Pragmatic Policy Development via Interpretable Behavior Cloning

Anton Matsson ANTMATS@CHALMERS.SE
Chalmers University of Technology and University of Gothenburg

Yaochen Rao YAOCHENR@CHALMERS.SE
Chalmers University of Technology and University of Gothenburg

Heather J. Litman HEATHER.LITMAN@THERMOFISHER.COM
Thermo Fisher Scientific

Fredrik D. Johansson FREDRIK.JOHANSSON@CHALMERS.SE
Chalmers University of Technology and University of Gothenburg

Abstract

Offline reinforcement learning (RL) holds great promise for deriving optimal policies from observational data, but challenges related to interpretability and evaluation limit its practical use in safety-critical domains. Interpretability is hindered by the black-box nature of unconstrained RL policies, while evaluation—typically performed off-policy—is sensitive to large deviations from the data-collecting behavior policy, especially when using methods based on importance sampling. To address these challenges, we propose a simple yet practical alternative: deriving treatment policies from the most frequently chosen actions in each patient state, as estimated by an interpretable model of the behavior policy. By using a tree-based model, which is specifically designed to exploit patterns in the data, we obtain a natural grouping of states with respect to treatment. The tree structure ensures interpretability by design, while varying the number of most common actions considered controls the degree of overlap with the behavior policy, enabling reliable off-policy evaluation. This pragmatic approach to policy development standardizes frequent treatment patterns, capturing the collective clinical judgment embedded in the data. Using real-world examples in rheumatoid arthritis and sepsis care, we demonstrate that policies derived under this framework can outperform current practice, offering interpretable alternatives to those obtained via offline RL.

Keywords: decision-making, reinforcement learning, off-policy evaluation, interpretability

Data and Code Availability We analyze two datasets: 1,565 rheumatoid arthritis (RA) patients from the PPD™ CorEvitas™ RA registry (Kremer, 2005) and 11,482 sepsis patients from the publicly available MIMIC-III database (Johnson et al., 2016). RA data are available from CorEvitas, LLC through a commercial agreement and are not publicly available. The code is available at https://github.com/Healthy-AI/ppdev.

Institutional Review Board (IRB) Research on de-identified data from MIMIC is exempt from IRB review under HIPAA. Approval for the study of treatment patterns within the CorEvitas RA registry was granted by the Swedish Ethical Review Authority (application no. 2021-06144-01).

1. Introduction

Observational data provide a valuable foundation for applying machine learning (ML) to derive treatment policies that enhance clinical decision-making (Chakraborty and Moodie, 2013). Recently, significant attention has been devoted to reinforcement learning (RL), a branch of ML focused on learning optimal decision-making strategies (Sutton and Barto, 2018). While classical RL relies on trial-and-error learning—ill-suited to high-stakes domains like healthcare—offline RL learns from previously collected data, offering the potential to turn static datasets into effective decision-making engines (Levine et al., 2020). However, applying offline RL in clinical settings presents several well-known challenges (Yu et al., 2021; Jayaraman et al., 2024), with the lack of interpretability and the difficulty of policy evaluation being among the most significant. These limitations raise the question of whether other

approaches may be better suited for practical clinical use.

Policy evaluation involves assessing the performance of a newly derived policy, often referred to as the target policy. In safety-critical domains, this evaluation—like policy learning—must be performed using an offline dataset, a problem known as off-policy evaluation (OPE). A large class of OPE methods rely on importance sampling (IS) (Precup et al., 2000), where outcomes from patient trajectories collected under the behavior policy—that is, the current, observed decision-making behavior—are weighted by how likely those trajectories would be under the target policy. When the target and behavior policies differ substantially, IS-based estimates of performance tend to exhibit high variance. While this issue can be mitigated—for example, by normalizing weights (Precup et al., 2000) or incorporating model-based components into the estimator (Jiang and Li, 2016; Thomas and Brunskill, 2016; Farajtabar et al., 2018)—reliable OPE generally requires that the policies are sufficiently similar (Gottesman et al., 2018; Voloshin et al., 2021).

The challenge of interpretability arises from the fact that much of RL's recent success is due to its integration with deep learning (deep RL), where black-box neural networks are used to represent policies (Mnih et al., 2015). Such target policies are typically difficult to interpret, and this lack of transparency can prevent domain experts from identifying errors or artifacts, potentially undermining trust in medical applications (Pace et al., 2022; Lipton, 2017). While there have been efforts to improve the interpretability of RL policies—either directly by defining interpretable policy classes (Silva et al., 2020; Verma et al., 2019; Hein et al., 2018), or indirectly by distilling interpretable policies from black-box models (Verma et al., 2018; Liu et al., 2019)—the prevailing view is that deep RL is not yet ready for deployment in high-stakes domains such as healthcare (Glanois et al., 2024).

Combining ideas from interpretable RL with robust offline RL, where the target policy is constrained to stay close to the behavior policy to improve evaluability (Fujimoto et al., 2019; Kostrikov et al., 2022; Kumar et al., 2020), offers a promising direction for addressing these practical limitations. However, before deploying the full RL machinery, it is reasonable to ask: can we develop interpretable policies that are amenable to reliable OPE in a simpler and more pragmatic way?

Contributions We propose using supervised learning of the behavior policy—also known as behavior cloning (Torabi et al., 2018)—to derive interpretable and evaluable target policies for clinical decision-making. Specifically, the proposed policy is constructed based on the most frequently chosen treatments in each patient state, as estimated by the behavior policy model, with the option to incorporate their observed outcomes to further guide treatment selection. As such, the policy exploits the collective clinical expertise embedded in the data. By varying the number of treatments considered, we control the degree of overlap with the behavior policy, facilitating reliable OPE. Furthermore, by choosing an interpretable model class for the behavior policy, the resulting target policy is interpretable by design. While several model classes are possible, we recommend using a tree-based structure, as it provides a natural partitioning of the patient space based on observed treatment patterns. To utilize a common situation in clinical decision-making—where patients often remain on the same treatment across decision points—we construct a meta-model that uses separate trees to predict whether a patient will switch treatments and, if so, which treatment they will switch to.

We refer to this approach as *pragmatic policy development*. While we cannot guarantee that such policies outperform current practice, they are explicitly designed to be amenable to OPE, enabling meaningful comparison. In experiments, we find that policies developed under this framework are, on average, estimated to have higher policy values than current practice in real-world examples from the management of rheumatoid arthritis (RA) and sepsis. In contrast, policies derived using offline RL yield estimates with high variance, raising questions about their practical relevance.

2. Leveraging Observational Data to Improve Clinical Decision-Making

The treatment of patients with chronic or acute diseases, such as RA and sepsis, can be formulated as a multi-stage decision process involving states $S_t \in \mathcal{S}$, actions $A_t \in \{1, \ldots, K\}$, and rewards $R_t \in \mathbb{R}$. At each stage $t = 1, \ldots, T$, the clinician selects a treatment A_t based on the patient's medical history, summarized in the state S_t; the reward R_t reflects the outcome of the chosen treatment. A patient's medical history comprises the sequence of covariates, actions,

and rewards observed up to stage t, where covariates $X_t \in \mathcal{X}$ include, for example, demographics, diagnostic test results, and comorbidities. Prior work by Matsson et al. (2024b) suggests that a sufficient state representation includes the current covariates X_t, the most recent action A_{t-1}, and the most recent reward R_{t-1}, along with simple aggregates of earlier history.

The process by which clinicians treat patients gives rise to trajectories of state-action-reward triplets, $\tau = S_1, A_1, R_1, \ldots, S_T, A_T, R_T$, which are assumed to be collected in a dataset $\mathcal{D}$. Patterns in how clinicians select treatments define the *behavior policy* μ. We quantify these treatment patterns by estimating $p_\mu(A_t \mid S_t)$, the probability of selecting treatment A_t given state S_t under current clinical practice, based on state-action pairs in $\mathcal{D}$. In other contexts, this is often referred to as propensity estimation (Abadie and Imbens, 2016), policy recovery (Deuschel et al., 2024) or behavior cloning (Torabi et al., 2018). We let $p_{\hat{\mu}}(A_t \mid S_t)$, or simply $\hat{\mu}$, denote a probabilistic model of the behavior policy.

A *target policy* π represents an alternative treatment strategy. We use $p_\pi(A_t \mid S_t)$ to denote the probability of taking action A_t in state S_t under this policy. Reinforcement learning is a classical approach to optimizing such policies (Sutton and Barto, 2018), where optimality is defined in terms of the expected cumulative reward, or policy value, V^π: $V^\pi := \mathbb{E}_\pi \left[\sum_{t=1}^{T} R_t\right]$. In a clinical context, the target policy must typically be derived solely from observations in $\mathcal{D}$—a setting referred to as offline RL (see, for example, Levine et al. (2020) for an overview). In principle, any off-policy RL algorithm, where the policy being learned differs from the data-collecting policy, can be applied to the fixed dataset $\mathcal{D}$. Most prior works in clinical settings have used methods based on Q-learning (Yu et al., 2021).

2.1. Interpretability and Evaluability

Before deploying a target policy π in practice, its performance must be evaluated relative to the behavior policy μ. In the offline setting, where no data collected under the target policy is available, this evaluation amounts to estimating the value of π using data collected under μ—a problem known as off-policy evaluation. OPE is inherently challenging due to the difference between the two policies, and the difficulty increases as the divergence between π and μ grows. This issue is especially pronounced in methods based on importance sampling (Precup et al., 2000), where observed rewards are weighted by the importance weights $\prod_t \frac{p_\pi(a_t|s_t)}{p_{\hat{\mu}}(a_t|s_t)}$, but exists for other methods as well. For example, value-based (Le et al., 2019) or model-based (Paduraru, 2013) approaches require extrapolation to account for the mismatch between the policies. As a result, the target policy must remain sufficiently close to the behavior policy to enable reliable offline evaluation (Gottesman et al., 2018).

A large class of offline RL methods tackles the distributional shift between target and behavior policies—which poses challenges not only for evaluation but also during training and deployment (Levine et al., 2020)—by explicitly constraining the target policy to remain close to the behavior policy during learning (Fujimoto et al., 2019; Kostrikov et al., 2022; Kumar et al., 2020). However, a key drawback of these methods—and RL algorithms more broadly—is their reliance on black-box neural networks, which limits their transparency to end users. For example, Raghu et al. (2017) learned a sepsis treatment policy using deep RL that recommended lower doses of vasopressors for severely ill patients compared to clinicians. The black-box nature of this policy made it difficult to interpret the rationale behind such recommendations. This level of opacity would be unacceptable in clinical practice, where interpretability is often considered a prerequisite for the effective implementation of machine learning (Stiglic et al., 2020).

Interpretable RL is an emerging field, with existing approaches exploring, for example, decision trees to represent the target policy (Roth et al., 2019; Silva et al., 2020). However, most work in this area relies on post-hoc explanations of policies represented by neural networks (Verma et al., 2018; Coppens et al., 2019; Bastani et al., 2018), in contrast to the preferable approach of constructing directly interpretable policies (Rudin, 2019). Moreover, few of these methods are specifically designed for the offline setting, underscoring the need for best practices in developing interpretable and reliably evaluable policies for clinical decision-making.

3. Pragmatic Policy Development via Interpretable Behavior Cloning

Motivated by the need for interpretability and evaluability, we propose a pragmatic approach to policy development based on interpretable modeling of the behavior policy. An interpretable behavior policy model enables explanation of current decision-making

behavior (Pace et al., 2022) and clarifies which types of alternate policies can be reliably evaluated with statistical support (Matsson and Johansson, 2022). Additionally, it facilitates reasoning about whether the state S_t accounts for confounding variables that causally affect both the treatment decision A_t and its outcome R_t. The absence of unmeasured confounding, along with overlapping support (that is, $p_\mu(a \mid s) > 0$ whenever $p_\pi(a \mid s) > 0$), are necessary assumptions for IS-based OPE (Namkoong et al., 2020).

We suggest deriving the target policy from the most frequently chosen treatments in each state, as estimated by the behavior policy model. The most direct form of this approach, which is generalized and extended in Section 3.1, defines the policy to always choose the most common treatment; that is, setting $p_\pi(A_t = a_t \mid S_t = s_t)$ to 1 if $a_t = \arg\max_a p_{\hat{\mu}}(a \mid S_t = s_t)$, and 0 otherwise. This yields a deterministic target policy that closely resembles the behavior policy as estimated by the model. Implementing such a policy in practice can be viewed as standardizing common treatment patterns, leveraging the collective expertise of clinicians represented in the data. Since the behavior policy model is interpretable, the resulting target policy is interpretable by design. From an OPE perspective, this construction ensures overlapping support with the behavior policy—that is, any action recommended by the target policy has nonzero probability under the behavior policy.

For the interpretable behavior policy model, we utilize decision trees for probabilistic classification. A decision tree partitions inputs into homogeneous groups, each representing a terminal (leaf) node. By fitting a decision tree to observed state-action pairs (s_t, a_t), each leaf node groups patients with similar propensity scores $p_\mu(A_t \mid S_t)$. Matching subjects on the propensity score is sufficient to adjust for confounding (Rosenbaum and Rubin, 1983). If the state S_t fully captures all confounding variables, the variation in A_t within each leaf of a sufficiently deep tree reflects practice variation among clinicians that does not stem from confounding.

Tree-based models are often used to identify subgroups within a population (Keramati et al., 2022), and behavior cloning with decision trees has been applied in domains such as robotic control (Cichosz and Pawełczak, 2014; Sheh et al., 2011). However, other classes of interpretable models can also be employed. For example, prototypical networks (Li et al., 2018; Ming et al., 2019), a form of case-based reasoning, perform a soft clustering of the state-action space when used for behavior policy modeling (Matsson and Johansson, 2022). The key desideratum is a model that naturally groups the data. As such, (generalized) linear models—while interpretable—are less suitable, as they do not inherently induce such groupings.

3.1. Generalizing and Extending the Idea

We recommend that researchers always consider the deterministic target policy for evaluation. However, depending on the amount and sparsity of the data, it may not be possible to evaluate such a policy with statistical confidence. In the general case, the target policy can be constructed based on the set of the k actions with the highest probability in state s_t under the behavior policy model $\hat{\mu}$, denoted as Top-$k(s_t; \hat{\mu})$. Formally, let

$$p_\pi(a_t \mid s_t) \coloneqq \begin{cases} \frac{p_{\hat{\mu}}(a_t \mid s_t)}{Z_k}, & \text{if } a_t \in \text{Top-}k(s_t; \hat{\mu}); \\ 0, & \text{otherwise,} \end{cases} \tag{1}$$

where the normalization constant $Z_k = \sum_{a \in \text{Top-}k(s_t;\hat{\mu})} p_{\hat{\mu}}(a \mid s_t)$ ensures that p_π defines a valid probability distribution. By setting $k > 1$, we obtain a stochastic policy that more closely resembles the behavior policy as k approaches K, thereby increasing the effective sample size in evaluation (Owen, 2013). In practice, implementing a target policy with $k > 1$ requires the clinician to choose which of the k treatment alternatives to administer to the patient. In this case, the probabilities defined in Equation (1) can help guide this decision.

We can naturally extend Equation (1) to incorporate observed outcomes. Specifically, for each leaf in the fitted tree, we compute the average outcome for each action. The target policy is then defined by selecting the treatment—among the k most common ones—that is associated with the highest outcome. Formally, let $O(s, a; \hat{\mu})$ map the state-action pair (s, a) to the average outcome observed for patients in the same leaf as s who received a, as determined by the (tree-based) behavior policy model $\hat{\mu}$. In state s_t, the outcome-guided target policy selects the action

$$\underset{a \in \text{Top-}k(s_t;\hat{\mu})}{\arg\max} \; O(s_t, a; \hat{\mu}). \tag{2}$$

A Note on Terminology We refer to the general approach as pragmatic policy development via

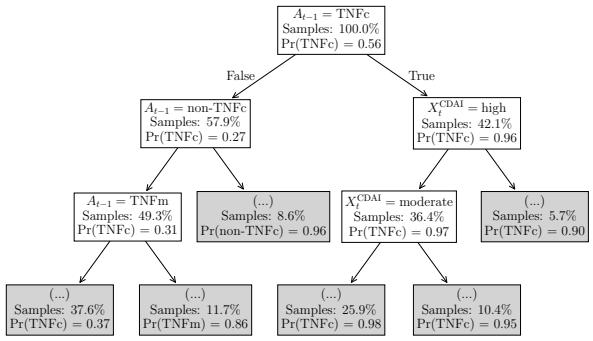


Figure 1: The topmost nodes of a standard decision tree fitted to a cohort of RA patients. The left side of the tree is dominated by the previous-treatment variable, A_{t-1}, leading to distinct subtrees for different treatment types. This structure results in repeated decision rules across subtrees, contributing to unnecessary model complexity.

interpretable behavior cloning to reflect its reliance on supervised learning of observed treatment patterns. However, the policies defined in Equations (1) and (2) differ from classical behavior cloning, which seeks to exactly replicate observed actions in each state—similar to the deterministic target policy introduced first in this section. A key assumption that distinguishes our approach from standard behavior cloning is that the dataset $\mathcal{D}$ contains examples from multiple practitioners, enabling us to exploit the collective knowledge embedded in the data.

4. Incorporating Known Structure

In clinical decision-making—particularly for chronic diseases—it is common for patients to receive the same treatment across multiple decision points. For example, if a treatment is effective, there is typically no need to change it. This presents a challenge when modeling the behavior policy using a decision tree: the model is often dominated by the tendency to continue with the previous treatment. As illustrated in Figure 1 for the case of RA management, the previous-treatment variable, A_{t-1}, dominates the left branch of the tree, effectively partitioning it into distinct subtrees for each treatment type. Many of the same decision rules are repeated across subtrees, leading to unnecessary model complexity.

We leverage this structure in the data to improve both the accuracy and interpretability of the behavior policy model. Specifically, we construct a meta-estimator that combines two decision tree classifiers: (1) a binary classifier that predicts whether a patient switches treatments, and (2) a multi-class classifier that predicts which treatment is chosen, trained only on treatment switch events. Let $p^s_{\hat{\mu}}(S_t)$ and $p^t_{\hat{\mu}}(S_t)$ denote the probabilistic output of classifier (1) and (2), respectively. Furthermore, let C_t be a binary random variable, where $C_t = 1$ represents a treatment change, and $C_t = 0$ indicates continuation of the current treatment.

Formally, we define $p^s_{\hat{\mu}}(S_t) \coloneqq p_{\hat{\mu}}(C_t = 1 \mid S_t)$ and $p^t_{\hat{\mu}}(k \mid S_t) \coloneqq p_{\hat{\mu}}(A_t = k \mid S_t)$. Let $\tilde{p}^t_{\hat{\mu}}(k \mid S_t)$ denote the probability of selecting treatment k in state S_t given that a treatment change occurs, $\tilde{p}^t_{\hat{\mu}}(k \mid S_t) \coloneqq p_{\hat{\mu}}(A_t = k \mid C_t = 1, S_t)$. We obtain this probability by excluding the probability of the previous treatment:

$$\tilde{p}^t_{\hat{\mu}}(k \mid S_t) = \frac{\mathbb{1}[k \neq a_{t-1}] p^t_{\hat{\mu}}(k \mid S_t)}{\sum_j \mathbb{1}[j \neq a_{t-1}] p^t_{\hat{\mu}}(j \mid S_t)}.$$

Finally, we obtain the probability $p_{\hat{\mu}}(A_t = k \mid S_t)$, which is the output of the meta-estimator, by incorporating the probability of staying on the same treatment:

$$\begin{aligned} p_{\hat{\mu}}(A_t = k \mid S_t) = \big(1 - p^s_{\hat{\mu}}(S_t)\big) \cdot \mathbb{1}[k = a_{t-1}] \\ + p^s_{\hat{\mu}}(S_t) \cdot \tilde{p}^t_{\hat{\mu}}(S_t, k). \end{aligned} \tag{3}$$

We use the meta-estimator (3) with decision trees for the switch and treatment classifiers to model the behavior policy in our experiments. We construct target policy candidates based on the most common treatments predicted by the model, as discussed in the previous section. To incorporate outcomes into the target policy, we compute the average observed outcome for each action within each leaf of the two trees. Let the state s fall into leaf i of the switch tree and leaf j of the treatment tree. For this state, the outcome associated with the previous treatment is the average outcome among patients in leaf i of the switch tree who remained on the same treatment ($c = 0$). The outcomes for other treatments are given by the average outcome observed among patients who switched treatments ($c = 1$) and belong to leaf j of the treatment tree.

5. Experiments

We demonstrate our proposed approach—deriving candidate target policies based on an interpretable behavior policy model— using two real-world clinical examples: the management of RA and sepsis. We begin by modeling the behavior policy to evaluate the structured learning approach introduced in the previous section. We then construct target policies from these models and assess their performance relative to those learned via RL.

RA Using data from the PPD™ CorEvitas™ RA registry (Kremer, 2005), we study RA treatment beginning with the initiation of the first biologic or targeted synthetic disease-modifying antirheumatic drug (b/tsDMARD). We identify a cohort of 1,565 patients who (i) have no history of b/tsDMARD use before registry enrollment, (ii) initiate at least one b/tsDMARD during registry participation, and (iii) have at least 2 years (720 days) of registry follow-up. As detailed in Appendix A, we allow the use of excluded data for behavior policy modeling. To simplify the action space, we group DMARDs into classes, resulting in eight distinct treatment options. We define the reward function as $R_t := 10 - I_{t+1}$, where I is the clinical disease activity index (CDAI). A CDAI of 10 marks the threshold between low and moderate-to-high disease activity.

Sepsis We use code provided by Komorowski et al. (2018) to collect and preprocess data for 11,482 patients meeting the international Sepsis-3 criteria (Singer et al., 2016) from the MIMIC-III database (Johnson et al., 2016). For each patient, we include up to 24 hours of measurements from estimated sepsis onset. Data are discretized into 4-hour time steps, with multiple measurements aggregated within each interval. Following Luo et al. (2024), we exclude patients with inconsistent time-series data. The action space includes 25 discrete actions, formed by combining doses of intravenous fluids and vasopressors, each binned into 5 levels. The reward is zero at all time steps except the final one, which is $+100$ if the patient survives and -100 otherwise.

Experimental Setup We split the data into separate training and testing sets. The training set is used to model the behavior policy and construct candidate target policies, while the testing set is used for OPE. A portion (20 %) of the training data is reserved for model validation and calibration. The entire procedure is repeated 50 times with different random seeds. Further details, including model architectures and hyperparameter selection, are provided in the following sections and Appendix B.

Table 1: Average test AUROC and SCE for different models of the behavior policy across 50 splits of each dataset. We compare a standard decision tree (DT) with the model described in Section 4, where a separate decision tree is used to predict treatment switching (DT-S). We also include a model that combines DT-S with a decision tree for classification at the first time step (DT-BLS).

	RA		**Sepsis**	
Model	AUROC	SCE	AUROC	SCE
DT	92.0	2.7	86.9	0.4
DT-S	92.8	2.6	86.0	0.5
DT-BLS	94.9	1.3	86.8	0.5
RNN	91.8	2.4	88.1	0.5

5.1. Behavior Policy Modeling

For behavior policy modeling, we include three types of tree-based models: (i) a standard decision tree (DT); (ii) the meta-model described in Section 4 (DT-S), which uses two decision trees to separately predict treatment switching and treatment assignment; and (iii) a variant (DT-BLS) that combines DT-S with a separate decision tree for treatment classification at the first time step (baseline). DT-BLS accounts for the fact that baseline treatment decisions are often more predictable than those made later. For example, in RA, most patients initiate their first b/tsDMARD therapy with a Tumor Necrosis Factor (TNF) inhibitor (Smolen et al., 2023). Finally, we include a recurrent neural network (RNN) model as a baseline to assess how well the behavior policy can be estimated when using the full patient history. For the tree-based models, a patient's state is represented using a combination of the most recent covariates and selected aspects of their history.

A good behavior policy model should achieve strong predictive performance while also producing well-calibrated probabilities. In Table 1, we report the average area under the receiver operating characteristic curve (AUROC) and static calibration error (SCE) (Nixon et al., 2019) for the different models across dataset splits. In the RA case, DT-S and DT-

BLS yield the highest predictive performance, with DT-BLS performing best. The RNN performs comparably to the standard decision tree but is outperformed by the meta-models, highlighting the advantages of structure-aware models in this context.

For sepsis, all tree-based models exhibit similar performance, which is not surprising. Unlike RA, where treatment involves follow-up visits scheduled several months apart, sepsis management requires continuous administration of fluids and vasopressors. As a result, it cannot be as naturally decomposed into a two-step decision process—first deciding whether to switch treatment, then determining what treatment to switch to—as is possible with RA.

To examine the behavior policy for RA, we fit DT-BLS to the full cohort dataset (1,565 patients), with hyperparameter selection performed using 3-fold cross-validation. At baseline, TNF-based therapies are the most common, aligning with clinical guideline recommendations (Smolen et al., 2023). In Figure 2(b), we show the decision tree fitted to predict whether a patient switches treatment. The tree captures a common pattern described by Matsson et al. (2024a): patients often switch treatment after a period without DMARDs or while on a conventional synthetic (cs) DMARD therapy. In such cases, higher CDAI scores increase the probability of switching therapy. Among the large group of patients in the leftmost subtree, the probability of switching treatment is generally low ($< 7\,\%$).

Figure 2(a) shows the decision tree fitted to cases where RA patients switched therapy. Each node specifies the probability of the most common treatment, which the deterministic target policy from Section 3 would recommend in these situations. The previous treatment, A_{t-1}, provides a strong signal for determining which therapy patients switch to. Patients previously on TNF monotherapy or csDMARD therapy are likely to switch to TNF combination therapy. For patients previously without DMARD treatment, non-TNF monotherapy is the most common choice. These transitions are consistent with the patterns described in Matsson et al. (2024a).

5.2. Policy Construction and Evaluation

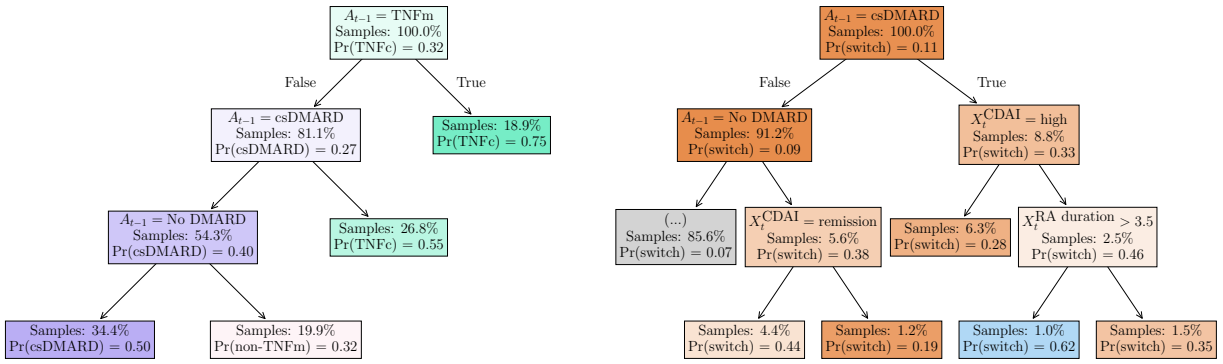
Next, we construct candidate target policies under the pragmatic policy development framework. We choose the DT-BLS behavior policy model, given its superior performance in the RA setting. We derive target policies based on the $k = \{1, 2, 3\}$ most common treatments (`MC`), including variants that incorporate observed outcomes (`MC+O`). When $k = 1$, these policies are equivalent.

For comparison, we also include a policy that recommends a randomly selected treatment, as well as four policies learned using reinforcement learning: standard Q-learning (QL), deep Q-learning (DQN) (Mnih et al., 2015), batch-constrained Q-learning (BCQ) (Fujimoto et al., 2019), and conservative Q-learning (CQL) (Kumar et al., 2020). Standard Q-learning is applied to a discrete Markov decision process based on clustered states, whereas DQN, BCQ, and CQL operate on continuous state representations. BCQ and CQL are specifically designed to address the distributional shift between behavior and target policies that arises in offline RL. See Appendix B for further details.

In Table 2, we present average OPE estimates of the policies using weighted importance sampling (WIS) (Precup et al., 2000), along with the corresponding average effective sample sizes (ESSs) (Owen, 2013). A small ESS relative to the number of trajectories indicates that a few weights dominate the weighted sum, limiting the reliability of the estimate. In RA, where the number of follow-up visits differs across patients, the estimated policy values are normalized accordingly.

For both RA and sepsis, we find that the deterministic `MC` policy ($k = 1$) is, on average, estimated to outperform the behavior policy. In the RA case, the confidence intervals for the value estimates of these policies are well separated. In both settings, the average ESS for the `MC` policy with $k = 1$ is substantially larger than that of the RL-based policies (406.1 vs. 3.0–7.3 for RA, and 64.1 vs. 1.7–19.3 for sepsis). The small ESSs for the RL policy estimates, together with their high variance, underscore the challenges of reliable OPE and highlight the importance of keeping target policies close to the behavior policy.

For the `MC` policies, adjusting k provides a direct way to control overlap with the behavior policy and, consequently, the variance of the OPE estimates. This effect is particularly evident in the sepsis case, where increasing k from 1 to 3 increases the ESS by a factor of 10. Interestingly, in the RA case, the variance remains largely unchanged when varying k in the same way. For the `MC+O` policies, which select the treatment with the highest estimated outcome among the k most common options, reducing k decreases the variance of the OPE estimates—although the overall variance remains high.



(*a*) Treatment prediction.

(*b*) Treatment switching prediction.

Figure 2: Decision trees from the meta-model described in Section 3, fitted to post-baseline events in RA: (a) treatment prediction and (b) treatment switching prediction. In (a), each node specifies the probability of the most common treatment.

By normalizing the estimated target policy values relative to the behavior policy in RA, we can interpret them as the average change in CDAI per patient and treatment stage if these policies were to replace the behavior policy. Specifically, negative values indicate a decrease in CDAI. As shown in Figure 3, the `MC` policies suggest a potential decrease in CDAI for $k = 1, 2, \ldots, 8$, at which point the behavior policy is recovered. The `MC+O` policies suggest an even larger potential reduction in CDAI, though the high variance of the estimates limits their reliability.

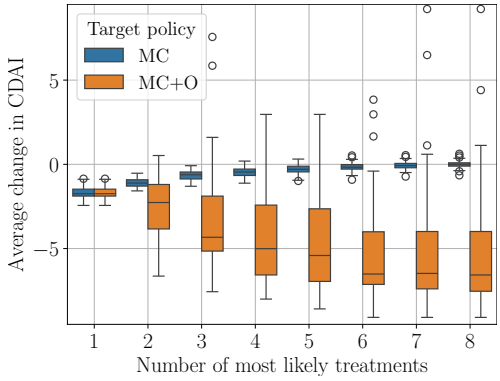


Figure 3: OPE using weighted importance sampling for RA target policies based on the most common treatments under the behavior policy. Boxes represent the interquartile range of the value distribution, normalized relative to the behavior policy.

6. Discussion

In this work, we proposed *pragmatic policy development*, a framework for deriving target policies for clinical decision-making through interpretable behavior cloning. Specifically, to support both interpretability and off-policy evaluation, we constructed target policies based on the most commonly selected treatments in each patient state, as identified by a tree-based model of the behavior policy. To improve behavior policy accuracy, we designed the model to capture inherent data structure, such as patients' tendency to remain on the same treatment across decision points. Our evaluation on real-world rheumatoid arthritis and sepsis treatment data demonstrated that interpretable policies based on the most common treatments offer a promising clinical approach, providing stronger statistical support in value estimates than reinforcement learning policies.

Intuitively, policies derived under this framework correspond to acting like the majority of clinicians in the observed data, leveraging collective expertise. Thus, such policies can be seen as standardizing frequent treatment patterns and reducing unwar-

Table 2: The average value estimate $\hat{V}$ using weighted importance sampling (WIS) and effective sample size (ESS) for different target policies in RA and sepsis. The behavior policy value is estimated as the average reward in the data. Confidence intervals show the interquartile range of each distribution. Both metrics should be considered together, as WIS estimates with low ESS reflect high variance and are unreliable (cf. the "Random" policy).

	RA		**Sepsis**	
Target policy	$\hat{V}^{\pi}_{\text{WIS}}$ (↑)	ESS (↑)	$\hat{V}^{\pi}_{\text{WIS}}$ (↑)	ESS (↑)
`MC` ($k=1$)	0.7 (0.4, 0.8)	406.1 (388.1, 415.6)	74.1 (66.8, 82.9)	64.1 (46.1, 80.8)
`MC` ($k=2$)	0.0 (−0.2, 0.2)	566.1 (553.3, 575.0)	74.6 (72.0, 79.4)	277.7 (250.6, 296.0)
`MC` ($k=3$)	−0.5 (−0.6, −0.2)	639.2 (624.6, 650.3)	75.1 (73.7, 76.6)	628.9 (604.9, 647.7)
`MC+O` ($k=1$)	0.7 (0.4, 0.8)	406.1 (388.1, 415.6)	74.1 (66.8, 82.9)	64.1 (46.1, 80.8)
`MC+O` ($k=2$)	1.2 (0.1, 2.7)	19.2 (7.4, 31.7)	80.7 (75.7, 93.9)	15.5 (7.8, 24.7)
`MC+O` ($k=3$)	3.2 (0.8, 4.1)	17.1 (9.0, 25.7)	85.3 (68.5, 95.5)	6.9 (3.0, 14.2)
RL (QL)	0.0 (−4.2, 3.7)	3.0 (2.1, 4.3)	86.0 (80.3, 92.3)	14.0 (6.8, 24.5)
RL (DQN)	0.0 (−4.3, 3.0)	5.3 (2.4, 9.5)	89.3 (70.9, 98.9)	1.7 (1.1, 8.0)
RL (BCQ)	−1.8 (−4.6, 0.4)	7.3 (4.4, 10.5)	83.0 (76.3, 88.2)	19.3 (12.7, 28.9)
RL (CQL)	−0.9 (−5.3, 1.2)	5.7 (2.8, 11.4)	67.3 (34.4, 85.2)	6.5 (2.7, 11.5)
Random	1.7 (−4.8, 5.1)	1.3 (1.0, 2.0)	97.9 (59.7, 99.6)	1.6 (1.1, 2.3)
Behavior policy	−1.1 (−1.2, −1.0)	779.0 (770.3, 788.8)	71.6 (70.7, 72.2)	2297.0 (2297.0, 2297.0)

ranted practice variation—a common issue in healthcare (Pace et al., 2022). While these policies may not represent the optimal strategies derivable from the data, their design ensures they can be evaluated meaningfully. After all, what cannot be evaluated has limited practical value. Additionally, the interpretability of these policies facilitates their use in clinical practice, where transparency is essential for validation and gaining trust from end users.

Our work has some limitations. First, we focused exclusively on off-policy evaluation using importance sampling. Doubly robust methods (Jiang and Li, 2016; Thomas and Brunskill, 2016) could help reduce the variance of value estimates (see Appendix C for related directions); however, all off-policy evaluation methods suffer to some extent from large discrepancies between target and behavior policies (Voloshin et al., 2021). Second, we assumed that the state S_t includes all confounding variables that causally affect both treatment and outcome. However, in practice, we were limited to the variables available in the observed data, meaning that unmeasured confounders could still be present. This assumption is particularly critical for the outcome-guided policy: if unmeasured confounders exist within the leaves of a fully grown tree, the estimated value of this policy may be biased upwards. Third, while previous research has shown that interpretable models such as sparse decision trees can effectively model policies for sequential decision-making (Matsson et al., 2024b), there may be cases where these models fit the data unsatisfactorily compared to more flexible models like neural networks. Prototype-based models may offer a better solution in such cases (Ming et al., 2019; Matsson and Johansson, 2022).

To conclude, we offer three recommendations for researchers and practitioners:

1. **Use interpretable models of the behavior policy.** Interpretable behavior policy models provide a basis for verifying assumptions and identifying candidate target policies that have sufficient support in the data.

2. **Exploit known structure to improve behavior policy modeling.** If there are known patterns in the data—e.g., differences in treatment assignment over time—models should be designed to account for these patterns, maintaining small model size while improving accuracy.

3. **Construct target policies with statistical support under the behavior policy.** Evaluation is critical for turning observational data into actionable policies. Therefore, policies should be developed with evaluation feasibility in mind.

Acknowledgments

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

The computations and data handling were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

References

Alberto Abadie and Guido W Imbens. Matching on the estimated propensity score. *Econometrica*, 84 (2):781–807, 2016.

Osbert Bastani, Yewen Pu, and Armando Solar-Lezama. Verifiable reinforcement learning via policy extraction. In *Advances in Neural Information Processing Systems*, volume 31, pages 2499–2509, 2018.

Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

Bibhas Chakraborty and Erica E Moodie. *Statistical methods for dynamic treatment regimes*, volume 2. Springer, 2013.

Pawel Cichosz and Łukasz Pawełczak. Imitation learning of car driving skills with decision trees and random forests. *International Journal of Applied Mathematics and Computer Science*, 24(3): 579–597, 2014.

Youri Coppens, Kyriakos Efthymiadis, Tom Lenaerts, Ann Nowé, Tim Miller, Rosina Weber, and Daniele Magazzeni. Distilling deep reinforcement learning policies in soft decision trees. In *Proceedings of the IJCAI 2019 Workshop on Explainable Artificial Intelligence*, pages 1–6, 2019.

Jannik Deuschel, Caleb Ellington, Yingtao Luo, Ben Lengerich, Pascal Friederich, and Eric P. Xing. Contextualized policy recovery: Modeling and interpreting medical decisions with adaptive imitation learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *PMLR*, pages 10642–10660, 2024.

Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. *Statistical Science*, 29(4):485–511, 2014.

Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *PMLR*, pages 1447–1456, 2018.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *PMLR*, pages 2052–2062, 2019.

Claire Glanois, Paul Weng, Matthieu Zimmer, Dong Li, Tianpei Yang, Jianye Hao, and Wulong Liu. A survey on interpretable reinforcement learning. *Machine Learning*, 113(8):5847–5890, 2024.

Omer Gottesman, Fredrik Johansson, Joshua Meier, Jack Dent, Donghun Lee, Srivatsan Srinivasan, Linying Zhang, Yi Ding, David Wihl, Xuefeng Peng, et al. Evaluating reinforcement learning algorithms in observational health settings. *arXiv preprint arXiv:1805.12298*, 2018.

Daniel Hein, Steffen Udluft, and Thomas A Runkler. Interpretable policies for reinforcement learning by genetic programming. *Engineering Applications of Artificial Intelligence*, 76:158–169, 2018.

Edward L Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.

Pushkala Jayaraman, Jacob Desman, Moein Sabounchi, Girish N Nadkarni, and Ankit Sakhuja. A primer on reinforcement learning in medicine for clinicians. *NPJ Digital Medicine*, 7 (1):337, 2024.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *PMLR*, pages 652–661, 2016.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3 (1):1–9, 2016.

Nathan Kallus and Masatoshi Uehara. Doubly robust off-policy value and gradient estimation for deterministic policies. In *Advances in Neural Information Processing Systems*, volume 33, pages 10420–10430, 2020.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Ramtin Keramati, Omer Gottesman, Leo Anthony Celi, Finale Doshi-Velez, and Emma Brunskill. Identification of subgroups with similar benefits in off-policy policy evaluation. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *PMLR*, pages 397–410, 2022.

Matthieu Komorowski, Leo A. Celi, Omar Badawi, Anthony C. Gordon, and A. Aldo Faisal. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, 2018.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit Q-learning. In *Proceedings of the 10th International Conference on Learning Representations*, 2022.

Joel Kremer. The CORRONA database. *Annals of the Rheumatic Diseases*, 64:iv37–iv41, 2005.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 1179–1191, 2020.

Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *PMLR*, pages 3703–3712, 2019.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 3530–3537, 2018.

Zachary C Lipton. The doctor just won't accept that! *arXiv preprint arXiv:1711.08037*, 2017.

G. Liu, O. Schulte, W. Zhu, and Q. Li. Toward interpretable deep reinforcement learning with linear model U-trees. In *Machine Learning and Knowledge Discovery in Databases*, volume 11052 of *Lecture Notes in Computer Science*, pages 414–430. Springer, Cham, 2019.

Zhiyao Luo, Yangchen Pan, Peter Watkinson, and Tingting Zhu. Position: Reinforcement learning in dynamic treatment regimes needs critical reexamination. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *PMLR*, pages 33432–33465, 2024.

Anton Matsson and Fredrik D. Johansson. Case-based off-policy evaluation using prototype learning. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence*, volume 180 of *PMLR*, pages 1339–1349, 2022.

Anton Matsson, Daniel H Solomon, Margaux M Crabtree, Ryan W Harrison, Heather J Litman, and Fredrik D Johansson. Patterns in the sequential treatment of patients with rheumatoid arthritis starting a biologic or targeted synthetic disease-modifying antirheumatic drug: 10-year experience from a us-based registry. *ACR Open Rheumatology*, 6(1):5–13, 2024a.

Anton Matsson, Lena Stempfle, Yaochen Rao, Zachary R. Margolin, Heather J. Litman, and Fredrik D. Johansson. How should we represent history in interpretable models of clinical policies? In *Proceedings of the 4th Machine Learning for Health Symposium*, volume 259 of *PMLR*, pages 714–734, 2024b.

Xiangrui Meng, Joseph Bradley, Burak Yavuz, Evan Sparks, Shivaram Venkataraman, Davies Liu, Jeremy Freeman, DB Tsai, Manish Amde, Sean Owen, et al. MLlib: Machine Learning in Apache

Spark. *Journal of Machine Learning Research*, 17 (34):1–7, 2016.

Yao Ming, Panpan Xu, Huamin Qu, and Liu Ren. Interpretable and steerable sequence learning via prototypes. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 903–913, 2019.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518 (7540):529–533, 2015.

Hongseok Namkoong, Ramtin Keramati, Steve Yadlowsky, and Emma Brunskill. Off-policy policy evaluation for sequential decisions under unobserved confounding. In *Advances in Neural Information Processing Systems*, volume 33, pages 18819–188313, 2020.

Jeremy Nixon, Mike Dusenberry, Ghassen Jerfel, Timothy Nguyen, Jeremiah Liu, Linchuan Zhang, and Dustin Tran. Measuring calibration in deep learning. *arXiv preprint arXiv:1904.01685*, 2019.

Art B Owen. *Monte Carlo theory, methods and examples*, chapter 9. https://artowen.su.domains/mc/, 2013.

Alizée Pace, Alex Chan, and Mihaela van der Schaar. POETREE: Interpretable policy learning with adaptive decision trees. In *Proceedings of the 10th International Conference on Learning Representations*, 2022.

Cosmin Paduraru. *Off-policy evaluation in Markov decision processes*. PhD thesis, McGill University, 2013.

Doina Precup, Richard S Sutton, and Satinder Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the 17th International Conference on Machine Learning*, pages 759–766, 2000.

Aniruddh Raghu, Matthieu Komorowski, Imran Ahmed, Leo Celi, Peter Szolovits, and Marzyeh Ghassemi. Deep reinforcement learning for sepsis treatment. *arXiv preprint arXiv:1711.09602*, 2017.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Aaron M Roth, Nicholay Topin, Pooyan Jamshidi, and Manuela Veloso. Conservative Q-improvement: Reinforcement learning for an interpretable decision-tree policy. *arXiv preprint arXiv:1907.01180*, 2019.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Raymond Sheh, Bernhard Hengst, and Claude Sammut. Behavioural cloning for driving robots over rough terrain. In *Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 732–737, 2011.

Andrew Silva, Matthew Gombolay, Taylor Killian, Ivan Jimenez, and Sung-Hyun Son. Optimization methods for interpretable differentiable decision trees applied to reinforcement learning. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *PMLR*, pages 1855–1865, 2020.

Mervyn Singer, Clifford S Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R Bernard, Jean-Daniel Chiche, Craig M Coopersmith, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). *JAMA*, 315(8):801–810, 2016.

Josef S. Smolen, Robert B.M. Landewé, Sytske Anne Bergstra, Andreas Kerschbaumer, Alexandre Sepriano, Daniel Aletaha, Roberto Caporali, Christopher John Edwards, Kimme L. Hyrich, Janet E. Pope, et al. EULAR recommendations for the management of rheumatoid arthritis with synthetic and biological disease-modifying antirheumatic drugs: 2022 update. *Annals of the Rheumatic Diseases*, 82(1):3–18, 2023.

Gregor Stiglic, Primoz Kocbek, Nino Fijacko, Marinka Zitnik, Katrien Verbert, and Leona Cilar. Interpretability of machine learning-based prediction models in healthcare. *WIREs Data Mining and Knowledge Discovery*, 10(5):e1379, 2020.

Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudík. Doubly robust off-policy evaluation with shrinkage. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *PMLR*, pages 9167–9176, 2020.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*, chapter 17.3. MIT press, 2018.

Adith Swaminathan and Thorsten Joachims. The self-normalized estimator for counterfactual learning. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *PMLR*, pages 2139–2148, 2016.

Faraz Torabi, Garrett Warnell, and Peter Stone. Behavioral cloning from observation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4950–4957, 2018.

Abhinav Verma, Vijayaraghavan Murali, Rishabh Singh, Pushmeet Kohli, and Swarat Chaudhuri. Programmatically interpretable reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *PMLR*, pages 5045–5054, 2018.

Abhinav Verma, Hoang Le, Yisong Yue, and Swarat Chaudhuri. Imitation-projected programmatic reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 32, pages 15752–15763, 2019.

Cameron Voloshin, Hoang Minh Le, Nan Jiang, and Yisong Yue. Empirical study of off-policy policy evaluation for reinforcement learning. In *Advances in Neural Information Processing Systems Datasets and Benchmarks*, volume 1, 2021.

Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys*, 55(1):1–36, 2021.

Appendix A. Datasets

Our experimental evaluation was based on two distinct datasets related to the management of rheumatoid arthritis (RA) and sepsis. Here, we provide additional details on the datasets, including the full set of covariates used to represent patient states. Table 3 summarizes key characteristics of these two decision-making processes.

A.1. Rheumatoid Arthritis

RA is a chronic autoimmune disease that causes joint inflammation, pain, and potential deformity. It is primarily managed with disease-modifying antirheumatic drugs (DMARDs), which include conventional synthetic (cs), biologic (b), and targeted synthetic (ts) DMARDs. Biologic DMARDs are further divided into Tumor Necrosis Factor (TNF) inhibitors and non-TNF inhibitors. Clinical guidelines for RA management recommend starting treatment with a csDMARD upon clinical diagnosis (Smolen et al., 2023). If the initial treatment fails, adding a bDMARD or tsDMARD—typically a TNF inhibitor—is advised for patients with poor prognostic factors such as high disease activity.

Our original dataset comprised 42,068 unique patients enrolled in the PPD™ CorEvitas™ RA registry (Kremer, 2005), a longitudinal clinical registry in the United States, between January 2012 and December 2021. According to the registry protocol, patients are recommended to complete follow-up visits every six months. We excluded records of visits where information related to therapy changes was missing, contradictory, or potentially erroneous, resulting in a cleaned dataset of 41,860 patients.[1] For example, visits where multiple b/tsDMARDs were prescribed simultaneously were excluded from the analysis, as such prescriptions are not clinically recommended and may represent a data reporting error.

We focused on RA treatment starting with the initiation of the first b/tsDMARD therapy, defined as the index visit (baseline). Specifically, we selected patients who (i) had no history of b/tsDMARD use before registry enrollment and (ii) initiated at least one b/tsDMARD therapy during registry participation. Sequences were truncated at visits where the subsequent visit lacked a clinical disease activity index (CDAI) measurement, necessary for reward computation. Additionally, follow-up visits were required to occur within intervals of 30 to 270 days, and patients needed a fixed follow-up duration of 2 years (720 days). Patients who did not meet these criteria were excluded from the cohort, resulting in 1,565 included patients.

Following Matsson et al. (2024a), we restricted the analysis to classes of DMARDs rather than individual drugs. Specifically, we divided bDMARDs into TNF inhibitors and non-TNF inhibitors, resulting in the following classes of drugs: csDMARDs, TNF inhibitor biologics, non-TNF inhibitor biologics, and Janus kinase (JAK) inhibitors (the main group of tsDMARDs). We included both monotherapies and combinations of a csDMARD and a b/tsDMARD; for example, a TNF combined with a csDMARD. In total, the action space consisted of eight therapies ($K = 8$): csDMARD therapy (`csDMARD`), TNF biologic monotherapy (`TNFm`), TNF biologic combination therapy (`TNFc`), non-TNF biologic monotherapy (`non-TNFm`), non-TNF biologic combination therapy (`non-TNFc`), JAK monotherapy (`JAKm`), JAK combination therapy (`JAKc`), and no DMARD therapy (`No DMARD`).

We constructed the state using the variables listed in Table 4, which also provides descriptive statistics at baseline for each variable. Additionally, we included variables indicating historical comorbidities and previous treatment history, including the most recent therapy selection.

A.2. Sepsis

The dataset of patients with sepsis, defined according to the international Sepsis-3 criteria (Singer et al., 2016), was sourced from the MIMIC-III database using code provided by Komorowski et al. (2018). We focused on the first 24 hours of measurements, whereas the original work used 72 hours of data. Following Luo et al. (2024), we excluded patients with irregularly sampled data. For each patient, the data was structured as a multivariate time series with a discrete time step of 4 hours. We constructed the state using a subset of the variables from the original work (see Table 5). In addition to the variables in Table 5, we included the doses of intravenous fluids and vasopressors administered during the previous 4-hour period.

1. Only the subtrajectory up to the point of an excluded visit was considered for off-policy evaluation.

Table 3: Characteristics of the decision-making processes studied in this work.

	RA	Sepsis
Patients, n	1,565	11,482
Age in years, median (IQR)	59.0 (51.0, 67.0)	66.0 (53.4, 77.8)
Female, n (%)	1,173 (75.1)	4,996 (43.5)
Number of state variables $\lvert\mathcal{S}\rvert$	33	20
Number of actions $\lvert\mathcal{A}\rvert$	8	25
Number of stages T, median (IQR)	4.0 (3.0, 5.0)	6.0 (6.0, 6.0)

Appendix B. Experimental Details

We split each dataset into separate training and testing sets, with the former used to fit, validate, and calibrate the behavior policy model and the latter used for off-policy evaluation (OPE). For sepsis, we applied a standard 80/20 split. For RA, the data was divided equally between training and testing, and the cohort-selection criteria defined in the previous section were applied to both the testing and validation sets (20 % of the original training data in both cases). All training data was used to estimate the behavior policy in RA. However, the proportion of observations collected before the first b/tsDMARD initiation and the fraction of observations from patients who started b/tsDMARD therapy before registry enrollment were treated as hyperparameters in the modeling process. These adjustments were implemented due to the limited number of patients meeting the cohort criteria for RA.

The preprocessing steps differed between datasets. For sepsis, we followed the approach described in Komorowski et al. (2018), standardizing normally distributed features and applying a log transformation followed by standardization for log-normally distributed features. For RA, missing values in numerical features were imputed using mean imputation, while categorical features were imputed using mode imputation and encoded using one-hot encoding. In both cases, the final model was calibrated using sigmoid calibration.

Below, we provide additional details for each experimental step: modeling the behavior policy, constructing target policy candidates, and evaluating each candidate using OPE. The entire procedure was repeated 50 times with different data splits to ensure statistical robustness of the results.

B.1. Behavior Policy Modeling

We used cross-validation to select the best model for each of four model types, each chosen from 30 candidates generated by randomly sampling hyperparameters. The model types were: a standard decision tree (DT); a variant with separate trees for switch and treatment prediction (DT-S, described in Section 4); an extension of DT-S with an additional tree for baseline treatment prediction (DT-BLS); and a recurrent neural network (RNN). Model selection was based on the area under the receiver operating characteristic curve (AUROC). AUROC scores were computed using one-vs-all macro-averaging. We used scikit-learn's implementation for the decision trees, whereas the RNN was implemented in PyTorch.

Hyperparameters for DT and RNN are listed in Table 6. For DT-S and DT-BLS, we used the same hyperparameters as for DT for each decision tree in these meta-models. As noted above, in the RA setting, we introduced two additional hyperparameters: the proportion of observations collected before the first b/tsDMARD initiation, and the fraction of observations from patients who initiated b/tsDMARD therapy prior to registry enrollment. Each of these was evaluated at $\{0, 0.25, 0.5, 0.75, 1\}$. The RNN was trained for up to 100 epochs using cross-entropy loss and the Adam optimizer, with early stopping based on validation performance and a patience of 10 epochs.

A Note on Scalability Our framework is primarily designed for settings with limited observational data, where robust offline evaluation and interpretability are paramount. In such contexts, the computational cost of training decision trees remains manageable. For larger-scale applications with extensive datasets, training decision trees could become computationally demanding. However, this challenge can be addressed through distributed training frameworks such as Apache Spark (Meng et al., 2016) or

Table 4: Rheumatoid arthritis. A summary of variables included in the state representation and their baseline statistics. N represents the number of patients with non-missing baseline information.

Variable	N	Statistics
Age in years, median (IQR)	1,562	59 (51, 67)
RA duration in years, median (IQR)	1,548	3 (1, 8)
Gender, n (%)	1,562	
Male		389 (24.9)
Female		1,173 (75.1)
BMI, n (%)	1,535	
Underweight		14 (0.9)
Healthy weight		356 (23.2)
Overweight		492 (32.1)
Obesity		673 (43.8)
Blood pressure, n (%)	1,561	
Elevated		247 (15.8)
Hypertension stage 1		531 (34.0)
Hypertension stage 2		390 (25.0)
Normal		393 (25.2)
Currently pregnant, n (%)	1,041	3 (0.3)
Pregnant since last visit, n (%)	875	4 (0.5)
Private insurance	1,565	1,131 (72.3)
Medicare insurance	1,565	529 (33.8)
Medicaid insurance	1,565	77 (4.9)
No insurance	1,565	31 (2.0)
CDAI, n (%)	1,557	
Remission		175 (11.2)
Low		370 (23.8)
Moderate		514 (33.0)
High		498 (32.0)
CCP outcome positive, n (%)	278	159 (57.2)
RF outcome positive, n (%)	304	191 (62.8)
Erosive disease, n (%)	1,183	98 (8.3)
Joint space narrowing, n (%)	368	218 (59.2)
Joint deformity, n (%)	361	65 (18.0)
Severe infections, n (%)	1,565	25 (1.6)
Tuberculosis outcome, n (%)	358	19 (5.3)
Comorbidities, n (%)	1,565	
Metabolic diseases		120 (7.7)
Cardiovascular diseases		172 (11.0)
Respiratory diseases		39 (2.5)
Cancer		37 (2.4)
GI and liver diseases		29 (1.9)
Musculoskeletal disorders		475 (30.4)
Other diseases		206 (13.2)

by leveraging recent advances in scalable tree-based methods (Ke et al., 2017).

B.2. Target Policy Construction

In addition to the target policies based on the most commonly selected treatments under the behavior policy, we learned policies using different reinforcement learning methods: standard Q-learning (QL), deep Q-learning (DQN) (Mnih et al., 2015), batch-constrained Q-learning (BCQ) (Fujimoto et al., 2019), and conservative Q-learning (CQL) (Kumar et al., 2020).

Table 5: Sepsis. A summary of variables included in the state representation and their baseline statistics. N represents the number of patients with non-missing baseline information.

Variable	N	Statistics
Age in years, median (IQR)	11,482	66.0 (53.5, 77.8)
Female, n (%)	11,482	4,996 (43.5)
Heart rate, median (IQR)	11,482	87.8 (76.0, 101.0)
SysBP, median (IQR)	11,482	118.0 (105.1, 133.9)
DiaBP, median (IQR)	11,482	57.0 (48.3, 66.0)
MeanBP, median (IQR)	11,482	77.2 (69.0, 87.2)
Shock index, median (IQR)	11,482	0.7 (0.6, 0.9)
Hemoglobin, median (IQR)	11,482	10.5 (9.3, 12.1)
Blood urea nitrogen, median (IQR)	11,482	23.1 (15.0, 39.0)
Creatinine, median (IQR)	11,482	1.1 (0.8, 1.7)
Total urine output, median (IQR)	11,482	0.0 (0.0, 300.8)
Base excess, median (IQR)	11,482	0.0 (−2.6, 3.0)
Lactate, median (IQR)	11,482	1.7 (1.2, 2.6)
pH, median (IQR)	11,482	7.4 (7.3, 7.4)
HCO3, median (IQR)	11,482	24.0 (21.0, 27.0)
PaO_2/FiO_2 ratio, median (IQR)	11,482	257.3 (167.5, 410.0)
Elixhauser, median (IQR)	11,482	4.0 (2.0, 5.0)
SOFA, median (IQR)	11,482	7.0 (5.0, 9.0)

For the Q-learning approach, we followed Komorowski et al. (2018) by clustering the raw states using k-prototypes clustering for RA and k-means clustering for sepsis. For each case, we tested multiple values for the number of clusters and selected the solution with the lowest clustering loss. We then formulated a tabular Markov decision process (MDP) by assigning each raw state to its nearest cluster centroid and estimating transition and reward matrices from the observed data. Finally, we learned a target policy by solving the MDP with Q-learning, using a discount factor of 0.99.

For the other methods—DQN, BCQ, and CQL—we used the implementation provided by Luo et al. (2024). To estimate the behavior policy, we employed a multi-layer perceptron for RA and a recurrent neural network for sepsis. This model was used to compute OPE estimates of the target policy during training, using truncated weighted importance sampling (Luo et al., 2024). For both the behavior model and each target policy, we randomly sampled 10 sets of hyperparameters. We refer to Luo et al. (2024) for the full set of hyperparameters considered. In all cases, training was conducted over 50 epochs. All RL policies, as well as the random policy, were made deterministic in the RA setting. For the sepsis case, the policies were softened by assigning each action a 1 % probability of being selected.

B.3. Off-Policy Evaluation

We performed off-policy evaluation using weighted importance sampling (WIS) (Precup et al., 2000), defined as

$$\hat{V}^{\pi}_{\mathrm{WIS}} = \frac{\sum_{i=1}^{n} w_i \sum_{t=1}^{T} r_t^{(i)}}{\sum_{i=1}^{n} w_i},$$

where the importance weights w_i are defined as

$$w_i = \prod_{t=1}^{T} \frac{p_{\pi}\left(A_t = a_t^{(i)} \mid S_t = s_t^{(i)}\right)}{p_{\hat{\mu}}\left(A_t = a_t^{(i)} \mid S_t = s_t^{(i)}\right)}.$$

Unlike the standard importance sampling (IS) estimator, which normalizes by the number of evaluation trajectories n, the WIS estimator introduces bias but typically achieves lower variance.

To quantify how many samples meaningfully contribute to the value estimate, we computed the effective sample size n_e (Owen, 2013):

$$n_e = \frac{\left(\sum_i w_i\right)^2}{\sum_i w_i^2}.$$

Appendix C. Variance Reduction in Off-Policy Evaluation

Beyond our policy-construction approach to variance reduction, several complementary strategies exist for improving OPE reliability. Doubly robust estimators combine direct method and importance sampling to achieve lower variance than either method alone, though they remain sensitive to policy overlap (Dudík et al., 2014; Jiang and Li, 2016). Recent advances in variance reduction include control variates that leverage auxiliary variables correlated with outcomes (Swaminathan and Joachims, 2015), self-normalized importance sampling that can reduce tail variance at the cost of introducing bias (Swaminathan and Joachims, 2015; Thomas and Brunskill, 2016), and truncated importance weights that explicitly bound variance by capping importance ratios (Bottou et al., 2013; Ionides, 2008). Hybrid approaches that adaptively combine multiple estimators based on estimated uncertainty show promise for balancing bias-variance tradeoffs in practice (Su et al., 2020; Kallus and Uehara, 2020). Future work could explore integrating these techniques with our interpretable policy framework to further enhance evaluation robustness while maintaining transparency for clinical decision-makers.

Appendix D. Adjusting the Probability of Treatment Switching

Using a separate model to predict whether a patient switches treatment—as described in the meta-model in Section 4—allows us to adjust the probability of switching when constructing target policy candidates. Specifically, let $\bar{p}^s_{\hat{\mu}} := p^s_{\hat{\mu}} + p_1$, where $p_1 \in \left[-p^s_{\hat{\mu}}, 1 - p^s_{\hat{\mu}}\right]$, be the adjusted probability of switching treatment. For example, by setting $p_1 > 0$, we can create target policies that more strongly encourage treatment switching compared to current practice. Note that we would still consider the most commonly selected treatments (see Equation (1)), but with an increased probability of selecting a different treatment.

Adjusting the probability of treatment switching in this way may lead to a target policy that partially violates the overlap assumption. Specifically, this means that $p_\pi(a \mid s) > 0$ when $p_\mu(a \mid s) \approx 0$ for some state-action pairs. In such cases, an IS-based OPE estimator could suffer from high variance, hindering accurate and reliable value estimation. Therefore, we recommend examining the modified switch probabilities for each leaf of the tree used to predict treatment switching: do they make clinical sense for each group of patients?

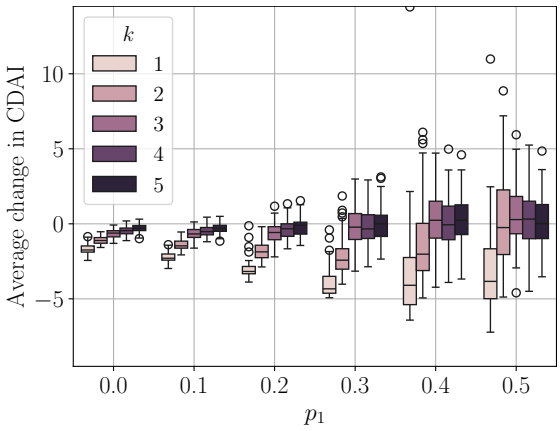


Figure 4: Off-policy evaluation of target policies based on the k most common treatments selected under the behavior policy, with the probability of treatment switching adjusted by adding a constant p_1. Value estimates are normalized relative to the behavior policy, representing the average decrease in CDAI per treatment stage if each target policy were used in place of the behavior policy.

In Figure 4, we examine the effect of increasing the probability of switching therapy in RA. Specifically, we vary the parameter p_1 from 0 to 0.5. As shown in the figure, this results in a decrease in CDAI for most values of k, suggesting that increasing treatment switching—when applied across patient groups defined by the behavior policy model—may be a promising strategy for RA treatment. However, as shown in the right panel, the variance of the estimates increases substantially with higher values of p_1, indicating that these results should be interpreted with caution.

Table 6: Hyperparameters of the decision tree (DT) and the recurrent neural network (RNN) along with their respective search spaces. For the meta-models DT-S and DT-BLS, each individual decision tree used the same hyperparameters as the standard DT. The same set of hyperparameters was applied to both RA and sepsis. However, for RA, we additionally considered two factors: (1) the proportion of observations collected before the first b/tsDMARD initiation and (2) the fraction of observations from patients who started b/tsDMARD therapy before registry enrollment. The following values were considered for each parameter: $\{0, 0.25, 0.5, 0.75, 1\}$.

Model	Hyperparameter	Values
DT	max depth	$\{2, 3, 4, 5, 6, 7, 8, 9\}$
	min fraction of samples per leaf	$\{0.01, 0.02, 0.03, 0.04, 0.05\}$
RNN	learning rate	$\{0.001, 0.01, 0.1\}$
	batch size	$\{32, 64, 128\}$
	encoding dimension	$\{16, 32, 64\}$