A Robust and Interpretable Framework for Sepsis Treatment Policy Optimization via Approximate Reinforcement Learning and Model-Assisted Off-Policy Evaluation

I. Introduction and Revised Mandate

Statement of Problem

Sepsis remains a leading cause of mortality in the Intensive Care Unit (ICU). Treatment involves a complex, high-stakes series of decisions regarding intravenous (IV) fluid and vasopressor administration. The optimal policy for this dynamic treatment is unknown. This research aims to derive an optimal treatment policy from large-scale, observational clinical data, specifically the MIMIC-III database. Consistent with the benchmark work by Komorowski et al., we formulate this problem as a Markov Decision Process (MDP), where the goal is to learn a policy that maximizes long-term patient survival.[1]

Addressing the Core Challenge

The primary challenge in this domain is not simply the application of a standard Reinforcement Learning (RL) algorithm. Any algorithm can produce a *policy*. The central difficulty lies in the rigorous and reliable *evaluation* of this newly learned evaluation policy (π_e) using a fixed, static, observational dataset. This dataset was generated by an unknown and likely sub-optimal behavior policy (π_b, i.e., the collective decisions of clinicians). This is the problem of Off-Policy Evaluation (OPE).

Critique Acknowledgment and Revised Focus

Standard OPE methods, such as Weighted Importance Sampling (WIS), are known to suffer from unacceptably high variance, particularly in problems with long time horizons (long patient trajectories) and high-dimensional state-action spaces, both of which are defining characteristics of medical data.[2] A policy value estimate with variance so high that its confidence interval spans from "worse than clinicians" to "significantly better" is scientifically unusable.

Proposed Contribution

The primary contribution is the development and application of a robust, low-variance, high-confidence OPE framework specifically for this clinical context. This framework is built upon two pillars: (1) a state-of-the-art, variance-reduced, model-assisted estimator, the **Weighted Doubly Robust (WDR) estimator** [5], to produce a reliable point estimate of the policy's value; and (2) a **non-parametric bootstrap methodology** [2] to generate statistically meaningful and reliable confidence intervals (CIs) for this estimate.

We will apply this rigorous evaluation framework to a policy learned via linear approximate Q-learning.[1] This revised proposal transforms all "Critical Problems" from the review into a set of interlocking methodological solutions. The seemingly disparate issues of feature engineering, OPE, and model validation are all facets of a single core challenge: the *curse of dimensionality*.[1] The high-dimensional feature space (148 features) exacerbates the variance of IS, necessitates regularization, and complicates validation. This proposal's unifying theme is a set of specific, robust solutions to manage this dimensionality at every stage of the RL pipeline: L2 regularization for *learning*, a linearity assumption for *representation*, and the WDR-OPE framework for *evaluation*.

II. Patient State, Action, and Reward Formulation (MDP Definition)

To apply RL, the problem of sepsis treatment must be formally defined as an MDP.[1]

A. State Space (S)

The patient state s at each 4-hour timestep is represented by a vector of 48 physiological and demographic variables. This set is adopted from the benchmark study and includes 34 continuous variables (e.g., vital signs, lab values) and 14 binary indicators (e.g., ventilation status, comorbidities).

B. Action Space (A)

The clinician's intervention a is discretized into a 25-dimensional action space. This is the Cartesian product of two components:

- **IV Fluids:** 5 bins representing the total volume administered in the 4-hour window.
- Vasopressors: 5 bins representing the maximum dose administered.
  An action a is thus a pair of (IV bin, vaso bin).

C. Reward Function (R)

The reward function $R(s, a, s')$ provides the immediate feedback signal. It is defined based on 4-hour changes in the patient's Sequential Organ Failure Assessment (SOFA) score, with a large terminal reward for survival:

- **Terminal Reward:** +15 for surviving hospital discharge.
- **Terminal Penalty:** -15 for in-hospital mortality.
- **Intermediate Reward:** -0.25 for each point increase in SOFA score (indicating worsening organ failure) and +0.1 for each point decrease (indicating improvement), with 0 for no change.

D. Discount Factor (γ)

A discount factor of $\gamma = 0.99$ will be used. This high value reflects the long-term, high-stakes nature of sepsis treatment, where short-term rewards are less important than the ultimate outcome of survival. This value is standard but will also be confirmed during the hyperparameter tuning phase.

E. Policy (Policy)

A policy $\pi(s)$ is a mapping from a given state s to a probability distribution over the 25 possible actions $a \in A$. The optimal policy π^* is the one that maximizes the expected sum of discounted future rewards, E.[1]

III. A Robust Pipeline for High-Dimensional Clinical Feature Engineering

The critique correctly identified that managing a high-dimensional feature vector is a non-trivial task. This section details our robust pipeline for feature construction, normalization, and validation, directly addressing this challenge.

A. Feature Vector Definition ($\mathbf{f}(s, a)$)

Our model will use linear function approximation, $Q(s,a) = \mathbf{w}^T \mathbf{f}(s,a)$ [1], where $\mathbf{f}(s, a)$ is a 148-dimensional feature vector constructed for each state-action pair. This vector is composed of:

1. **State Features (48):** The 48 base physiological variables.
2. **Action Features (10):** The 2 discrete action components (IV bin, vaso bin) are expanded into a 10-dimensional one-hot encoded vector (5 bins + 5 bins).
3. **State-Action Interaction Features (90):** To capture the differential effect of actions based on patient state, we create interaction terms. These are generated by the element-wise product of each of the 45 *non-demographic* state features with each of

the 2 *primary* action components (e.g., lactate * iv_dose_bin). This results in $45 \times 2 = 90$ features. *Correction from outline: We use 90, not 96, interactions, as demographic features do not interact with actions. The total feature count is $48 + 10 + 90 = 148$.*

B. Scaling, Normalization, and Regularization

To manage this 148-dimensional vector, we will employ a rigorous pre-processing and regularization strategy.

1. **Normalization:** The 48 base state features have vastly different scales (e.g., heart rate $\sim$100, lactate $\sim$2). All continuous state variables will be normalized using **z-scoring** (transforming to mean 0, standard deviation 1). The mean and standard deviation for this transformation will be computed *only* from the 70% training set and then persistently applied to the validation and test sets to prevent data leakage.
2. Regularization: With 148 features, our model is highly susceptible to overfitting. We will mitigate this by incorporating L2 regularization (Ridge regression) into our learning algorithm. The Q-learning update is a form of stochastic gradient descent (SGD).1 The standard update rule for a weight w_i is 1:

   $$w_i \leftarrow w_i + \alpha \cdot (sample - Q(s,a)) \cdot f_i(s,a)$$

   where $sample = r + \gamma \max_{a'} Q(s', a')$.
   We will modify this update to include an L2 weight decay term:

   $$w_i \leftarrow w_i(1 - \alpha\lambda) + \alpha \cdot (sample - Q(s,a)) \cdot f_i(s,a)$$

   The hyperparameter λ controls the strength of the regularization, penalizing large weights and promoting a simpler, more generalizable model. λ will be tuned via the validation process (see Section V).

C. Validating the Linear Separability Assumption

The critique that our "interaction features" approach assumes linear separability is valid and represents a significant methodological risk. We cannot prove this assumption *a priori*. However, we can empirically quantify the potential value lost by this simplification.

Our linear approximate Q-learning model, $Q_w(s,a)$, is computationally efficient and relatively interpretable. A more complex, non-linear model, $Q_{NL}(s,a)$ (e.g., a Gradient Boosting Tree or a Multi-Layer Perceptron), could, in theory, capture complex interactions that our feature set misses.

To validate our assumption, we will perform the following:

1. Train a non-linear "challenger" model (e.g., a simple MLP) on the 70% training set to learn $Q_{NL}(s,a)$.
2. Extract the greedy policy π_{NL} from this model.
3. Using the robust WDR-OPE method (detailed in Section IV), we will evaluate both our final linear policy π_{linear} and the non-linear policy π_{NL} on the 15% test set.
4. The resulting policy values, $V(\pi_{linear})$ and $V(\pi_{NL})$, will be compared. If $V(\pi_{NL})$ is found to be significantly greater than $V(\pi_{linear})$, this provides a quantitative measure of the value "left on the table" by our linearity assumption. This transforms the unvalidated assumption into a testable hypothesis.

Table 1: Feature Vector Specification (Representative Subset)

Feature Name	Base Variable(s)	Type	Pre-processing
Lactate	lactate	State (Continuous)	Z-score
MechVent	mechvent	State (Binary)	None (0/1)
IV_Bin_1	iv_dose_bin	Action (Categorical)	One-hot
...	...	...	...
IV_Bin_5	iv_dose_bin	Action (Categorical)	One-hot
Vaso_Bin_1	vaso_dose_bin	Action (Categorical)	One-hot
...	...	...	...
Vaso_Bin_5	vaso_dose_bin	Action (Categorical)	One-hot
Lactate * IV_Dose	lactate, iv_dose_bin	Interaction	None
Lactate * Vaso_Dose	lactate, vaso_dose_bin	Interaction	None

SPO2 * IV_Dose	spo2, iv_dose_bin	Interaction	None

IV. High-Confidence Off-Policy Evaluation Strategy

This section addresses the most critical flaw in the original proposal: the superficial treatment of Off-Policy Evaluation. We retract the casual mention of WIS and instead propose a state-of-the-art, variance-reducing OPE framework designed for high-confidence estimation in medical datasets.

A. Moving Beyond Standard WIS to Per-Decision WIS (W-PDIS)

Standard Importance Sampling (IS) and its weighted variant (WIS) are known to have variance that grows exponentially with trajectory length. In an ICU setting where trajectories can be hundreds of 4-hour timesteps long, this variance renders the estimator useless.

As a robust baseline, we will implement **Per-Decision Weighted Importance Sampling (W-PDIS)**. PDIS computes the importance weight as a product of per-decision ratios, $w_H = \prod_{t=1}^T \frac{\pi_e(a_t|s_t)}{\pi_b(a_t|s_t)}$. This approach is known to have significantly lower variance than standard IS in long-horizon tasks and is a more credible baseline.[5]

B. Primary OPE: Variance Reduction via Weighted Doubly Robust (WDR) Estimation

The critique of high variance in IS methods is fundamentally correct.[2] Even W-PDIS variance may be too high to draw firm conclusions. Therefore, our primary OPE method will be the **Weighted Doubly Robust (WDR) estimator**.[5] This model-assisted method provides the best-known tradeoff between bias and variance.

The WDR estimator combines a model-based OPE estimate with an importance-sampling-based correction. The "doubly robust" property stems from its construction:

1. We will first train a *dynamics model* $\hat{T}(s, a, s')$ and *reward model* $\hat{R}(s,a)$ on the 70% training set. This model allows us to compute a purely model-based estimate of the policy's value, $V_{\hat{T}}$. This estimate is low-variance but potentially high-bias (if the model is wrong).[1]
2. The WDR estimator then uses W-PDIS to estimate and correct for the *bias* (or *model*

error) of this model-based estimate.
3. The estimator is "doubly robust" because it is guaranteed to be unbiased and consistent if *either* the transition/reward models ($\hat{T}, \hat{R}$) are correctly specified *or* the importance weights (derived from π_b and π_e) are correct.

This is a perfect fit for our problem. We *know* our linear Q-function is an approximation (Sec III.C), and we *know* our learned $\hat{T}$ will also be an approximation. WDR gives us two chances to be right, providing the best-known protection against the high bias of a pure model-based approach and the high variance of a pure IS approach.[5]

C. Justification for Policy Softening (99%/1%)

The 99%/1% "softening" of the clinician's behavior policy π_b is a methodological requirement for *all* importance sampling methods.

- **Problem (Lack of Common Support):** IS requires that any action our new policy π_e might take ($\pi_e(a|s) > 0$) must have had a non-zero probability of being taken by the clinician ($\pi_b(a|s) > 0$). If $\pi_b(a|s) = 0$ for an action a that π_e wants to take, the IS ratio $\rho = \pi_e / \pi_b$ becomes infinite.
- **Solution:** We cannot guarantee this from the data. Therefore, we "soften" the learned π_b to $\pi_b'(a|s) = (1-\epsilon)\pi_b(a|s) + \epsilon(1/|A|)$. The 1% (i.e., $\epsilon=0.01$) probability floor ensures $\pi_b'(a|s) > 0$ for all actions, guaranteeing a finite IS ratio. This introduces a small, known bias into the estimate in exchange for preventing the catastrophic, infinite variance of unconstrained IS.

D. Generating Reliable Confidence Intervals via Non-Parametric Bootstrap

The critique that CIs from high-variance estimators are unreliable is correct. However, the reliability of a bootstrap CI is a direct function of the stability and variance of the *underlying estimator* being bootstrapped.[2]

By using the low-variance WDR estimator (from IV.B) as our statistic, we create a stable foundation for a reliable bootstrap analysis. Our procedure will be as follows:

1. A **non-parametric bootstrap** will be performed on the 15% test set. This involves resampling *entire patient trajectories* with replacement to generate $B=1,000$ new bootstrap samples (datasets).
2. For each of the B bootstrap samples, we will run our full WDR-OPE analysis. This will yield B independent estimates of the policy value: $\{V_1, V_2,..., V_B\}$.
3. This empirical distribution $\{V_b\}$ represents the uncertainty in our policy value

estimate.

4. The 95% Confidence Interval will be extracted directly from the 2.5th and 97.5th percentiles of this distribution. This method is robust, non-parametric, and its reliability is grounded in the low-variance WDR estimator, directly addressing the critique.[4]

V. Protocol for Hyperparameter Tuning and Model Validation

This section provides the concrete methodological details for model selection and validation that were absent from the original proposal.

A. Data Splitting Strategy

The original 80/20 split was insufficient for robust model development. We will adopt a **70/15/15 Train/Validation/Test split**.

- **Training Set (70%):** Used exclusively to train the model parameters: the approximate Q-function weights ($\mathbf{w}$) and the transition/reward models ($\hat{T}, \hat{R}$) required for the WDR estimator.
- **Validation Set (15%):** A held-out set used *only* for hyperparameter tuning. The model never trains on this data; it is only used to evaluate different model configurations via OPE.
- **Test Set (15%):** This set is "locked" until the end of the project. It is used *only once* to generate the final, unbiased OPE results (Table 3) for the single best policy selected during the validation phase.

B. Hyperparameter Tuning Procedure

We will perform a **Random Search** over the hyperparameter space, which is more efficient than a grid search for a large number of parameters. The objective metric to be optimized is the **WDR-OPE policy value (from Sec IV.B) as computed on the 15% Validation Set**. We will run 50-100 random trials. The trial that yields the model with the highest WDR-OPE value on the validation set will be selected as the final model.

Table 2: Hyperparameter Tuning Protocol

Hyperparameter (Symbol)	Description	Search Space	Search Strategy
α (Learning Rate)	Rate of Q-value SGD update [1]	log-uniform(1e-5, 1e-1)	Random
γ (Discount Factor)	Future reward importance [1]	[0.9, 0.95, 0.99]	Random
λ (L2 Regularization)	Weight decay penalty (Sec III.B)	log-uniform(1e-6, 1e-2)	Random
N_{epochs} (Epochs)	Max training iterations	[100, 200,..., 500]	See Sec VI.A

VI. Computational Plan and Convergence Criteria

This section addresses the practical implementation of the training process, including stopping criteria and resource estimation.

A. Convergence Criteria

The vague "N epochs" statement is replaced by a formal, dual-criterion approach to prevent overfitting:

1. **Primary Criterion (Early Stopping):** At the end of each training epoch, the current model weights will be saved. We will then evaluate the WDR-OPE value of the policy derived from these weights on the 15% *validation set*. The final model selected for testing will be the one from the epoch that achieved the **maximum validation-set OPE value**.
2. **Stop Condition:** The training process will be terminated if (a) the validation-set OPE value has not improved for 20 consecutive epochs, or (b) a maximum of $N=500$ epochs is reached.

B. Computational Resources

The approximate Q-learning training (SGD) is computationally efficient. The primary computational burden is the OPE, which must be run at the end of every epoch on the validation set. The most intensive component is the final 1,000-iteration bootstrap analysis

(Sec IV.D) on the test set.

We estimate that the full pipeline (training the dynamics model $\hat{T}$, running 50-100 hyperparameter trials with early stopping, and performing the final 1,000-sample bootstrap analysis) will require approximately 72 hours of computation on a high-performance computing (HPC) cluster node. This is well within the project's allocated resources.

VII. Revised Evaluation Plan and Statistical Significance

This section replaces the weak evaluation plan with a rigorous, statistically-grounded framework.

A. OPE Script

The "adapt the existing OPE script" plan is retracted. We will develop, validate, and open-source a new Python implementation of the **Weighted Doubly Robust (WDR) estimator** and the **non-parametric bootstrap CI** method as described in Section IV.

B. Statistical Significance Testing

The bootstrap procedure (Sec IV.D) provides the direct mechanism for significance testing. When we generate the $B=1,000$ bootstrap samples, we will compute the WDR-OPE value for three policies on *each* sample:

1. Our new policy (π_{new})
2. The benchmark policy (π_{bench})
3. The observed clinician policy ($\pi_{clinician}$)

This process yields 1,000-sample empirical distributions for V_{new}, V_{bench}, and $V_{clinician}$. To test if our policy is a statistically significant improvement over the benchmark, we will compute the distribution of the difference:
$\Delta_i = V_{new, i} - V_{bench, i}$ for $i=1...1000$.
We will then report the 95% CI of this *difference*. If the lower bound of this CI is greater than zero (e.g., a CI of [+0.5, +2.1]), it represents a statistically significant improvement (corresponding to a one-sided $p < 0.025$).

C. Contingency Plan for Negative Results

A result where our policy performs worse than the benchmark or clinicians is not a failure, but a valuable scientific finding. We have the following contingency plan for interpreting such

results:

1. **Scenario 1: $V_{new} < V_{bench}$** (Our policy is worse than the benchmark). This would suggest that our linear function approximation (Sec III.C) and feature set failed to capture the value present in the benchmark's k-means-based state abstraction. The finding would be that a simple linear model is insufficient for this task.
2. **Scenario 2: V_{new} has unacceptably large CIs** (e.g., [-5.0, +7.0]). This would be a finding about the *limits* of OPE itself. It would imply that even with WDR, the transition dynamics are too stochastic or the behavior policy is too far from any optimal policy to draw reliable conclusions from this dataset.
3. **Scenario 3: V_{new} performs well on OPE, but *fails* clinical validation (Sec VIII)**. This is the most critical scientific finding. It would imply that the SOFA-based reward function R (Sec II.C) is misaligned with true clinical utility. This would challenge the validity of the benchmark itself and suggest that future work must focus on reward function design.

Table 3: Final Off-Policy Evaluation Results (Target Structure)

Policy	**Policy Value (WDR-OPE Estimate)**	**95% Confidence Interval (Bootstrap)**	**p-value (vs. πclinician)**	**p-value (vs. πbench)**
Clinician Policy ($\pi_{clinician}$)	[Value]	[CI]	---	---
Benchmark Policy (π_{bench})	[Value]	[CI]	[p-val]	---
Proposed Policy (π_{new})	[Value]	[CI]	[p-val]	[p-val]

VIII. A Pragmatic Framework for Clinical Interpretability

This section addresses the critique that the interpretability claims of a 148-feature model were "overblown."

A. Acknowledging Limitations

The critique is valid. "High interpretability" is an overstatement for a 148-feature linear model. We revise this claim to **"actionable policy interpretability."** We are not claiming to interpret the *model's parameters*; we are claiming to interpret the *model's behavior*.

B. Validation Strategy (Beyond Weight Inspection)

Inspecting 148 feature weights (e.g., $w(lactate \times iv\_dose) = -0.85$) is clinically meaningless. The critique that the benchmark's "State 451" is a strawman is also fair; one could analyze the centroids of their clusters.

Our validation strategy will be more direct and clinically relevant. A clinician does not care about model weights; they care about the *recommendation for a specific patient*. Our learned Q-function, $Q(s,a)$, provides a ranked list of all 25 actions for any patient state s. We will use this to perform a **policy-level sensitivity analysis**.

1. **Procedure:** We will define several "prototypical" patient states s based on common clinical presentations (e.g., an "average" patient, a "high lactate" patient).
2. **Analysis:** For a given prototype s, we will systematically vary a single, clinically critical input variable (e.g., Lactate or SPO2) across its plausible range.
3. **Visualization:** We will generate **"policy-sensitivity plots."** For example, a plot will have "Lactate Level" (from 1 to 10) on the x-axis and the "Optimal Vasopressor Dose" (Bin 1-5) recommended by our policy π_{new} on the y-axis.

These plots directly answer the actionable clinical question: "What does your model do when the patient's lactate goes up?" We will present these plots—not feature weights—to our clinical collaborators for validation. This method provides a rigorous, non-trivial validation of the policy's *clinical sensibility* and is far more useful than inspecting either raw weights or abstract cluster centroids.

Works cited

1. mdps1.pdf
2. Bootstrapping with Models: Confidence Intervals for Off-Policy Evaluation, accessed November 11, 2025, https://ojs.aaai.org/index.php/AAAI/article/view/11123/10982
3. [1606.06126] Bootstrapping with Models: Confidence Intervals for Off-Policy Evaluation, accessed November 11, 2025, https://arxiv.org/abs/1606.06126
4. Bootstrapping with Models: Confidence Intervals for Off-Policy Evaluation - IFAAMAS, accessed November 11, 2025,

https://www.ifaamas.org/AAMAS/aamas2017/proceedings/pdfs/p538.pdf
5. Bootstrapping with Models: Confidence Intervals for Off-Policy Evaluation - UT Computer Science, accessed November 11, 2025, https://www.cs.utexas.edu/~jphanna/posters/hanna2017bootstrapping.poster.pdf