

# HRRT: A Holistic Renal Replacement Therapy Decision-Making Support System Using Hierarchical Reinforcement Learning

**Qianyi Xu**

National University of Singapore

**Feng Wu**

National University of Singapore

**Zi Yi Christopher Thong**

National University Hospital

**Mark Sen Liang Goh**

National University Hospital

**Pengpeng Chen**

Sir Run Run Shaw Hospital

**Jie Yang**

Sir Run Run Shaw Hospital

**Chen Huang**

Sir Run Run Shaw Hospital

**Zhongheng Zhang**

Sir Run Run Shaw Hospital

**Yucai Hong**

Sir Run Run Shaw Hospital

**Kay Choong See**

[kaychoongsee@nus.edu.sg](mailto:kaychoongsee@nus.edu.sg)

National University Hospital

**Mengling Feng**

National University of Singapore

---

**Article**

**Keywords:**

**Posted Date:** December 11th, 2025

**DOI:** <https://doi.org/10.21203/rs.3.rs-8266196/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.  
[Read Full License](#)

**Additional Declarations:** No competing interests reported.

---

---

# HRRT: A HOLISTIC RENAL REPLACEMENT THERAPY DECISION-MAKING SUPPORT SYSTEM USING HIERARCHICAL REINFORCEMENT LEARNING

---

**Qianyi Xu**

Saw Swee Hock School of Public Health  
National University of Singapore  
xuqianyi@u.nus.edu

**Feng Wu**

Saw Swee Hock School of Public Health  
National University of Singapore  
wufeng@u.nus.edu

**Zi Yi Christopher Thong**

Department of Medicine  
National University Hospital  
christopher.thong@mohh.com.sg

**Mark Sen Liang Goh**

Department of Paediatrics  
National University Hospital  
mark.goh@mohh.com.sg

**Pengpeng Chen**

Department of Emergency Medicine  
Sir Run Run Shaw Hospital  
3317054@zju.edu.cn

**Jie Yang**

Department of Emergency Medicine  
Sir Run Run Shaw Hospital  
faraday1025@zju.edu.cn

**Chen Huang**

Department of Nursing  
Sir Run Run Shaw Hospital  
3410082@zju.edu.cn

**Zhongheng Zhang**

Department of Emergency Medicine  
Sir Run Run Shaw Hospital  
zh\_zhang1984@zju.edu.cn

**Yucai Hong**

Department of Emergency Medicine  
Sir Run Run Shaw Hospital  
realhealth@zju.edu.cn

**Kay Choong See \***

Department of Medicine  
National University Hospital  
kaychoongsee@nus.edu.sg

**Mengling Feng**

Saw Swee Hock School of Public Health  
National University of Singapore  
ephfm@nus.edu.sg

## ABSTRACT

Renal replacement therapy (RRT) is a critical intervention for patients with acute kidney injury (AKI). However, clinical decision-making regarding the timing of initiation, modality selection, optimal ultrafiltration rate, and weaning criteria remains highly complex and exhibits significant practice variation. While RCTs have demonstrated that indiscriminate high-intensity RRT offers no benefit in unselected populations, they fail to guide adaptive strategies, leaving dynamic decision-making heavily reliant on empirical experience. Consequently, a substantial gap persists in delivering patient-specific recommendations, particularly for dynamically adjusting treatment in response to clinical progression. Therefore, we develop **H**ierarchical **R**einforcement Learning for **R**enal **R**eplacement **T**herapy (HRRT), a holistic clinical decision support system (CDSS) that covers the full decision-making process in RRT. 2,467 Intensive Care Unit (ICU) stays of 1,439 patients, within a cohort of patients with AKI from a US hospital, were used for training and internal testing of the model. The model's performance was evaluated on two external validation sets, where we selected 1085 ICU stays of 1085 patients from Netherlands and 1230 stays of 845 patients from China. The estimated mortality rate decreased by 6.1 percentage points from 47.7% (95%CI: 45.2 – 50.0) to 41.6% (95%CI: 35.6 - 47.2) compared to clinician-led outcomes.

---

\*Corresponding author

# 1 Introduction

Acute Kidney Injury (AKI) is common in critically ill patients and often triggers the need for Renal Replacement Therapy (RRT, also termed kidney replacement therapy). It is widely known that up to half of patients in the ICU develop AKI, and mortality rates can reach 80% in those with Stage 3 AKI [1]. While prior work has largely focused on predicting the onset of AKI, determining the optimal time to initiate RRT, and selecting the treatment modality, it has mostly produced mixed clinical benefits [7]. Early initiation of RRT has not been shown to consistently improve survival, but delayed initiation can lead to fluid overload and metabolic complications, affecting morbidity[4]. Moreover, conventional biochemical and urine output monitoring, does not reflect the complex and dynamic physiology of critically ill patients, limiting their ability to guide discontinuation of therapy[10].

Beyond decisions on treatment initiation, weaning timing, and modality selection, the debate surrounding the optimal dosing intensity for RRT remains particularly contentious. Large randomized controlled trials such as the Veteran Affairs/National Institutes of Health (VA/NIH) Acute Renal Failure Trial Network[20] and The Randomized Evaluation of Normal vs Augmentin Level (RENAL) Replacement Therapy Study[8] both showed that more intensive RRT regimens do not reduce mortality or accelerate renal recovery compared with less-intensive strategies[3]. More recently, Faria et al[5] demonstrated on the real-world ICU data that across a wide range of fluid removal rates (5-40 mL/kg/hr), there was no strong association with mortality. These findings collectively suggest that fixed ultrafiltration targets—whether high or low—do not significantly influence patient outcomes. Given the dynamic clinical course of critically ill patients, the ideal RRT dose is likely not a static value but a variable target that must be continuously adapted in response to evolving hemodynamic and metabolic conditions. The challenge lies in the cognitive burden that complex decision making places on clinicians, who must continuously synthesize high-dimensional physiological data to make precise adjustments without robust, evidence-based guidelines.

Incorporating artificial intelligence (AI) for RRT offers great potential for improving AKI management through individualized treatment strategies[21]. Among current AI technologies, Reinforcement Learning (RL) is especially suitable for healthcare domain due to its ability of sequential decision-making[17, 12]. While previous studies have explored the use of RL for recommending RRT modalities, they have not adequately addressed the dynamic adjustment of appropriate ultrafiltration targets tailored to individual patient needs[31]. Hierarchical Reinforcement Learning (HRL) is uniquely designed to decompose complex problems into multi-level decision processes and has been successfully applied to manage conditions such as multi-organ dysfunction[22]. In this study, we leverage HRL to develop a holistic decision-support framework that captures the entire therapeutic process of RRT. This process requires clinicians to sequentially determine when to initiate therapy, select the appropriate modality, establish suitable ultrafiltration goals and fluid removal rates, adjust these parameters based on the patient’s evolving condition, and ultimately decide when to discontinue treatment. We hypothesize that this comprehensive, multidimensional, and individualized approach will outperform standard clinician-driven protocols, moving beyond a “one-size-fits-all” paradigm to improve mortality outcomes in critically ill patients with AKI.

## 2 Results

### 2.1 Patients

The study includes a total of 4,782 patient stays that received RRT, derived from three datasets (Figure 1 and Table 1). The derivation cohort, sourced from the Medical Information Mart for Intensive Care (MIMIC-IV) dataset[9], comprises 2,467 stays. Patients in this cohort had a median age of 66 years (IQR 56-75), and 59.8% were male, the ICU mortality rate was 46.5%. The external validation cohort consisted of two datasets: AmsterdamUMCdb[28] and Dr.ECC. The AmsterdamUMCdb dataset, which included 1,085 patient stays, with a similar median age of 65 years (IQR 55-75) and a higher percentage of male patients (67.0%) and an ICU mortality rate 44.4%. The Dr.ECC dataset contained 1,230 stays, with an older median age of 69 years (IQR 56-78), less male patients (29.9%) and a lower ICU mortality rate of 10.5%.

Race was incorporated as a feature to enhance comparability. The MIMIC-IV dataset includes diverse ethnicity, whereas the validation dataset lacks this information, hence we assumed a single ethnicity due to data limitations. The external validation cohorts show slightly higher median blood pressure readings compared to the derivation cohort. The Dr.ECC cohort shows comparably healthier features, such as higher PaO<sub>2</sub>/FiO<sub>2</sub> ratio, urine outputs, and much lower mortality rate. The derivation cohort contains more utilization of intermittent hemodialysis (IHD) only treatment compared to the other two datasets, while AmsterdamUMCdb and Dr.ECC contain significantly more continuous renal replacement therapy (CRRT) only patients.

Table 1: Patient characteristics at ICU admission levels.

| Characteristic                                         | Derivation cohort     | External validation cohorts |                       |
|--------------------------------------------------------|-----------------------|-----------------------------|-----------------------|
|                                                        | MIMIC-IV              | AmsterdamUMCdb              | Dr.ECC                |
| <b>Summary of datasets</b>                             |                       |                             |                       |
| Enrollment period                                      | 2008–2022             | 2003–2016                   | 2017–2024             |
| No. of included ICU stays                              | 2,467                 | 1,085                       | 1,230                 |
| Location                                               | Boston                | Amsterdam                   | Hangzhou              |
| <b>Patient characteristics</b>                         |                       |                             |                       |
| Age, median (IQR), y                                   | 66 (56–75)            | 65 (55 – 75)                | 69 (56 – 78)          |
| Weight, mean (SD), kg                                  | 82.2 (15.8)           | 78.7 (18.4)                 | 73.9 (1.6)            |
| Gender, n(%)                                           |                       |                             |                       |
| Male                                                   | 1476 (59.8)           | 727 (67.0)                  | 368 (29.9)            |
| Female                                                 | 991 (40.2)            | 358 (33.0)                  | 862 (70.1)            |
| Race, n(%)                                             |                       |                             |                       |
| White                                                  | 1392 (56.4)           | NA <sup>b</sup>             | NA <sup>b</sup>       |
| Black                                                  | 424 (17.2)            | NA <sup>b</sup>             | NA <sup>b</sup>       |
| Asian                                                  | 61 (2.5)              | NA <sup>b</sup>             | NA <sup>b</sup>       |
| Others <sup>a</sup>                                    | 590 (23.9%)           | NA <sup>b</sup>             | NA <sup>b</sup>       |
| Dialysis type, n(%)                                    |                       |                             |                       |
| IHD only                                               | 750 (30.4)            | 40 (3.7)                    | 0 (0)                 |
| CRRT only                                              | 1330 (53.9)           | 981 (90.4)                  | 1230 (100)            |
| Both IHD and CRRT                                      | 387 (15.7)            | 64 (5.9)                    | 0 (0)                 |
| AKI Stage at ICU Admission, n(%)                       |                       |                             |                       |
| Stage 1                                                | 53 (2.1)              | NA <sup>b</sup>             | NA <sup>b</sup>       |
| Stage 2                                                | 30 (1.2)              | NA <sup>b</sup>             | NA <sup>b</sup>       |
| Stage 3                                                | 2384 (96.6)           | NA <sup>b</sup>             | NA <sup>b</sup>       |
| Systolic blood pressure, median (IQR), mmHg            | 110.0 (99.0 – 125.0)  | 127.0 (111.0 – 147.0)       | 128.0 (110.0 – 140.0) |
| Mean blood pressure, median (IQR), mmHg                | 71.8 (64.0 – 81.0)    | 81.3 (72.3 – 92.3)          | 82.0 (70.0 – 90.0)    |
| Diastolic blood pressure, median (IQR), mmHg           | 55.0 (47.5 – 64.0)    | 58.0 (51.0 – 67.0)          | 62.0 (50.0 – 70.0)    |
| SpO <sub>2</sub> , median (IQR), %                     | 98.0 (96.0 – 100.0)   | 98.0 (96.0 – 99.0)          | 96.0 (94.0 – 100.0)   |
| Urea, median (IQR), mmol/L                             | 13.2 (7.9 – 21.1)     | 15.0 (10.2 – 22.8)          | 15.0 (8.0 – 25.0)     |
| Creatinine, median (IQR), ummol/L                      | 327.1 (176.8 – 521.6) | 170.0 (112.0 – 348.0)       | 172.7 (108.0 – 275.3) |
| pH, median (IQR)                                       | 7.4 (7.3 – 7.4)       | 7.4 (7.3 – 7.5)             | 7.3 (7.2 – 7.4)       |
| Heart Rate, median (IQR), bpm                          | 90.2 (78.2 – 103.0)   | 90.1 (78.3 – 103.4)         | 90.2 (80.0 – 110.0)   |
| Bicarbonate, median (IQR), mmol/L                      | 25.7 (22.4 – 28.9)    | 25.7 (22.4 – 28.9)          | 23.5 (18.0 – 25.0)    |
| Potassium, median (IQR), mmol/L                        | 4.2 (3.9 – 4.5)       | 4.2 (3.9 – 4.5)             | 4.1 (3.8 – 4.6)       |
| Sodium, median (IQR), mmol/L                           | 137.0 (134.0 – 140.0) | 138.0 (135.0 – 143.0)       | 138.0 (135.0 – 142.0) |
| PaO <sub>2</sub> /FiO <sub>2</sub> ratio, median (IQR) | 216.0 (145.0 – 292.5) | 217.7 (171.1 – 274.1)       | 287.0 (180.0 – 450.0) |
| Urine Output (6hr), median (IQR), mL                   | 205.0 (75.0 – 465.0)  | 185.0 (35.0 – 595.0)        | 487.5 (400.0 – 679.2) |
| Patient Outcome, n(%)                                  |                       |                             |                       |
| ICU mortality                                          | 1148 (46.5)           | 483 (44.4)                  | 129 (10.5)            |

<sup>a</sup> Includes Asian, Native American, Pacific Islander, or missing data.<sup>b</sup> Data not recorded in the dataset.

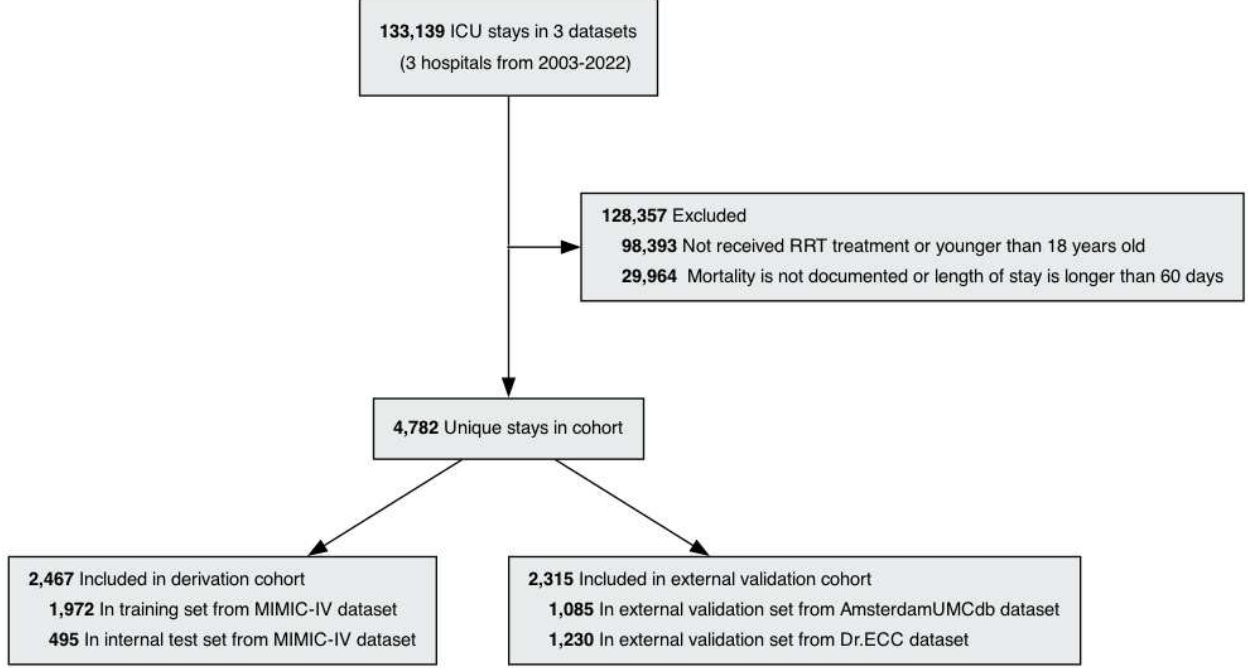


Figure 1: Flowchart of dataset construction.

## 2.2 HRRT framework

We propose the HRRT (Figure 2), a two-step hierarchical decision-making framework designed to optimize RRT decisions. Fundamentally different from monolithic models that struggle with complex hybrid action spaces, HRRT decomposes the problem into a Hierarchical Markov Decision Process (HMDP). This enables HRRT to integrate specialized offline RL algorithms at each stage, thereby effectively addressing the distinct challenges of hybrid action space associated with modality selection and fluid management task.

HRRT takes patient clinical data and historical clinician actions from the Electronic Health Records (EHR) database as input (Figure 2A). The patient data include demographics, vital signs, and lab test results, which form the state. Clinician action data include whether the patient is on dialysis for each time step, and if so, which modality. The action data also include details on IHD ultrafiltrate volumes (discrete actions) and CRRT effluent rates (continuous actions). HRRT treats the patient state trajectories as the environment. The high-level action is selecting a modality (No RRT, IHD, or CRRT), and the subsequent low-level action is selecting a specific fluid management dosage. Thus, RRT optimization is framed as an HMDP of high-level modality selection followed by low-level dosage determination.

Central to HRRT is a two-level policy represented by specialized neural networks (Figure 2B). The first step Modality Selection (Figure 2B left) involves a high-level policy ( $\pi_{\text{modality}}$ ) that selects one of three discrete options. To train this policy, we employ discrete Batch-Constrained Q-Learning (BCQ)[6] to mitigate extrapolation error, a critical failure mode in offline RL where the agent overestimates the value of out-of-distribution actions. It achieves this by using an imitation head (I) to learn the clinician’s behavior, which then creates a plausibility mask. This mask constrains the Q head (Q), forcing it to only select the best action from a safe subset of actions that are well-supported by the data, thus preventing a policy collapse.

Upon selection of a modality, HRRT activates one of two specialized, low-level agents for Step 2. Fluid Management (Figure 2B right). For the discrete-action IHD task that selects an ultrafiltrate volume, we utilize Conservative Q-Learning (CQL) [14]. CQL addresses overestimation by adding a pessimistic penalty term to the Bellman loss, pushing down the Q-values of unseen actions and stabilizing the policy. For the continuous-action CRRT task to determine an effluent rate, we employ Implicit Q-Learning (IQL)[13]. IQL avoids extrapolation error by implicitly constraining the policy to actions well-supported by the observed data. By applying the well-chosen algorithm to each component, HRRT learns a robust, multi-level policy from the static offline clinical dataset.

In the evaluation phase (Figure 2C), the performance of the full HRRT policy is compared against the clinician baseline using multiple evaluation methods, including Off-Policy Evaluation (OPE) via Weighted Importance Sampling (WIS), mortality vs. Q-value, and analyses of renal function and treatment timing.

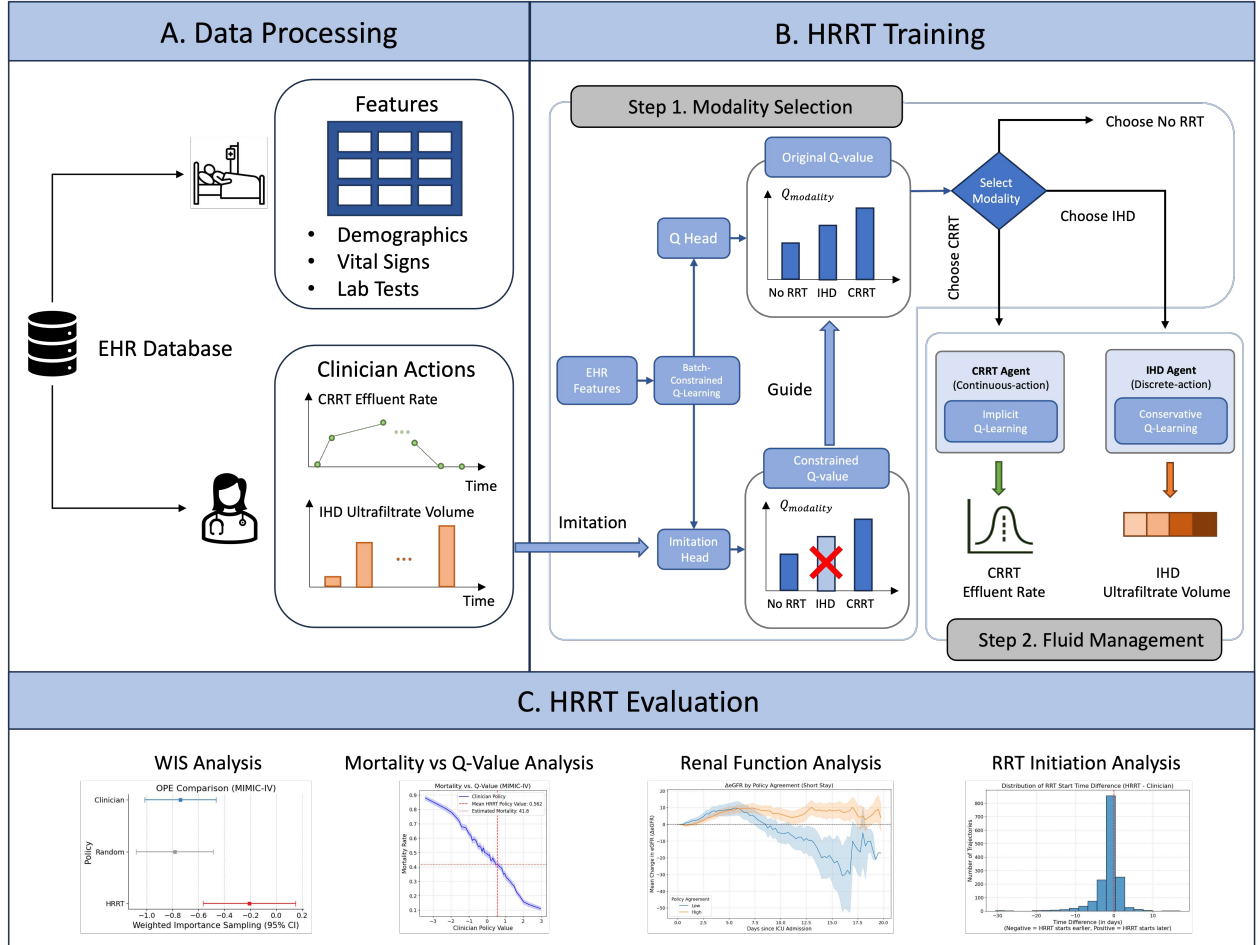


Figure 2: Overview of HRRT workflow. A. We extracted patient states and clinician actions from MIMIC-IV dataset, the IHD modality data are oversampled to mitigate the data imbalance issue for modality agent training. B. Overview of the hierarchical training process. Step 1 is the modality selection that uses BCQ to choose whether to initiate dialysis and which modality. It learns the feature representations through the baseline networks, and then predicts the q values through the q head and the i head will imitate the clinicians’ actions to mask out the less possible actions. In Step 2 Fluid Management the respective IHD or CRRT agent will learn to recommend exact fluid dosage or rate. C. Evaluation of HRRT framework.

### 2.3 HRRT shows superior performance on modality decision-making over clinicians

To evaluate the HRRT policy, we conducted a series of offline analyses comparing its performance to the historical clinician policy, as summarized in Figure 3. First, to evaluate the policy’s performance without deploying it in a live clinical setting, we employed OPE using the WIS method (Figure 3A). The WIS metric consistently confirms HRRT performance superiority across both internal and external cohorts. In the MIMIC-IV internal validation cohort, the HRRT policy achieved a substantially higher expected return of -0.21 (95%CI: -0.56 – -0.15), clearly separating its performance from the baseline clinician -0.74 (95%CI: -1.01 – -0.46) and random policies -0.78 (95%CI: -1.08 – -0.48)). This superior performance is robustly validated in the AmsterdamUMCdb external cohort. Since Dr.ECC dataset does not contain information about IHD, we only used AmsterdamUMCdb for external validation. The HRRT policy obtained a mean WIS return of -0.47 (95% CI: -0.64 – -0.29), which represents a marked improvement over the clinician policy, measured at -0.65 (95% CI: -0.83 – -0.46), and the random baseline, which performed similarly at -0.67 (95% CI: -0.83 – -0.49). The distinct separation of the HRRT confidence interval from the confidence intervals

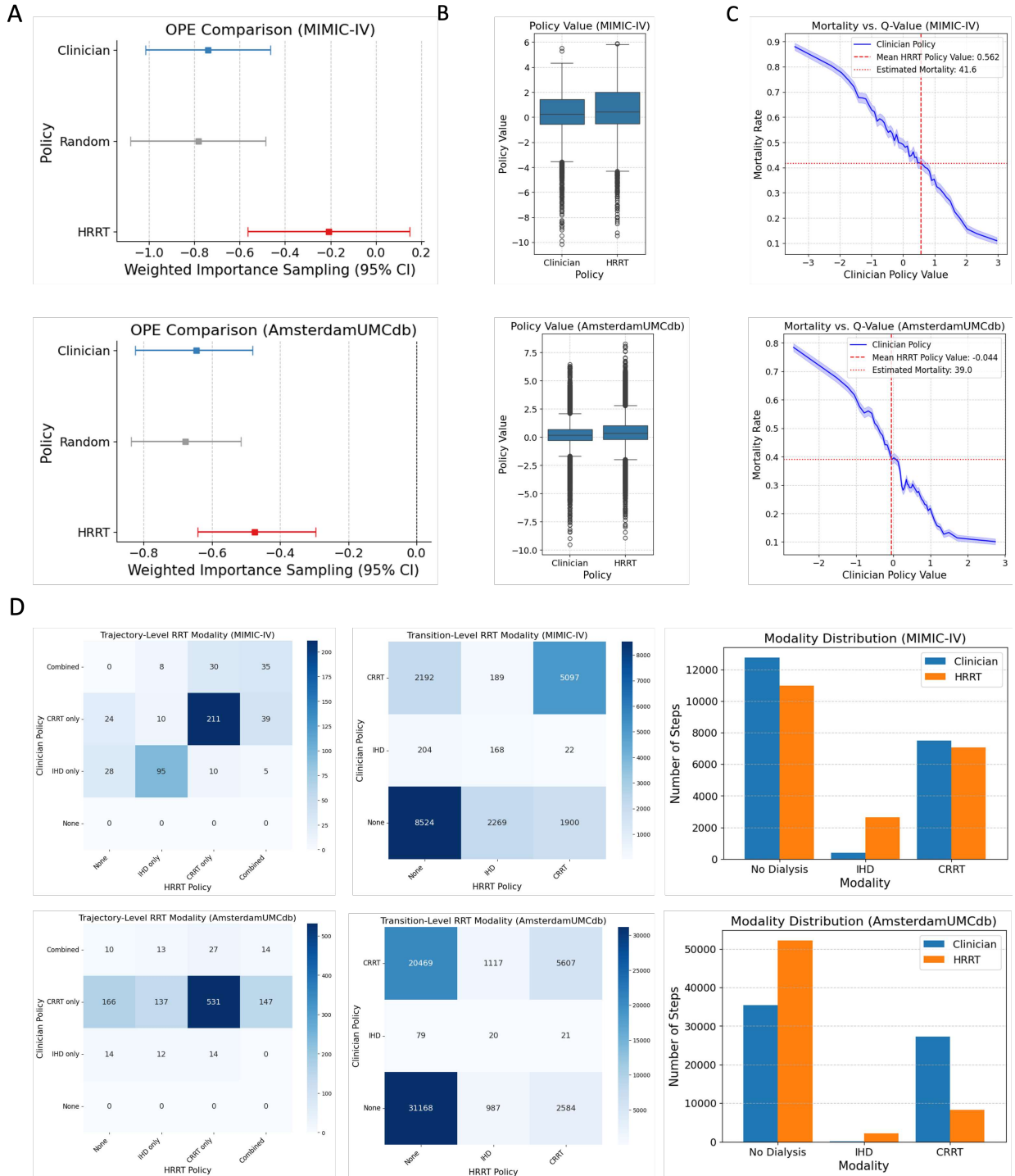


Figure 3: Performance comparison on the modality selection task for internal (MIMIC-IV) and external (AmsterdamUMCdb) validation set. A. OPE via WIS for clinician, random, and HRRT policy. B. Box plots of Q-Values where the decisions disagree between clinicians and HRRT. C. Mortality vs. Q-Value Analysis. The blue line is the clinician policy value and its corresponding mortality rate. The mean Q-value of the HRRT policy (red dashed line) maps to an estimated mortality rate. D. The modality comparison between clinician and HRRT policy on trajectory-level and transition-level and the bar plot of modality distributions.

of both baseline policies provides strong evidence that the learned policy successfully generalizes to unseen patient populations and consistently achieves higher expected long-term outcomes than the standard of care.

We further analyzed the agent’s learned Q-values (Figure 3B), which represent its belief in the long-term value of a given action. The box plot compares the Q-values of the actions chosen by the clinician versus those chosen by the HRRT policy for the exact same states where the two policies disagree. This plot reveals that in cases of disagreement, the HRRT policy consistently assigns a higher Q-value to its own recommended action despite the imitation constraint. It shows that even within the safe set of actions that the clinician also used, the agent learned to successfully distinguish the high-value safe actions from the low-value safe actions.

To validate that these Q-values are clinically meaningful, we plotted the actual mortality rate against the Q-values of the clinician’s actions (Figure 3C). This plot demonstrates a strong inverse correlation, confirming that higher Q-values are indeed associated with lower mortality. We then calculated the mean Q-value for the HRRT policy and mapped it to this curve, resulting in an estimated mortality of 41.6% (95%CI: 35.6 – 47.2) for MIMIC-IV compared to the mortality under clinicians’ action 47.7% (95%CI: 45.2 – 50.0) for the internal validation set averaging across 5 seeds. For the external validation set, we achieved an estimated mortality rate of 39.0% (95%CI: 36.7 – 45.1) compared to 44.4% (95%CI: 38.4 – 46.3). This provides strong evidence that the HRRT policy’s superior Q-values translate to tangible improvements in predicted patient survival.

Finally, we analyzed the behavioral differences between the policies (Figure 3D), examining trajectory-level and transition-level concordance, and aggregate modality distributions. First, the Trajectory-Level RRT Modality matrices (Figure 3D Left) reveal how the agent manages entire patient episodes compared to the clinician. On the MIMIC-IV training set, there is notable concordance in the CRRT only (211 trajectories) and IHD only (95 trajectories) categories, indicating the agent frequently agrees with the clinician’s choice of modality consistency. On the unseen AmsterdamUMCdb validation set, the agent demonstrates robust generalization by maintaining high concordance on the most critical cases. The policy validates the clinician’s decision in the majority of CRRT only trajectories (531 cases). In instances where the policy differs, it exercises prudent judgment by recommending IHD only (137 trajectories) or No RRT (166 trajectories), effectively identifying opportunities to minimize continuous invasive therapy and avoid potential over-treatment in an out-of-distribution setting. This de-escalation is further detailed in the Transition-Level matrices (Figure 3D Middle). In the AmsterdamUMCdb cohort, the agent demonstrates a robust, safety-driven policy shift. This indicates that while the agent may agree with the need for dialysis at the trajectory level, it often recommends pausing or terminating CRRT sessions earlier than the clinician. The Modality Distribution bar charts (Figure 3D Right) summarize these aggregate effects. In the MIMIC-IV training set, the HRRT policy is characterized by a distinct preference for IHD, showing a significantly higher volume of IHD steps compared to the clinician. In the AmsterdamUMCdb domain, the policy modulation is evident: the agent’s distribution is dominated by No RRT, effectively reducing the total volume of renal replacement therapy compared to the clinician to avoid over-optimistic interventions on out-of-distribution patients.

## **2.4 HRRT shows superior performance on fluid management over clinicians**

### **2.4.1 IHD agent**

The IHD agent demonstrates a reasonable alignment with the clinician’s discrete ultrafiltration volume levels on the MIMIC-IV training set (Figure 4A). As IHD is rarely documented in ICU EHRs and the treatment strategy varies between different regions, validation was restricted to the internal dataset. The confusion matrix reveals that the agent’s policy is concentrated around the intermediate volume levels (Level 2 and 3: 1.4L to 3L), which corresponds to common clinical practice. While the agent and clinician do not always select the exact same level, the off-diagonal elements are clustered tightly around the diagonal, indicating that the agent’s deviations are generally conservative and within one level of the clinician’s choice. The mortality analysis (Figure 4B) further supports the safety of this policy: the mortality rate remains stable when the absolute difference between the agent’s and clinician’s action level is small. A rise in mortality is only observed at the largest deviation (a difference of 3), suggesting the agent has successfully learned to mimic the low-mortality actions from the expert distribution.

### **2.4.2 CRRT agent**

The continuous-action CRRT agent shows excellent policy generalization across multiple domains. We discretized the continuous effluent rates into levels by dividing 10, where rates  $< 10$  mL/kg/hr or  $> 60$  mL/kg/hr were categorized into extreme classes. The confusion matrices for effluent rate levels (Figure 5A) display a strong diagonal correlation not only on the MIMIC-IV training set but also on the external validation sets (AmsterdamUMCdb and Dr.ECC). This confirms that the policy successfully identifies and recommends appropriate dosage levels for unseen patient states in varying hospital systems. The policy exhibits a distinct preference for Levels 2 and 3 (corresponding to 20-30

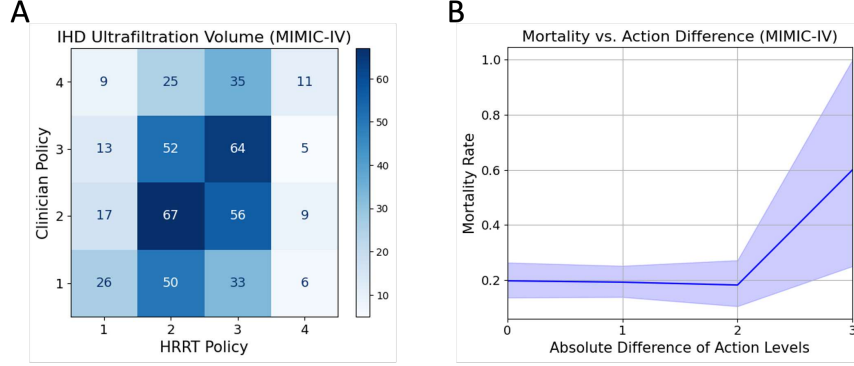


Figure 4: Comparison of treatment fluid decisions made by clinicians vs. HRRT on internal and external validation sets for IHD agent  $\pi_{IHD}$ . A. Heatmap showing the alignment between clinician-selected IHD ultrafiltration volumes (y-axis) and HRRT-recommended ultrafiltration levels (x-axis). For IHD fluid management, ultrafiltration volume was chosen as the action variable and discretized into four levels based on the observed data distribution:  $<1.4$  L,  $1.4 - 2.0$  L,  $2.0 - 3.0$  L, and  $>3.0$  L. Each cell represents the number of treatment instances corresponding to that pair. B. Mortality rate (y-axis) as a function of the absolute difference in ultrafiltration level between clinician practice and the HRRT policy (x-axis). Larger action differences indicate greater deviation from the HRRT-recommended ultrafiltration volume. Shaded regions denote the 95% confidence interval.

mL/kg/hr), aligning with standard clinical guidelines for safety and efficacy. Notably, the distribution of recommended effluent rates for the Dr.ECC cohort is shifted toward slightly higher levels (Levels 3 and 4) compared to MIMIC-IV and AmsterdamUMCdb. This distribution shift is likely attributable to the lower acuity of the Dr.ECC population. The policy’s tendency to recommend higher effluent rates in the Dr.ECC dataset reflects the agent’s sensitivity to the underlying patient severity. As the Dr.ECC cohort represents a comparatively healthier population with better physiological stability, the agent identifies that these patients can tolerate higher effluent flows for more efficient solute clearance, whereas it correctly constrains dosages in the higher-mortality MIMIC-IV and AmsterdamUMCdb cohorts to moderate solute shifts and preserve hemodynamic stability. The mortality analysis (Figure 5B) provides a critical safety insight that holds true across all three datasets: patient mortality is consistently at its lowest when the absolute difference between the HRRT policy and the clinician’s action is small (0-1 levels). As the agent’s recommended action diverges significantly from the clinician’s (differences  $>2$ ), the mortality rate trends upward. This rising mortality curve validates the learned policy, demonstrating that the agent has successfully identified a therapeutic range for CRRT dosage that correlates with the best patient outcomes.

## 2.5 Time of initiation and cessation shift between clinicians and HRRT

To evaluate the temporal characteristics of our learned HRRT policy, we compared its recommendations for dialysis initiation and cessation against the recorded clinician policy on the internal validation set. The results are visualized in Figure 6. In Figure 6A, the distribution for the HRRT policy (top) is highly concentrated within the first 2-3 days post-ICU admission, peaking at nearly 1,000 patients. This contrasts with the clinician policy (middle), which shows a broader distribution with a lower peak (around 500 patients) and a longer tail. This discrepancy is explicitly quantified in the bottom plot, the distribution is clearly shifted to the left of zero, with a mean in the negative values. This demonstrates that our learned policy is systematically more aggressive, tending to initiate RRT 1-3 days sooner than the clinician policy. Conversely, Figure 6B reveals a strong consensus between the two policies regarding the cessation of RRT. The distributions of time from ICU admission to the last dialysis stop are highly similar, both exhibiting a right-skewed distribution peaking around 4-7 days. The bottom plot, showing the distribution of the stop time difference, confirms this observation. The distribution is sharp, symmetric, and centered precisely at zero. This indicates that for the vast majority of patient trajectories, there is no systematic difference between the HRRT policy and the clinician policy on when to terminate dialysis. In summary, the primary divergence between the policies lies in treatment initiation, where the HRRT policy is notably more proactive, while both policies converge on the optimal timing for treatment conclusion. The time of initiation and cessation shows similar pattern on the external validation set; we report the results in Supplementary Figure11.

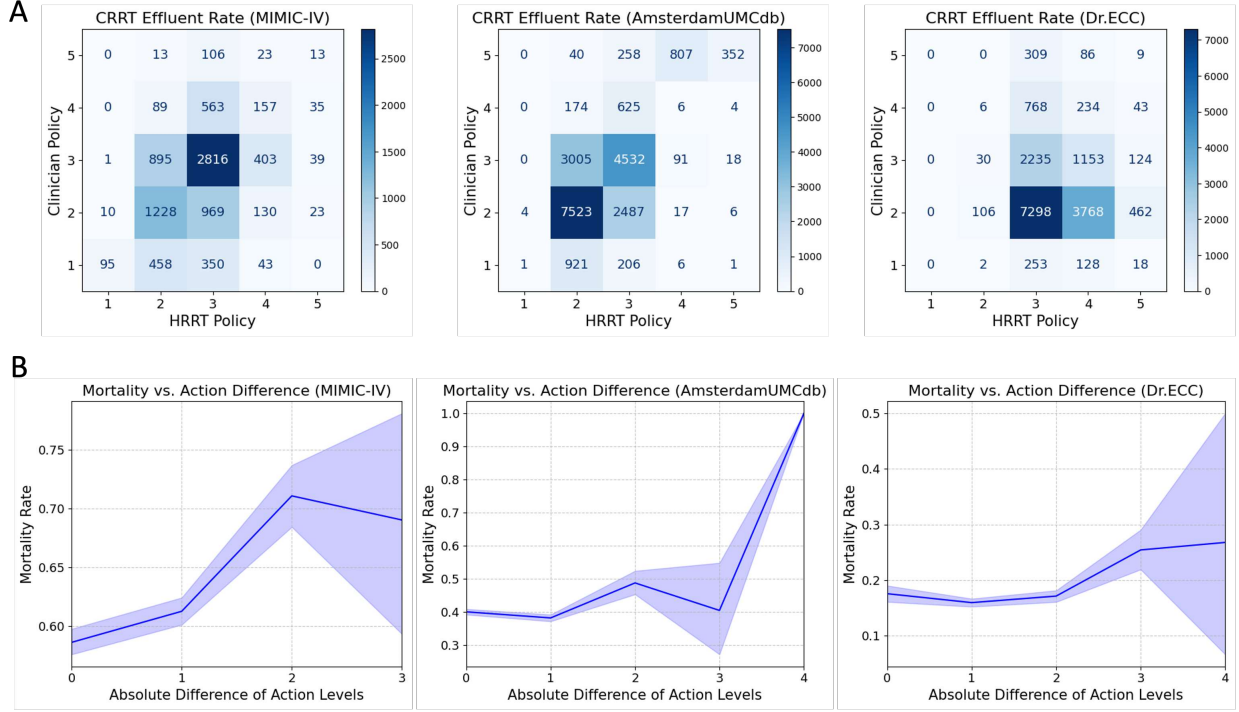


Figure 5: Comparison of treatment fluid decisions made by clinicians vs. HRRT on internal and external validation sets for CRRT agent  $\pi_{CRRT}$ . A. Heatmaps showing the agreement between clinician-chosen CRRT effluent rates (y-axis) and HRRT-recommended effluent rates (x-axis) across three datasets (MIMIC-IV, AmsterdamUMCdb, Dr.ECC). Continuous effluent rates were discretized into 10 levels; values  $<10$  mL/kg/hr or  $>60$  mL/kg/hr were grouped into extreme bins. Each cell indicates the number of treatment decisions falling into the corresponding (clinician level, HRRT level) pair. B. Relationship between the absolute action difference between clinician and HRRT policies (x-axis) and the observed mortality rate (y-axis). The x-axis reflects how far clinical practice deviates from the HRRT-recommended effluent rate level; the y-axis shows the corresponding mortality among those treatment instances. Shaded regions represent 95% confidence intervals.

## 2.6 Subgroup analysis based on length of ICU stay

To assess the policy’s effect on different patient cohorts, we performed a subgroup analysis based on the length of ICU stay (Figure 7), which serves as a proxy for illness severity. We stratified patients from the internal validation cohort into short stay ( $<20$  days, low severity, Figure 7A) and long stay ( $>20$  days, high severity, Figure 7B) groups. For each group, we evaluated the mean change in estimated glomerular filtration rate ( $\Delta eGFR$ ) [16] over time. We calculated the agreement which is the percentage of HRRT recommended actions that are the same as clinician actions along the trajectory. We plotted the  $\Delta eGFR$  against time for the trajectories with high versus low agreement. In the Short Stay cohort (Figure 7A), a clear divergence in renal function is observed. Patients with high policy agreement demonstrate a consistent improvement in renal function. The mean  $\Delta eGFR$  becomes positive within the first few days and stabilizes at approximately  $+10$  mL/min/1.73m<sup>2</sup> for the duration of the stay. Conversely, the group with low agreement shows an initial slight improvement followed by a steady decline in renal function after day 5. The mean  $\Delta eGFR$  trends progressively negative, indicating worsening renal function. This suggests that for low-severity patients, adherence to the policy is associated with renal recovery, while non-adherence is associated with renal function deterioration.

The Long Stay cohort (Figure 7B) shows a different temporal pattern due to patient severity, where the high agreement group maintains a stable renal function, with the mean  $\Delta eGFR$  hovering around zero for the first 40 days, followed by a trend towards improvement in the later stages of the ICU stay. Low policy agreement group also maintains a stable renal function for the first 40-43 days, mirroring the high-agreement group. However, after this point, these patients experience a precipitous drop in eGFR, signifying a rapid and severe worsening of renal function. For high-severity patients, the protective effect of policy adherence becomes most evident in the very late stages of critical illness, where non-adherence is linked to a severe decline in renal function. Across both low- and high-severity subgroups, high agreement with the learned HRRT policy is associated with superior renal function outcomes—either improvement

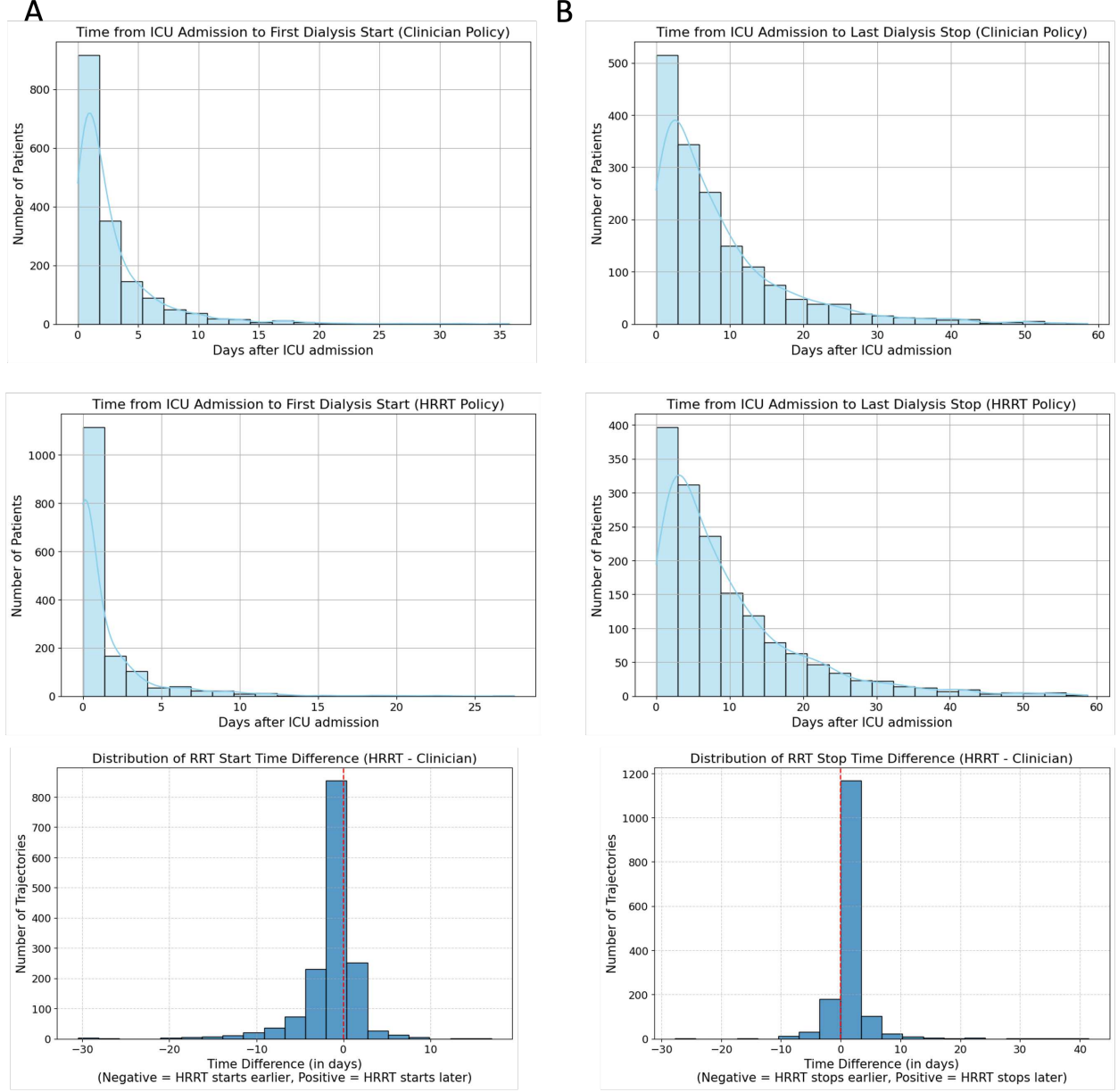


Figure 6: Temporal validation of RRT initiation and termination on the internal MIMIC-IV cohort. A. Time from ICU admission to the initiation of the first RRT session. The histograms compare the temporal distributions of the Clinician (top) and HRRT policy (middle). The difference distribution (bottom) shows the trajectory-level time difference ( $t_{HRRT} - t_{Clinician}$ ), where negative values indicate the HRRT policy initiates treatment earlier than the clinician. B. Time from ICU admission to the termination of the last RRT session. The difference histogram (bottom) demonstrates high concordance between the agent and clinician regarding the cessation of therapy, centered tightly around zero.

(Short Stay) or maintenance followed by recovery (Long Stay). In contrast, low policy agreement is consistently associated with renal function deterioration, a decline that is particularly abrupt and severe in the long-stay, high-severity cohort.

## 2.7 Sensitivity analysis

To evaluate the robustness of our reinforcement learning framework, we performed a sensitivity analysis regarding the discretization of the decision time step. Our primary analysis utilizes a 6-hour interval, a choice motivated by the

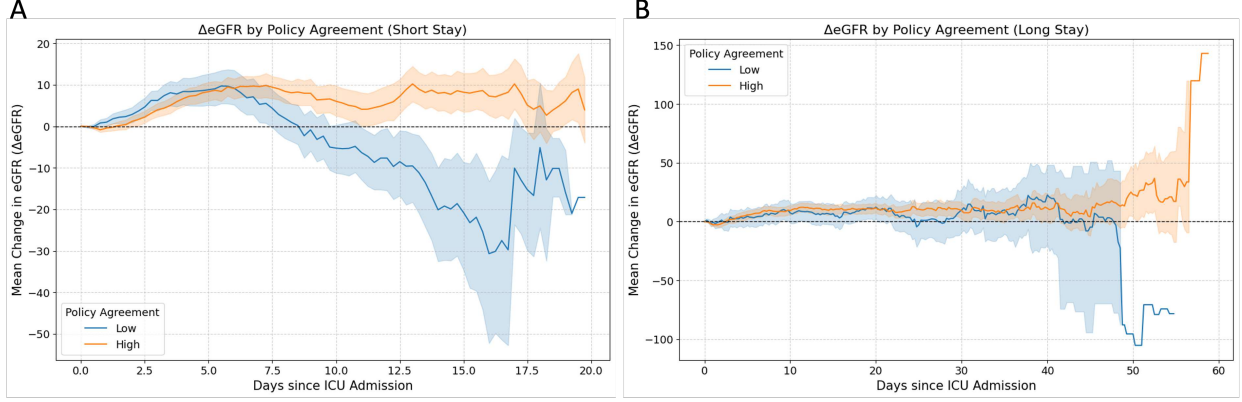


Figure 7: Subgroup analysis of mean  $\Delta eGFR$  based on ICU stay length (severity). A. Short Stay (Low Severity) cohort. B. Long Stay (High Severity) cohort. Blue and orange lines represent patients with low and high policy agreement, respectively. Shaded areas represent the 95% confidence intervals.

granularity of urine output measurements in the MIMIC-IV dataset and the physiological reality that IHD sessions frequently exceed 4 hours. We excluded a 1-hour interval as it necessitates an unrealistically high frequency of re-evaluation for renal replacement therapy, which is practically infeasible in a clinical setting. To test for stability, we retrained and evaluated the policy using a 4-hour interval (Suppl. Figure 12). The results indicate that the policy’s performance is robust to this variation. The WIS for the 4-hour interval confirms that the HRRT policy ( $-0.27$ , 95% CI:  $-0.57 - 0.03$ ) continues to significantly outperform the clinician policy ( $-1.10$ , 95% CI:  $-1.36 - -0.85$ ). Furthermore, the behavioral characteristics—including the strong negative correlation between Q-values and mortality (Figure 12C) and the preference for IHD in the modality distribution (Figure 12D) remain consistent with the 6-hour baseline, demonstrating that the learned policy is stable and not an artifact of the specific temporal discretization.

## 2.8 Reward function design calibration

Since the reward signal implicitly defines the agent’s objective, its careful design is essential for effectively guiding the policy learning toward clinically optimal strategies. Previous sensitivity analyses in medical RL have highlighted that learned policies can be dangerously sensitive to reward formulations[19]. Past papers on treatment recommendation typically consider two primary settings: terminal(long-term) rewards based solely on patient outcomes[27], or combined with intermittent (short-term) rewards that integrate patient state changes[12, 18]. Therefore, we conducted experiments that compare mortality only rewards  $R_{mort}$  and a mixed reward  $R_{mix}$  using Sequential Organ Failure Assessment (SOFA)[15] score. We selected the SOFA score as the basis for the intermediate reward because it provides a holistic evaluation of patient status across multiple organ systems, reflecting genuine physiological improvement rather than simple solute clearance. Critically, we excluded direct filtration markers such as urea and creatinine from the reward function. Since the serum concentrations of these small solutes are mechanically controlled by the CRRT settings, including them would introduce a spurious feedback loop. This analysis allows us to understand how nuanced treatment effects and outcomes are related to the applicability of our results. The reward functions are defined as follows:

$$R_{mort} = \begin{cases} +a, & \text{if } s_{t+1} \in S_T \cap S_{sur} \\ -a, & \text{if } s_{t+1} \in S_T \cap S_{dec} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

$$R_{mix} = \begin{cases} +a, & \text{if } s_{t+1} \in S_T \cap S_{sur} \\ -a, & \text{if } s_{t+1} \in S_T \cap S_{dec} \\ R_{inter}(s_t, s_{t+1}), & \text{otherwise} \end{cases} \quad (2)$$

$$R_{inter} = -c \cdot (SOFA_{t+1} - SOFA_t) - b \cdot \mathbb{I}(SOFA_{t+1} = SOFA_t > 0) \quad (3)$$

where  $S_T$  means set of terminal states,  $S_{sur}$  means the set of survived patients and  $S_{dec}$  for deceased patients.  $a$ ,  $b$ ,  $c$  are hyperparameters of weights for each component. In our analysis, we explored various scenarios to assess the impact of different reward structures

- The effect of parameter  $a$  on  $R_{mort}$ .
- The effect of parameters  $b, c$  on  $R_{mix}$ .
- The comparison between  $R_{mort}$  and  $R_{mix}$ .

### 2.8.1 The effect of parameter $a$ on $R_{mort}$

First, we assessed the effect of the terminal reward magnitude by comparing  $R_{mort}$  with  $a = 10$  and  $a = 100$ . As shown in Table 2, the WIS values for the static baselines (Clinician and Random) scale linearly with the reward parameter (exactly 10 times), as these policies are fixed evaluations of the test data. In contrast, the HRRT policy scales approximately, but not linearly (ratio  $\approx 8.6$ ). This divergence occurs because HRRT is a learned policy parameterized by a neural network; increasing the reward scale alters the magnitude of the Q-values during training. Due to the non-linear Softmax activation in the BCQ imitation head, larger Q-values result in a sharper, more deterministic probability distribution compared to the  $a = 10$  setting. Consequently, the learned policy  $\pi_\theta$  varies slightly between the two settings. Despite this numerical variance in scaling, the policy’s relative advantage remains robust. The  $a = 10$  setting achieved a 19.09% relative improvement in WIS, while the  $a = 100$  setting achieved a 2.03% improvement. Importantly, both reward scales drove the agent toward a similar mortality reduction (13.42% and 11.74%, respectively), confirming that the optimal policy is stable regardless of the absolute magnitude of the terminal reward  $a$ . Our results suggest that for a mortality-only objective, the HRRT policy is robust to the absolute scale of the terminal reward signal.

### 2.8.2 The effect of parameters $b, c$ on $R_{mix}$

We further analyzed the relative weights of the intermediate reward parameters  $b$  and  $c$  (Table 2). Operating under the prerequisite that the terminal reward must significantly outweigh cumulative intermediate signals to prioritize patient survival, we first tested the uncalibrated setting where  $b = c = 1$ . As expected, this equal-weight configuration proved misaligned with the clinical objective. The large intermediate weights create a cumulative penalty that effectively overpowers the terminal reward signal ( $a = 10$ ). Consequently, the agent pathologically optimizes for these intermediate proxies at the expense of the primary survival goal. This misalignment explains the discrepancy in our results: while the policy achieved a substantial relative WIS improvement (+40.03%), this optimization translated into a negligible reduction in mortality (4.19%). Subsequently, we calibrated the parameters to set the expected total intermediate reward  $|R_{intermediate}| \approx 2.0$ , utilizing a theoretical analysis based on average trajectory lengths (Suppl. Section D). By ensuring the intermediate signal remains secondary to the survival goal, we achieved the most robust policy performance. This configuration yielded the highest relative WIS gain (+71.62%) and, crucially, the most significant reduction in patient mortality (12.79%). These findings confirm that calibrated intermediate rewards effectively guide the agent toward safer, survival-oriented strategies without distracting from the ultimate therapeutic objective.

### 2.8.3 Comparison Between $R_{mort}$ and $R_{mix}$

Finally, we evaluated the trade-off between pure outcome optimization ( $R_{mort}, a = 10$ ) and holistic patient management ( $R_{mix}, b = 0.033, c = 0.25$ ) (Table 2). Theoretically, the policy optimizing exclusively for the terminal signal should achieve the absolute minimum mortality rate. Our empirical results align with this theory: the  $R_{mort}$  policy achieved a marginally lower estimated mortality (41.3%) compared to the optimal  $R_{mix}$  (41.6%). Despite this slight numerical advantage for the mortality-only policy, we posit that the calibrated  $R_{mix}$  represents the superior clinical strategy. In healthcare, the therapeutic objective is not merely survival, but the maintenance of patient well-being and the minimization of organ failure during the ICU stay. The calibrated  $R_{mix}$  achieves a mortality reduction comparable to the theoretical optimum (12.79% vs 13.42%) but does so by actively minimizing the SOFA score at every time step. This aligns with recent findings that intermediate physiological guidance can produce more robust policies by ensuring the agent maintains the patient in a stable condition throughout the trajectory, rather than potentially allowing physiological degradation to satisfy a sparse terminal signal [32].

## 3 Discussion

In this multicohort retrospective validation study spanning three international ICUs, we developed and evaluated a hierarchical reinforcement learning framework to provide end-to-end management and optimization of RRT for ICU patients with AKI, including initiation and weaning timing, modality selection, and dynamic adjustment of ultrafiltration. The model was trained and internally validated using the MIMIC-IV database, with external validation performed on two independent cohorts. Together, these three datasets represented 4,782 stays of more than 3,000 critically ill patients across different countries and health systems, enabling robust evaluation of model generalizability. As seen in Figure 3–5, HRRT achieved better performance compared to the clinician policy on both the internal and external validation sets for all the tasks.

Table 2: Analysis on different reward function weights with relative OPE metric changes (%). The percentage for WIS is the relative improvement over the clinician, while for Mortality it is the relative reduction.

| Reward                                 | Method    | WIS (% $\uparrow$ ) | Mortality (% $\downarrow$ ) |
|----------------------------------------|-----------|---------------------|-----------------------------|
| $R_{mort}(a = 100)$                    | Clinician | 5.91                | 47.7                        |
|                                        | Random    | 3.73                | –                           |
|                                        | HRRT      | 6.03 (2.03)         | 42.1 (11.74)                |
| $R_{mort}(a = 10)$                     | Clinician | 0.59                | 47.7                        |
|                                        | Random    | 0.37                | –                           |
|                                        | HRRT      | 0.70 (18.64)        | 41.3 (13.42)                |
| $R_{mix}(a = 10, b = 1, c = 1)$        | Clinician | -27.13              | 47.7                        |
|                                        | Random    | -22.62              | –                           |
|                                        | HRRT      | -16.27 (40.03)      | 45.7 (4.19)                 |
| $R_{mix}(a = 10, b = 0.033, c = 0.25)$ | Clinician | -0.74               | 47.7                        |
|                                        | Random    | -0.78               | –                           |
|                                        | HRRT      | -0.21 (71.62)       | 41.6 (12.79)                |

Using the HRRT framework to provide a customized dose of RRT can improve outcomes through several distinct mechanisms. Firstly, the temporal analysis (Figure 6 and Suppl. Figure 11) reveals a distinct shift in treatment duration. The HRRT policy tends to initiate therapy earlier than the clinician, suggesting a proactive approach to prevent toxic accumulation before it becomes critical. Conversely, the policy tends to terminate treatment later, ensuring that renal support is maintained until physiological stability is fully consolidated, thereby reducing the risk of premature weaning and subsequent relapse. Secondly, the policy demonstrates a context-aware selection of modality (Figure 3). Rather than adhering to a rigid protocol, the agent promotes the use of IHD when efficient clearance is feasible while strictly limiting CRRT in sensitive or out-of-distribution scenarios to maximize patient safety. Thirdly, from Figure 5-5, we demonstrated that individualized effluent dosing better matches metabolic demands. This dynamic adjustment avoids the risks of underdialysis while simultaneously preventing the hemodynamic instability often caused by aggressive, static overdialysis.

Beyond optimizing treatment timing and selecting therapy modalities, our findings highlight the potential of RL to move beyond the long-standing debate of the optimal dose of RRT towards individualized and dynamic treatment strategies. Large randomized controlled trials such as the VA/NIH acute renal failure trial[20] and RENAL study[8] have already established that higher-intensity dosing did not improve mortality compared with standard therapy. More recently, real world analysis did not demonstrate a consistent dose response relationship across a range of effluent intensities. These results demonstrate that there is no universal dosing threshold and underscores the importance of individualized titration. However, current practice is heterogeneous, usually guided by clinician experience, institutional protocols, or even isolated laboratory values rather than the illness trajectory. The HRRT framework aims to address this gap by learning sequential treatment policies that account for evolving patient physiology, with the goal of aligning the dose delivery of RRT closer to the principles of individualized care. This study reports the first use of applying RL to RRT dosing to sharpen clinical decision-making for ICU acquired AKI patients.

The HRRT model employs a hierarchical decision-making approach that integrates the identification of key decision nodes with personalized adjustments based on the dynamic physiological evolution of patients during RRT. This framework addresses a critical limitation in most clinical AI applications: the fragmentation of decision support. AI systems that provide decision support only at isolated points in the continuously evolving patient journey may undermine trust among clinicians and patients. In contrast, HRRT offers an end-to-end coverage, evaluating all potentially eligible patients and guiding every relevant decision phase. Consequently, our model advances beyond narrow-scope applications to establish a comprehensive, full-spectrum CDSS for continuous patient management.

A notable strength of this study is the comprehensive validation using three independent datasets across North America, Europe, and Asia. Despite significant heterogeneity in baseline demographics, practice patterns, and resource availability, the HRRT policy consistently outperformed clinician-observed policies in off-policy evaluations. The model demonstrated robust transferability, effectively generalizing across cohorts with distinct modality distributions: a balanced mix of IHD and CRRT in the MIMIC-IV training set versus a predominantly CRRT-driven practice in the AmsterdamUMCdb and Dr.ECC cohort. Consequently, while local calibration and contextual adaptation remain important for implementation, the underlying decision-making framework is transferable across diverse global populations.

The interpretability analysis using SHAP (Figure 9) further validates the clinical relevance of the HRRT policy’s decision-making logic. While the clinician’s policy heavily emphasizes fluid balance metrics such as urine output rate and serum sodium, reflecting traditional indicators for volume management, the HRRT agent demonstrates a sophisticated prioritization of metabolic markers. Specifically, the agent identifies creatinine, urea, and bicarbonate as the most critical features driving intervention. This prioritization aligns with the pathophysiological goals of renal replacement therapy: the clearance of uremic toxins and the correction of metabolic acidosis. By elevating these solute-based indicators, the reinforcement learning model effectively targets the underlying metabolic derangements that necessitate renal support, suggesting that its policy is grounded in robust, physiologically sound signals essential for patient survival and recovery.

The primary limitation of this study lies in its retrospective design based on offline data. Although we employed robust off-policy evaluation methods to estimate policy value, these techniques cannot definitively establish efficacy. Consequently, future work is essential to bridge the gap between algorithmic validation and clinical deployment. Our next steps focus on prospective validation through RCTs to rigorously confirm the safety and tangible clinical benefits of the HRRT model in real-time decision-making[29].

## 4 Methods

### 4.1 Database and cohort construction

For the derivation cohort, the data for training and internal validation were sourced from the public MIMIC-IV v3.1 database[9], a comprehensive, de-identified health dataset for 94,458 stays of 364,627 patients collected from the ICUs at the Beth Israel Deaconess Medical Center from 2008-2022. We included adult patients ( $\geq 18$  years old) who received IHD or CRRT treatment during their ICU stay (length of stay should be less than 60 days) and mortality was documented. For external validation, we used 2 datasets, AmsterdamUMCdb[28] and Dr.ECC. AmsterdamUMCdb contains 23,106 admissions of 20,109 patients admitted to Amsterdam University Medical Center from 2003 to 2016. AmsterdamUMCdb is a public critical care database comprising de-identified health data from patients admitted to the ICUs of the Amsterdam University Medical Centers in the Netherlands. Dr.ECC is a non-public database of emergency critically ill patients led and established by Sir Run Run Shaw Hospital that includes more than 90,000 patients with emergency visit records from 2017 to 2024, covering data of patients from outpatient and emergency visits, hospitalization, to intensive care unit treatment.

### 4.2 Data extraction and preprocessing

Details of the cohort selection process are documented in Figure1. 133,139 ICU stays from three datasets were evaluated for potential inclusion. After applying the exclusion criteria, 4,782 unique patient stays were included in our study, there were 2,467 stays for MIMIC-IV dataset where we used a 8:2 split for training and internal validation set. 2,315 stays were included for external validation in total, where 1,085 stays were from AmsterdamUMCdb and 1,230 stays were included in Dr.ECC. Each stay was considered as a single patient trajectory and we used 6-hour interval to discretize the data. The detailed sensitivity analysis about the binning interval is detailed in 2.7. Each patient stay is described by 18 clinical features (Suppl. Table4), which encompasses patient demographics, vital signs, lab test results, and urine outputs. For IHD fluid management, we selected the ultrafiltration volume as the action variable and discretized it into four levels based on the data distribution ( $< 1.4$  L,  $1.4-2.0$  L,  $2.0-3.0$  L, and  $> 3.0$  L). For CRRT fluid management, we calculated the effluent rate based on CRRT subtypes, the definition of effluent rate for different subtypes is in Suppl. Table5. The missing rates for features are reported in Suppl. Table3. To handle missing features, we employed forward and backward filling for general features within the same ICU stay. For fluid dosages, we imputed values based on RRT sessions, defining a new session whenever the gap between two dialysis records exceeded 8 hours.

### 4.3 Hierarchical markov decision process

Standard Markov Decision Processes (MDPs) model tasks satisfying the Markov property[23] can struggle with long-horizon problems or complex, multi-stage decisions. To address this, we formulate our problem as a HMDP, which decomposes the decision-making process into multiple levels of policies[26]. In our two-level hierarchy, a top-level policy,  $\pi_h$ , does not select primitive actions. Instead, it selects from a set of options or sub-policies,  $o \in \mathcal{O}$ . An option is a complete policy in itself that runs for multiple time steps until it reaches a termination condition. Once an option  $o$  is selected, a lower-level policy,  $\pi_o(a | s)$ , executes the primitive actions  $a$  at each subsequent time step. This framework allows the agent to reason at a higher level of abstraction. The top-level policy learns the long-term strategy of which sub-policy to invoke, while the lower-level policies learn the fine-grained, step-by-step actions required to implement that sub-policy. Formally, the environment’s dynamics are still defined by the primitive actions, where  $p(s', r | s, a)$  is

the probability of transitioning to state  $s'$  with reward  $r$  after taking primitive action  $a$  in state  $s$ :

$$p(s', r | s, a) = \Pr\{S_{t+1} = s', R_{t+1} = r | S_t = s, A_t = a\}. \quad (4)$$

However, the agent’s policies operate at different levels of this hierarchy to manage the complexity of the task. The MDP components for our study are defined as follows:

- **Policy  $\pi$ :** A policy defines the learning agent’s behavior by mapping the perceived environmental state to a specific action. Given the complexity of determining both the therapeutic modality and the precise dosage, we formulated the task using a hierarchical structure. The decision-making burden is distributed across three distinct sub-policies: a high-level modality policy  $\pi_{modality}$ , and two low-level policies  $\pi_{IHD}$  and  $\pi_{CRRT}$  respectively.
- **State  $S$ :** At each time step  $t$ , the agent receives a state vector  $s_t$  representing the patient’s clinical condition. To facilitate the optimization of a dynamic treatment regime, this representation incorporates patient demographics, vital signs, laboratory results, and other relevant biomarkers [11, 19]. The complete list of features is detailed in Table 4, this feature set remains consistent across all hierarchical subtasks.
- **Action  $A$ :** At each time  $t$ , the agent executes an action  $a$ . We utilize a hybrid action space: actions for the modality and IHD policies are discrete, whereas the CRRT policy action is continuous. Specifically, the modality policy selects from {No RRT, IHD, CRRT}; the IHD policy selects from four discrete levels of ultrafiltration volume; and the CRRT policy outputs a continuous effluent flow rate.
- **Reward  $r$ :** The reward signal provides feedback to the agent, defining positive and negative therapeutic events. Our framework incorporates both short-term rewards, which provide immediate feedback on physiological state changes, and long-term rewards associated with patient outcomes. The methodology for calibrating this reward function is elaborated in Section 2.8.

#### 4.4 Analysis on oversampling

A significant challenge in the dataset was the severe class imbalance within the dialysis action space. In the original data, IHD was recorded in only 2% of observed actions. This sparsity presents a major obstacle for offline RL algorithms like BCQ, which rely on sufficient data coverage to estimate values and constrain the policy to the behavioral distribution. When trained on the original imbalanced data (Figure 8 No-Oversampling), the model failed to learn the value of IHD entirely, resulting in a degenerate policy that never recommended IHD. To address this, we employed an oversampling strategy to incrementally increase the prevalence of IHD samples (4%, 6%, 8%, and 10%) in the training buffer. At lower ratios (4% and 6%), the agent began to identify IHD but remained hesitant. For example, at 6%, the agent matched the clinician’s IHD decision in only 83 cases. Conversely, aggressive oversampling at 10% caused the policy to deviate significantly from clinical norms, recommending IHD in 3,350 instances where the clinician deemed no RRT necessary, an unrealistic and potentially unsafe escalation of care. Since the objective of BCQ is to optimize decisions within a safe, clinically plausible range, we identified 8% as the optimal configuration. In this setting, the agent effectively learned the relative value of IHD, achieving a balanced distribution that improved recognition of IHD needs without the excessive false positives observed at higher ratios. This balanced policy yielded the highest estimated WIS and a significantly improved estimated mortality rate.

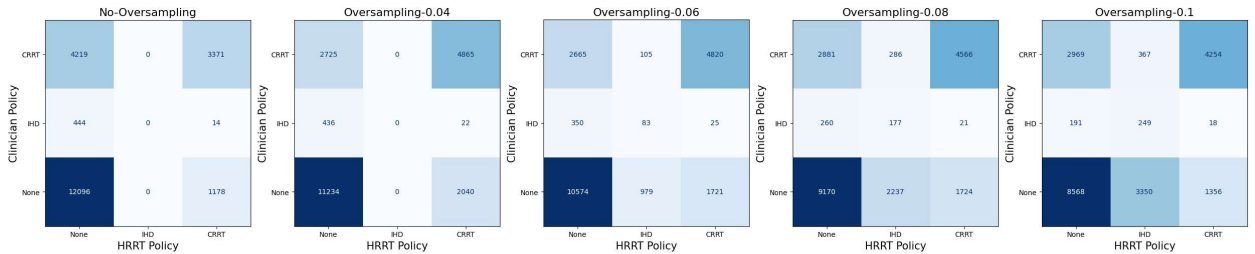


Figure 8: Impact of IHD oversampling on policy action distribution.

#### 4.5 Offline reinforcement learning

A primary challenge in developing a clinical decision-support agent from existing medical records is the static, offline nature of the data. Standard online reinforcement learning algorithms, such as Double Dueling Deep Q-Networks (D3QN)[30], are not designed for this setting. They are exploration-based since usually they can interact with the

environment to correct mistaken beliefs. When applied to an offline dataset, they are highly susceptible to extrapolation error: the agent learns to exploit out-of-distribution (OOD) actions that were rare in the dataset but which, due to spurious correlations or function approximation errors, appear to have unrealistically high Q-values. This exact failure mode was observed when a standard D3QN agent was first applied to our high-level modality task. The policy quickly collapsed, learning to recommend IHD almost exclusively. The root cause is a spurious correlation in the data: IHD, while the rarest modality, is associated with a less severe patient cohort and thus lower mortality. The D3QN agent incorrectly learned that the IHD action caused this low mortality and developed an explosive, over-optimistic Q-value for it, leading to a total policy collapse. To mitigate this critical extrapolation error, our HRRT framework employs a unified hierarchy of specialized offline algorithms. Each algorithm is specifically chosen to address the unique data challenges such as class imbalance, data sparsity, or action space complexity at its stage of the hierarchy.

For the high-level modality agent, we selected BCQ[6]. The modality task is characterized by severe class imbalance, with IHD actions being significantly rarer than CRRT or No RRT. BCQ solves the overestimation problem by using a hard constraint. It has an imitation head to learn the clinician’s policy, and the RL agent is only permitted to select actions that fall within this learned, safe distribution. This directly prevents the agent from exploiting the IHD action and ensures its high-level decisions remain robust and safe. For the low-level IHD agent, a similar issue was observed. Even when training data was balanced, the total number of IHD samples was still small, and a standard D3QN again collapsed. We therefore employed CQL[14]. Instead of a hard constraint, CQL introduces a soft constraint by adding a conservative penalty to the loss function. This penalty pushes down the Q-values of all OOD actions, forcing the agent to be pessimistic about any action not well-supported by the data, thereby achieving stability. Finally, for the low-level CRRT agent, we need to handle the continuous action space of the effluent rate. For this, we employed IQL[13], a state-of-the-art offline algorithm that is particularly effective for continuous control. Unlike BCQ’s hard constraint or CQL’s soft penalty, IQL works by implicitly constraining the policy. It learns the Q-function by performing regression only on actions observed in the dataset, which prevents the agent from querying or overestimating the value of OOD actions. This approach provides robust stability for the continuous control task and completes our fully offline framework. Please refer to Suppl. SectionA for the details about each algorithm and why we chose a decoupled method instead of end-to-end training.

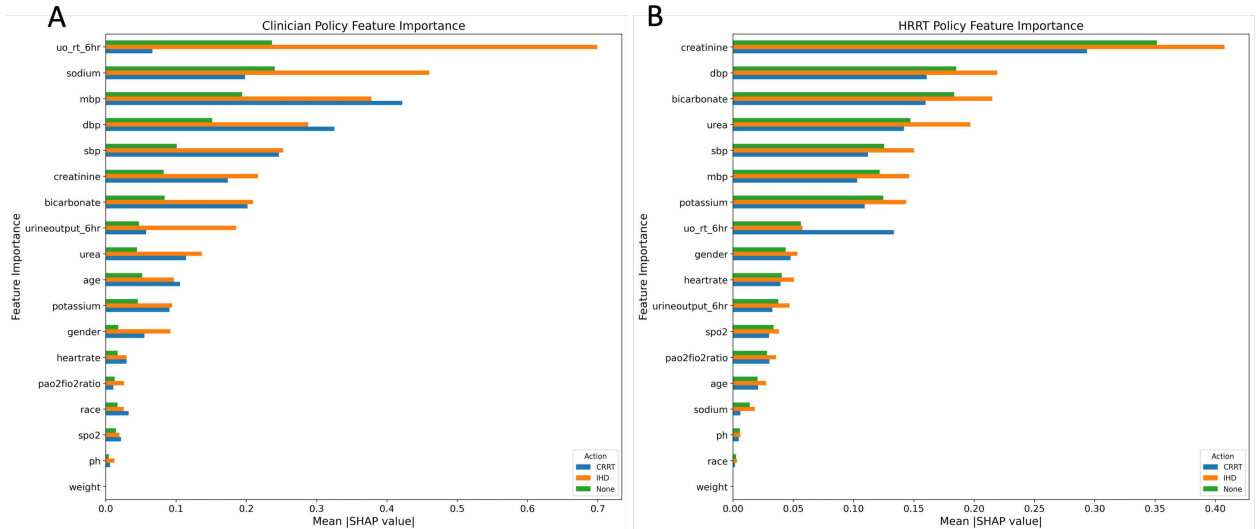


Figure 9: Feature importance for both clinician policy and HRRT policy.

#### 4.6 Evaluation of reinforcement learning model

We conducted a comprehensive evaluation of the trained HRRT agent using various qualitative and quantitative metrics designed to assess both model performance and clinical safety without risks towards the patients. To estimate policy value, we employed WIS (Figure 3A), which re-weights observed retrospective returns to calculate the expected cumulative reward of the HRRT policy relative to clinician and random baselines. This was complemented by an analysis of  $Q$ -values (Figure 3C), where we constructed a curve linking clinician values to observed ICU mortality, thereby validating the reward function as a proxy for survival and allowing us to estimate the agent’s potential mortality benefit. To understand the agent’s strategic behavior, we utilized confusion matrices (Figure 3D) at both the trajectory level, to assess overall modality agreement, and the transition level, to distinguish between fundamental disagreements

in necessity versus temporal shifts in initiation. Safety was further rigorously evaluated for the low-level IHD and CRRT agents by plotting observed mortality against the absolute deviation between the agent’s recommended dosage and the clinician’s action(Figure 4–Figure 5), verifying that the policy operates within a low-mortality therapeutic window. We also validated the temporal plausibility of the policy by comparing the distributions of RRT initiation and cessation times against clinical practice to ensure the agent respects physiological constraints(Figure 6, Suppl. Figure 11). Finally, we analyzed renal functional outcomes by stratifying patients based on policy agreement (Figure 7), which demonstrates that patients with high concordance to the HRRT policy exhibited sustained improvements in eGFR throughout their ICU stay.

#### 4.7 Statistical analysis

Python 3.11.0 (Python Software Foundation, Wilmington, DE, USA) was used for signal preprocessing, model development and validation, statistical testing, and visualization. Continuous variables were described as mean(SD) or median (IQR) and categorical variables as counts (percentage). For global outcome metrics such as mortality where patient-level variance is aggregated into a single policy value, uncertainty was quantified across 5 independent experimental runs initialized with different random seeds.

#### 4.8 Model interpretability

To explain the decision-making logic of the HRRT agent compared to human experts, we performed a feature importance analysis using SHAP (Shapley Additive Explanations) values (Figure 9). The clinician policy (Figure 9A) is characterized by a heavy reliance on immediate fluid status and hemodynamic markers, with the 6-hour urine output rate and serum sodium acting as the dominant drivers for RRT initiation. This aligns with traditional clinical heuristics that prioritize oliguria and volume management. In contrast, the HRRT policy (Figure 9B) exhibits a paradigm shift toward metabolic and solute-based indicators. The agent identifies creatinine, urea, and bicarbonate as the most critical features, assigning them significantly higher importance than urine output. This indicates that the reinforcement learning policy has moved beyond reactive volume management to a more proactive, solute-oriented strategy, prioritizing the clearance of uremic toxins and the correction of metabolic acidosis over simple urine volume metrics. Furthermore, while both policies recognize the importance of hemodynamic stability, the HRRT agent’s specific focus on filtration biomarkers suggests a more precise targeting of renal metabolic demand rather than just hemodynamic support.

### 5 Data availability

MIMIC-IV is publicly available via <https://physionet.org/content/mimiciv/3.1/> and AmsterdamUMCdb is publicly available via <https://amsterdammedicaldatascience.nl/>. Dr.ECC dataset is not publicly available due to institutional restrictions on data privacy.

### 6 Code availability

The codes used for generating the results of this study are available at: <https://github.com/xuqianyi/RL-RRT>.

### References

- [1] Mark Andonovic, Jamie P Traynor, Martin Shaw, Malcolm AB Sim, Patrick B Mark, and Kathryn A Puxty. Short-and long-term outcomes of intensive care patients with acute kidney disease. *EClinicalMedicine*, 44, 2022.
- [2] Pierre-Luc Bacon, Jean Harb, and Doina Precup. The option-critic architecture. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.
- [3] Rinaldo Bellomo, Claudio Ronco, Ravindra L Mehta, Pierre Asfar, Julie Boisramé-Helms, Michael Darmon, Jean-Luc Diehl, Jacques Duranteau, Eric AJ Hoste, Joannes-Boyau Olivier, et al. Acute kidney injury in the icu: from injury to recovery: reports from the 5th paris international conference. *Annals of intensive care*, 7(1):49, 2017.
- [4] Bruce A Cooper, Pauline Branley, Liliana Bulfone, John F Collins, Jonathan C Craig, Margaret B Fraenkel, Anthony Harris, David W Johnson, Joan Kesselhut, Jing Jing Li, et al. A randomized, controlled trial of early versus late initiation of dialysis. *New England Journal of Medicine*, 363(7):609–619, 2010.

- [5] Maria Luiza Medeiros Faria and Alexandre Braga Libório. Continuous kidney replacement therapy dosage and mortality in critically ill patients: a retrospective cohort study using marginal structural model. *Shock*, pages 10–1097, 2023.
- [6] Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*, pages 2052–2062. PMLR, 2019.
- [7] Nada Hammouda and Javier A Neyra. Can artificial intelligence assist in delivering continuous renal replacement therapy? *Advances in chronic kidney disease*, 29(5):439–449, 2022.
- [8] RENAL Replacement Therapy Study Investigators. Intensity of continuous renal-replacement therapy in critically ill patients. *New England Journal of Medicine*, 361(17):1627–1638, 2009.
- [9] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.
- [10] Lorenz Kapral, Mohammad Mahdi Azarbeik, Richard Weiss, Razvan Bologheanu, Clemens Heitzinger, and Oliver Kimberger. Optimal timing for renal replacement therapy in critically ill patients using reinforcement learning algorithms. *Journal of Critical Care*, 86:154964, 2025.
- [11] Kia Khezeli, Scott Siegel, Benjamin Shickel, Tezcan Ozrazgat-Baslanti, Azra Bihorac, and Parisa Rashidi. Reinforcement learning for clinical applications. *Clinical Journal of the American Society of Nephrology*, 18(4):521–523, 2023.
- [12] Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018.
- [13] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021.
- [14] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in neural information processing systems*, 33:1179–1191, 2020.
- [15] Simon Lambden, Pierre Francois Laterre, Mitchell M Levy, and Bruno Francois. The sofa score—development, utility and challenges of accurate assessment in clinical trials. *Critical Care*, 23(1):374, 2019.
- [16] Andrew S Levey, Lesley A Stevens, Christopher H Schmid, Yaping Zhang, Alejandro F Castro III, Harold I Feldman, John W Kusek, Paul Eggers, Frederick Van Lente, Tom Greene, et al. A new equation to estimate glomerular filtration rate. *Annals of internal medicine*, 150(9):604–612, 2009.
- [17] Siqi Liu, Kay Choong See, Kee Yuan Ngiam, Leo Anthony Celi, Xingzhi Sun, and Mengling Feng. Reinforcement learning for clinical decision support in critical care: comprehensive review. *Journal of medical Internet research*, 22(7):e18477, 2020.
- [18] Siqi Liu, Qianyi Xu, Zhuoyang Xu, Zhuo Liu, Xingzhi Sun, Guotong Xie, Mengling Feng, and Kay Choong See. Reinforcement learning to optimize ventilator settings for patients on invasive mechanical ventilation: retrospective study. *Journal of Medical Internet Research*, 26:e44494, 2024.
- [19] MingYu Lu, Zachary Shahn, Daby Sow, Finale Doshi-Velez, and H Lehman Li-wei. Is deep reinforcement learning ready for practical applications in healthcare? a sensitivity analysis of duel-ddqn for hemodynamic management in sepsis patients. In *AMIA Annual Symposium Proceedings*, volume 2020, page 773, 2021.
- [20] VA/NIH Acute Renal Failure Trial Network. Intensity of renal support in critically ill patients with acute kidney injury. *New England Journal of Medicine*, 359(1):7–20, 2008.
- [21] Es-saad Ounci, Abdellah Darraz, Rajae Alkhouh, Youssef Motiaa, and Smael Labib. Timing of kidney replacement therapy in critically ill patients: A call to shift the paradigm in the era of artificial intelligence. *Science Progress*, 108(2):00368504251338927, 2025.
- [22] Dilruk Perera, Siqi Liu, and Mengling Feng. Demystifying complex treatment recommendations: a hierarchical cooperative multi-agent rl approach. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10. IEEE, 2023.
- [23] Martin L Puterman. Markov decision processes. *Handbooks in operations research and management science*, 2:331–434, 1990.
- [24] Aniruddh Raghu, Matthieu Komorowski, Imran Ahmed, Leo Celi, Peter Szolovits, and Marzyeh Ghassemi. Deep reinforcement learning for sepsis treatment. *arXiv preprint arXiv:1711.09602*, 2017.
- [25] Aniruddh Raghu, Matthieu Komorowski, and Sumeetpal Singh. Model-based reinforcement learning for sepsis treatment. *arXiv preprint arXiv:1811.09602*, 2018.

- [26] Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.
- [27] Shengpu Tang, Aditya Modi, Michael Sjoding, and Jenna Wiens. Clinician-in-the-loop decision making: Reinforcement learning with near-optimal set-valued policies. In *International Conference on Machine Learning*, pages 9387–9396. PMLR, 2020.
- [28] Patrick J Thorat, Jan M Peppink, Ronald H Driessen, Eric JG Sijbrands, Erwin JO Kompanje, Lewis Kaplan, Heatherlee Bailey, Jozef Kesecioglu, Maurizio Cecconi, Matthew Churpek, et al. Sharing icu patient data responsibly under the society of critical care medicine/european society of intensive care medicine joint data science collaboration: the amsterdam university medical centers database (amsterdamumcdb) example. *Critical care medicine*, 49(6):e563–e577, 2021.
- [29] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.
- [30] Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Hasselt, Marc Lanctot, and Nando Freitas. Dueling network architectures for deep reinforcement learning. In *International conference on machine learning*, pages 1995–2003. PMLR, 2016.
- [31] Haoyuan Zhang, Minqi Xiong, Tongyue Shi, Wentie Liu, Haowei Xu, Huiying Zhao, and Guilan Kong. Reinforcement learning-based decision-making for renal replacement therapy in icu-acquired aki patients. In *Artificial Intelligence and Data Science for Healthcare: Bridging Data-Centric AI and People-Centric Healthcare*, 2024.
- [32] Tianyi Zhang, Yimeng Qu, Deyong Wang, Ming Zhong, Yunzhang Cheng, and Mingwei Zhang. Optimizing sepsis treatment strategies via a reinforcement learning model. *Biomedical Engineering Letters*, 14(2):279–289, 2024.

## Funding Declaration

This work is partially supported by the AI for Public Health Program (25-1447-A0001) in Saw Swee Hock School of Public Health, National University of Singapore and the NUHS Innovation Grant (NUH-SRO/2025/088/RO5+5/Innovation/LOA03).

## Author contributions

Q.X., K.C.S, M.F., Z.Y.C.T and M.S.L.G conceived the idea and study design. Y.H. and Z.Z. coordinated data acquisition and provided essential resources. Q.X., P.C, J.Y, and C.H. extracted and curated data. Q.X. and F.W. performed methodology design, data analysis and provided the figures/tables. Q.X. and Z.Y.C.T drafted the manuscript and M.F., K.C.S and M.S.L.G revised it critically for important intellectual content. M.F.,K.C.S, Y.H. and Z.Z. supervised the project or provided clinical interpretation. All authors gave critical review and approved the final submitted manuscript.

## Competing Interests

The authors declare no competing interests.

## Appendix

### A Building HRRT policy

We employed a hierarchical reinforcement learning framework to model the complex decision-making process in RRT. This framework involved three distinct agents. The modality agent was responsible for high-level decisions, choosing between IHD, CRRT, or no RRT. Once a modality was selected, lower-level agents were activated to determine the specific dosage. The IHD agent decided among four levels of ultrafiltration volume, while the CRRT agent recommended a continuous effluent rate. The primary outcome of our study was ICU mortality. All agents aimed to decrease mortality by maximizing future rewards, learning to make optimal decisions through state-action pairs. The reward function was directly influenced by patient mortality outcomes and SOFA scores. We also explored an alternative end-to-end learning approach using the Option-Critic architecture[2], which theoretically allows for simultaneous optimization of both the high-level modality policy and the low-level dosage policies within a unified framework. However, this approach proved to be unstable in practice. The difficulty of coordinating multiple hierarchical agents concurrently led to severe convergence issues, where the high-level policy struggled to stabilize while the low-level options were still exploring or under-optimized. This lack of coordination resulted in policy values that failed to exhibit a consistent or meaningful correlation with patient mortality, undermining the safety and interpretability required for clinical deployment. Consequently, we adopted the decoupled hierarchical approach, which ensures robust and independent optimization of each decision layer.

A primary challenge in developing a clinical agent from existing medical records is the static, offline nature of the data. Standard online algorithms like D3QN are unsuitable as they suffer from extrapolation error, learning to exploit out-of-distribution (OOD) actions that were rare in the dataset but appear overly-optimistic. This was observed in our initial experiments, where D3QN would collapse to recommending IHD, a rare action associated with a less severe cohort, mistaking this correlation for causation. To mitigate this, our framework employs specialized offline algorithms for each agent. For the high-level modality agent, we selected BCQ[6]. BCQ addresses extrapolation error by learning a policy  $\pi$  that is constrained to the data distribution. It uses an imitation model to generate actions  $\{a_i\}$  similar to the behavior policy  $\beta$ , and the agent is restricted to selecting the action that maximizes the Q-value from this safe set:

$$\pi(s) = \arg \max_{a_i \sim G(s)} Q(s, a_i)$$

This prevents the agent from choosing OOD actions like IHD and ensures robust, safe decisions.

For the low-level IHD agent, which also suffered from data sparsity, we employed CQL[14]. CQL is a soft constraint method that adds a conservative penalty to the standard Bellman loss. This penalty pushes down the Q-values of OOD actions while pushing up the Q-values for actions present in the data:

$$\mathcal{L}(\theta) = \alpha \mathbb{E}_{s \sim \mathcal{D}, a \sim \pi(\cdot|s)} [Q(s, a)] - \mathbb{E}_{s, a \sim \mathcal{D}} [Q(s, a)] + \mathcal{L}_{\text{Bellman}}(\theta)$$

By tuning the weight  $\alpha$ , we force the agent to be pessimistic about unknown actions, preventing policy collapse.

For the low-level CRRT agent, the dataset was abundant, but the action space was continuous. We therefore employed IQL[13], a state-of-the-art offline algorithm for continuous control. IQL avoids querying OOD actions by learning the Q-function implicitly. It first learns a state-value function  $V$  using an expectile regression loss:

$$\mathcal{L}_V(\psi) = \mathbb{E}_{(s,a) \sim \mathcal{D}} [L_2^\tau(Q_{\hat{\theta}}(s, a) - V_\psi(s))]$$

where  $L_2^\tau(\cdot)$  is the expectile loss, which asymmetrically penalizes errors to estimate the upper-expectile of the Q-values. The Q-function is then updated by regressing towards the observed rewards and the learned value function:

$$\mathcal{L}_Q(\theta) = \mathbb{E}_{(s,a,r,s') \sim \mathcal{D}} [(r + \gamma V_\psi(s') - Q_\theta(s, a))^2]$$

Finally, the policy is extracted using a weighted advantage function  $A(s, a) = Q_{\hat{\theta}}(s, a) - V_\psi(s)$ , which implicitly constrains the policy to actions with high advantages. This approach provides robust stability for the continuous control task and completes our fully offline framework.

### B Hyperparameter tuning

Hyperparameter tuning is a crucial aspect of refining algorithm performance. To achieve optimal performance, we utilized a grid search approach, systematically testing various combinations of hyperparameters such as learning rates, hidden layers, batch sizes, and hidden sizes. This comprehensive exploration helps identify configurations that provide the most effective outcomes. For the Modality agent, a learning rate of 0.001, a batch size of 1024, a hidden size of

8, and 3 hidden layers are utilized. The IHD agent uses a learning rate of 0.001, batch size of 256, hidden size of 32, with 3 hidden layers. Meanwhile, the CRRT agent employs a learning rate of 0.001, a smaller batch size of 128, a hidden size of 16, and 4 hidden layers. These configurations ensure tailored learning dynamics and adaptability for each specific task. In the process of hyperparameter tuning, we use a 5-fold cross-validation to ensure the generalizability of the model and reduce overfitting.

## C Behavior policy

Since we need to simulate clinician’s policy during OPE evaluation, and the accuracy of the behavior policy decides how well OPE metrics can reflect the performance of our trained RL policy. We build a surrogate behavior model  $\pi_b$  based on multi-layer perceptron (MLP), which simulates how clinicians view current patient features and make decisions. We used a batch size of 512, hidden size of 256, and the hidden layers are 4. We trained the model with learning rate of 0.005 for 100 epochs and achieved an accuracy of 80% on test set.

## D Theoretical scaling of intermediate rewards

The central challenge in designing  $R_{mix}$  is to scale the intermediate reward parameters  $b$  and  $c$  such that they provide a useful secondary signal without overpowering the primary mortality signal  $a$ . To determine this scaling a priori, we first analyzed the cumulative impact of  $R_{inter}$  over a full trajectory. The cumulative penalty from  $c$  (Equation 3) approximates a telescoping sum, dependent only on the net SOFA change, while the penalty from  $b$  accumulates with each stagnant step:

$$R_{inter\_total} = \sum_t R_{inter}(s_t, s_{t+1}) \approx -c \cdot (\text{SOFA}_{\text{final}} - \text{SOFA}_{\text{initial}}) - b \cdot (N_{\text{stagnant\_steps}})$$

Based on the distribution of trajectory lengths (Figure 10), we estimated a typical long-stay trajectory to have approximately  $T = 600$  hours (100 steps with 6-hour interval). We posited that for a patient who reaches a terminal state, a net SOFA change of 4 points and 30 stagnant steps (30% of the time) were reasonable heuristics.

Our primary design goal was to ensure the intermediate reward magnitude remained a small nudge compared to the terminal reward (using  $|a| = 10$ ). We set a target for  $|R_{intermediate\_total}|$  to be approximately 20% of  $|R_{terminal}|$ , or 2.0. We also enforced a balance between the two penalty components ( $4c \approx 30b$ ) according to previous study[24, 25]. This led to a system of equations:

$$\begin{aligned} 4c + 30b &\approx 2.0 \\ 4c &\approx 30b \end{aligned}$$

Solving this system by substituting  $c \approx 7.5b$  yielded our candidate parameters for the optimal setting:  $b \approx 0.033$  and  $c \approx 0.25$ .

## E Off-policy evaluation metrics

To evaluate the RL policy, unlike applications in robotics or gaming, where we can collect samples from real environment, in healthcare, we are unable to collect samples from real patients. Therefore, we evaluate policies using OPE metrics due to ethical and safety concerns.

Suppose we have RL policy  $\pi$  and  $\pi_b$ , the Importance Sampling (IS) metric is defined as:

$$\text{IS} = \frac{1}{N} \sum_{i=1}^N \rho_i R_i, \quad (5)$$

where  $\rho_i$  is the importance weight for episode  $i$ , defined as:

$$\rho_i = \prod_{t=1}^T \frac{\pi(a_t | s_t)}{\pi_b(a_t | s_t)}, \quad (6)$$

The Weighted Importance Sampling (WIS) metric is expressed as:

$$\text{WIS} = \frac{\sum_{i=1}^N \rho_i R_i}{\sum_{i=1}^N \rho_i}, \quad (7)$$

which normalizes the importance weights for balanced evaluation.

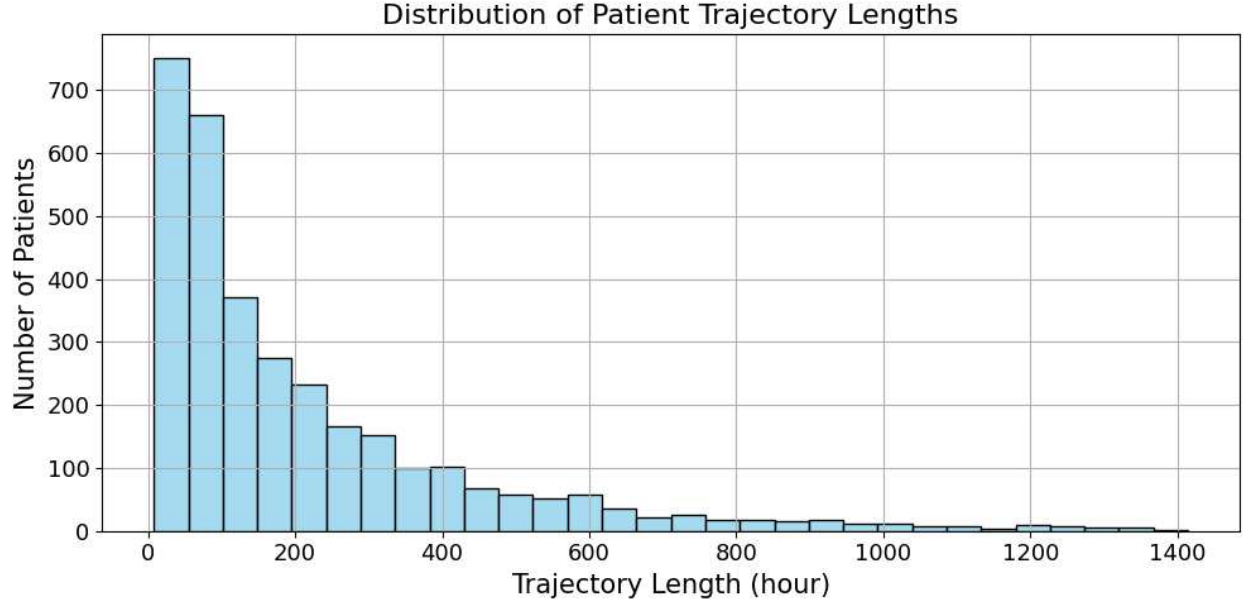


Figure 10: Distribution of patient trajectory length for MIMIC-IV dataset.

## F Implementation of IHD oversampling

To implement oversampling, we first calculate the original probability of each dialysis action. We then compute a specific sampling weight for each class using the formula  $w_{\text{class}} = P_{\text{target}}/P_{\text{original}}$ , where  $P_{\text{target}}$  is our desired distribution. To correct for the bias introduced by this new distribution during training, we also calculate and store a normalized Importance Sampling (IS) weight for each sample,  $w_{\text{IS}} = (P_{\text{original}}/P_{\text{target}})^\beta$ , which is then applied during the model's loss calculation.

## G CRRT effluent rate calculation

The total effluent dose (mL/kg/h) for CRRT is defined by the following equations:

### Continuous Veno-Venous Hemofiltration (CVVH/CVVHF)

$$\text{Dose} = \frac{Q_{\text{rep,pre}} + Q_{\text{rep,post}} + Q_{\text{net}}}{W}$$

### Continuous Veno-Venous Hemodialysis (CVVHD)

$$\text{Dose} = \frac{Q_{\text{dial}} + Q_{\text{net}}}{W}$$

### Continuous Veno-Venous Hemodiafiltration (CVVHDF)

$$\text{Dose} = \frac{Q_{\text{rep,pre}} + Q_{\text{rep,post}} + Q_{\text{net}} + Q_{\text{dial}}}{W}$$

### Notation:

- $Q_{\text{rep,pre}}$  = Pre-Filter Replacement Fluid Rate (mL/h)
- $Q_{\text{rep,post}}$  = Post-Filter Replacement Fluid Rate (mL/h)
- $Q_{\text{dial}}$  = Dialysate Fluid Rate (mL/h)
- $Q_{\text{net}}$  = Fluid Removal Rate (mL/h)
- $W$  = Patient Current Weight (kg)

Table 3: Feature missing rates(%) of 3 datasets.

| Feature                            | MIMIC-IV | AmsterdamUMCdb | Dr.ECC |
|------------------------------------|----------|----------------|--------|
| Age                                | 0        | 0              | 0      |
| Gender                             | 0        | 0              | 0      |
| Race                               | 0        | 100            | 100    |
| Weight                             | 67.99    | 7.8            | 21.81  |
| SBP                                | 3.46     | 4.25           | 12.62  |
| DBP                                | 3.43     | 4.23           | 12.62  |
| MBP                                | 3.34     | 4.25           | 12.70  |
| Heart Rate                         | 1.83     | 1.96           | 12.62  |
| SpO <sub>2</sub>                   | 4.13     | 4.32           | 12.96  |
| pH                                 | 88.35    | 72.94          | 47.26  |
| Urea                               | 65.53    | 93.68          | 42.09  |
| Creatinine                         | 65.56    | 91.64          | 42.08  |
| Sodium                             | 63.05    | 72.06          | 42.07  |
| Potassium                          | 62.70    | 71.65          | 42.07  |
| Bicarbonate                        | 63.62    | 73.02          | 47.26  |
| PaO <sub>2</sub> /FiO <sub>2</sub> | 74.31    | 73.03          | 47.26  |
| Urine Output (6hr)                 | 61.00    | 17.76          | 12.76  |
| Urine Output Rate (6hr)            | 66.04    | 17.76          | 24.22  |

Table 4: The list of features for RRT.

| Category     | Feature Name                                                                              |
|--------------|-------------------------------------------------------------------------------------------|
| Demographics | Age, Gender, Weight, Race                                                                 |
| Vital Signs  | SBP, DBP, MBP, SpO <sub>2</sub> , Heart Rate                                              |
| Lab Tests    | Urea, Creatinine, Potassium, Sodium, pH, PaO <sub>2</sub> /FiO <sub>2</sub> , Bicarbonate |
| Urine Output | Total urine output in the past 6 hours, Change rate of urine output in the past 6 hours   |

Abbreviations - SBP: Systolic blood pressure; MBP: Mean blood pressure; DBP: Diastolic blood pressure; SpO<sub>2</sub>: Peripheral oxygen saturation; PaO<sub>2</sub>/FiO<sub>2</sub>: Arterial oxygen partial pressure to fractional inspired oxygen ratio.

Table 5: Variables used for modality and fluid management decisions.

| Level                     | Action Type                 | Variable                                 | Unit            |
|---------------------------|-----------------------------|------------------------------------------|-----------------|
| <b>Modality Selection</b> | Modality choice             | IHD and CRRT (CVVH, CVVHD, CVVHDF, SCUF) | Discrete {0, 1} |
| <b>IHD Parameters</b>     | Ultrafiltration volume      | Total ultrafiltrate                      | mL              |
|                           | Dialysate rate              | Dialysate rate                           | mL/hr           |
|                           | Prefilter replacement rate  | Prefilter rate                           | mL/hr           |
| <b>CRRT Parameters</b>    | Postfilter replacement rate | Postfilter rate                          | mL/hr           |
|                           | Fluid removal rate          | Hourly fluid removal or goal             | mL/hr           |

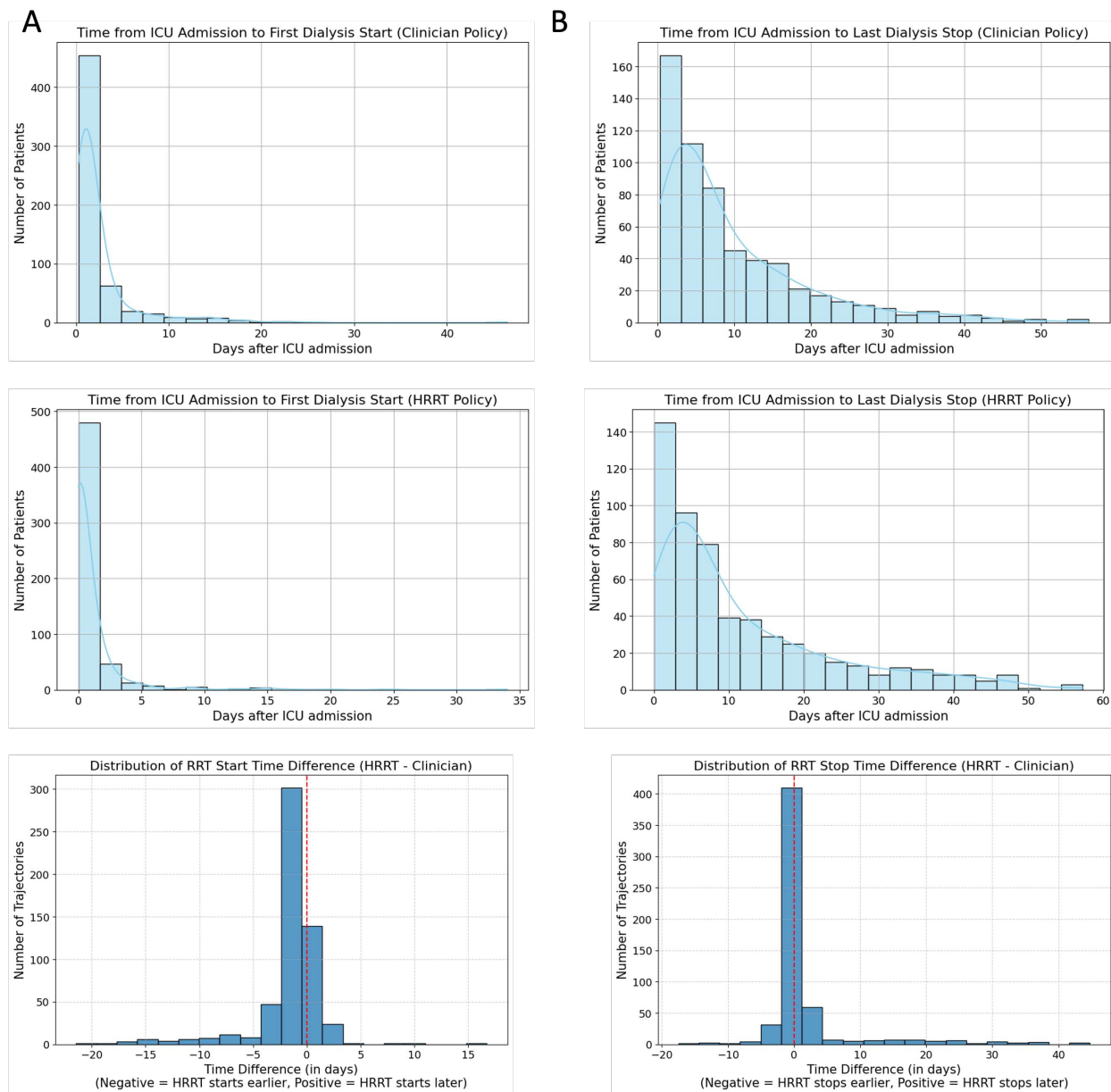


Figure 11: Temporal validation of RRT initiation and termination on the external AmsterdamUMCdb cohort.

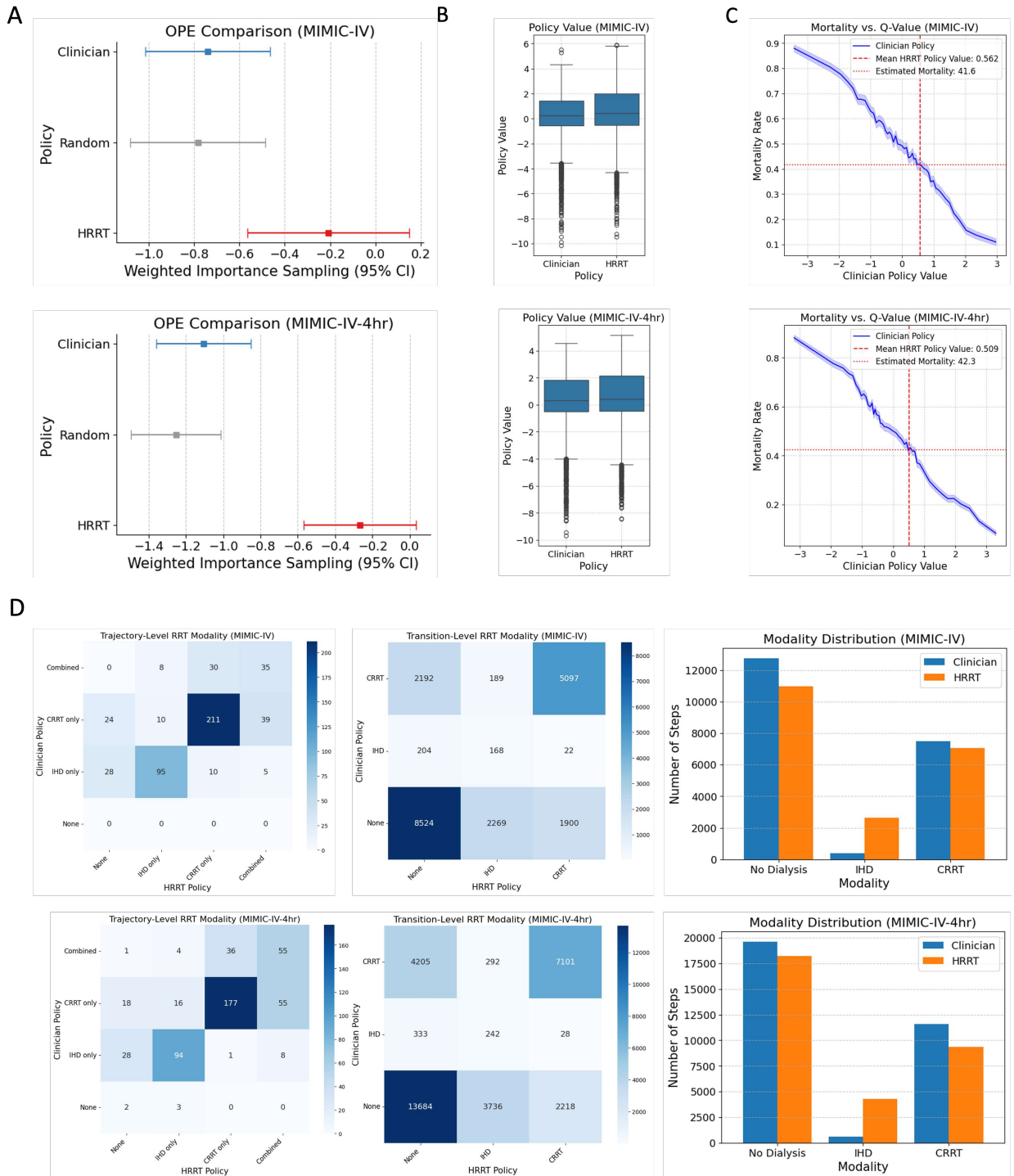


Figure 12: Sensitivity analysis on the binning interval. For each panel, the top row uses 6 hour interval while the bottom rows uses 4 hour interval.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementary.pdf](#)