Extended Abstract

Motivation. Sepsis remains one of the leading causes of in-hospital mortality, yet optimal treatment strategies are still unclear. Because prospective experimentation on critically ill-patients is ethically constrained, we must learn directly from retrospective electronic health-record data. Classic "predict-then-optimize" pipelines often produce clinically unsafe policies when the predictive loss is misaligned with the downstream decision objective. We therefore investigate "decision-focused learning" (DFL) for offline reinforcement learning (RL), explicitly training models to improve treatment decisions rather than forecast accuracy.

Method. We couple a neural predictor $\phi_{\boldsymbol{w}}$ of transition dynamics with a Double Deep-Q Network (DDQN) planner, forming a differentiable end-to-end pipeline. Policy values are estimated with the recent counterfactual-augmented importance-sampling estimator that can incorporate imperfect expert annotations while retaining unbiasedness. Gradients of the estimated value are back-propagated through both the planner and predictor, aligning model updates with expected patient outcomes.

Implementation. Data are synthesized with the GYMIC sepsis simulator, yielding $20{,}000$ logged trajectories over a discretized 18×18 grid of SOFA scores and lactate levels (i.e., state) and a 5×5 grid of vasopressor/IV-fluid doses (i.e., action). We train for 10 epochs, batch size 128, using Adam (10^{-4}) and $\gamma = 0.99$. Counterfactual sets of $0 - 20$k trajectories are generated by replaying each patient with a random alternative first action; Gaussian noise $\mathcal{N}(2, 1)$ is added to create imperfect annotations.

Results. Compared with an online DDQN oracle (average return 12.1), DFL with weighted importance sampling (Weighted-IS) attains 11.6, whereas DFL with Counterfactual-IS plus "perfect" counterfactuals reaches 11.4. Without counterfactuals, Counterfactual-IS still achieves 10.7, and with noisy annotations 9.9. Crucially, Counterfactual-IS reduces evaluation variance by 40% relative to Weighted-IS, confirming its robustness.

Discussion. The results show that perfect counterfactuals act like safe exploration data, closing most of the gap to the online oracle, while even biased annotations improve stability. Smaller variance is vital in clinical deployment, where over-optimistic value estimates can jeopardize patient safety.

Conclusion Our decision-focused offline RL framework learns near-optimal sepsis treatment policies from logged data alone, and counterfactual augmentation markedly improves reliability. The approach is readily extensible to other high-stakes domains that provide limited but targeted expert feedback.

Decision-Focused Offline Deep Reinforcement Learning for Healthcare Policy Optimization

Praneet Bhoj
Department of Computer Science
Stanford University
praneet@stanford.edu

Ali Eshragh
Carey Business School
Johns Hopkins University
Ali.Eshragh@jhu.edu

Yuexing Li
Carey Business School
Johns Hopkins University
Yuexing.Li@jhu.edu

Abstract

We present a decision-focused offline reinforcement-learning framework for optimizing sepsis treatment in intensive-care units. A neural model predicts patient-specific transition dynamics, which are solved with a Double Deep-Q Network whose gradients are back-propagated to align learning with downstream clinical reward. Policy value is estimated via the counterfactual-augmented importance-sampling off-policy estimator, enabling the use of noisy expert annotations to reduce variance without sacrificing unbiasedness. Experiments on $20{,}000$ synthetic trajectories from the GYMIC sepsis simulator show that our approach matches the performance of fully online deep RL (average return 12.1 vs. 11.4) when $2{,}000$ perfect counterfactual trajectories are provided, and outperforms prior weighted-importance-sampling DFL while exhibiting 40% lower evaluation variance. Even with biased annotations, Counterfactual-IS method remains robust, underscoring its practicality for high-stakes clinical decision support. These findings highlight the promise of combining decision-focused training with counterfactual reasoning to deliver reliable, data-driven policies in safety-critical healthcare settings.

1 Introduction

Healthcare decision-making often unfolds sequentially over time, making Markov decision processes (MDPs) a natural modeling framework for capturing the dynamic interactions between patients and treatments (Puterman, 2005). However, in many clinical applications, the true underlying parameters of the MDP, such as transition probabilities and reward functions, are unknown and must be estimated from offline health record data. This introduces substantial challenges, as conventional machine learning pipelines tend to prioritize predictive accuracy, often at the expense of decision quality. Consequently, such approaches can lead to unreliable or even unsafe clinical decisions, particularly in high-stakes environments like intensive care units (ICUs) (Wang et al., 2021).

This project focuses on learning optimal treatment strategies for sepsis patients in ICUs, where decisions must be made sequentially under uncertainty and time pressure. Sepsis is the third leading cause of death worldwide and the primary cause of in-hospital mortality, yet its optimal treatment strategy remains uncertain (Komorowski et al., 2018). It is a life-threatening condition characterized by complex, high-dimensional state spaces and delayed effects of treatment, making it an ideal, yet challenging, use case for reinforcement learning (RL) in healthcare (Raghu et al., 2017). Our goal is to build a clinically grounded, decision-focused framework for offline policy learning, aimed at improving patient outcomes from historical health record data, as depicted in Figure 1.

Specifically, we address three key challenges in this setting:

(i) **Learning Environment Dynamics with Decision-Focused Learning:** Rather than treating the learning of MDP parameters (e.g., transition probabilities) as a purely predictive task,

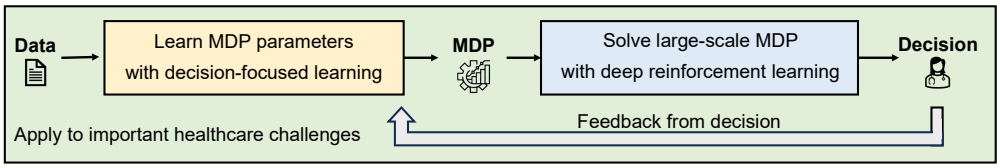

Figure 1: Integrated decision-focused framework that learns MDP parameters from offline data, optimizes large-scale policies with deep RL, and applies the resulting decisions to healthcare challenges.

we adopt a *decision-focused learning* (DFL) approach that tunes parameter estimates to improve the quality of downstream decision-making. This aligns model training with the ultimate policy objectives, enhancing clinical relevance.

(ii) **Scalable Policy Optimization via Deep Offline Reinforcement Learning:** To solve the learned MDP efficiently, especially in large-scale settings, we integrate deep Q-networks into an offline DRL pipeline. This approach is particularly important given the curse of dimensionality and the presence of continuous state spaces, such as vital signs (e.g., blood pressure, heart rate) and lab values (e.g., lactate levels, creatinine), commonly observed in sepsis patients. Our method enables the computation of optimal policies using logged trajectories, without requiring additional online exploration, which is often infeasible or unethical in healthcare.

(iii) **Counterfactual-Augmented Off-Policy Evaluation:** To ensure reliability and robustness in policy evaluation, we incorporate counterfactual expert annotations within a robust off-policy evaluation (OPE) framework. This approach reduces variance and maintains unbiasedness under suitable conditions on the annotations, providing more stable and trustworthy estimates of policy value.

Together, these components form a unified, plug-and-play pipeline that transforms retrospective healthcare data into high-quality treatment policies without requiring on-patient experimentation. Our implementation leverages the GYMIC sepsis simulator (Kiani et al., 2019) and includes both perfect and imperfect expert counterfactual annotations to rigorously evaluate policy performance (Tang and Wiens, 2023). The results demonstrate that our framework achieves policy quality comparable to fully online learning methods while offering significant reductions in variance and greater robustness under real-world imperfections.

In addition to advancing the theoretical and practical foundations of decision-focused learning in sequential settings, our approach is broadly extensible to other healthcare domains, including chronic disease management, triage systems, and beyond. All components have been implemented and validated using open-source tools, ensuring reproducibility and adaptability across use cases.

2 Related Work

Decision-making under uncertainty is central to many high-stakes domains, from supply chains to healthcare. A common machine learning approach is to first build predictive models of system dynamics or rewards, and then use these predictions to optimize decisions, a paradigm known as "predict-then-optimize". However, this separation between prediction and decision-making often leads to suboptimal outcomes, especially when the predictive loss function is misaligned with the downstream decision objective. To address this challenge, the *decision-focused learning* (DFL) framework has emerged as a principled alternative that integrates prediction and optimization in a single learning loop. Rather than minimizing prediction error, DFL trains models to directly improve decision quality (Mandi et al., 2024). Wang et al. (2021) extend this paradigm to sequential decision problems where MDP parameters are unknown, introducing a sample-based, differentiable RL loss and scalable second-order optimization techniques. Their method enables robust policy learning in offline settings and demonstrates strong generalization on unseen MDPs.

Healthcare has become a prominent application area for deep reinforcement learning (DRL) due to the abundance of logged data and the ethical constraints on live experimentation. Raghu et al. (2017) were among the first to model sepsis treatment as a sequential decision problem using deep Q-networks (DQNs). Their architecture incorporates dueling networks and prioritized experience

replay, with actions representing discretized dosages of vasopressors and IV fluids derived from MIMIC-III data. The reward function blends physiological signals, such as SOFA score and lactate levels, with survival outcomes to shape meaningful policies. The learned policy showed alignment with clinical guidelines and indicated improved survival estimates.

However, subsequent analyses revealed critical limitations of DQN in offline settings, including a tendency to overestimate Q-values and difficulties in evaluating deterministic policies due to high-variance importance sampling. To address these challenges, Kumar et al. (2020) proposed Conservative Q-Learning (CQL), which penalizes Q-values for actions insufficiently supported by the logged data. By encouraging the Q-function to remain close to the empirical distribution, CQL mitigates overestimation and prevents spurious value propagation in data-scarce regimes. This makes CQL particularly well-suited for healthcare applications, where ensuring robustness from observational data is essential for learning safe treatment policies.

Evaluating a learned policy without online deployment is essential in high-stakes environments such as healthcare. Off-policy evaluation (OPE) techniques like importance sampling and doubly robust estimators provide unbiased estimates but often suffer from large variance and require strong support assumptions between behavior and target policies (Jiang and Li, 2016). Tang and Wiens (2023) introduced a novel counterfactual-augmented importance sampling (Counterfactual-IS) method that allows expert annotations on a number of counterfactual trajectories. These annotations are optimally integrated with observed data to reduce the variance of the estimator while maintaining unbiasedness under reasonable assumptions. Their theoretical analysis proves that even noisy annotations can yield variance reduction, provided the bias is bounded—making Counterfactual-IS especially attractive for domains where targeted expert input is feasible.

Together, these strands of work form a robust foundation for building high-quality decision-making systems in data-limited and safety-critical settings. First, decision-focused learning ensures that model training is aligned with decision objectives rather than surrogate prediction losses. Second, counterfactual-augmented OPE improves policy evaluation reliability, especially when expert knowledge is available to guide counterfactual reasoning. This project integrates both threads: we adopt the sequential DFL framework of Wang et al. (2021) to learn from offline trajectories and leverage the Counterfactual-IS estimator, proposed by Tang and Wiens (2023), for policy evaluation, thereby advancing the state of the art in decision-focused offline RL for healthcare.

3 Method

We model every patient trajectory as a Markov decision Process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \boldsymbol{P}, \boldsymbol{R}, \gamma)$. The state space $\mathcal{S}$ is discrete and can be high-dimensional, indicating the patient's health conditions such as the Sequential Organ Failure Assessment (SOFA) scores and lactate levels. The action space $\mathcal{A}$ is the discrete set of all possible treatment options. The true transition probabilities $\boldsymbol{P}$ are unknown. However, it is correlated with certain features $\boldsymbol{x}$ associated with the patient, for instance, the patient's demographic information and medical history. The reward $\boldsymbol{R}$ depends on the current state and action, and is assumed to be known, defined by the deterministic function of SOFA scores and lactate levels proposed in Raghu et al. (2017). The parameter γ is the discount factor.

We have access to a set of trajectories generated under behavior policies denoted by π_{b}, which may vary across clinicians. Each trajectory corresponds to a single patient and is associated with that patient's features $\boldsymbol{x}$. The transition probabilities depend on the features through an unknown functional relationship. In addition to these trajectories, we also have access to a set of counterfactual annotations provided by human domain experts. These annotations specify the rewards that would have been observed had alternative actions been taken in the same clinical states.

Our goal is to learn a model to predict the MDP parameters $\boldsymbol{P}$ from features $\boldsymbol{x}$ using decision-focused learning (cf. Figure 1). Specifically, we train a neural network $\phi_{\boldsymbol{w}}$ that maps a given feature $\boldsymbol{x}$ to the predicted MDP parameters, and solve the predicted MDP using the Double Deep-Q Network (DDQN), yielding a policy $\pi(\phi_{\boldsymbol{w}}(\boldsymbol{x}))$. We then evaluate the policy using the counterfactual-augmented off-policy evaluation (OPE) method in Tang and Wiens (2023). The value of the policy is then backpropagated through the planner layer that solves the predicted MDP and the prediction layer to update the weights $\boldsymbol{w}$ in the predictive neural network.

4 Data

The dataset we use for our work is generated using GYMIC[1], which is an OpenAI Gym environment designed to simulate sepsis treatment for ICU patients based on the MIMIC dataset (Johnson et al., 2016). We randomly select actions at each state to generate $20{,}000$ rollouts with the simulator, where each rollout represents a patient's treatment path over the course of their ICU stay. Each rollout contains about 10 transitions on average before reaching a terminal state. The states in the simulated trajectories contain 46 normalized features tracking various vital signs and lab measurements of the patient. At every state, there are 25 possible actions corresponding to a 5×5 grid of vasopressor and IV fluid doses (as in Raghu et al. (2017)) that a doctor can provide to the patient. The terminal states are defined as either patient death (with a corresponding reward of -15) or patient release from the hospital (with a corresponding reward of $+15$). Though all the other states initially had a reward of 0, we modify the reward function within the environment to align with the reward function in Raghu et al. (2017). In order to make our work more computationally feasible, we consider the two most important (in terms of indicating sepsis severity) features for our state analysis: the SOFA score and lactate measurement. We postprocess our simulated trajectories by isolating these two features and discretizing them into 18 bins each. The postprocessed trajectories then serve as our primary dataset for our work.

In addition to the $20,000$ original trajectories described above, we also collect $20,000$ counterfactual trajectories with perfect and imperfect annotations. For each trajectory in our initial dataset, we reset our environment to the trajectory's start state, then randomly select an action that is different from the initial action in our original dataset. After taking this new action, we rollout the rest of the trajectory, and track the new rewards that we receive. For the perfect annotations, the rewards are left exactly as they are returned from the simulator. For the imperfect annotations, we add Gaussian noise $\mathtt{N}(2, 1)$ to the true rewards to simulate a biased expert who overestimates rewards.

5 Experimental Setup

For every experiment, we train the specified framework for 10 epochs. Each experiment therefore takes approximately 12 to 14 hours.

5.1 Framework Comparison

The first set of experiments we run in our work compares the efficacy of five different setups to solve for an optimal sepsis treatment strategy.

Online DQN. As a baseline framework, we train a DDQN on the sepsis simulator environment in a fully online manner, such that the network can explore any states and actions it wants to in order to develop a strong estimation of the values over the state-action space. This effectively simulates a scenario where a doctor can experiment with various treatment strategies on their patients and observe how those treatments impact patient recovery over time. Of course, this is infeasible in the real world, so our work seeks to develop an offline strategy that can come close to this fully online baseline where the doctor can experiment with patient treatment however they like.

DFL with Weighted-IS OPE. Our second setup is a decision-focused learning implementation that uses the original OPE formulation from (Wang et al., 2021). We will refer to this OPE formulation as Weighted-IS.

DFL with Counterfactual-IS OPE and No Counterfactuals. Our third setup is a decision-focused learning implementation that uses the new OPE formulation from (Tang and Wiens, 2023). We will refer to this OPE formulation as Counterfactual-IS. For this setup, we do not include any of our counterfactual data in the training process.

DFL with Counterfactual-IS OPE and Perfect Counterfactuals. Our fourth setup is similar to the third setup (decision-focused learning with Counterfactual-IS), however for this setup, we use 2000 trajectories from our perfect counterfactual data in the training process.

[1] https://github.com/akiani/gym-sepsis

DFL with Counterfactual-IS OPE and Imperfect Counterfactuals. Our fifth setup is similar to the third setup (decision-focused learning with Counterfactual-IS), however for this setup, we use 2,000 trajectories from our imperfect counterfactual data in the training process.

5.2 Counterfactual Dataset Size Comparison

The second set of experiments in our work is designed to explore the impact of counterfactual dataset size on the policy values. For these experiments, we use the decision-focused learning with Counterfactual-IS framework and we vary the amount of counterfactual data that is used during training. We use 0, 2000, 5000, 10000, 15000, and 20000 trajectories randomly selected from our set of perfect counterfactual trajectories for this set of experiments.

6 Results

After training the predictive model and learning a corresponding policy using decision-focused learning, we evaluate the policy on another 2,000 trajectories not included in our training set. To illustrate the value of incorporating counterfactual annotations (CF), we compare the performance of several policies based on various forms of OPE with and without counterfactual annotations in terms of the mean and variance of the evaluation reward.

Experiment	Average Evaluation Reward
Online learning	12.1
DFL w/ Weighted-IS OPE	11.6
DFL w/ Counterfactual-IS OPE	10.7
DFL w/ perfect CF & Counterfactual-IS OPE	11.4
DFL w/ imperfect CF & Counterfactual-IS OPE	9.9

Table 1: Average cumulative rewards of the policies learned in our first set of experiments (described in Section 5.1), each evaluated over 2,000 evaluation rollouts.

Based on our first set of experiments, we observe that decision-focused learning using the original Weighted-IS OPE from Wang et al. (2021) is able to produce policies that achieve performance closest to that of the fully online learning approach. However, as seen in Table 1, we also find that decision-focused learning using the Counterfactual-IS OPE from Tang and Wiens (2023) with perfect counterfactuals achieves comparable performance (with an evaluation reward of 11.4) to both decision-focused learning with Weighted-IS OPE and online learning. This aligns with our intuition that perfect counterfactual data effectively serves as additional exploration data, and with broader coverage of the state and action space, the model more accurately reflects the underlying ground-truth dynamics and value structure. Furthermore, the average rewards show that decision-focused learning using Counterfactual-IS OPE with no or imperfect counterfactuals (with evaluation rewards of 10.7 and 9.9, respectively) are not too much worse than the other experimental setups.

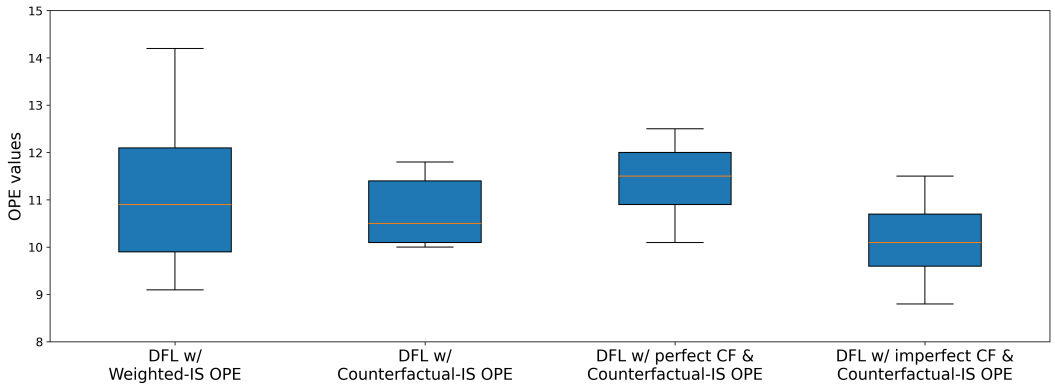

Figure 2: Distributions of the OPE values on 2,000 evaluation rollouts in our first set of experiments described in Section 5.1.

More notably, Figure 2 shows that decision-focused learning with Counterfactual-IS OPE has much smaller variance compared to decision-focused learning with Weighted-IS OPE regardless of the presence and quality of counterfactual data. This suggests that Counterfactual-IS OPE yields more reliable performance evaluation that is robust even to imperfect (biased) annotations on input trajectory data.

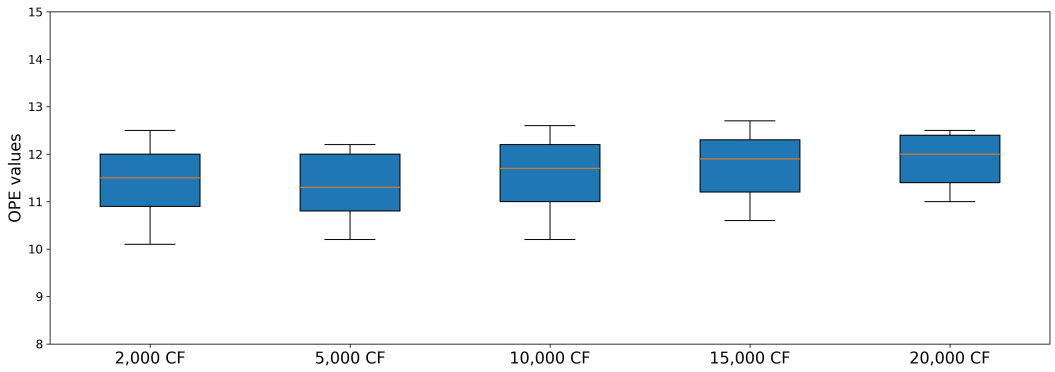

Figure 3: Distributions of the OPE values on 2000 evaluation rollouts in our second set of experiments described in Section 5.2.

In our second set of experiments, we analyze the impact of counterfactual dataset size on the value of the resulting policies. As seen in Figure 3, aside from the small dip in values when changing the number of counterfactual trajectories from $2{,}000$ to $5{,}000$, increasing the number of counterfactual trajectories generally leads to higher-value policies. Importantly, this improvement occurs without compromising evaluation stability—the spread of OPE values remains consistent as more counterfactual data is added. This serves as further evidence supporting our intuition that increased perfect counterfactual data has a similar effect as added exploration, thus yielding better policies.

7 Discussion and Future Work

Our findings open several promising directions for future investigation aimed at enhancing the robustness and realism of decision-focused learning in healthcare settings.

One important avenue is to explore alternative value-based offline RL algorithms, particularly Conservative Q-Learning (CQL) (Kumar et al., 2020) and Implicit Q-Learning (IQL) (Kostrikov et al., 2021). While CQL explicitly penalizes overestimated Q-values to address the overestimation bias common in off-policy learning, IQL takes a different approach by avoiding direct policy improvement steps and instead optimizing an implicit policy that matches high-value actions. This makes IQL potentially more robust in settings where offline data is limited or heavily biased. Comparing the two in our decision-focused pipeline could reveal tradeoffs between conservatism and expressiveness in the learned policies, especially in the presence of counterfactual annotations that may amplify overestimation or underestimation.

Second, we see value in expanding the scope of counterfactual annotations by introducing greater variety and imperfection in the feedback data. By simulating annotations with different levels of overestimation and underestimation, we can better capture the heterogeneity of expert opinions and behavioral policies. Such diversity would allow us to test the generalization of our models across a broader range of assumptions about expert quality and bias, making our framework more robust in practice.

Third, a natural extension is to move beyond discretized representations of clinical variables by modeling continuous states within a Partially Observable Markov Decision Process (POMDP) framework. This would enable us to treat rich, high-frequency data from simulators like GYMIC as noisy emissions from latent physiological states, thus preserving the fidelity of the input data. Avoiding coarse discretization could lead to more accurate state inference and better downstream policy decisions.

Overall, these directions aim to address key limitations in current decision-focused learning pipelines by reducing estimation bias and increasing robustness to annotation noise. We believe they will be critical in scaling DFL approaches to more complex and uncertain real-world domains.

8 Conclusion

This project presents a decision-focused offline reinforcement learning framework tailored for optimizing treatment strategies for sepsis in ICU patients. By integrating a predictive model for MDP parameters with a deep Q-network planner and counterfactual-augmented off-policy evaluation (OPE), we demonstrate the effectiveness of aligning model learning with downstream decision quality. Our experiments using the GYMIC simulator show that incorporating perfect counterfactual annotations can significantly improve policy performance while maintaining low variance in evaluation. Even with imperfect annotations, the proposed Counterfactual-IS OPE method remains robust, offering reliable policy evaluation in settings with biased or noisy expert input. These results highlight the potential of decision-focused learning combined with counterfactual reasoning to support safe, effective treatment planning in high-stakes healthcare applications.

9 Team Contributions

The contributions to this work from each team member can be summarized as follows:

- **Praneet Bhoj:** Implements the technical contribution, conducts numerical experiments, drafts reports (proposal/milestone/final) and poster, and proofreads all submissions.
- **Ali Eshragh:** Provides a literature review, develops theoretical insights, drafts reports (proposal/milestone/final) and poster, and proofreads all submissions.
- **Yuexing Li:** Outlines the technical contribution, develops and refines theoretical insights, drafts reports (proposal/milestone/final) and poster, and proofreads all submissions.

9.1 Changes from Proposal

We did not deviate from our proposed team contributions. We found that, based on our individual strengths, it would be most effective for us to follow our initial planned contributions with Ali and Yuexing driving theoretical discussion and Praneet focusing on technical implementation.

References

N. Jiang and L. Li. 2016. Doubly robust off-policy value evaluation for reinforcement learning. In *International conference on machine learning*. PMLR, New York, NY, USA, 652–661.

A.E.W. Johnson, T.J. Pollard, L. Shen, L.H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R.G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3, 1 (2016), 1–9.

A. Kiani, C. Wang, and A. Xu. 2019. Sepsis World Model: A MIMIC-based OpenAI Gym "World Model" Simulator for Sepsis Treatment. arXiv preprint arXiv:1912.07127.

M. Komorowski, L. Celi, O. Badawi, A.C. Gordon, and A.A. Faisal. 2018. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine* 24, 11 (2018), 1716–1720.

I. Kostrikov, A. Nair, and S. Levine. 2021. Offline reinforcement learning with implicit Q-learning. arXiv preprint arXiv:2110.06169.

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

J. Mandi, J. Kotary, S. Berden, M. Mulamba, V. Bucarey, T. Guns, and F. Fioretto. 2024. Decision-Focused Learning: Foundations, State of the Art, Benchmark and Future Opportunities. *Journal of Artificial Intelligence Research* 81 (2024), 623–1701.

M.L. Puterman. 2005. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Hoboken, NJ.

A. Raghu, M. Komorowski, I. Ahmed, L. Celi, P. Szolovits, and M. Ghassemi. 2017. Deep reinforcement learning for sepsis treatment. arXiv preprint arXiv:1711.09602.

S. Tang and J. Wiens. 2023. Counterfactual-Augmented Importance Sampling for Semi-Offline Policy Evaluation. In *Advances in Neural Information Processing Systems*, Vol. 36. Curran Associates, Inc., Red Hook, NY, 11394–11429.

K. Wang, S. Shah, H. Chen, A. Perrault, F. Doshi-Velez, and M. Tambe. 2021. Learning MDPs from Features: Predict-Then-Optimize for Sequential Decision Making by Reinforcement Learning. In *Advances in Neural Information Processing Systems*, Vol. 34. Curran Associates, Inc., Red Hook, NY, USA, 8795–8806.