Reconciling Set-Valued Policy & Dead-End Discovery in Healthcare Reinforcement Learning: An Empirical Analysis

Victor Li, Sixing Wu, and Shengpu Tang†

Department of Computer Science, Emory University, Atlanta, Georgia 30322, USA

†*Email: shengpu.tang@emory.edu*

Standard reinforcement learning aims to learn a policy that maps each state to a single optimal action. Recent work has proposed alternative formulations inspired by healthcare applications, including Set-Valued Policies (SVP), which map each state to multiple near-optimal actions to support clinician-in-the-loop decision making, as well as Dead-End Discovery (DeD), which eliminates high-risk actions in order to avoid undesirable outcomes. While SVP and DeD appear complementary, such that the actions not chosen by SVP could correspond to the same actions eliminated by DeD, and vice versa, the consistency of their recommendations has not been systematically studied. In this work, we empirically evaluate the consistency of SVP and DeD on clinically inspired grid-world simulations, as well as real-world ICU sepsis data. Our results show that the most natural closed-form condition on the hyperparameters, requiring the near-optimality margin and death threshold to sum to at most one, does not guarantee consistency. However, in real ICU data at clinically reasonable thresholds, conflicts are rare and arise in the most severe patient states. Analyzing consistency at the policy and state levels, we trace these conflicts to the differences between SVP's relative near-optimality criterion and DeD's absolute risk threshold. Building on this, we propose reconciling conflicting recommendations into a per-state partial ordering of the action set, intended to support rather than replace clinical judgment.

Keywords: AI for Healthcare, Decision Making, Reinforcement Learning, Clinician-in-the-Loop

1. Introduction

Reinforcement learning (RL) has traditionally been framed as the problem of solving for the optimal policy, which identifies one action for each state that yields the highest discounted expected cumulative reward.[1,2] This is done both for simplicity of algorithm implementation and because such deterministic policies are sufficient for optimally solving fully observable environments.[3] However, under this standard formulation, the non-optimal actions are entirely ignored. Among recent developments, Set-Valued Policies (SVP) by Tang et al.[4] and Dead-End Discovery (DeD) by Fatemi et al.[5] are two notable approaches that proposed alternative RL paradigms, both motivated by applications in healthcare. SVP produces a set of near-optimal actions for each state that lead to similar returns in the worst case rather than a single best action that leads to the highest return, which enables clinician-in-the-loop decision making.[6]

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

In contrast, DeD adopts a risk-sensitive perspective by proactively identifying irreversible states (e.g., severe deterioration or mortality) and eliminating actions with a high probability of leading to such states.[7,8]

While both SVP and DeD move beyond the rigid single-action policy of standard RL, they tackle the problem from different perspectives. SVP is local and relative: for each state, it recommends every action whose Q-value falls within a predefined "near-optimality margin" (e.g., 5%) of the local optimum. DeD is global and absolute: using all available data, it estimates the probability of death for each state-action pair and eliminates, per state, those exceeding a predefined "death threshold." Both SVP and DeD can be seen collectively as a mechanism to partition the action space into a desirable action set and an undesirable action set for each state. Since they share a similar goal of communicating action recommendations in a user-friendly format, it seems natural to allow the two algorithms to work together under one unified framework. However, it is unclear whether the recommendations from SVP and DeD are guaranteed to be consistent with each other. That is, in an ideal scenario, SVP should never recommend actions that DeD flags as high-risk. Understanding if and when these two approaches produce conflicting recommendations can guide us to better leverage their strengths to arrive at a coherent recommendation to present to clinicians. We posit that the distinction between the mathematical definitions of SVP and DeD creates a conflict in the most severe patient states, where SVP's local notion of "best" may favor an action that DeD's global risk perspective flags as exceeding its risk threshold.

In this paper, we answer this question via empirical evaluations on two RL domains: a synthetic LifeGate[5] environment and a real-world sepsis cohort from the MIMIC-III ICU database.[9] Investigating the consistency of recommendations for SVP and DeD across these two domains enables us to first systematically experiment with different conditions that aim at ensuring consistency and then confirm our hypotheses in a real-world example. Specifically, we consider the condition that the "near-optimality margin" and "death threshold" sum to at most 1. Our results show that SVP and DeD can indeed produce conflicting recommendations, and no simple, generalizable hyperparameter condition guarantees consistency. We therefore suggest reconciling conflicts in favor of SVP's recommendations because conflicting actions, though high-risk under global expectations, are often the best and sometimes only remaining clinically viable option, yielding the per-state partial ordering: non-conflicting SVP-recommended actions > conflicting actions > neutral actions > non-conflicting DeD actions.

2. Background

Reinforcement learning formalizes sequential decision-making as a Markov decision process (MDP), defined by the tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $P(s'|s,a)$ is the transition probability from state s to s' under action a, $R(s,a)$ is the expected reward under action a given state s, and $\gamma \in [0,1]$ is the discount factor. In the standard RL formulation, the goal is to learn a policy $\pi(a|s)$ that produces the highest expected discounted return. The state-action value function $Q^\pi(s,a)$ estimates the expected return from taking action a in state s and following policy π thereafter. In this work, we consider two approaches that move beyond standard RL policy formulation, namely SVP[4] and DeD.[5]

2.1. *Set-Valued Policies*

In set-valued policies (SVP), a policy maps each state to a set of near-optimal actions. Formally,

$$\pi_{\text{SVP}}(s) = \{a \in \mathcal{A} \,|\, Q^{\pi_{\text{SVP}}}(s,a) \geq (1-\zeta)V^*(s)\}\,,$$

where $\zeta \in [0,1]$ is a hyperparameter that specifies the "near-optimality margin." We learn this policy through the near-greedy value iteration originally proposed by Tang et al.[4] Specifically, $Q^{\pi_{\text{SVP}}}(s,a)$ represents the worst-case cumulative reward Q for actions recommended by SVP. Any actions that are within the "near-optimality margin" ζ of $V^*(s)$, the best cumulative reward achieved at the given state s, are selected by SVP. Therefore, all actions in the near-optimal set have a cumulative reward that is ζ-close to the optimal value $V^*(s)$. When SVP finds no actions within the margin of near-optimality, the algorithm falls back to the greedy action. By this formulation, SVP determines which actions to recommend based on a localized, relative measurement of optimality, computed for each state s.

2.2. *Dead-End Discovery*

In dead-end discovery (DeD), a policy maps each state to a set of high-risk actions. Formally,

$$\pi_{\text{DeD}}(s) = \{a \in \mathcal{A} \,|\, -Q^{\pi_{\text{DeD}}}(s,a) \geq \theta_D\}\,,$$

where $\theta_D \in [0,1]$ is a hyperparameter that specifies the "death threshold" or maximum probability of inevitable death the algorithm allows. We learn this policy through value iteration on the "death MDP" originally proposed by Fatemi et al.[5] Specifically, DeD estimates a separate value function $Q^{\pi_{\text{DeD}}}(s,a) \in [-1,0]$ from the "death MDP," such that $-Q^{\pi_{\text{DeD}}}(s,a)$ represents the best-case probability of inevitable death after taking action a in state s, assuming optimal future actions. The DeD action sets are created by thresholding these $-Q^{\pi_{\text{DeD}}}$ against θ_D.

2.3. *Relation to Other Safety-Aware RL*

SVP and DeD are both post hoc, set-valued offline RL algorithms built on standard value functions, which distinguishes them from methods that directly build safety into the learning objective itself. Previous work in risk-sensitive offline RL has optimized a tail statistic of the return rather than its expectation,[10] and within the dead-end formulation itself, Killian et al.[8] replaced the expected-value criterion with a risk-sensitive one. Additionally, conservative and pessimistic offline RL methods instead penalize actions that the data support poorly,[11] and multi-objective formulations return policies over explicit trade-offs.[12,13] Each of these methods yields an approximately optimal policy under its own criterion, which can be adapted to produce SVP and DeD policies; however, to our knowledge, we are the first to raise and study the question of consistency between SVP and DeD under standard value functions, which has not been well understood.

3. Experimental Setup

We conduct our experiments across two domains: (1) the LifeGate environment,[5] a synthetic grid-world domain designed to simulate the clinical setting; and (2) MIMIC-III,[9] a real-world

dataset of ICU patients, from which we use a cohort diagnosed with sepsis. While the synthetic setting enables us to systematically vary environmental factors to test conditions and determine sources of inconsistency, the real-world setting enables us to ground our conclusions with actual patients and concrete measurements of health conditions. Both SVP and DeD are trained using value iteration on their respective reward functions to obtain V^*. We use $\gamma = 1$ for both algorithms so that values are undiscounted outcome probabilities.[7] Any $\gamma < 1$ would make both criteria dependent on the time to outcome rather than the outcome itself.[7,14] We then run near-greedy value iteration[4] to construct the SVP policy, and threshold[5] $Q^{\pi_{\text{DeD}}}$ to construct the DeD policy. Since our setting is a finite-horizon episodic task with sparse terminal rewards and no intermediate rewards, using $\gamma = 1$ is mathematically valid and follows the undiscounted formulation of DeD.[5,7]

Although SVP's near-greedy value iteration is guaranteed to converge for directed acyclic MDPs with non-negative rewards,[4] both of our domains violate these conditions. In practice, the iteration frequently cycles among a finite set of candidate policies rather than converging. We therefore handle cycles explicitly. After every full sweep over $\mathcal{S}$, we recompute the near-greedy policy $\pi^{(k)}$ and store it. We declare convergence when $\pi^{(k)} = \pi^{(k-1)}$ and $\max_s |V^{(k)}(s) - V^{(k-1)}(s)| < 10^{-10}$. We declare a cycle when $\pi^{(k)}$ equals any earlier policy $\pi^{(i)}$ and output the intersection of the policies in the cycle, $\pi_{\text{cycle}}(s) = \bigcap_{k \in \text{cycle}} \pi^{(k)}(s)$. The intersection is deliberately conservative, recommending only actions endorsed by every policy in the cycle. A state whose intersection is empty receives the greedy optimal action.

3.1. *Consistency Metrics*

Given policies $\pi^{(h)}_{\text{SVP}}$ and $\pi^{(h)}_{\text{DeD}}$ trained under hyperparameters $h \in \mathcal{H}$, we say they are inconsistent at state $s \in \mathcal{S}$ if $\pi^{(h)}_{\text{SVP}}(s) \cap \pi^{(h)}_{\text{DeD}}(s) \neq \emptyset$. We aggregate this criterion in two ways to quantify conflict frequency. At the policy level, we fix h and compute the fraction of states that produce conflicting SVP and DeD action sets, formally:

$$\text{Conflict Fraction}(h) = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \mathbf{1}\Big[\, \pi^{(h)}_{\text{SVP}}(s) \cap \pi^{(h)}_{\text{DeD}}(s) \neq \emptyset \,\Big]$$

where $\mathbf{1}[\cdot]$ is the indicator function that equals 1 if true and 0 if false. At the state level, we instead fix the state s and measure the normalized frequency with which it produces a conflict over the hyperparameter space $\mathcal{H}$, formally:

$$\text{Norm Freq of Conflicts}(s) = \frac{1}{|\mathcal{H}|} \sum_{h \in \mathcal{H}} \mathbf{1}\left[\pi^{(h)}_{\text{SVP}}(s) \cap \pi^{(h)}_{\text{DeD}}(s) \neq \emptyset\right]$$

To measure how severe these conflicts are when they do occur, we additionally compute the intersection-over-union (IOU) between π_{SVP} and π_{DeD} across states. This represents an assessment of conflict severity at the state level, given either a specific hyperparameter setting or averaged across $\mathcal{H}$, formally:

$$\text{IOU}(s, h) = \frac{\left|\pi^{(h)}_{\text{SVP}}(s) \cap \pi^{(h)}_{\text{DeD}}(s)\right|}{\left|\pi^{(h)}_{\text{SVP}}(s) \cup \pi^{(h)}_{\text{DeD}}(s)\right|}$$

where we define the ratio to be 0 whenever $\pi_{\text{SVP}}(s) \cup \pi_{\text{DeD}}(s) = \emptyset$.

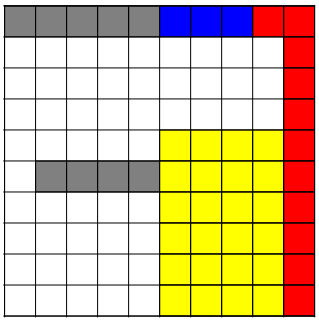


Fig. 1. LifeGate environment with barriers in gray, dead-end states in yellow, death states in red, recovery states in blue, and neutral states in white.

3.2. *LifeGate*

In the LifeGate environment,[5] an agent's action set comprises moving up, down, right, left, or staying in place. We visualize the LifeGate environment in Figure 1, which includes the following states. Barriers: redirect the agent back to its current position with no reward or penalty. Dead-end States: irreversible failure conditions. Upon entry, the agent remains in place with 30% probability and is pushed one cell right by a "death-drag" with 70% probability, simulating deterioration that cannot be avoided even with intervention. Death States: immediate termination as a failure to recover. Recovery States ("life gates"): immediate termination as a successful recovery. Neutral States: 53 states later used for analysis that comprise the majority of the environment. They allow navigation without immediate reward or penalty but are subject to a natural drag pushing the agent right with 40% probability regardless of the chosen action, a tunable proxy for inevitable patient deterioration.

We implemented the environment with two different reward functions tailored to the goals of SVP and DeD. For SVP, the agent receives a reward of $+1$ when reaching a recovery zone and -1 when reaching a death zone, with zero reward for all intermediate transitions to ensure the value function reflects both long-term success and failure.[4] For DeD, the agent receives a reward of -1 only when entering a death zone, while all other transitions yield zero reward.[5] This ensures the value function reflects the probability of reaching eventual death.

3.3. *MIMIC-III Sepsis*

The MIMIC-III[9] dataset comprises ICU patients, including 19,626 diagnosed with sepsis, a life-threatening organ dysfunction caused by a dysregulated host response to infection[15] and a leading cause of hospital mortality.[16] Patient data are represented as time series recorded every 4 hours, spanning from 24 hours before to 48 hours after the onset of sepsis. Following the preprocessing procedure detailed by Komorowski et al.,[16] we cluster samples into 750 discrete health states using K-Means and add two terminal states, corresponding to death and recovery. We then discretize the total amount of IV fluids and vasopressors administered into 25 discrete actions following Tang et al.,[4] a well-established choice in the sepsis RL literature[16–20] that matches both how clinicians conceptualize dosing as discrete levels and the discrete action sets over which SVP and DeD inherently operate.

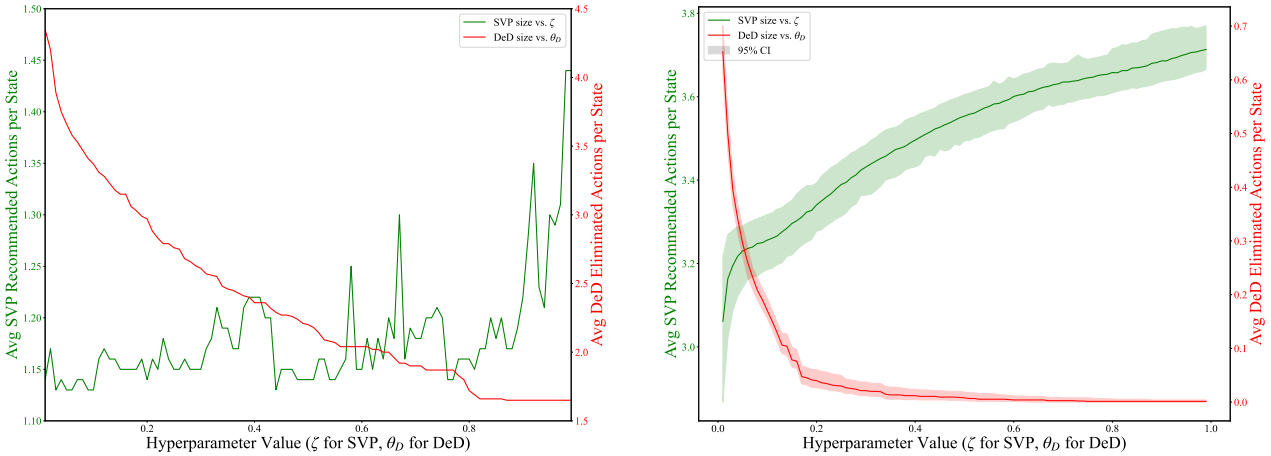


Fig. 2. Plots of average SVP (green, left axis) and DeD (red, right axis) action set sizes vs. hyperparameters ζ and θ_D for the LifeGate (left plot) and MIMIC-III (right plot, 95% CI) experiments.

We use reward setups similar to those in LifeGate: for SVP, we assign rewards of $+1$ for recovery, -1 for death, and 0 for intermediate transitions; for DeD, we assign -1 for death, and 0 for all other steps. Additionally, we mask out state-action pairs that are observed fewer than 6 times. This prevents the learned policies from extrapolating to actions that are rarely taken by clinicians. The masked pairs carry 8.8% of the recorded transitions, compared with 4.0% and 13.6% at thresholds of 3 and 10.

4. Results

Our experiments aim to investigate the consistency of SVP and DeD across the two experimental settings. In each setting, we evaluate consistency at the policy and state levels. At the policy level, we first examine how SVP and DeD action set sizes vary with hyperparameters ζ and θ_D. Then, we analyze how consistency changes as we sweep the hyperparameter space. Specifically, we train 99 SVP policies with $\zeta \in \{0.01, 0.02, \ldots, 0.99\}$ and 99 DeD policies with $\theta_D \in \{0.01, 0.02, \ldots, 0.99\}$. We then evaluate the frequency of consistency for every pair (ζ, θ_D) to create a matrix of conflict fractions. Afterwards, we zoom in on the state level to analyze which states are most vulnerable to conflicts and subsequently discuss ways to reconcile these conflicting states. For LifeGate, value iteration on the fixed MDP is deterministic, so no error bounds apply. For MIMIC-III policy-level results, we report the bootstrap median with a 95% confidence interval over 100 resamples of the training-set ICU stays drawn with replacement.

4.1. *LifeGate Results*

Policy-Level. Looking at the LifeGate environment at the policy level, we start by investigating how ζ and θ_D affect SVP size and DeD size, respectively. In Figure 2, we plot the average number of actions per state across the range of hyperparameters. As expected, a higher ζ generally tends to produce a larger SVP set, since more actions are considered near-optimal. On the other hand, increasing θ_D results in fewer actions being eliminated by DeD, as the threshold for risk becomes more tolerant. However, the relationship between DeD size and θ_D

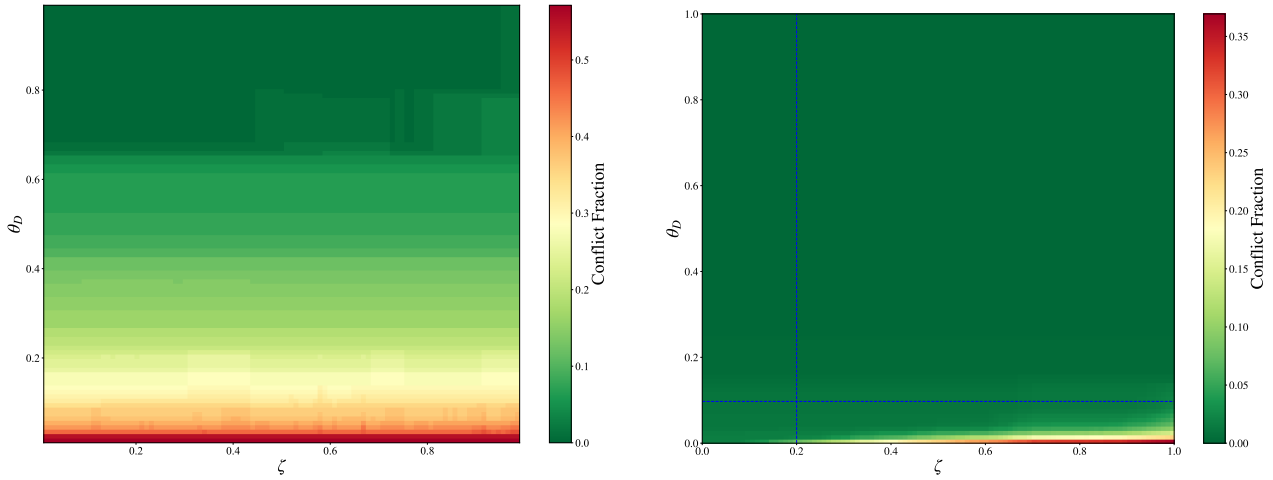


Fig. 3. Heatmaps of SVP and DeD inconsistency at the policy level, measured as the conflict fraction, under different hyperparameter combinations of (ζ, θ_D) for the LifeGate (left) and MIMIC-III (right) experiments. Dark green indicates no conflicts. The blue dashed lines in the MIMIC-III heatmap indicate the selected hyperparameters for state-level experiments ($\zeta = 0.2, \theta_D = 0.0973$).

is monotonic, whereas that between SVP size and ζ is not. This can be attributed to the LifeGate environment's design, in which the agent can cycle indefinitely. Furthermore, we observe that SVP and DeD always recommend and eliminate at least one treatment on average for all pairs of (ζ, θ_D), confirming that all policies are non-trivial.

Next, Figure 3 visualizes policy-level inconsistency as a heatmap over hyperparameter settings (ζ, θ_D). Consistency is not guaranteed, as a large proportion of configurations produce conflicts. Only the dark green top-left region, where the most conservative settings sit, is fully conflict-free, while the extreme opposite bottom-right corner reflects the most conflicts. Fixing ζ, the conflict fraction increases as θ_D decreases, tracking DeD's monotonically growing set size. Fixing θ_D, it also tends to increase with ζ, but less pronouncedly and non-monotonically, mirroring the non-monotonic SVP set sizes in Figure 2.

No prior work proposes a condition under which SVP and DeD are consistent, and the most intuitive condition is $\zeta + \theta_D \leq 1$. Since ζ controls the tolerance of SVP for near-optimal actions and θ_D controls the tolerance of DeD for high-risk actions, we might reasonably expect the SVP and DeD action sets to be exact complements of each other when $\zeta + \theta_D = 1$. However, looking at the corresponding lower-left region of the heatmap in Figure 3, it is clear that this condition does not guarantee consistency. Out of 4,950 parameter combinations satisfying $\zeta + \theta_D \leq 1$, only 496 (10.02%) resulted in consistent policies. Furthermore, many combinations that sum to > 1 do achieve consistency. We identify a dense band of hyperparameter pairs that always yield consistent policies: $\zeta \in [0.01, 0.44]$ and $\theta_D \in [0.69, 0.99]$. Thus, the condition fails in both directions, indicating that the relationship between ζ and θ_D is not additive at all. Furthermore, given the irregular contours we observed in the heatmap, we could not find theoretical consistency guarantees between SVP and DeD for the LifeGate environment.

State-Level. Looking closer at the state-level inconsistencies between SVP and DeD, we compute the frequency and severity of conflicts between SVP and DeD action sets at each

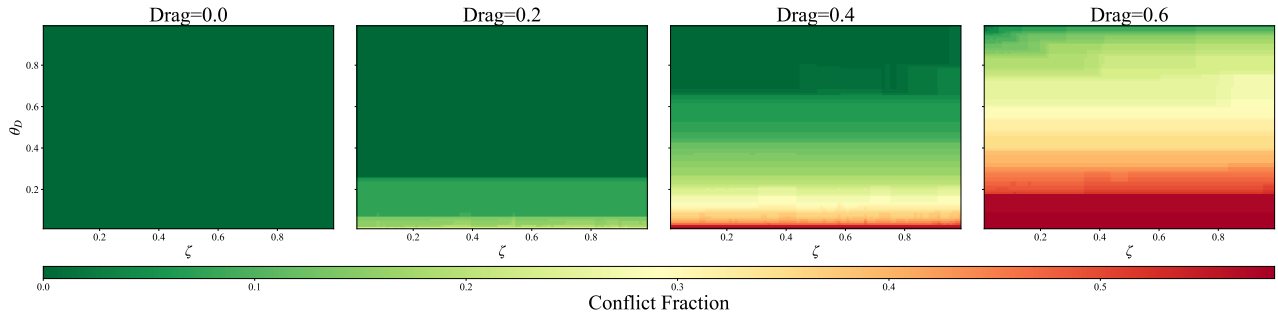


Fig. 4. Comparison of policy-level inconsistency heatmaps with natural drag ranging from 0.0 to 0.6, measured as the conflict fraction across different hyperparameter combinations (ζ, θ_D).

state by computing the normalized frequency of inconsistencies and intersection-over-union (IOU), averaged across hyperparameter pairs (ζ, θ_D). Exactly one state, the leftmost cell of the middle row of Figure 1, is conflict-free under every hyperparameter setting tested. Additionally, the two measures are strongly correlated across states (Spearman $\rho = 0.965$), so states that conflict more often also conflict more severely. Specifically, conflict frequency and IOU are highest for states with at least one neighbor in the death or dead-end set (mean frequency 0.501, mean IOU 0.112) and lower for the remaining states (mean frequency 0.136, mean IOU 0.0302). This suggests some correlation between the proximity to death termination (i.e., severity of a patient's health) and the frequency and severity of conflicts in SVP and DeD policies.

Furthermore, the one conflict-free state is also the only state that is invulnerable to natural drag, as it is immediately left of a barrier state, suggesting a relationship between the probability of inevitable death and inconsistency. We analyze this directly by regenerating the policy-level heatmaps with varying natural drag in Figure 4. For each plot, we configure the LifeGate environment with natural drag ranging from 0.0 to 0.6. At a natural drag of 0.0, there is an exact optimal action for each state that will always lead to survival, and SVP and DeD are fully consistent for 100% of the 9,801 hyperparameter pairs. As natural drag increases toward 0.6, the probability of inevitable death increases alongside the conflict fractions across all hyperparameter pairs.

4.2. *MIMIC-III Results*

Policy-Level. Similarly, we start investigating the MIMIC-III dataset by examining SVP and DeD set sizes across the hyperparameter space. As shown in Figure 2, we observe the same trend where DeD set size decreases monotonically as θ_D increases. Interestingly, we see that SVP sizes also appear to increase almost monotonically. We attribute this near monotonicity to the real clinical setting, where the MDP cycles common in the LifeGate environment are significantly reduced. In addition, the average DeD sets are smaller and the SVP sets larger than in LifeGate, as this clinical setting offers more plausible routes to recovery and fewer actions posing irreversible mortality risk. Consequently, although the average DeD set sizes flatten out to 0, these empty sets are not trivial; they indicate that no available action explicitly pushes the patient toward death, not that DeD fails to flag any.

In Figure 3, we create the same policy-level heatmap visualization of inconsistency across hyperparameter combinations (ζ, θ_D). As expected from the nearly monotonic SVP set sizes in Figure 2, we see a much more pronounced and monotonic increase in the conflict fraction as both θ_D decreases and ζ increases. Unlike in LifeGate, inconsistent recommendations are much rarer. Whenever $\theta_D \geq 0.34$, the two algorithms are always consistent regardless of ζ, though the $\zeta + \theta_D \leq 1$ condition still fails here. Cross-referencing Figure 2, DeD sizes are nearly zero when $\theta_D \geq 0.34$, indicating that these are trivial hyperparameter pairs: consistency there stems from standard clinical care in this cohort already preventing any action from exceeding a 34% probability of inevitable death.

State-Level. To further explore consistency at the state level, we start by proposing an applicable hyperparameter pairing. We set θ_D to the cohort's empirical mortality rate (9.73%) and $\zeta = 0.2$. We do not claim this is the pair an ICU team would choose, but it acts as a natural reference point, and the conclusions we draw from it are not sensitive to these choices. The region around this choice of hyperparameters, which is shown by the blue dashed lines in Figure 3, indicates no sharp fluctuations in conflict fraction. In fact, the number of conflicting states stays between 3 and 9 across $\zeta \in [0.1, 0.3]$ and $\theta_D \in [0.05, 0.2]$ and never changes by more than one state between adjacent settings. For comparison, the levels chosen by Fatemi et al.[5] for their MIMIC-III sepsis cohort are $\theta_D = 0.15$ and 0.25, both of which correspond to fewer conflicting states.

Using the DeD state values $V^{\pi_{\text{DeD}}}$, whose negation is a state's best-case probability of inevitable death, we obtain a concrete measure of a patient's health severity. By ranking patient states according to $V^{\pi_{\text{DeD}}}$, we can empirically investigate the relationship between health severity and consistency. To examine only the SVP-chosen outputs, we also experiment with removing the greedy fallback detailed in the SVP algorithm, returning a non-greedy policy with an average policy size of 4.162 and conflict fraction of 0.001 (1 of 750 states). For comparison, a non-greedy loose SVP policy at a much higher $\zeta = 0.8$ has an average SVP policy size of 4.543 and a conflict fraction of 0.004 (3 of 750 states).

In Figure 5, we plot a heatmap of the DeD Q-values over the 10 most clinically severe states. These states are sorted by $V^{\pi_{\text{DeD}}}$, with the most severe states at the top. We then overlay the strict SVP policy (the chosen setting, $\zeta = 0.2$) and the DeD policy on top of the heatmap using blue circles and red crosses, respectively. In the conflicting states, SVP's recommendations correspond to the actions with the smallest best-case probability of inevitable death, $-Q^{\pi_{\text{DeD}}}(s, a)$. Seven of the eight conflicts arise solely from the greedy fallback when $V^*(s) \geq 0$ because even the action deemed best by both algorithms exceeds $\theta_D = 0.0973$ (between 0.125 and 0.333). The single remaining conflict in State 271 results from there being only one action available, so the relative criterion is trivially satisfied.

At the much looser margin $\zeta = 0.8$, two additional conflicts appear outside the severe states, where DeD does not eliminate all actions and SVP is not forced into selecting a greedy action. To contextualize these conflicts, we examine the empirical distributions of the normalized MIMIC-III features for the two conflicting states. Specifically, State 47 shows relatively low normalized systolic, diastolic, and mean blood pressure (BP) of -1.1, -0.5, and -0.8 (97.8, 46.1, and 63.6 raw), respectively. This indicates hypotension, and the conflicting Action 0

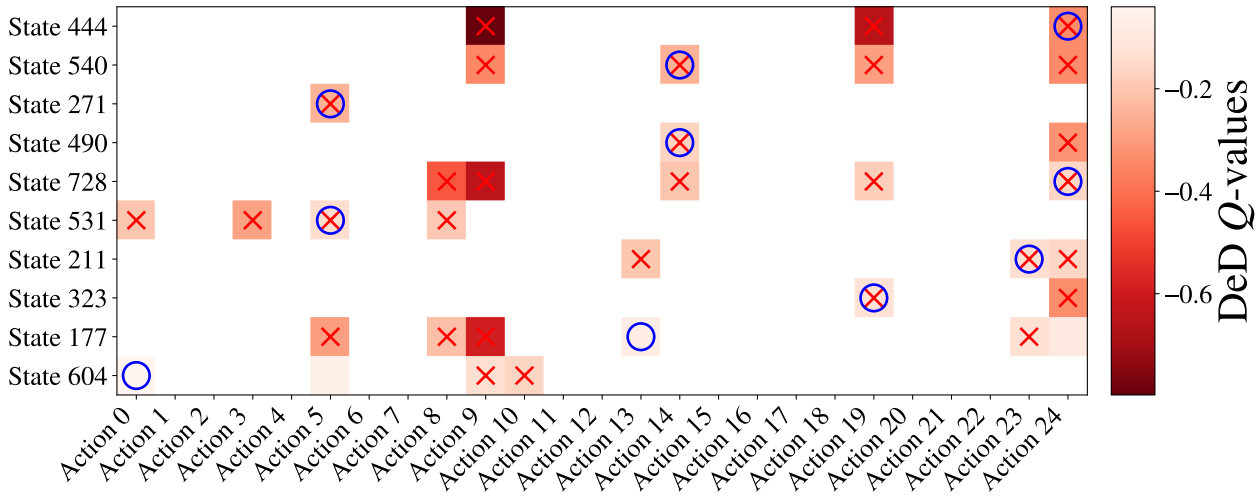


Fig. 5. Heatmap of DeD Q-values plotted against SVP-chosen (state, action) pairs for the MIMIC-III dataset at the chosen setting $(\zeta, \theta_D) = (0.2, 0.0973)$. Each row is one patient state, ordered from most (top) to least severe by $V^{\pi_{\text{DeD}}}(s)$. Blue circles indicate SVP-chosen actions and red crosses indicate DeD-removed actions. White cells indicate masked-out pairs that occur fewer than 6 times in the dataset.

corresponds to administering neither IV fluids nor vasopressors. The overly permissive "near-optimality margin" considers the patient healthy enough that administering no treatment is comparable to other treatments, while DeD flags it as too high risk. This corresponds to a high-risk, high-reward option frequently taken by clinicians (34 of 160 visits). On the other hand, State 546 exhibits an extremely low normalized Glasgow Coma Scale (GCS) score of −2.1 (5.7 raw), indicating severely impaired consciousness, and the conflicting Action 7 corresponds to a small amount of IV fluids with moderate-low vasopressors. However, Action 7 was observed only 6 times, which we read as poorly supported extrapolation rather than a frequently taken high-risk, high-reward action.

5. Discussion

In this work, we have investigated the consistency between SVP and DeD, two recent approaches that move beyond the standard RL formulation of single-action optimal policies. Although intuitively complementary, with SVP retaining "good" actions and DeD eliminating "bad" ones, their recommendations are not always consistent, even under the most natural conditions one might impose. Our observations demonstrate that these inconsistencies are grounded in the relative versus absolute perspectives from which the two algorithms are defined. This is not specific to SVP and DeD: any pairing of a relative optimality criterion with an absolute risk constraint can create these inconsistencies. States carrying a non-zero probability of inevitable death emerge as the underlying trigger, and conflict frequency correlates with the severity of a patient's health.

While the one condition we found that guarantees consistency depends on a property of the environment we can only set synthetically, and we found no theoretical guarantee

for choosing hyperparameters, these conflicts are not as detrimental as they seem. Under reasonable hyperparameter settings chosen via methods similar to ours, conflicts arise only for the most extreme patients, typically coinciding with SVP's greedy selections, which still correspond to the smallest available best-case probability of inevitable death. In other words, when no action clears the near-optimality margin, SVP defaults to the best remaining choice, the one that minimizes the probability of inevitable death. In these extreme cases, DeD's flag should be read as what it is, a statement that every available action leaves the patient above the death threshold, rather than as a reason to suppress every action, leaving the clinician without decision support precisely where it is most needed.

Reconciled Ordering. For this reason, we propose a practical reconciliation: when SVP and DeD conflict, the algorithm should prioritize SVP's recommendations and interpret DeD as contextualized risk rather than a strict exclusion. Because these treatments are inherently high-risk, high-reward and require consideration in critical scenarios, we rank them above the neutral actions that are neither endorsed by SVP nor eliminated by DeD. In turn, we create a partial ordering from best to worst actions in the action space for each patient: non-conflicting SVP-recommended actions > conflicting actions > neutral actions > non-conflicting DeD actions. This also aligns with the methodology of clinician-in-the-loop decision making, where expert judgment can identify high-risk actions as potentially life-saving instead of strictly recommending and eliminating actions outright.

To evaluate this ordering, we compute the clinical mortality rate for each type of action in the reconciled ordering. For each conflicting state across all hyperparameter pairs, we run a one-step model for every action and then compute the mortality rate of the subsequent state. These mortality rates are first averaged within each of the four action types in a state and then across states, with a 95% confidence interval computed from cluster bootstrapping. Over all 277 conflicting states, actions SVP recommends and DeD does not flag give an estimated one-step mortality of 14.91% (13.81%–16.08%), conflicting actions at 17.88% (16.31%–19.60%), neutral actions at 25.60% (20.47%–31.19%), and actions DeD flags and SVP does not recommend at 29.87% (26.79%–33.06%). These estimates fall in exactly the order we propose, and the conflicting actions sit far closer to the SVP recommendations (2.97 percentage points above) than to the actions DeD flags alone (11.99 points below).

We present this ordering as a hypothesis supported by preliminary empirical evidence, not as a validated design. The estimates above are consistent with treating the DeD flag in conflicting states as contextualized risk rather than grounds for removing the action. However, they come from a one-step model rather than from clinical use, and a structured review with critical-care clinicians is needed to validate the ordering. Until then, any clinician consulting our reconciled ordering should read it as a whole rather than acting on a single tier. Ultimately, these models are intended to support clinician decision making rather than replace it.

Limitations. We acknowledge that our analysis is empirical rather than theoretical. Based on the irregular consistency patterns observed across our hyperparameter sweep, we believe a clean theoretical guarantee is unlikely to exist. Additionally, while continuous state and action representations could yield more accurate results, our use of discrete representations is a deliberate choice that enables a clean investigation of consistency between SVP and DeD, which is

the central focus of this work. Our MIMIC-III results also inherit the known limitations[16,21,22] of the Komorowski-style pipeline: clustered states discard within-cluster heterogeneity, the fluid and vasopressor action space omits antimicrobials and ventilation, and transitions are estimated from a single center. Our masking threshold is modeled on the action-restriction rule of Tang et al.,[4] who keep only the actions observed at least five times in each state. The effect of the chosen threshold is not neutral across treatments, as pairs involving the two highest vasopressor doses make up 35.9% of the observed pairs and only 16.5% of those that survive masking, so the rarer high-dose combinations are pruned more often than typical care. However, we expect our analysis to generalize, since the same patterns appear in LifeGate, where the dynamics are known exactly. Finally, our cycle-handling rule, intersecting the policies visited in a cycle, is a conservative choice rather than a validated one, and we do not compare it against alternative resolutions. Its effect on the measured conflict rate is unknown, though we expect our downstream analysis to still hold.

Future Work. Future work should aim at developing a more rigorous solution for resolving conflicts in the extreme edge cases, building on the results of our experiments. Cycling itself is also worth addressing directly: SVP's convergence guarantee covers only directed acyclic MDPs with non-negative rewards, and neither of our domains satisfies those conditions, nor do clinical MDPs estimated from observational data in general. Extending the guarantee to cyclic settings, or characterizing what the cycle-intersection rule converges to, would put these states on firmer ground than the conservative solution we adopt. Another promising direction is a hybrid framework[12,13] that jointly trains the two algorithms rather than treating them as independent components. Pursuing these directions may lead to safer and more adaptable decision-making support, particularly in high-risk domains such as healthcare, where both flexibility and risk-awareness are essential.[23,24]

Code Release. All code to reproduce every result and figure in this paper is released at github.com/Veeeeeee7/Reconciling-SVP-DeD. The LifeGate data were generated through the simulation included in the code, but the MIMIC-III data require credentialed PhysioNet access.

References

1. R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction* (The MIT Press, Cambridge, MA, 1998).
2. R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd edn. (The MIT Press, Cambridge, MA, 2018).
3. M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming* (John Wiley & Sons, New York, 1994).
4. S. Tang, A. Modi, M. W. Sjoding and J. Wiens, Clinician-in-the-loop decision making: Reinforcement learning with near-optimal set-valued policies, in *Proceedings of the 37th International Conference on Machine Learning*, 2020.
5. M. Fatemi, T. W. Killian, J. Subramanian and M. Ghassemi, Medical dead-ends and learning to identify high-risk states and treatments, in *Advances in Neural Information Processing Systems*, 2021.
6. M. Milani Fard and J. Pineau, Non-deterministic policies in markovian decision processes, *Journal of Artificial Intelligence Research* **40**, 1 (2011).

7. M. Fatemi, S. Sharma, H. van Seijen and S. E. Kahou, Dead-ends and secure exploration in reinforcement learning, in *Proceedings of the 36th International Conference on Machine Learning*, 2019.
8. T. W. Killian, S. Parbhoo and M. Ghassemi, Risk sensitive dead-end identification in safety-critical offline reinforcement learning, *Transactions on Machine Learning Research* (2023).
9. A. E. W. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi and R. G. Mark, Mimic-iii, a freely accessible critical care database, *Scientific Data* **3**, p. 160035 (2016).
10. N. Armengol Urpí, S. Curi and A. Krause, Risk-averse offline reinforcement learning, in *International Conference on Learning Representations*, 2021.
11. A. Kumar, A. Zhou, G. Tucker and S. Levine, Conservative q-learning for offline reinforcement learning, in *Advances in Neural Information Processing Systems*, 2020.
12. D. J. Lizotte and E. B. Laber, Multi-objective markov decision processes for data-driven decision support, *Journal of Machine Learning Research* **17**, 1 (2016).
13. H. Satija, P. S. Thomas, J. Pineau and R. Laroche, Multi-objective spibb: Seldonian offline policy improvement with safety constraints in finite mdps, in *Advances in Neural Information Processing Systems*, 2021.
14. Y. Shi, M. Lafrance and S. Tang, Between life and death: Examining sparse reward designs in healthcare rl, in *RLC 2025 Workshop on Practical Insights into RL for Real Systems (RL4RS)*, 2025.
15. J. E. Gotts and M. A. Matthay, Sepsis: pathophysiology and clinical management, *BMJ* **353**, p. i1585 (2016).
16. M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon and A. A. Faisal, The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care, *Nature Medicine* **24**, 1716 (2018).
17. A. Raghu, M. Komorowski, L. A. Celi, P. Szolovits and M. Ghassemi, Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach, in *Proceedings of Machine Learning for Healthcare*, 2017.
18. X. Peng, Y. Ding, D. Wihl, O. Gottesman, M. Komorowski, L.-w. H. Lehman, A. Ross, A. Faisal and F. Doshi-Velez, Improving sepsis treatment strategies by combining deep and kernel-based reinforcement learning, *AMIA Annual Symposium Proceedings* **2018**, 887 (2018).
19. K. Choudhary, D. Gupta and P. S. Thomas, Icu-sepsis: A benchmark mdp built from real medical data, *Reinforcement Learning Journal* **4**, 1546 (2024).
20. S. Tang, J. Yao, J. Wiens and S. Parbhoo, Off by a beat: the effects of temporal misalignment in reinforcement learning for sepsis treatment, *npj Digital Medicine* **9**, p. 360 (2026).
21. R. Jeter, C. Josef, S. Shashikumar and S. Nemati, Does the "artificial intelligence clinician" learn optimal treatment strategies for sepsis in intensive care?, *arXiv preprint arXiv:1902.03271* (2019).
22. O. Gottesman, F. Johansson, M. Komorowski, A. Faisal, D. Sontag, F. Doshi-Velez and L. A. Celi, Guidelines for reinforcement learning in healthcare, *Nature Medicine* **25**, 16 (2019).
23. S. Gu, L. Yang, Y. Du, G. Chen, F. Walter, J. Wang and A. Knoll, A review of safe reinforcement learning: Methods, theories, and applications, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **46**, 11216 (2024).
24. V. Sivaraman, L. A. Bukowski, J. Levin, J. M. Kahn and A. Perer, Ignore, trust, or negotiate: Understanding clinician acceptance of ai-based treatment recommendations in health care, in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023.