DeepEN: A Deep Reinforcement Learning Framework for Personalized Enteral Nutrition in Critical Care

Daniel Jason Tan[a], Jiayang Chen[b], Dilruk Perera[c], Kay Choong See[†b], Mengling Feng[†,*a]

[a] *Institute of Data Science and Saw Swee Hock School of Public Health, National University of Singapore, Singapore*
[b] *National University Hospital, Singapore*
[c] *Saw Swee Hock School of Public Health, National University of Singapore, Singapore*

Abstract

Objective: Current ICU enteral feeding remains sub-optimal due to limited personalization and ongoing uncertainty about appropriate calorie, protein, and fluid targets—particularly in the context of rapidly changing metabolic demands and heterogeneous responses to therapeutic interventions. This study introduces *DeepEN*, a novel reinforcement learning (RL)–based framework designed to dy-namically personalize enteral nutrition (EN) dosing for critically ill patients using electronic health record data.

Methods: *DeepEN* was trained on data from over 11,000 ICU patients in the MIMIC-IV database to generate 4-hourly, patient-specific targets for caloric, protein, and fluid intake. The model's state space integrates demographics, comorbidities, vital signs, laboratory measurements, and recent interventions considered relevant to nutritional management. The reward function was de-signed with domain expertise to balance short-term physiological and nutrition-related goals with long-term survival outcomes, reflecting real-world clinical priorities. The framework employs a dueling double deep Q-network with Con-servative Q-Learning regularization to ensure safe and reliable policy learning from retrospective data. Model performance was benchmarked against both clinician-derived and guideline-based policies.

Results: *DeepEN* outperformed both clinician and guideline-based policies, achieving a 3.7 ± 0.17 percentage-point absolute reduction in estimated mor-

tality compared with the clinician policy (18.8% vs 22.5%) and higher expected returns relative to the gold-standard guideline policy (11.89 vs 8.11). Control of key nutritional biomarkers was also improved under the learned policy. U-shaped associations were observed between deviations from clinician dosing and mortality for caloric, protein, and fluid recommendations, suggesting that the learned policy aligns with clinician actions associated with higher survival while diverging from suboptimal behaviors.

Conclusion: *DeepEN* demonstrates the feasibility and potential of conservative offline RL for safe, individualized enteral nutrition therapy in critical care. These findings suggest that data-driven personalization may enhance patient outcomes beyond traditional guideline- or heuristic-based approaches.

Keywords: Reinforcement learning, critical care, enteral nutrition, personalized medicine

---

† *Jointly supervised this work.*

* *Corresponding author: ephfm@nus.edu.sg*

1. Introduction

Enteral nutrition (EN), also known as tube feeding or enteral feeding, refers to the supply of nutrients such as carbohydrates, proteins, and fluids directly through the gastrointestinal tract. EN is typically initiated within 48 hours of Intensive Care Unit (ICU) admission for patients who are unable to meet their nutritional requirements through oral intake [1]. This can be due to increased metabolic demand or impairments in swallowing, nutrient absorption, or appetite regulation. Such conditions include mechanical ventilation, severe head injury, altered mental status, and gastrointestinal dysfunction [2]. The provision of adequate calories and protein is essential for such patients, but it remains clinically challenging.

Evidence based guidelines such as from the American Society for Parenteral and Enteral Nutrition (ASPEN) provide weight and condition specific daily targets for calories and protein intake, considering factors such as obesity, renal

replacement therapy, and hypermetabolic states [3]. However, these guidelines would continue to evolve as new clinical evidence emerge. Recent randomized controlled trials comparing aggressive verses conservative nutritional strategies in critically ill patients have reported conflicting results, ranging from improved outcomes to even potential harm [3, 4, 5]. These mixed findings underscore persistent uncertainties around the optimal dose and timing of enteral nutrition in the ICU setting.

Moreover, there is a well-documented gap between guidelines and clinical practice. Surveys of ICU physicians and nurses reveal that lack of familiarity with both international and local nutrition guidelines, coupled with perceived gaps in the supporting evidence base, significantly hinders the consistent implementation of recommended enteral nutrition practices [6, 7, 8, 9] . Furthermore, nutritional decisions in the ICU are typically guided by heuristic-based practices or one-size-fits-all protocols, which fail to account for the dynamic and patient-specific trajectories of critically ill patients. Moreover factors such as evolving organ function, variable tolerance to feeding, competing interventions (e.g., vasopressors, sedation, dialysis) and rapidly varying metabolic demands make nutritional needs highly individualized and temporally variable. As a result, intervention inconsistencies such as underfeeding, overfeeding and delayed initiation remain common, contributing to increased infection risk, prolonged mechanical ventilation, muscle wasting, and mortality [10].

Consequently, there is a pressing need address these limitations using adaptive, data-driven decision support systems that can effectively recommend personalized nutrition therapy in real time. Reinforcement learning (RL), which is a class of machine learning algorithms that learn optimal decision-making policies through interaction with complex environments, offers a promising framework for this task. RL models can learn from retrospective clinical data to recommend patient-specific, time-dependent feeding strategies that optimize long-term outcomes such as survival and recovery. Unlike static and population-wide guidelines, RL approaches can continuously adapt to a patient's evolving physiological state, balance competing clinical goals, and offer individualized recommenda-

tion at each time point. While RL has shown promise in healthcare settings such as in sepsis [11, 12] and mechanical ventilation management [13], to our knowledge, it has yet to be utilized in the domain of enteral nutrition in critical care.

Our main contributions are as follows:

- We introduce DeepEN, the first RL-based framework for EN management in critical care. DeepEN generates dynamic, patient-specific recommendations for caloric, protein, and fluid targets, optimizing long term clinical outcomes. DeepEN conditions its decisions on a comprehensive list of patient features including diagnoses, labs, vitals, and past treatments to make personalized recommendations. The solution uses conservative RL techniques to ensure clinically safe and plausible recommendations when learning from retrospective data.

- We conduct a comprehensive evaluation of DeepEN's performance against current clinical practice and established clinical guidelines. We evaluate the learned policies using well-established qualitative and quantitative off-policy evaluation methods, and showcase DeepEN's ability to support more consistent, individualized, and outcome-driven nutritional care in critical care.

- We introduce a domain-specific reward function tailored to the context of critical care nutrition. This reward function balances both long- and short- term clinical goals by integrating intermediate signals with terminal outcomes, allowing more clinically aligned policy optimization in the complex, high-stakes setting of the ICU.

2. Background and Related Work

2.1. ***Reinforcement Learning***

Reinforcement Learning (RL) is a framework used to optimize sequential decision-making processes. It is typically modeled as a Markov Decision Process

(MDP), which is defined by the tuple $(\mathcal{S}, \mathcal{A}, r, P, \gamma)$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition probability function, and $\gamma \in (0, 1)$ is the discount factor.

At each discrete time step t, an agent observes its current state $s_t \in \mathcal{S}$, selects an action $a_t \in \mathcal{A}$, and transitions to a new state s_{t+1} according to the transition dynamics $P(s_{t+1} \mid s_t, a_t)$, while receiving a reward $r_t = r(s_t, a_t)$.

The overarching goal in reinforcement learning is to discover a policy $\pi : \mathcal{S} \to \mathcal{A}$ that maximizes the agent's expected cumulative discounted return, often defined as the sum of future rewards: $\sum_{t=0}^{T} \gamma^t r_t$, where T is the time horizon.

2.2. ***Q-learning and Offline RL***

Q-learning [14] is one of the foundational algorithms in reinforcement learn-ing (RL) and is among the most commonly used approaches in healthcare RL.[15] The method aims to learn the value of taking an action a in a given state s, known as the Q-value Q(s, a). At each time step t, the agent updates this estimate based on the reward r_t received and the expected future return, using the Bellman update rule:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \eta \left(r_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t) \right),$$

where η is the learning rate and $\gamma \in (0, 1)$ is the discount factor applied to future rewards.

To handle large or continuous state spaces, Deep Q-Networks (DQN) [16] approximate the Q-function using a neural network. While this leverages the expressive power of neural networks, it also introduces sensitivity to the distribution of training data. In particular, the presence of the maximization operator $\max_{a'} Q(s_{t+1}, a')$ in the update rule can cause upward bias in the target, especially when Q-values are based on noisy or overfitted estimates, leading to overestimation errors.

This issue becomes more profound in offline (or *batch*) reinforcement learning, where the agent is limited to learning from a fixed dataset $\mathcal{D}$ collected by

a behavior policy [17]. In healthcare RL, offline RL is the standard approach since direct interaction with the environment (i.e., directly acting on real patients) is unsafe and ethically impermissible. Without access to environmental interaction (as in online RL), the algorithm must generalize from the static dataset. Consequently, state-action pairs that are rare or missing in $\mathcal{D}$ may result in out-of-distribution (OOD) estimates, exacerbating overestimation error. In high-stakes domains like healthcare, such inaccurate Q-values can lead to unsafe or clinically inappropriate treatment recommendations if not properly mitigated.

2.3. ***Addressing overestimation bias with D3QN and CQL in offline RL***

To mitigate the overestimation bias inherent in standard DQN, the Dueling Double DQN (D3QN) [18, 19, 20] algorithm was proposed as an enhancement. D3QN incorporates two key enhancements. First, it employs the Double Q-learning approach to decouple action selection and evaluation. The action is selected using the online network as $a' = \arg\max_a Q(s_{t+1}, a; \theta)$, and evaluates using a separate target network as $Q(s_{t+1}, a'; \theta^-)$. Second, D3QN incorporates a dueling network architecture that separately estimates the state-value function and the advantage function. This structure allows the agent to more effectively learn which states are valuable independently of the chosen actions, improving learning efficiency and stability. While these augmentations reduce the overestimation bias caused by the maximization operator over noisy Q-estimates, D3QN still does not fully resolve challenges in offline reinforcement learning, where distributional shift and insufficient coverage of state-action pairs remain significant limitations.

To further combat overestimation in offline reinforcement learning, Conservative Q-Learning (CQL) [21] was proposed. CQL addresses the risks posed by out-of-distribution (OOD) state-action pairs by explicitly penalizing overestimated Q-values for unseen or rarely seen actions. Specifically, CQL augments the Q-learning error with the term $\mathbb{E}_{s_t \sim \mathcal{D}, a_t \sim \mathcal{A}}[Q(s_t, a_t)]$, which penal-

izes large Q-values for all possible actions at a given state, including those not sampled in the dataset. This discourages the agent from overvaluing out-of-distribution actions. In addition, CQL also includes the negative expectation
$-\mathbb{E}_{(s_t,a_t)\sim\mathcal{D}}[Q(s_t,a_t)]$, which promotes higher Q-values for the actions that are actually observed in the dataset. Together, these terms ensure that the learned Q-function remains pessimistic about unsupported actions while maintaining high values for well-supported actions, implicitly steering the learned policy towards in-distribution ('clinician-like') behavior. We combine both techniques of D3QN and CQL to avoid unsafe policy recommendations driven by overestimated Q-values in underrepresented areas of the state-action space.

2.4. ***RL for Enteral Nutrition Management***

In the context of clinical decision support, RL has been widely explored for applications such as fluid and vasopressor dosing in sepsis patients, [11, 12], mechanical ventilation management, [13], and sedation titration [15]. However, to the best of our knowledge, RL has not yet been applied to the management of enteral nutrition (EN) in the ICU. Existing AI efforts in this domain have largely focused on predictive tools for malnutrition screening [22, 23] or detecting intolerance to EN [24, 25], rather than personalized nutritional support. Developing an RL-based nutrition management system presents both novel opportunities and unique challenges. Unlike pharmacologic or mechanical interventions (e.g., vasopressors or ventilator settings) that typically produce immediate and measurable physiological responses, nutritional interventions yield more gradual and diffuse effects, and often confounded by concurrent therapies and patient heterogeneity. This complicates the construction of a reward function that meaningfully encodes clinical outcomes such as mortality, recovery, or nutritional adequacy.

Furthermore, as EN remains a largely unexplored use case for RL, the design of the state representation and reward function requires careful consideration. These components must be constructed by leveraging both clinical expertise and empirical analysis to ensure the agent can learn relevant dynamics and

make safe, effective recommendations. Our work represents a first step toward leveraging offline RL for safe and effective personalization of nutritional support in critical care.

3. Methodology

This section details our methodology, including RL problem formulation, algorithm design, cohort construction and experiment setup. Figure 1 illustrates an overview of the solution pipeline.

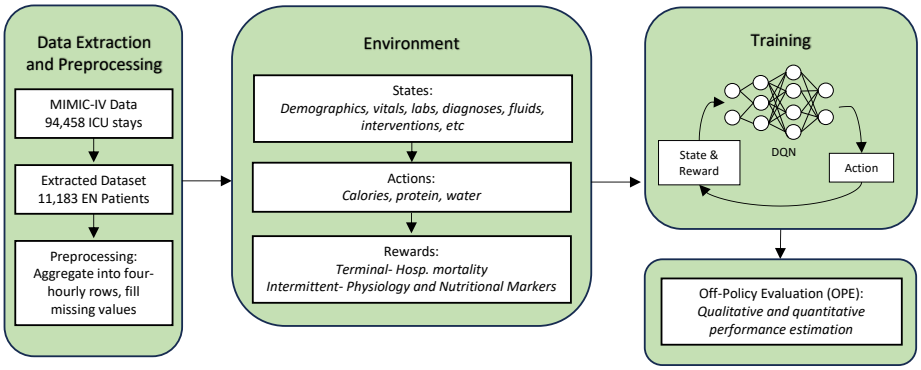

Figure 1: Overview of the DeepEN solution pipeline.

3.1. RL Problem Definition

3.1.1. States

Our state space (S) comprises 102 variables consisting 63 base variables outlined in table 1, and an additional 39 'rate-of-change' variables.1 The rate-of-change variables are the average rate of change calculated across the previous three time-steps for 39 chosen base variables. The chosen variables have dynamics that reflect clinically important trends. These were included to capture additional contextual information on the patient's health trajectory and disease progression.

Table 1: State Space Variables

Category	Variables
Demographics (4)	Age, Gender, Weight*, ICU readmission
Diagnoses and Comorbidities (6)	Burns, CKD, Diabetes, Sepsis, Trauma, Elixhauser score
Vitals (13)*	HR, SBP, MBP, DBP, Resp. rate, Temperature, $PaCO_2$, PaO_2, PF ratio, SpO_2, SOFA, GCS, Shock index
Labs (24)*	Albumin, pH, Calcium, Glucose, Hemoglobin, Magnesium, WBC, Creatinine, Bicarbonate, Sodium, Lactate, Chloride, Platelets, Potassium, PTT, PT, AST, ALT, BUN, INR, Ionised calcium, Total bilirubin, Base excess, Phosphate
Feeding Related (5)	Previous calories, Previous protein, Previous water, Cumulative calories, Cumulative protein
Treatments and Interventions (8)	Mechanical ventilation, FiO_2, CRRT, IV fluids, Vasopressor dose, Propofol dose, Insulin dose, 24-h cumulative insulin dose
Others (3)	Urine output (4-hourly)*, Total output, Time since EN initiation

* Variables whose average rate of change was also included.

Abbreviations: CKD = chronic kidney disease; DBP = diastolic blood pressure; GCS = Glasgow Coma Scale; HR = heart rate; ICU = intensive care unit; MBP = mean blood pressure; PTT = partial thromboplastin time; Resp. = respiratory rate; SBP = systolic blood pressure; WBC = white blood cell.

3.1.2. Actions

The action space is defined using three main components of enteral nutrition: calories, protein and water. For each patient and time window, we consider the weight-adjusted quantities of these component, each discretized into four levels based on empirical quantiles. In line with clinical guidelines, we define total energy and fluid intake as the sum of EN and non-EN sources (e.g., IV fluids and propofol). Although non-EN sources are included for completeness, EN make up the vast majority of these nutritional components.

The action space A is the Cartesian product of the set of these three compo-

nents, resulting in a theoretical total of 64 (4 × 4 × 4) possible action combina-tions. However, due to inherent correlation among calories, protein and water administration, only 51 of the combinations are observed in the dataset. Hence these combinations are considered as the functional action space. Note that the quantities of all three components are strictly negative, as we focus only on periods where enteral nutrition was actively administered. Further details on the action composition and discretization process is detailed in the Appendix (Section 2: *Action Information*).

3.1.3. Rewards

We define a composite reward function that integrates both intermediate physiological indicators and long-term patient outcomes. The formulation is inspired by well-established reward structures used in healthcare RL research [11], and tailored to the clinical context of enteral nutrition.

Let $R_t \in R$ denote the scalar reward assigned at time step t. For each patient trajectory, rewards are defined as:

$$R_t = \begin{cases} R_{\text{term}}(m), & \text{if } d_t = 1 \\ R_{\text{im}}(t), & \text{if } d_t = 0 \end{cases}$$

where $d_t \in \{0, 1\}$ is an indicator of whether the current state is terminal, $m \in \{0, 1\}$ denotes ICU mortality (1 for death, 0 for survival), s_t and s_{t+1} are the current and next state vectors, $R_{\text{term}}(m)$ is the terminal reward assigned based on mortality outcome, and $R_{\text{im}}(t)$ are the intermediate reward components based on physiological and biomarker changes, respectively.

Terminal Reward ($R_{term}(m)$):. The terminal reward is a scalar assigned at the final time step and is defined as:

$$R_{\text{term}}(m) = \begin{cases} +r_T, & \text{if } m = 0 \\ -r_T, & \text{if } m = 1 \end{cases}$$

Consistent with established reward formulations in prior work in RL-based dynamic treatment regimes [11, 12], we set $r_T = 15$, which preserves the dom-

inance of the terminal outcome while remaining compatible with the scale of intermediate shaping signals.

Intermediate Rewards ($R_{im}(m)$):. To mitigate the challenge of sparse rewards and enable efficient credit assignment, we introduce intermediate rewards at each non-terminal timestep. This design facilitates learning by providing frequent and interpretable signals about the clinical trajectory as follows:

$$R_{\text{im}}(t) = R_{\text{phys}}(s_t, s_{t+1}) + R_{\text{bio}}(s_t, s_{t+1})$$

By shaping rewards using short-term physiological responses and penalizing biomarker deviations, the agent is encouraged to learn policies consistent with safe, clinically appropriate trajectories, even when terminal outcomes are not immediately observable.

Intermediate Physiological Reward (R_{phys}): The physiological component helps guide clinical improvement in organ function and perfusion. Based on prior work in critical care RL, it is designed to encourage immediate improvement in SOFA score and blood lactate as follows:

$$\begin{aligned} R_{\text{phys}} = & -c_0\, \mathbf{1}\{\,\Delta\text{SOFA} = 0 \;\wedge\; \text{SOFA}_t > 0\,\} && \text{(stagnation)} \\ & - c_1(\text{SOFA}_{t+1} - \text{SOFA}_t) && \text{(SOFA change)} \\ & - c_2\, \tanh(\text{Lactate}_{t+1} - \text{Lactate}_t) && \text{(lactate penalty)} \end{aligned} \tag{1}$$

where $c_0 = 0.025$, $c_1 = 0.125$, and $c_2 = 2.0$ are fixed scalar weights chosen based on prior validation. The formulation rewards reductions in severity of organ dysfunction and lactate accumulation while penalizing stagnation or worsening trends.

Intermediate Biomarker Deviation Reward (R_{bio}): We further incorporate two biomarkers — blood glucose and serum phosphate — due to their nutritional relevance and responsiveness to carbohydrate and protein intake [26, 27, 28, 29]. These biomarkers serve as strong signals indicative of feeding tolerance and general safety-related clinical outcomes [30, 31]. Interventions such as insulin, which can directly alter serum glucose levels, were included in the state space

to account for their confounding effects on these biomarkers. The composite biomarker reward is defined as:

$$R_{\text{bio}} = \lambda_g R_g + \lambda_p R_p$$

where R_g and R_p are individual reward terms for glucose and phosphate, respectively. Both weighting coefficients were set to $\lambda_g = \lambda_p = 1$ in the final policy to balance the biomarkers and constrain the reward's magnitude. This was done to reduce the risk of short-term biomarker fluctuations overshadowing critical long-term outcomes while maintaining fidelity to domain priorities.

Each component reward is calculated using a smooth sigmoid shaping function and a bonus term for improvement:

$$R_x = f_x(x_{t+1}) + \varepsilon \left(\delta_x^{\text{pre}} - \delta_x^{\text{post}}\right)_+,$$

where $(\cdot)_+$ denotes the positive part operator, ensuring the bonus is non-negative. The term ε (set to 0.2) weights the bonus for decreasing deviation, and

$$\delta_x = \frac{\max\big(0,\, x - x_{\max}\big) + \max\big(0,\, x_{\min} - x\big)}{x_{\max} - x_{\min}}$$

is the normalized deviation from the clinical target range.

The shaping function f_x is implemented as a difference of two logistic curves centred at the lower and upper bounds of the target range:

$$f_x(x) = \frac{2}{1 + \exp\big(-(x - x_{\min})\big)} - \frac{2}{1 + \exp\big(-(x - x_{\max})\big)}$$

This function yields a smooth plateau when x lies inside $[x_{\min}, x_{\max}]$ and decays gradually outside the interval, encouraging maintenance of the biomarker within the desired range rather than penalising small fluctuations.

For glucose, the target range is $[140, 180]$ mg/dL [32], and for phosphate, the target range is $[2.5, 4.5]$ mg/dL [33]. This formulation rewards both the absolute proximity of each biomarker to its target and any improvement over time.

3.2. Experimental Setup

3.2.1. Data Extraction and Preprocessing

We extracted a cohort of enteral feeding patients from the MIMIC-IV database [34] which contains data for over 60,000 ICU patients from the Beth Israel Dea-

coness Mystical Center (Boston, MA, USA) from between 2008 to 2019. Inclu-sion criteria contained following conditions: 1) patient is >18 years of age, 2) has at least 12 hours of enteral feeding data, and 3) has at least one recorded weight measurement. In total, data for 11,378 patients were included. For each patient, we collect data on demographics, vital signs, lab-values, fluids, specific diagnoses, enteral feeding data, and other relevant variables detailed in section 3.2. We only include data from periods where EN is administered, and did not consider periods where EN is absent or interrupted, nor did we consider data from parenteral feeding. In line with clinical evidence, we use enteral feeding data only from the first 10 days post-ICU admission to focus on the acute phase of critical illness (PCI) [35]. Data for calories, protein, and water intake were extracted from MIMIC-IV's *inputevents* table. We aggregated the patient tra-jectories into four-hourly windows using mean or sum as appropriate, and fill any missing values using linear interpolation. More details on the patient cohort can be found in the Appendix (see Section 1: *Cohort Information*).

3.2.2. Baselines

To benchmark the performance of our reinforcement learning (RL) policy, we compare it against four distinct baselines: a random dosage policy, a clinician policy, a behavior cloning (BC) policy, and a expert guideline-based policy.

1. **Random Dosage Policy:** The random policy selects actions uniformly at random from the discrete action space. Formally, it is defined as $\pi(a) = \frac{1}{M}$, where M is the total number of possible actions. This baseline serves as a critical sanity check, establishing a worst-case scenario or lower performance bound. It allows us to assess whether the learned policy outperforms naive, uninformative behavior.
2. **Clinician Policy:** The clinician policy directly reflects the observed actions taken by human clinicians in the MIMIC-IV dataset. It represents real-world practice and serves as an empirical reference point for evaluating the clinical plausibility of the learned policy.

3. **Behavior Cloning (BC) Policy:** The BC policy is a supervised learning model trained to imitate clinician decisions. Specifically, it is implemented as a neural network that minimizes the cross-entropy loss between the predicted actions and those recorded in the dataset. Although it may not replicate clinician behavior perfectly, its performance is expected to be similar to the clinician policy. This baseline helps evaluate whether the use of reinforcement learning yields meaningful improvements over simpler supervised learning approaches.
4. **Expert Guidelines (EG) Policy:** This rule-based policy encodes clinical recommendations drawn from the American Society for Parenteral and Enteral Nutrition (ASPEN) guidelines [3], in combination with more recent evidence-based literature [36, 37], to reflect up-to-date, evidence-informed recommendations for enteral nutrition in critically ill patients. The policy is deterministic and constructed according to expert-derived nutritional targets and clinical rules. A full specification of this policy is provided in the Appendix (Section 3: *Expert Guidelines Policy Definition*).

3.2.3. Training and Hyperparameters

The patient trajectories were randomly split into 80% training and 20% test-ing sets. Our deep RL model is based on the D3QN framework coupled with CQL regularization. A grid search was conducted over CQL scaling coefficient $\alpha \in \{0.01, 0.1, 0.5, 1\}$, and the discount factor $\gamma \in \{0.75, 0.9, 0.95, 0.99\}$. For the network architecture, we evaluated different combinations of ReLU and sig-moid activation functions with 1-4 hidden layers, with layer widths drawn from {64, 128, 256, 512}. All models were trained with a learning rate of 1e-4, batch size of 500.

The final model was selected based on offline evaluation performance, result-ing the optimal hyperparameters to be $\alpha = 0.5$, $\gamma = 0.99$, ReLU activation on 3-layer network with decreasing dimensions (256, 128, and 64).

3.2.4. Off-Policy Evaluation

In online reinforcement learning (RL), policies are learned and evaluated via direct interaction with the environment. However, in the healthcare RL context—where the "environment" corresponds to real patients—any real-time exploration is infeasible. Instead, we evaluate the performance of policies using various quantitative and qualitative off-policy evaluation (OPE) metrics tailored specifically to our target clinical outcomes. These do not require a simulator and instead adopt a well-established approach using offline learning on retrospective healthcare data.

1. **Average Returns:** We used Consistent Weighted Per-Decision Impor-tance Sampling (CWPDIS) [38] to estimate expected returns from each policy. As importance sampling based OPE methods require a behavior policy, we used our trained BC model baseline to estimate the behavior policy. A higher V_{CWPDIS} corresponds to higher estimated returns and indicates a more effective policy.
2. **Mortality vs. Estimated Returns:** To evaluate the clinical alignment, We analyzed the correlation between estimated returns and observed mor-tality under the learned policies. An effective model should display a strong negative correlation, where higher expected returns translate to lower mortality. This indicates that the model learns to associate actions leading to lower returns with higher mortality, and vice versa.
3. **Estimated Mortality Rate:** We further computed the estimated mor-tality rates for the learned policies as follows. First, we categorized ex-pected returns of patient trajectories into distinct bins and used mortality occurrences within each bin to calculate the average mortality rate. The resulting relationship between expected returns and mortality is then used to estimate mortality rates for all policies based on their expected returns.
4. **Dosage Deviation vs. Mortality, and Biomarker Deviation:** We qualitatively evaluated the performance of our learned policy by plotting the relationship between outcome metrics (mortality rate or next-state

biomarker deviations) and dosage differences between clinician-chosen actions and policy-chosen actions. We estimated this relationship by categorizing quantile-level dosage differences into bins and computing mortality rates or deviation units for each bin. An effective policy should align with clinician dosages that resulted in positive outcomes (x-axis = 0), and increasingly deviate from those associated with negative outcomes, ideally forming a U-shaped curve centered at 0. For biomarker deviation analysis, we focused on comparing calorie dosage differences against glucose deviations, and protein dosage differences against phosphate deviations for a clinically centered evaluation.

4. Results

We evaluated DeepEN against four baselines: clinician behavior, a random policy, behavior cloning (BC), and the expert guidelines policy. Our analysis spans both quantitative and qualitative dimensions, incorporating return-based metrics, estimated mortality, biomarker alignment, and policy behavior.

4.1. Quantitative Evaluation: Returns and Mortality

Table 2: Evaluation of Policies by Mortality Rate and CWPDIS

Policy	**Mortality Rate (%)**	**CWPDIS**
Clinician	22.5	5.87
Random	26.8 ± 0.49	1.62
BC	22.3 ± 0.15	5.95
EG	20.5 ± 0.29	8.11
DeepEN	$\mathbf{18.8 \pm 0.17}$	**11.89**

DeepEN achieved the lowest estimated mortality among all tested policies, with a 3.7 ± 0.17 percent lower mortality rate compared to the clinician policy (18.8% vs. 22.5%) (Table 2). This corresponds to approximately 37 ± 1.7 fewer deaths per 1000 ICU patients. DeepEN also resulted the highest CWPDIS score among the policies, indicating superior expected returns under the learned value

function.DeepEN also resulted the highest CWPDIS score among the policies, indicating superior expected returns under the learned value function (Table 2).

Note that, the guideline-based ASPEN policy shows improvement over clinicians, and DeepEN shows even further gains suggesting that offline RL can augment even expert-derived rules in enteral nutrition management.

4.2. Return-Mortality Correlation: Clinical Alignment

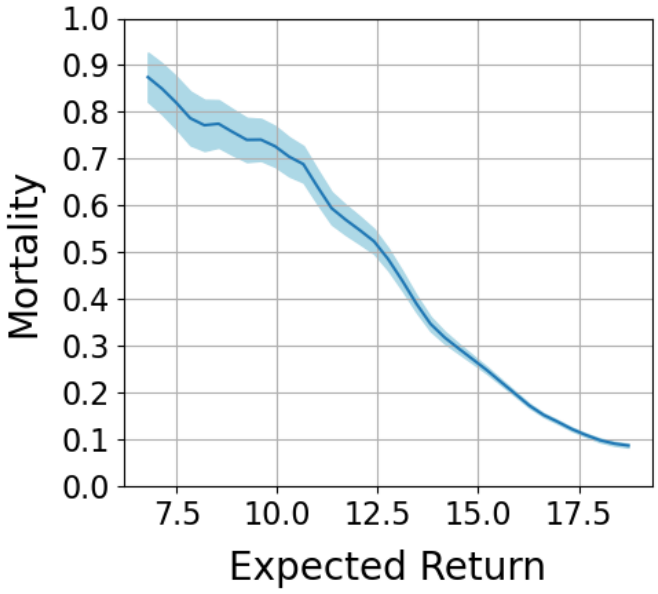

Figure 2: Mortality vs. expected returns. The shaded area represents the confidence interval.

Figure 2 displays a strong negative correlation between expected returns and mortality ($p < 0.001$). This indicates that the policy learns to associate actions leading to lower returns with higher mortality and vice versa.

We analyzed how policy-recommended dosages differ from clinician choices, and their associated outcomes. Figure 3 shows that DeepEN achieves the closest approximation to the desired U-shaped relationship between dosage difference and mortality across all three components (calories, protein, and water). The lowest mortality is achieved when the policy recommendation closely matches the clinician's choice (i.e., dosage difference ≈ 0), and mortality increase with

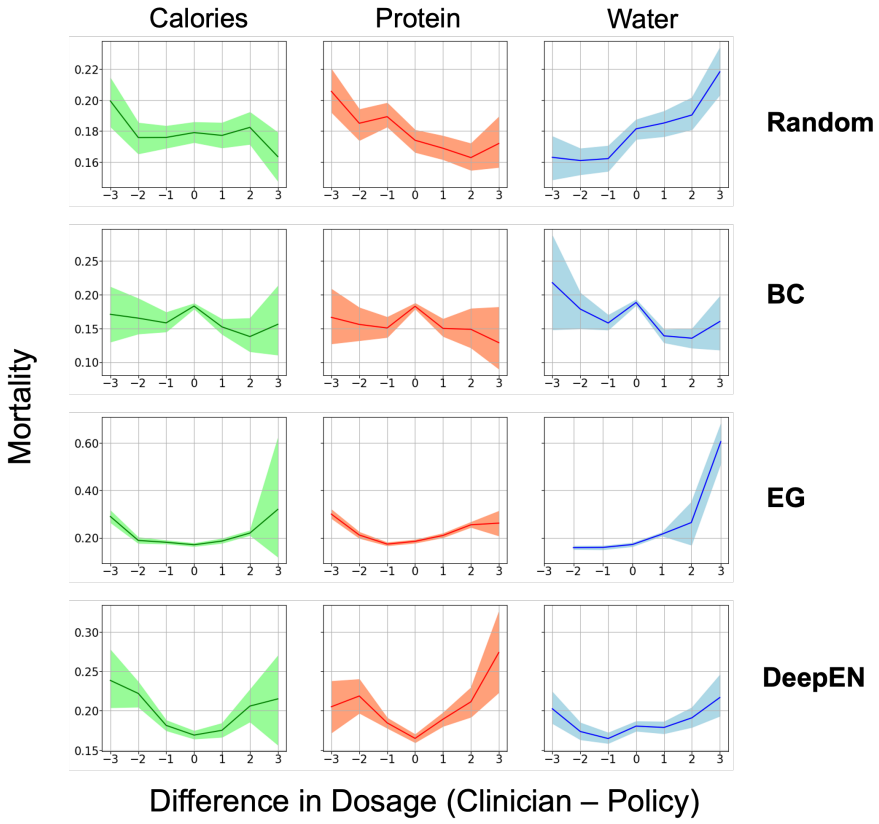

Figure 3: Dosage differences (x-axis) versus mortality (y-axis) for all policies. The shaded area indicates the confidence interval.

larger differences. This is evident that the DeepEN policy agrees with clinician behavior in favorable cases, and it deviates when clinician appears to recommend sub-optimal actions.

4.3. Dosage vs. Biomarker Deviations: Physiological Alignment

In Figure 4, we analyzed the relationship between dosage differences and next-state biomarker deviation to evaluate short-term physiological alignment. For both glucose and phosphate, DeepEN again showed the desired U-shaped pattern with minimal deviation when policy- and clinician- recommended dosages are similar.

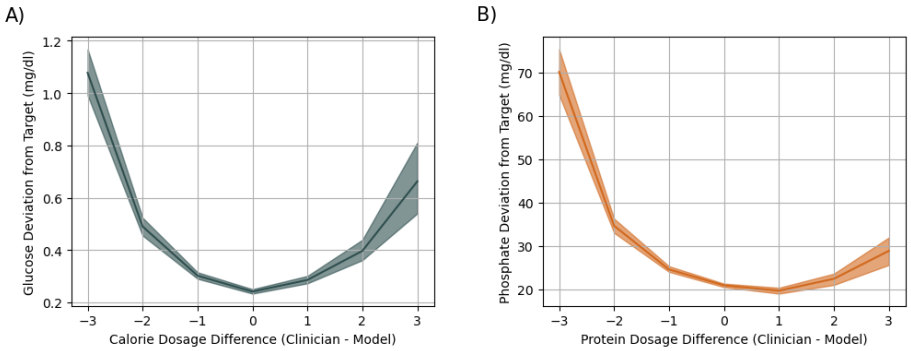

Figure 4: Dosage differences (x-axis) versus biomarker deviation (y-axis) for DeepEN. (A) Glu-cose deviation vs. calories dosage difference and (B) Phosphate deviation vs. protein dosage.

This further validates that the intermediate biomarker rewards used during training were able to effectively guide the target toward actions that stabilize key metabolic indicators. Moreover, the model appears to effectively penalize both over- and under-nutrition, consistent with established physiological responses to EN.

5. Discussion and Conclusions

DeepEN applies deep RL to dynamically personalize enteral nutrition for criti-cally ill patients, with the overarching goal of improving hospital survival. The framework is trained using an expert-informed state space and a physiologically grounded reward function that integrates long-term outcomes with short-term nutrition-relevant signals. Together, these design choices enable DeepEN to ef-fectively distinguish actions associated with higher mortality from those linked to better survival, producing a policy that aligns with high-value clinician de-cisions while appropriately deviating from suboptimal ones. To ensure clinical plausibility, the framework incorporates a conservative Q-learning penalty that constrains the learned policy to remain within the support of observed clinical practice.

DeepEN significantly outperforms both clinician-derived and guideline-based policies across qualitative and quantitative evaluation metrics. The learned pol-

icy's action distribution demonstrates closer adherence to established guideline recommendations than clinician practice, while still exhibiting meaningful and data-driven divergence (Appendix Section 4: *Comparison of Action Distributions*). Moreover, its superior performance relative to the behavior cloning policy indicates that the added complexity of reinforcement learning yields tangible benefits over simpler supervised learning approaches.

Although DeepEN's observed mortality reduction of 3.7% (from 22.5% to 18.8%) is smaller than the improvements typically reported in healthcare RL studies focused on fluid and vasopressor dosage optimization [39, 40], this effect remains clinically meaningful within the context of enteral nutrition therapy. The multifactorial nature of critical illness and the historically modest survival benefits demonstrated in large-scale nutrition trials [41, 3] underscore the significance of even incremental gains. Unlike hemodynamic interventions, which exert immediate effects on survival through cardiovascular stabilization, enteral nutrition confers its benefits more gradually through metabolic regulation, maintenance of gut integrity, and reduction of malnutrition-related complications. These mechanisms, though less acutely life-saving, are essential for supporting physiological recovery and long-term outcomes in critically ill patients. [1]

Taken together, these findings demonstrate that personalized, reinforcement learning–based treatment strategies offer a distinct advantage over one-size-fits-all, guideline-based approaches, highlighting the potential of data-driven personalization to improve the safety and efficacy of nutritional care in the ICU.

While we train and evaluate DeepEN on a relatively large patient cohort, the data we use comes from a single hospital system. This may limit the generalizability of our model, as clinical practices and patient populations can vary substantially across institutions and regions. Validation on external cohorts will therefore be essential to confirm the robustness of our findings. However, the lack of detailed EN data in other large open-source critical care databases currently precludes this effort. We must also note that DeepEN's reward function—while based heavily on well-established rewards functions in healthcare

RL [11, 12, 39]—is handcrafted. Consequently, it may not optimally capture all relevant aspects of nutritional adequacy or clinical recovery. The exploration and development of alternative rewards functions that better integrate and capture short- and long-term goals relevant to EN, represents an important future direction.

Other promising directions for future research include the incorporation of multimodal data—particularly clinical text—to enhance contextual understanding of enteral nutrition (EN) decision making. Clinical notes can contain information critical to understanding feeding tolerance and complications, such as episodes of gastrointestinal (GI) intolerance (e.g., vomiting, diarrhea, abdominal distension) or risks of aspiration; this can be leveraged to improve the model's ability to distinguish appropriate feeding interruptions from suboptimal underfeeding. Moreover, beyond dosage optimization, the model could be extended into a more holistic EN decision support tool capable of predicting optimal EN initiation timing, as well as nutrition-related complications such as refeeding syndrome, aspiration risk, or delayed gastric emptying.

6. Ethical Statement

In line with prior work on reinforcement learning–driven clinical decision support systems (CDSSs) for conditions such as sepsis [11], our proposed framework adopts a human-in-the-loop approach. The system is designed to assist, rather than replace, the role of clinicians by offering data-driven recommendations derived from past clinical trajectories. Importantly, the ultimate responsibility for treatment decisions remains with the clinician, preserving professional autonomy and ensuring that medical expertise governs final care plans. Our framework enables the integration of machine intelligence with clinical judgment, leveraging the strengths of both to improve patient outcomes without undermining ethical standards.

Additionally, our models were developed and validated solely on de-identified, publicly available retrospective data, with no involvement of real-time clinical

interventions or patient contact. As such, the work poses no direct risk to patients and aligns with ethical research practices, minimizing concerns regarding safety, accountability, or consent in the current study phase.

7. Data Availability

The open-source MIMIC-IV 2.0 data used in this study is available at `https://physionet.org/content/mimiciv/2.0/`.

8. Code Availability

The underlying code for this study is available on Github via `https://github.com/danjst/en_rl`.

Acknowledgment

The authors would like to acknowledge the contributions of their team members and mentors from the Singapore IMAGINE AI 2024 Datathon.

References

[1] A. Adeyinka, A. S. Rouster, M. Valentine, Enteric feedings, Journal Name (2018).

[2] J.-C. Preiser, Y. M. Arabi, M. M. Berger, M. Casaer, S. McClave, J. C. Montejo-González, S. Peake, A. Reintam Blaser, G. Van den Berghe, A. van Zanten, et al., A guide to enteral nutrition in intensive care units: 10 expert tips for the daily practice, Critical Care 25 (1) (2021) 424.

[3] C. Compher, A. L. Bingham, M. McCall, J. Patel, T. W. Rice, C. Braunschweig, L. McKeever, Guidelines for the provision of nutrition support therapy in the adult critically ill patient: The american society for parenteral and enteral nutrition, Journal of Parenteral and Enteral Nutrition 46 (1) (2022) 12–41.

[4] J. L. Bels, S. Thiessen, R. J. van Gassel, A. Beishuizen, A. D. B. Dekker, V. Fraipont, S. Lamote, D. Ledoux, C. Scheeren, E. De Waele, et al., Effect of high versus standard protein provision on functional recovery in peo-ple with critical illness (precise): an investigator-initiated, double-blinded, multicentre, parallel-group, randomised controlled tial in belgium and the netherlands, The Lancet 404 (10453) (2024) 659–669.

[5] D. K. Heyland, J. Patel, C. Compher, T. W. Rice, D. E. Bear, Z.-Y. Lee, V. C. González, K. O'Reilly, R. Regala, C. Wedemire, et al., The effect of higher protein dosing in critically ill patients with high nutritional risk (effort protein): an international, multicentre, pragmatic, registry-based randomised trial, The Lancet 401 (10376) (2023) 568–576.

[6] M. Mirhosiny, M. Arab, P. M. Shahrbabaki, How do physicians and nurses differ in their perceived barriers to effective enteral nutrition in the intensive care unit?, Acute and critical care 36 (4) (2021) 342–350.

[7] S. Friesecke, A. Schwabe, S.-S. Stecher, P. Abel, Improvement of enteral nutrition in intensive care unit patients by a nurse-driven feeding protocol, Nursing in Critical Care 19 (4) (2014) 204–210.

[8] R. J. Jarden, L. J. Sutton, A practice change initiative to improve the provision of enteral nutrition to intensive care patients, Nursing in critical care 20 (5) (2015) 242–255.

[9] C. S. Ellis, Improving nutrition in mechanically ventilated patients, Journal of Neuroscience Nursing 47 (5) (2015) 263–270.

[10] T. Ramaswamy, M. P. DeWane, H. S. Dashti, M. Lau, P. E. Wischmeyer, A. Nagrebetsky, J. Sparling, Nine myths about enteral feeding in critically ill adults: an expert perspective, Advances in Nutrition (2024) 100345.

[11] A. Raghu, M. Komorowski, I. Ahmed, L. Celi, P. Szolovits, M. Ghas-semi, Deep reinforcement learning for sepsis treatment, arXiv preprint arXiv:1711.09602 (2017).

[12] D. J. Tan, Q. Xu, K. C. See, D. Perera, M. Feng, Advancing multi-organ disease care: A hierarchical multi-agent reinforcement learning framework, arXiv preprint arXiv:2409.04224 (2024).

[13] A. Peine, A. Hallawa, J. Bickenbach, G. Dartmann, L. B. Fazlic, A. Schmeink, G. Ascheid, C. Thiemermann, A. Schuppert, R. Kindle, et al., Development and validation of a reinforcement learning algorithm to dynamically optimize mechanical ventilation in critical care, NPJ digital medicine 4 (1) (2021) 32.

[14] C. J. Watkins, P. Dayan, Q-learning, Machine learning 8 (1992) 279–292.

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

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

[17] S. Levine, A. Kumar, G. Tucker, J. Fu, Offline reinforcement learn-ing: Tutorial, review, and perspectives on open problems, arXiv preprint arXiv:2005.01643 (2020).

[18] H. Van Hasselt, A. Guez, D. Silver, Deep reinforcement learning with dou-ble q-learning, in: Proceedings of the AAAI conference on artificial intelli-gence, Vol. 30, 2016.

[19] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, N. Freitas, Dueling network architectures for deep reinforcement learning, in: International conference on machine learning, PMLR, 2016, pp. 1995–2003.

[20] M. Hessel, J. Modayil, H. Van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, D. Silver, Rainbow: Combining improvements in deep reinforcement learning, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 32, 2018.

[21] A. Kumar, A. Zhou, G. Tucker, S. Levine, Conservative q-learning for offline reinforcement learning, Advances in neural information processing systems 33 (2020) 1179–1191.

[22] M. Besculides, M. Mazumdar, S. Phlegar, R. Freeman, S. Wilson, H. Joshi, A. Kia, K. Gorbenko, et al., Implementing a machine learning screening tool for malnutrition: insights from qualitative research applicable to other machine learning–based clinical decision support systems, JMIR Formative Research 7 (1) (2023) e42262.

[23] O. Raphaeli, L. Statlender, C. Hajaj, I. Bendavid, A. Goldstein, E. Robin-son, P. Singer, Using machine-learning to assess the prognostic value of early enteral feeding intolerance in critically ill patients: a retrospective study, Nutrients 15 (12) (2023) 2705.

[24] K. Hu, X. lei Deng, L. Han, S. Xiang, B. Xiong, L. Pinhu, et al., De-velopment and validation of a predictive model for feeding intolerance in intensive care unit patients with sepsis, Saudi Journal of Gastroenterology 28 (1) (2022) 32–38.

[25] Q. Chen, Y. Chen, H. Wang, J. Huang, X. Ou, J. Hu, X. Yao, L. Guan, Development and validation of a predictive model for diarrhea in icu pa-tients with enteral nutrition, Journal of Parenteral and Enteral Nutrition 47 (4) (2023) 563–571.

[26] A. R. Gosmanov, G. E. Umpierrez, Management of hyperglycemia during enteral and parenteral nutrition therapy, Current diabetes reports 13 (1)(2013) 155–162.

[27] J. Uribarri, Phosphorus homeostasis in normal health and in chronic kidney disease patients with special emphasis on dietary phosphorus intake., in: Seminars in dialysis, Vol. 20, 2007.

[28] K. Kalantar-Zadeh, L. Gutekunst, R. Mehrotra, C. P. Kovesdy, R. Bross, C. S. Shinaberger, N. Noori, R. Hirschberg, D. Benner, A. R. Nissenson,

et al., Understanding sources of dietary phosphorus in the treatment of patients with chronic kidney disease, Clinical Journal of the American Society of Nephrology 5 (3) (2010) 519–530.

[29] T. A. Ikizler, N. J. Cano, H. Franch, D. Fouque, J. Himmelfarb, K. Kalantar-Zadeh, M. K. Kuhlmann, P. Stenvinkel, P. TerWee, D. Teta, et al., Prevention and treatment of protein energy wasting in chronic kidney disease patients: a consensus statement by the international society of renal nutrition and metabolism, Kidney international 84 (6) (2013) 1096–1107.

[30] H. M. Mehanna, J. Moledina, J. Travis, Refeeding syndrome: what it is, and how to prevent and treat it, Bmj 336 (7659) (2008) 1495–1498.

[31] N. Nguyen, K. Ching, R. Fraser, M. Chapman, R. Holloway, The relationship between blood glucose control and intolerance to enteral feeding during critical illness, Intensive care medicine 33 (12) (2007) 2085–2092.

[32] 16. diabetes care in the hospital: standards of care in diabetes—2024, Diabetes Care 47 (Supplement_1) (2024) S295–S306.

[33] D. A. Geerse, A. J. Bindels, M. A. Kuiper, A. N. Roos, P. E. Spronk, M. J. Schultz, Treatment of hypophosphatemia in the intensive care unit: a review, Critical Care 14 (4) (2010) R147.

[34] A. E. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, T. J. Pollard, S. Hao, B. Moody, B. Gow, et al., Mimic-iv, a freely accessible electronic health record dataset, Scientific data 10 (1) (2023) 1.

[35] T. J. Iwashyna, C. L. Hodgson, D. Pilcher, N. Orford, J. D. Santamarial, M. Bailey, R. Bellomo, Towards defining persistent critical illness and other varieties of chronic critical illness, Critical Care and Resuscitation 17 (3) (2015) 215–218.

[36] P. E. Wischmeyer, D. E. Bear, M. M. Berger, E. De Waele, J. Gunst, S. A. McClave, C. M. Prado, Z. Puthucheary, E. J. Ridley, G. Van den

Berghe, et al., Personalized nutrition therapy in critical care: 10 expert recommendations, Critical Care 27 (1) (2023) 261.

[37] M. Stoian, A. Andone, S. R. Bândilă, D. Onișor, D.-F. Babă, R. Niculescu, A. Stoian, L. Azamfirei, Personalized nutrition strategies for patients in the intensive care unit: A narrative review on the future of critical care nutrition, Nutrients 17 (10) (2025) 1659.

[38] P. Thomas, E. Brunskill, Data-efficient off-policy policy evaluation for reinforcement learning, in: International conference on machine learning, PMLR, 2016, pp. 2139–2148.

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

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

[41] D. E. Bear, L. Wandrag, J. L. Merriweather, B. Connolly, N. Hart, M. P. Grocott, E. R. A. C. I. P. G. E. investigators, The role of nutritional support in the physical and functional recovery of critically ill patients: a narrative review, Critical Care 21 (1) (2017) 226.

Appendix

1. Cohort Information

Table 3: Cohort details

Cohort	% Female	Mean Age (years)	Mean ICU Stay (hours)	Total Population (n)
Overall	41.63	64.92	241	11378
Non-Survivors	43.19	68.76	258	2556
Survivors	41.18	63.81	237	8822

2. Action Information

3. Expert Guidelines Policy Definition

These nutritional targets are based on the 2021 ASPEN requirements for enteral nutrition [3]. Original protein targets have been altered slightly with more recent evidence-based literature recommending a low-dosing period during the early-acute phase of critical illness [36, 37].

3.1. Calories

- For patients with BMI < 30: provide **25 kcal/kg/day**.
- For patients with BMI between 30 and 50: provide **22 kcal/kg/day**.
- For patients with BMI > 50: provide **11 kcal/kg/day**.
- For all patients, only **70%** of the above amounts are provided during the **first 3 days** of enteral feeding to simulate hypocaloric underfeeding.

3.2. Protein

3.2.1. Early Acute Phase (ICU Day 1-4)

- Day 1-2: **0.8 g/kg/day** of protein.
- Day 3-4: **1.0 g/kg/day** of protein.

Table 4: Distribution of Actions in Dataset

ID	Cal	Pro	Water	Count	ID	Cal	Pro	Water	Count
0	1	1	1	21268	26	3	2	2	3147
1	4	4	4	15932	27	2	3	3	3128
2	4	4	3	13241	28	4	3	2	2980
3	2	2	1	11755	29	3	2	1	2352
4	4	4	2	9322	30	4	4	1	2337
5	3	3	2	9179	31	2	1	3	2232
6	1	1	2	9000	32	1	2	2	1955
7	2	2	2	8683	33	2	3	4	1774
8	1	1	4	8392	34	2	1	2	1578
9	1	1	3	8093	35	1	2	3	1384
10	3	3	3	7969	36	4	3	1	1218
11	2	2	3	6763	37	2	1	1	1205
12	3	3	1	5740	38	3	4	1	1179
13	2	2	4	5536	39	1	2	4	1062
14	3	3	4	5458	40	4	2	4	1037
15	4	3	4	5169	41	3	1	4	535
16	3	4	2	4777	42	2	4	1	385
17	2	3	2	4695	43	2	4	2	368
18	3	4	3	4644	44	4	2	3	324
19	1	2	1	4325	45	2	4	3	311
20	4	3	3	4253	46	1	3	1	272
21	3	2	4	4098	47	2	4	4	249
22	2	3	1	3993	48	3	1	3	163
23	3	2	3	3485	49	4	2	2	162
24	2	1	4	3434	50	1	3	2	147
25	3	4	4	3266					

Table 5: Action Quantile Thresholds (4 hourly doses)

Quantile	**1**	**2**	**3**	**4**
Calories (kcal/kg)	0–1.91	1.91–3.05	3.05–4.13	>4.13
Protein (g/kg)	0–0.08	0.08–0.14	0.14–0.19	>0.19
Water (ml/kg)	0–3.61	3.61–5.40	5.40–8.15	>8.15

3.2.2. Stable Phase (ICU Day >4)

- For patients with BMI < 30: provide **1.2 g/kg/day** of protein.
- For patients with BMI between 30 and 40: provide **2 g/kg ideal body weight/day**.
- For patients with BMI > 40: provide **2.5 g/kg ideal body weight/day**.
- Patients with **burns** receive **2 g/kg/day** of protein, regardless of BMI.
- Patients on **CRRT** receive **2.5 g/kg/day** of protein, regardless of BMI.

3.3. Water

- 1.5 ml per 1 kcal of calories administered.

4. Comparison of Action Distributions

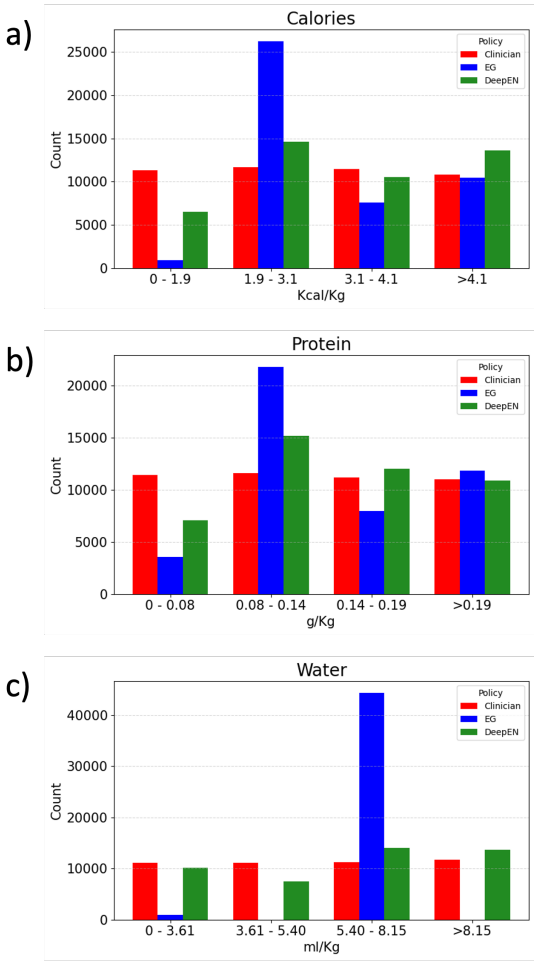

Figure 5: Action Distributions of a) Calories, b) Protein, and c) Water for Clinician, EG, and DeepEN policies.