeonhoon et al. “Development and validation of a reinforcement learning model for ventilation control during emergence from general anesthesia”. In: *npj Digit. Medicine* (2023). URL: <https://www.nature.com/articles/s41746-023-00893-w.pdf>.
- [150] Shirali Ali, Schubert Alexander, and Alaa A. “Pruning the Way to Reliable Policies: A Multi-Objective Deep Q-Learning Approach to Critical Care”. In: *IEEE journal of biomedical and health informatics* (2023).
- [151] Yitao Lu et al. “Deep reinforcement learning identifies personalized intermittent androgen deprivation therapy for prostate cancer”. In: *Briefings in Bioinformatics* 25.2 (Jan. 2024). URL: <http://dx.doi.org/10.1093/bib/bbae071>.
- [152] Dipesh Niraula et al. “Quantum deep reinforcement learning for clinical decision support in oncology: application to adaptive radiotherapy”. In: *Scientific Reports* 11.1 (Dec. 2021). URL: <http://dx.doi.org/10.1038/s41598-021-02910-y>.
- [153] Matteo Tortora et al. “Deep Reinforcement Learning for Fractionated Radiotherapy in Non-Small Cell Lung Carcinoma”. In: *Artificial Intelligence in Medicine* 119 (Sept. 2021), p. 102137. URL: <http://dx.doi.org/10.1016/j.artmed.2021.102137>.
- [154] Chirath Hettiarachchi et al. “G2P2C — A modular reinforcement learning algorithm for glucose control by glucose prediction and planning in Type 1 Diabetes”. In: *Biomedical Signal Processing and Control* 90 (Apr. 2024), p. 105839. URL: <http://dx.doi.org/10.1016/j.bspc.2023.105839>.
- [155] Fox Ian et al. “Deep Reinforcement Learning for Closed-Loop Blood Glucose Control”. In: *Machine Learning in Health Care* (2020).
- [156] Seunghyun Lee et al. “Toward a Fully Automated Artificial Pancreas System Using a Bioinspired Reinforcement Learning Design: In Silico Validation”. In: *IEEE Journal of Biomedical and Health Informatics* 25.2 (Feb. 2021), pp. 536–546. URL: <http://dx.doi.org/10.1109/JBHI.2020.3002022>.
- [157] Maria Cecilia Serafini, Emilia Fushimi, and Fabricio Garelli. “Reinforcement Learning Adjustment of Conventional Insulin Therapy for People with Type 1 Diabetes”. In: *2024 IEEE Biennial Congress of Argentina (ARGENCON)*. IEEE, Sept. 2024, pp. 1–6. URL: <http://dx.doi.org/10.1109/ARGENCON62399.2024.10735891>.

- [158] Zhu Taiyu, Li Kezhi, and Georgiou P. “A Dual-Hormone Closed-Loop Delivery System for Type 1 Diabetes Using Deep Reinforcement Learning”. In: *arXiv.org* (2019).
- [159] Phuwadol Viroonluecha, Esteban Egea-Lopez, and Jose Santa. “Evaluation of blood glucose level control in type 1 diabetic patients using deep reinforcement learning”. In: *PLOS ONE* 17.9 (Sept. 2022). Ed. by Pasquale Palumbo, e0274608. URL: <http://dx.doi.org/10.1371/journal.pone.0274608>.
- [160] Zihao Wang et al. “Reinforcement Learning-Based Insulin Injection Time And Dosages Optimization”. In: *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE. July 2021, pp. 1–8. URL: <http://dx.doi.org/10.1109/IJCNN52387.2021.9533957>.
- [161] Taiyu Zhu et al. “An Insulin Bolus Advisor for Type 1 Diabetes Using Deep Reinforcement Learning”. In: *Sensors* 20.18 (Sept. 2020), p. 5058. URL: <http://dx.doi.org/10.3390/s20185058>.
- [162] Taiyu Zhu et al. “Basal Glucose Control in Type 1 Diabetes Using Deep Reinforcement Learning: An *In Silico* Validation”. In: *IEEE Journal of Biomedical and Health Informatics* 25.4 (Apr. 2021), pp. 1223–1232. URL: <http://dx.doi.org/10.1109/JBHI.2020.3014556>.
- [163] Taiyu Zhu, Kezhi Li, and Pantelis Georgiou. “Offline Deep Reinforcement Learning and Off-Policy Evaluation for Personalized Basal Insulin Control in Type 1 Diabetes”. In: *IEEE Journal of Biomedical and Health Informatics* 27.10 (Oct. 2023), pp. 5087–5098. URL: <http://dx.doi.org/10.1109/JBHI.2023.3303367>.
- [164] Raghu Aniruddh et al. “Continuous State-Space Models for Optimal Sepsis Treatment: a Deep Reinforcement Learning Approach”. In: *Machine Learning in Health Care* (2017).
- [165] D. Futoma Joseph et al. “Learning to Treat Sepsis with Multi-Output Gaussian Process Deep Recurrent Q-Networks”. In: (2018).
- [166] Song Ju et al. “To Reduce Healthcare Workload: Identify Critical Sepsis Progression Moments through Deep Reinforcement Learning”. In: *2021 IEEE International Conference on Big Data (Big Data)*. IEEE. Dec. 2021, pp. 1640–1646. URL: <http://dx.doi.org/10.1109/BigData52589.2021.9671407>.
- [167] Mollura Maximiliano et al. “A Reinforcement Learning Application for Optimal Fluid and Vasopressor Interventions in Septic ICU Patients”. In: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (2022).
- [168] Mollura Maximiliano et al. “Optimal Fluid And Vasopressor Interventions In Septic ICU Patients Through Reinforcement Learning Model”. In: *2022 Computing in Cardiology (CinC)* (2022).
- [169] Dilruk Perera et al. “Smart Imitator: Learning from Imperfect Clinical Decisions”. In: *Journal of the American Medical Informatics Association* (Jan. 2025). URL: <http://dx.doi.org/10.1093/jamia/ocae320>.
- [170] Lyu R. “Improving Treatment Decisions for Sepsis Patients by Reinforcement Learning”. In: (2020).

- [171] Bologheanu Răzvan et al. “Development of a Reinforcement Learning Algorithm to Optimize Corticosteroid Therapy in Critically Ill Patients with Sepsis”. In: *Journal of Clinical Medicine* (2023). URL: <https://www.mdpi.com/2077-0383/12/4/1513/pdf?version=1676429618>.
- [172] Job Simi et al. “Optimal Treatment Strategies for Critical Patients with Deep Reinforcement Learning”. In: *ACM Transactions on Intelligent Systems and Technology* (2024).
- [173] HuanHsin Tseng et al. “Deep reinforcement learning for automated radiation adaptation in lung cancer”. In: *Medical Physics* 44.12 (Nov. 2017), pp. 6690–6705. URL: <http://dx.doi.org/10.1002/mp.12625>.
- [174] Olga Teplytska et al. “Machine Learning Methods for Precision Dosing in Anticancer Drug Therapy: A Scoping Review”. In: *Clinical Pharmacokinetics* 63.9 (Aug. 2024), pp. 1221–1237. URL: <http://dx.doi.org/10.1007/s40262-024-01409-9>.
- [175] Oroojeni Mohammad Javad Mahsa et al. “A Reinforcement Learning–Based Method for Management of Type 1 Diabetes: Exploratory Study”. In: *JMIR Diabetes* (2019). URL: <https://doi.org/10.2196/12905>.
- [176] Mahsa Oroojeni Mohammad Javad et al. “Reinforcement Learning Based Method for Managing Type 1 Diabetes (Preprint)”. In: (Nov. 2018). URL: <http://dx.doi.org/10.2196/preprints.12905>.
- [177] Gupta Akash, T. Lash Michael, and Nachimuthu S. “Optimal Sepsis Patient Treatment using Human-in-the-loop Artificial Intelligence”. In: *Expert systems with applications* (2020).
- [178] Aaron Boussina et al. “Representation Learning and Spectral Clustering for the Development and External Validation of Dynamic Sepsis Phenotypes: Observational Cohort Study”. In: *Journal of Medical Internet Research* 25 (June 2023), e45614. URL: <http://dx.doi.org/10.2196/45614>.
- [179] Grolleau F. et al. “CO10.3 - Personalizing renal replacement therapy initiation in the intensive care unit: a statistical reinforcement learning-based dynamic strategy with external validation on the AKIKI randomized controlled trials”. In: *Revue d'épidémiologie et de santé publique* (2023).
- [180] Lawrence Hunter. *Faculty Opinions recommendation of The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care*. Oct. 2018. URL: <http://dx.doi.org/10.3410/f.734264991.793552129>.
- [181] Hyewon Jeong et al. “Identifying Differential Patient Care Through Inverse Intent Inference”. In: *arXiv* (2024). URL: <https://arxiv.org/abs/2411.07372>.
- [182] Komorowski M. et al. “The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care”. In: *Nature Network Boston* (2018).
- [183] Romain Pirracchio et al. “Assessment of Machine Learning to Estimate the Individual Treatment Effect of Corticosteroids in Septic Shock”. In: *JAMA Network Open* 3.12 (Dec. 2020), e2029050. URL: <http://dx.doi.org/10.1001/jamanetworkopen.2020.29050>.

- [184] D. Futoma Joseph, A. Masood M., and Doshi-Velez F. “Identifying Distinct, Effective Treatments for Acute Hypotension with SODA-RL: Safely Optimized Diverse Accurate Reinforcement Learning”. In: *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science* (2020).
- [185] Aamer Abdul Rahman et al. “Empowering Clinicians with Medical Decision Transformers: A Framework for Sepsis Treatment”. In: *arXiv.org* (2024). URL: <https://arxiv.org/abs/2407.19380>.
- [186] Alistair EW Johnson et al. “MIMIC-III, a freely accessible critical care database”. In: *Scientific data* 3.1 (2016), pp. 1–9.
- [187] Russell Jeter et al. “Does the " Artificial Intelligence Clinician" learn optimal treatment strategies for sepsis in intensive care?” In: *arXiv preprint arXiv:1902.03271* (2019).
- [188] Omer Gottesman et al. “Evaluating reinforcement learning algorithms in observational health settings”. In: *arXiv preprint arXiv:1805.12298* (2018).
- [189] Aniruddh Raghu et al. “Deep reinforcement learning for sepsis treatment”. In: *arXiv preprint arXiv:1711.09602* (2017).
- [190] Aniruddh Raghu, Matthieu Komorowski, and Sumeetpal Singh. “Model-based reinforcement learning for sepsis treatment”. In: *arXiv preprint arXiv:1811.09602* (2018).
- [191] Xuefeng Peng et al. “Improving sepsis treatment strategies by combining deep and kernel-based reinforcement learning”. In: *AMIA Annual Symposium Proceedings*. Vol. 2018. 2018, p. 887.
- [192] Scott L Weiss et al. “Defining pediatric sepsis by different criteria: discrepancies in populations and implications for clinical practice”. In: *Pediatric Critical Care Medicine* 13.4 (2012), e219–e226.
- [193] C-C Chen et al. “Risk stratification of severe sepsis patients in the emergency department”. In: *Emergency medicine journal: EMJ* 23.4 (2006), p. 281.
- [194] Peter MC Klein Klouwenberg et al. “Likelihood of infection in patients with presumed sepsis at the time of intensive care unit admission: a cohort study”. In: *Critical care* 19 (2015), pp. 1–8.
- [195] Aldo Robles Arévalo et al. “Data-driven curation process for describing the blood glucose management in the intensive care unit”. In: *Scientific Data* 8.1 (2021), p. 80.
- [196] Jill J Hall et al. “Thiazide diuretic-induced change in fasting plasma glucose: a meta-analysis of randomized clinical trials”. In: *Journal of general internal medicine* 35 (2020), pp. 1849–1860.
- [197] Edoardo Casiglia and Valérie Tikhonoff. *Long-standing problem of β -blocker-elicited hypoglycemia in diabetes mellitus*. 2017.
- [198] Héctor Eloy Tamez-Pérez et al. “Steroid hyperglycemia: prevalence, early detection and therapeutic recommendations: a narrative review”. In: *World journal of diabetes* 6.8 (2015), p. 1073.
- [199] A Penforinis and S Kury-Paulin. “Immunosuppressive drug-induced diabetes”. In: *Diabetes & metabolism* 32.5 (2006), pp. 539–546.

- [200] Diana Grajales, Vitor Ferreira, and Ángela M Valverde. “Second-generation antipsychotics and dysregulation of glucose metabolism: beyond weight gain”. In: *Cells* 8.11 (2019), p. 1336.
- [201] Roger S McIntyre et al. “The effect of antidepressants on glucose homeostasis and insulin sensitivity: synthesis and mechanisms”. In: *Expert opinion on drug safety* 5.1 (2006), pp. 157–168.
- [202] Kimia Honarmand et al. “Society of critical care medicine guidelines on glycemic control for critically ill children and adults 2024”. In: *Critical care medicine* 52.4 (2024), e161–e181.
- [203] Janet L Kelly. “Continuous insulin infusion: when, where, and how?”. In: *Diabetes Spectrum: a Publication of the American Diabetes Association* 27.3 (2014), p. 218.
- [204] Alpesh Goyal et al. “A practical guidance on the use of intravenous insulin infusion for management of inpatient hyperglycemia: intravenous insulin infusion for management of inpatient hyperglycemia”. In: *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 15.5 (2021), p. 102244.
- [205] Philip Thomas and Emma Brunskill. “Data-efficient off-policy policy evaluation for reinforcement learning”. In: *International Conference on Machine Learning*. PMLR. 2016, pp. 2139–2148.
- [206] Masatoshi Uehara, Chengchun Shi, and Nathan Kallus. “A review of off-policy evaluation in reinforcement learning”. In: *arXiv preprint arXiv:2212.06355* (2022).
- [207] Aniruddh Raghu et al. “Behaviour policy estimation in off-policy policy evaluation: Calibration matters”. In: *arXiv preprint arXiv:1807.01066* (2018).
- [208] Yoshihiko Ozaki et al. “Multiobjective tree-structured parzen estimator for computationally expensive optimization problems”. In: *Proceedings of the 2020 genetic and evolutionary computation conference*. 2020, pp. 533–541.
- [209] Benjamin Ribba et al. “A tumor growth inhibition model for low-grade glioma treated with chemotherapy or radiotherapy”. In: *Clinical Cancer Research* 18.18 (2012), pp. 5071–5080.
- [210] Hanis Nasir and Auni Aslah Mat Daud. “Population models of diabetes mellitus by ordinary differential equations: a review”. In: *Mathematical Population Studies* 29.3 (2022), pp. 95–127.
- [211] Jinsung Yoon, Daniel Jarrett, and Mihaela Van der Schaar. “Time-series generative adversarial networks”. In: *Advances in neural information processing systems* 32 (2019).
- [212] Eoin Brophy et al. “Generative adversarial networks in time series: A systematic literature review”. In: *ACM Computing Surveys* 55.10 (2023), pp. 1–31.
- [213] Georgios Douzas and Fernando Bacao. “Effective data generation for imbalanced learning using conditional generative adversarial networks”. In: *Expert Systems with applications* 91 (2018), pp. 464–471.
- [214] Taiyu Zhu et al. “Basal glucose control in type 1 diabetes using deep reinforcement learning: An in silico validation”. In: *IEEE Journal of Biomedical and Health Informatics* 25.4 (2020), pp. 1223–1232.
- [215] Ian Fox et al. “Deep reinforcement learning for closed-loop blood glucose control”. In: *Machine Learning for Healthcare Conference*. PMLR. 2020, pp. 508–536.

- [216] Kritib Bhattarai et al. “Using artificial intelligence to learn optimal regimen plan for Alzheimer’s disease”. In: *Journal of the American Medical Informatics Association* 30.10 (2023), pp. 1645–1656.
- [217] Adrià Puigdomènech Badia et al. “Agent57: Outperforming the atari human benchmark”. In: *International conference on machine learning*. PMLR. 2020, pp. 507–517.
- [218] Caglar Gulcehre et al. “RI unplugged: A suite of benchmarks for offline reinforcement learning”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 7248–7259.
- [219] Yuxuan Jiang et al. “A Reinforcement Learning Benchmark for Autonomous Driving in General Urban Scenarios”. In: *IEEE Transactions on Intelligent Transportation Systems* (2023).
- [220] Inkyung Ahn and Jooyoung Park. “Drug scheduling of cancer chemotherapy based on natural actor-critic approach”. In: *BioSystems* 106.2-3 (2011), pp. 121–129.
- [221] Michael Oberst and David Sontag. “Counterfactual off-policy evaluation with gumbel-max structural causal models”. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 4881–4890.
- [222] A Ghaffari, B Bahmaie, and M Nazari. “A mixed radiotherapy and chemotherapy model for treatment of cancer with metastasis”. In: *Mathematical methods in the applied sciences* 39.15 (2016), pp. 4603–4617.
- [223] Chiara Dalla Man et al. “The UVA/PADOVA type 1 diabetes simulator: new features”. In: *Journal of diabetes science and technology* 8.1 (2014), pp. 26–34.
- [224] Vladimir A Kuznetsov et al. “Nonlinear dynamics of immunogenic tumors: parameter estimation and global bifurcation analysis”. In: *Bulletin of mathematical biology* 56.2 (1994), pp. 295–321.
- [225] Lisette G De Pillis and Ami Radunskaya. “The dynamics of an optimally controlled tumor model: A case study”. In: *Mathematical and computer modelling* 37.11 (2003), pp. 1221–1244.
- [226] Lisette G de Pillis and Ami Radunskaya. “A mathematical model of immune response to tumor invasion”. In: *Computational fluid and solid mechanics 2003*. Elsevier, 2003, pp. 1661–1668.
- [227] Lisette G de Pillis, Weiqing Gu, and Ami E Radunskaya. “Mixed immunotherapy and chemotherapy of tumors: modeling, applications and biological interpretations”. In: *Journal of theoretical biology* 238.4 (2006), pp. 841–862.
- [228] Boris P Kovatchev et al. *In silico preclinical trials: a proof of concept in closed-loop control of type 1 diabetes*. 2009.
- [229] Chiara Fabris, Stephen D Patek, and Marc D Breton. “Are risk indices derived from CGM interchangeable with SMBG-based indices?” In: *Journal of diabetes science and technology* 10.1 (2016), pp. 50–59.
- [230] Greg Brockman et al. *OpenAI Gym*. 2016. eprint: [arXiv:1606.01540](https://arxiv.org/abs/1606.01540).
- [231] Mark Towers et al. *Gymnasium*. Mar. 2023. URL: <https://zenodo.org/record/8127025> (visited on 07/08/2023).

- [232] Jiayi Weng et al. “Tianshou: A Highly Modularized Deep Reinforcement Learning Library”. In: *Journal of Machine Learning Research* 23.267 (2022), pp. 1–6. URL: <http://jmlr.org/papers/v23/21-1127.html>.
- [233] Takuya Akiba et al. “Optuna: A Next-Generation Hyperparameter Optimization Framework”. In: *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019, pp. 2623–2631.
- [234] Ying Liu et al. “Deep reinforcement learning for dynamic treatment regimes on medical registry data”. In: *2017 IEEE international conference on healthcare informatics (ICHI)*. IEEE. 2017, pp. 380–385.
- [235] Rahul Kidambi et al. “MOREL: Model-Based Offline Reinforcement Learning”. In: *Advances in Neural Information Processing Systems*. 2020.
- [236] Mila Nambiar et al. “Deep Offline Reinforcement Learning for Real-world Treatment Optimization Applications”. In: *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2023, pp. 4673–4684.
- [237] Yong Huang, Rui Cao, and Amir Rahmani. “Reinforcement learning for sepsis treatment: A continuous action space solution”. In: *Machine Learning for Healthcare Conference*. PMLR. 2022, pp. 631–647.
- [238] Flemming Kondrup et al. “Towards safe mechanical ventilation treatment using deep offline reinforcement learning”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 2023, pp. 15696–15702.
- [239] XiaoDan Wu et al. “A value-based deep reinforcement learning model with human expertise in optimal treatment of sepsis”. In: *npj Digital Medicine* 6.1 (2023), p. 15.
- [240] Lu Wang et al. “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*. 2018, pp. 2447–2456.
- [241] H Bryant Nguyen et al. “Early lactate clearance is associated with improved outcome in severe sepsis and septic shock”. In: *Critical care medicine* 32.8 (2004), pp. 1637–1642.
- [242] Narani Sivayoham et al. “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”. In: *Annals of medicine* 53.1 (2021), pp. 1863–1874.
- [243] Lisa Mellhammar et al. “NEWS2 is superior to qSOFA in detecting sepsis with organ dysfunction in the emergency department”. In: *Journal of clinical medicine* 8.8 (2019), p. 1128.
- [244] Sepp Hochreiter and Jürgen Schmidhuber. “Long short-term memory”. In: *Neural computation* 9.8 (1997), pp. 1735–1780.
- [245] Scott Fujimoto and Shixiang Shane Gu. “A minimalist approach to offline reinforcement learning”. In: *Advances in neural information processing systems* 34 (2021), pp. 20132–20145.
- [246] Chenjun Xiao et al. “The In-Sample Softmax for Offline Reinforcement Learning”. In: *International Conference on Learning Representations*. 2023.

- [247] Chuan Guo et al. “On calibration of modern neural networks”. In: *International conference on machine learning*. PMLR. 2017, pp. 1321–1330.
- [248] Philip Thomas, Georgios Theodorou, and Mohammad Ghavamzadeh. “High-confidence off-policy evaluation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2015.
- [249] Philip Thomas, Georgios Theodorou, and Mohammad Ghavamzadeh. “High confidence policy improvement”. In: *International Conference on Machine Learning*. PMLR. 2015, pp. 2380–2388.
- [250] Omer Gottesman et al. “Interpretable off-policy evaluation in reinforcement learning by highlighting influential transitions”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 3658–3667.
- [251] Yao Liu, Yannis Flet-Berliac, and Emma Brunskill. “Offline policy optimization with eligible actions”. In: *Uncertainty in Artificial Intelligence*. PMLR. 2022, pp. 1253–1263.
- [252] Steven J. Bradtke and Andrew G. Barto. “Linear Least-Squares algorithms for temporal difference learning”. In: *Machine Learning* (1996).
- [253] Richard S. Sutton, A. Rupam Mahmood, and Martha White. “An Emphatic Approach to the Problem of Off-policy Temporal-Difference Learning”. In: *Journal of Machine Learning Research* (2016).
- [254] Hamid Reza Maei. “Gradient temporal-difference learning algorithms”. In: *PhD thesis, University of Alberta Education and Research Archive* (2011).
- [255] Yangchen Pan, Adam White, and Martha White. “Accelerated gradient temporal difference learning”. In: *AAAI Conference on Artificial Intelligence* (2017), pp. 2464–2470.
- [256] Yangchen Pan, Erfan Sadeqi Azer, and Martha White. “Effective sketching methods for value function approximation”. In: *Conference on Uncertainty in Artificial Intelligence* (2017).
- [257] James Stephen Krinsley. “Association between hyperglycemia and increased hospital mortality in a heterogeneous population of critically ill patients”. In: *Mayo Clinic Proceedings*. Vol. 78. 12. Elsevier. 2003, pp. 1471–1478.
- [258] Moritoki Egi et al. “Hypoglycemia and outcome in critically ill patients”. In: *Mayo Clinic Proceedings*. Vol. 85. 3. Elsevier. 2010, pp. 217–224.
- [259] Greet Van Den Berghe et al. “Intensive insulin therapy in critically ill patients”. In: *New England journal of medicine* 345.19 (2001), pp. 1359–1367.
- [260] Philip A Goldberg et al. “Implementation of a safe and effective insulin infusion protocol in a medical intensive care unit”. In: *Diabetes care* 27.2 (2004), pp. 461–467.
- [261] Andrew Zale and Nestoras Mathioudakis. “Machine learning models for inpatient glucose prediction”. In: *Current diabetes reports* 22.8 (2022), pp. 353–364.
- [262] Oisín Fitzgerald et al. “Incorporating real-world evidence into the development of patient blood glucose prediction algorithms for the ICU”. In: *Journal of the American Medical Informatics Association* 28.8 (2021), pp. 1642–1650.

- [263] Jacob M Desman et al. “A distributional reinforcement learning model for optimal glucose control after cardiac surgery”. In: *npj Digital Medicine* 8.1 (2025), pp. 1–12.
- [264] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).
- [265] Edward Choi et al. “Retain: An interpretable predictive model for healthcare using reverse time attention mechanism”. In: *Advances in neural information processing systems* 29 (2016).
- [266] Boris P Kovatchev et al. “Evaluation of a new measure of blood glucose variability in diabetes”. In: *Diabetes care* 29.11 (2006), pp. 2433–2438.
- [267] Volodymyr Mnih et al. “Playing atari with deep reinforcement learning”. In: *arXiv preprint arXiv:1312.5602* (2013).
- [268] Scott Fujimoto, David Meger, and Doina Precup. “Off-Policy Deep Reinforcement Learning without Exploration”. In: *International Conference on Machine Learning*. 2019, pp. 2052–2062.
- [269] Will Dabney et al. “Distributional reinforcement learning with quantile regression”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. 1. 2018.
- [270] Xue Bin Peng et al. “Advantage-weighted regression: Simple and scalable off-policy reinforcement learning”. In: *arXiv preprint arXiv:1910.00177* (2019).
- [271] Josiah Hanna, Peter Stone, and Scott Niekum. “Bootstrapping with models: Confidence intervals for off-policy evaluation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31. 1. 2017.
- [272] Yi Su et al. “Doubly robust off-policy evaluation with shrinkage”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 9167–9176.
- [273] Jan Malte Lichtenberg et al. “Double clipping: Less-biased variance reduction in off-policy evaluation”. In: *arXiv preprint arXiv:2309.01120* (2023).
- [274] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. “Axiomatic attribution for deep networks”. In: *International conference on machine learning*. PMLR. 2017, pp. 3319–3328.
- [275] Mark Cutler and Jonathan P How. “Efficient reinforcement learning for robots using informative simulated priors”. In: *2015 IEEE international conference on robotics and automation (ICRA)*. IEEE. 2015, pp. 2605–2612.
- [276] Dhruva Tirumala et al. “Behavior priors for efficient reinforcement learning”. In: *Journal of Machine Learning Research* 23.221 (2022), pp. 1–68.
- [277] Kshitij Judah et al. “Reinforcement learning via practice and critique advice”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 24. 1. 2010, pp. 481–486.
- [278] Suhana Bedi et al. “A systematic review of testing and evaluation of healthcare applications of large language models (LLMs)”. In: *medRxiv* (2024), pp. 2024–04.
- [279] Wenxuan Wang et al. “A survey of llm-based agents in medicine: How far are we from baymax?” In: *arXiv preprint arXiv:2502.11211* (2025).

- [280] Shengnan An et al. “Make your llm fully utilize the context”. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 62160–62188.
- [281] Wenqi Shi et al. “Ehragent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records”. In: *arXiv preprint arXiv:2401.07128* (2024).
- [282] Aske Plaat et al. “Reasoning with large language models, a survey”. In: *arXiv preprint arXiv:2407.11511* (2024).
- [283] Takeshi Kojima et al. “Large language models are zero-shot reasoners”. In: *Advances in neural information processing systems* 35 (2022), pp. 22199–22213.
- [284] Jason Wei et al. “Emergent abilities of large language models”. In: *arXiv preprint arXiv:2206.07682* (2022).
- [285] Paul Hager et al. “Evaluation and mitigation of the limitations of large language models in clinical decision-making”. In: *Nature medicine* 30.9 (2024), pp. 2613–2622.
- [286] Qwen Team. “Qwen2 technical report”. In: *arXiv preprint arXiv:2412.15115* (2024).
- [287] Aaron Grattafiori et al. “The llama 3 herd of models”. In: *arXiv preprint arXiv:2407.21783* (2024).
- [288] Matthew Renze. “The effect of sampling temperature on problem solving in large language models”. In: *Findings of the association for computational linguistics: EMNLP 2024*. 2024, pp. 7346–7356.
- [289] Berivan Isik et al. “Scaling laws for downstream task performance of large language models”. In: *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*. 2024.