How Should We Represent History in Interpretable Models of Clinical Policies?

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

Lena Stempfle STEMPFLE@CHALMERS.SE
Chalmers University of Technology and University of Gothenburg

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

Zachary R. Margolin ZMARGOLIN@COREVITAS.COM
CorEvitas, LLC

Heather J. Litman HLITMAN@COREVITAS.COM
CorEvitas, LLC

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

Abstract

Modeling policies for sequential clinical decision-making based on observational data is useful for describing treatment practices, standardizing frequent patterns in treatment, and evaluating alternative policies. For each task, it is essential that the policy model is interpretable. Learning accurate models requires effectively capturing a patient's state, either through sequence representation learning or carefully crafted summaries of their medical history. While recent work has favored the former, it remains a question as to how histories should best be represented for interpretable policy modeling. Focused on model fit, we systematically compare diverse approaches to summarizing patient history for interpretable modeling of clinical policies across four sequential decision-making tasks. We illustrate differences in the policies learned using various representations by breaking down evaluations by patient subgroups, critical states, and stages of treatment, highlighting challenges specific to common use cases. We find that interpretable sequence models using learned representations perform on par with black-box models across all tasks. Interpretable models using handcrafted representations perform substantially worse when ignoring history entirely, but are made competitive by incorporating only a few aggregated and recent elements of patient history. The added benefits of using a richer representation are pronounced for subgroups and in specific use cases. This underscores the importance of evaluating policy models in the context of their intended use.

Keywords: decision-making, interpretable policy modeling, history representation

Data and Code Availability. We use four medical datasets to model sequential clinical decision-making across various conditions. The Alzheimer's disease dataset comprises 1,605 patients from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (https://adni.loni.usc.edu/). Data on 4,391 patients with rheumatoid arthritis (RA) are sourced from the CorEvitas RA registry (Kremer, 2005). Sepsis and chronic obstructive pulmonary disease datasets are derived from the MIMIC-III and MIMIC-IV databases (Johnson et al., 2016, 2023), containing data on 20,932 and 7,977 patients, respectively. ADNI and MIMIC are publicly available to researchers. RA data are available from CorEvitas, LLC through a commercial subscription agreement and are not publicly available. The code is available at https://github.com/Healthy-AI/inpole.

Institutional Review Board (IRB). Research on de-identified data from ADNI and MIMIC is exempt from IRB review under HIPAA. Approval for the investigation of treatment patterns within the

CorEvitas RA registry was granted by the Swedish Ethical Review Authority (application no. 2021-06144-01).

1. Introduction

Sequential decision-making is central to the treatment of many medical conditions, both acute and chronic (Chakraborty and Moodie, 2013; Gottesman et al., 2019). The patterns in how treatment decisions depend on available information, aggregated over physicians and patients, are commonly referred to as the *behavior policy*. Modeling these patterns has several use cases: **Explanation:** A behavior policy model can provide insights into current treatment strategies and support the development of new clinical guidelines (Pace et al., 2022; Hüyük et al., 2021; Deuschel et al., 2024); **Implementation:** Identifying common treatment patterns and standardizing them can reduce practice variation and leverage the collective expertise of many clinicians (Esteva et al., 2017; Hannun et al., 2019); **Evaluation:** Most approaches to off-policy evaluation of a new policy, such as importance weighting (Precup, 2000), rely on accurate probabilistic models of the behavior policy.

All three use cases of behavior policy modeling benefit from an *interpretable* model. For explanation and implementation, interpretability is even crucial for gaining the trust of end users (Stiglic et al., 2020). In off-policy evaluation, interpretability allows for verifying the model fit, reasoning about omitted confounding variables—variables that affect both treatment decisions and outcomes—and comparing the behavior policy to a target policy representing new clinical guidelines (Matsson and Johansson, 2022). Failing to account for confounders in off-policy evaluation may result in biased estimates of the value of the target policy (Namkoong et al., 2020). To accurately model sequential clinical decision-making and mitigate bias in downstream tasks, it is important to account for the patient's medical history, including previous contexts, treatments, and outcomes (Gottesman et al., 2019). Despite this, several studies neglect historical context, focusing solely on present observations when formulating clinical policies (Javad et al., 2019; Asoh et al., 2013; Utomo et al., 2018; Lin et al., 2018; Weng et al., 2017). This raises the question: How should we represent history in interpretable models of clinical policies?

The literature suggests two primary approaches to summarizing patient history: learned sequence representations and hand-crafted summary features. Recent work in interpretable policy modeling favors the former, employing techniques such as recurrent decision trees (Pace et al., 2022) or recurrent neural networks (Deuschel et al., 2024) to create abstract history representations. For individual patients, these models provide policy descriptions based on the most recent patient information, individualized through the encoded history. Other representation learning methods identify prototypes (Li et al., 2018; Ming et al., 2019)—patients that represent larger groups of individuals—providing a compact description of the overall policy (Matsson and Johansson, 2022).

On the other hand, summarizing history using hand-crafted features is useful for fitting simple, interpretable models such as linear or rule-based classifiers. These are arguably more interpretable—at least more transparent (Lipton, 2018)—than representation learning methods, as they explicitly show how historical information influences the current decision. Naturally, the best summary is specific to the problem at hand (Gottesman et al., 2019), but several strategies frequently appear in the literature, such as aggregating patient information across time (Raghu et al., 2017; Komorowski et al., 2018; Guez et al., 2018), using a fixed-sized window of the most recent observations (Bertsimas et al., 2022; Escandell-Montero et al., 2014), or incorporating indicators for past decisions (Bertsimas et al., 2022).

Despite the popularity of these approaches to summarizing history, they have not been compared systematically in the context of interpretable policy learning. Notably, prior work using representation learning (Pace et al., 2022; Deuschel et al., 2024) has not explored the impact of different history representations during evaluation. Moreover, their methods are primarily evaluated on two small datasets, focusing solely on binary decisions.

Contributions. We compare diverse methods for representing history in interpretable modeling of clinical policies, asking: How does the quality of the model fit depend on the representation method and the level of detail in history summaries? What factors explain variations in performance across different representation methods? And how does the choice of representation affect common use cases such as explanation, implementation and evaluation? We collect evidence to address these questions by comparing eight history representations across distinct decision-making tasks, featuring both binary and multi-class

action spaces: 1) selecting therapy for patients with rheumatoid arthritis, 2) conducting an MRI scan for patients with suspected Alzheimer's disease, and managing 3) sepsis and 4) acute exacerbations of chronic obstructive pulmonary disease in the ICU. In all tasks, interpretable models, whether using representation learning or hand-crafted representations, perform similarly to black-box models in aggregate, suggesting that interpretable policy modeling is viable. By analyzing patient subgroups, temporal patterns, and critical states, we find that methods with similar average performance differ in the kinds of errors they make and the stages at which these errors occur, which can potentially have major consequences for specific use cases such as off-policy evaluation.

2. Interpretable Policy Modeling

In sequential clinical decision-making, a behavior policy μ represents the treatment patterns physicians generate when making decisions for patients. Our task is to estimate an unknown behavior policy μ from observational data consisting of n patient sequences of observations (contexts) $X_t \in \mathcal{X}$ and medical decisions (actions) $A_t \in \mathcal{A} = \{1, \dots, K\}$, recorded at each stage $t = 1, \dots, T$ of treatment.[1] We let $H_t := (X_1, A_1, \dots, X_{t-1}, A_{t-1}, X_t)$ represent the history of contexts and actions up until the current stage t.[2] We refer to the basis for a physician's choice of treatment as the state $S_t \in \mathcal{S}$ of a patient, assumed to be an unknown function of the history H_t (Sutton and Barto, 2018). In other words, all direct causes of A_t are assumed to be contained in S_t and in H_t.

We quantify treatment patterns by estimating $p_\mu(A_t \mid S_t)$, the probability of choosing an action A_t in a state S_t under the behavior policy μ. In other contexts, this is called *propensity estimation* (Abadie and Imbens, 2016), *policy recovery* (Deuschel et al., 2024) or *behavior cloning* (Torabi et al., 2018). Particular use cases of estimates $\hat{p}_\mu(A_t \mid S_t)$ may impose additional constraints on the model. To explain sequential clinical decision-making, the model should ideally be fully interpretable, allowing humans to understand its calculations in their entirety. When implementing the policy in clinical practice, it may be sufficient to understand the model's predictions for individual patients to detect errors or unexpected behavior. For off-policy evaluation, the objective is to estimate the importance ratio ρ of a target policy p_π and the behavior policy model $\hat{p}_\mu$, $\rho = \frac{p_\pi(A_t|S_t)}{\hat{p}_\mu(A_t|S_t)}$ (Precup, 2000). In this case, an interpretable model of the behavior policy allows for, e.g., understanding differences between the policies and detecting violations of policy overlap (Matsson and Johansson, 2022).

Following common practice, we construct the state S_t either as a hand-crafted summary or a learned representation of the history, $S_t = f(H_t)$. Regardless of method and use case, a key challenge in policy estimation is to ensure that S_t retains sufficient information to predict the action A_t (Gottesman et al., 2019). This is especially true for interpretable models which aim to have small, transparent policy descriptions. In particular, the state must account for confounding variables that have a causal effect on both the treatment decision and its outcome. What constitutes a sufficient state depends on the problem. For example, in the treatment of patients with rheumatoid arthritis, is the choice of treatment A_t based solely on the current context X_t, including patient demographics, disease activity measures, and presence of comorbidities? Do previous treatments or their outcomes matter? What about their mutual order? Next, we discuss two common approaches to summarizing a patient's medical history: learned sequence representations and hand-crafted features formed by history truncation and history aggregation. In our experiments, we use hand-crafted features as building blocks to explore different state constructions.

2.1. Sequence Representation Learning

Since the space of possible histories H_t grows exponentially with time, the history quickly becomes unwieldy. Sequence models such as recurrent neural networks (RNNs) can be used to learn compact summaries of patient histories to use as the state S_t. For example, Wang et al. (2018) used a long short-term memory RNN to summarize the history for dynamic treatment recommendation in intensive care. An RNN is known to be opaque but it is possible to open the black-box by introducing a prototype layer into the architecture (Li et al., 2018; Ming et al., 2019) or using it to parameterize an interpretable model (Deuschel et al., 2024). Another approach is to represent the history using recurrent decision trees Pace et al. (2022). However, such models require post-processing to enable human interpretation.

1. The total number of stages, T, is a finite random variable that indicates the end of the course of medication.
2. We assume that the observed outcome of a treatment choice is part of the next set of patient observations.

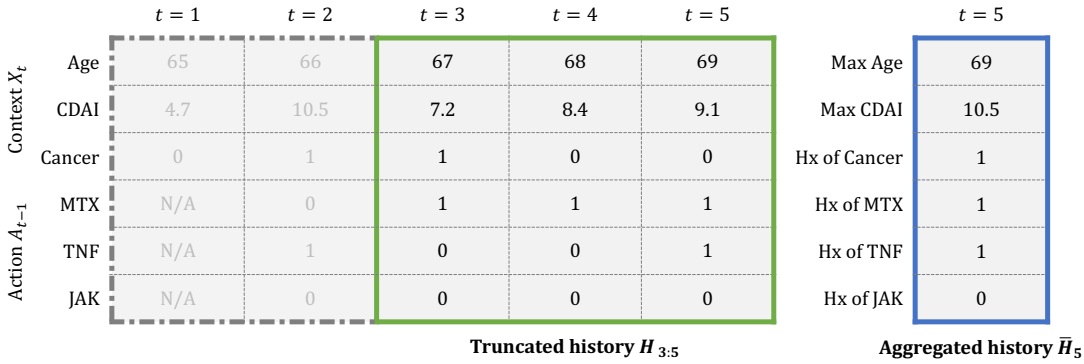

Figure 1: History truncation and history aggregation using the `max` operator applied to the history of a patient with rheumatoid arthritis. A rolling window of size three is used for the history truncation. The context X_t is a vector with three components, X_t^1, X_t^2, and X_t^3, representing the patient's age, clinical disease activity index (CDAI), and co-existence of cancer. The simplified action space consists of three therapies and their combinations: methotrexate (MTX), tumor necrosis factor (TNF) inhibitor, and Janus kinase (JAK) inhibitor.

2.2. History Truncation

History truncation involves selecting a fixed-sized portion of the most recent history, or parts of it, assuming that distant historical events have limited impact on the current decision. Formally, let $H_{(t-k):t} := (X_{t-k}, A_{t-k}, \ldots, X_{t-1}, A_{t-1}, X_t)$, where $k \geq 0$, be the truncated history until stage t. For $k = 2$, as illustrated in Figure 1, $H_{(t-2):t}$ includes the current context X_t and contexts from the two preceding stages, along with the actions taken at stage $t-1$, $t-2$ and $t-3$. In our experiments, we apply the history window only to variables for which previous observations are assumed to potentially influence the current decision. For example, in Figure 1, assuming regular follow-up visits, the patient's age at stage $t-2$ and $t-3$ is redundant given the current age.

Truncating sequence data is a common preprocessing step in natural language processing and bioinformatics. In medical applications, Bertsimas et al. (2022) defined the state based on the most recent heart rate observations to learn a message delivery policy for mobile health. Escandell-Montero et al. (2014) optimized anemia treatment by formulating a state based on the treatment dose at stage $t-1$, $t-2$ and $t-3$. While truncating history allows for constructing a compact state that captures recent historical events, this representation has two obvious disadvantages. First, truncating the history at a specific time step may exclude critical past information. Second, when $t \leq k$, the absence of earlier history requires some form of imputation, especially if the behavior policy model, $\hat{p}_\mu$, expects an input of fixed size, which is the case for several models in this work.

2.3. History Aggregation

History aggregation is applied under the assumption that the temporal order of historical events holds little significance. This method aggregates historical information, such as previous treatment assignments, across time, creating a rough summary of the history. Let X_t^i be a component of the context vector X_t. The observations $X_1^i, \ldots, X_t^i$ are combined into a single variable $\bar{X}_t^i$ according to $\bar{X}_t^i = \text{agg}_t\, X_t^i$, where the aggregation operator can be, e.g., `sum`, `max` or `mean`. Aggregations of binarized actions are defined analogously, $\bar{A}_t^i = \text{agg}_t\, A_t^i$, and the aggregated history is the set of aggregated observations and actions, $\bar{H}_t = \{\bar{X}_t, \bar{A}_{t-1}\}$. We apply this operation to variables for which the aggregate is assumed to provide different information than the current observation alone. Again, the age of a patient is an example of a variable for which aggregation provides no extra information. In such cases, we set $\bar{X}_t^i = X_t^i$.

The meaning of history aggregation depends on the aggregation operator and variable type, as illustrated in Figure 1 using the `max` operator. For numerical variables like CDAI, the aggregate corresponds to the highest observed value. For categorical vari-

ables, such as the presence of cancer or previously administered therapies, the aggregate indicates whether the patient has ever experienced the event. Other examples are found in the literature. For example, Komorowski et al. (2018) and Raghu et al. (2017) accumulated fluids outputs ($\text{sum}_t\, X_t^i$) for learning optimal policies for the management of sepsis. Bertsimas et al. (2022) counted the number of messages previously sent ($\text{sum}_t\, A_t^i$) while developing their message delivery policy. Guez et al. (2018) applied `mean` and `max` transformations to EEG signals to optimize stimulation policies for the treatment of epilepsy.

3. Experiments

We study interpretable models of clinical policies in a series of experiments, aiming to answer the questions raised in Section 1: How does the quality of the model fit depend on the representation method and the level of detail in the state S_t? What factors explain variations in performance across different representation methods? And how does the choice of representation affect common use cases? Recognizing that interpretation is strongly tied to domain knowledge, we do not aim to evaluate the degree of interpretability of the different models. Instead, we seek to understand how the model classes differ in their fit. We compare learning using diverse states based on sequence representation learning and hand-crafted features within decision processes related to four medical conditions: Alzheimer's disease, rheumatoid arthritis, sepsis, and chronic obstructive pulmonary disease.

3.1. Datasets

Our datasets, as detailed in the "Data and Code Availability" statement, illustrate the diversity of sequential clinical decision-making tasks. For instance, the treatment of Alzheimer's disease and rheumatoid arthritis (RA) spans several years, with regularly scheduled follow-up visits to slow disease progression. In contrast, managing sepsis and chronic obstructive pulmonary disease (COPD) in the ICU requires continuous administration of treatments like intravenous fluids, vasopressors, and sedative drugs to preserve the patient's life. In all cases except for ADNI, where decisions are binary, clinicians are faced with multiple treatment options at each stage of care. See Table 1 for brief characteristics of the datasets. Details are included in Appendix A.

3.2. Models

We include three types of interpretable models based on sequence representation learning: prototype-based models designed for sequential data (PSN) (Ming et al., 2019), recurrent decision trees (RDT) (Pace et al., 2022), and models leveraging the recent contextualized policy recovery framework (CPR) (Deuschel et al., 2024). The latter is developed for binary actions and thus only used for ADNI. For comparison, we learn generalized linear models and rule-based models using hand-crafted history representations. Generalized linear models, particularly logistic regression (LR), are widely used as propensity score models for estimating treatment effects in observational studies (Feng et al., 2012; Spreeuwenberg et al., 2010). Rule-based models, such as decision trees (DT), are commonly employed in clinical decision support systems (Banerjee et al., 2019; Chrimes et al., 2023). For ADNI, we also include risk scores (RS) (Ustun and Rudin, 2019), i.e., scoring systems enabling probabilistic predictions. In addition, we include a multilayer perceptron (MLP) and a recurrent neural network (RNN) in the form of a long short-term memory. These models serve as benchmarks to demonstrate the potential accuracy of policy modeling based on the available data. Table 2 provides an overview of the included models.

3.3. Experimental Setup

We divide each dataset—ADNI, RA, Sepsis, and COPD—into training and testing subsets using an 80/20 split, with 20 % of the training dataset aside for model validation. For the ADNI, RA and COPD datasets, we apply one-hot encoding to categorical features. Depending on the type of model, we standardize continuous variables or discretize them into five equally-sized partitions. Missing values are primarily imputed on a patient level by propagating the last valid observation, secondarily using mean imputation or frequent category imputation. The Sepsis dataset is preprocessed as described in Komorowski et al. (2018), with normally distributed data standardized and log-normally distributed data log-transformed before standardization.

Models based on sequence representation learning are trained used the full history as input, $S_t = H_t$. For other models, we consider the following state representations: X_t, A_{t-1}, $\{X_t, A_{t-1}\}$, $\bar{H}_t$, $\{X_t, A_{t-1}, \bar{H}_t\}$, $\{H_{(t-1):t}, \bar{H}_t\}$, and $\{H_{(t-2):t}, \bar{H}_t\}$. Note that $\{X_t, A_{t-1}\} = H_{(t-0):t}$. To simplify no-

Table 1: Characteristics of the ADNI, RA, Sepsis, and COPD datasets.

	ADNI	RA	Sepsis	COPD
Patients, n	1,605	4,391	20,932	7,977
Age in years, median (IQR)	73.9 (69.3, 78.8)	58.0 (49.0, 66.0)	66.1 (53.7, 77.9)	67.0 (56.0, 77.0)
Female, n (%)	715 (44.5)	3355 (76.5)	9,250 (44.2)	3,472 (43.5)
Patient observations X_t, n	6	33	18	37
Actions A_t, n	2	8	25	25
Stages T, median (IQR)	3.0 (3.0, 3.0)	5.0 (3.0, 8.0)	13.0 (10.0, 17.0)	18.0 (18.0, 18.0)

Table 2: An overview of the models used in our experiments.

Model	Interpretable policy	Accepts $\lvert\mathcal{A}\rvert > 2$	Accepts H_t
Risk scores (RS)	✓	✗	✗
Logistic regression (LR)	✓	✓	✗
Decision tree (DT)	✓	✓	✗
Multilayer perceptron (MLP)	✗	✓	✗
Contextualized policy recovery (CPR)	✓	✗	✓
Prototypical sequence network (PSN)	✓	✓	✓
Recurrent decision tree (RDT)	✓	✓	✓
Recurrent neural network (RNN)	✗	✓	✓

tation, we let $H_{(k)} \coloneqq H_{(t-k):t}$. For truncation, we compare three operators: `sum`, `max`, and `mean`.

For each dataset, state representation and relevant model type, we train five candidate models using randomly sampled hyperparameters. Final models are selected based on the highest area under the receiver operating characteristic curve (AUROC) score achieved on the validation set. These models are then evaluated on the held-out test set with respect to AUROC and calibration error. The entire process is repeated for five different splits of the dataset, and we report 95 % confidence intervals based on the bootstrap distribution of the respective metric. Further details on the experimental setup, including hyperparameter selection, can be found in Appendix B.

3.4. General and Stratified Performance

In Table 3, we report average test AUROC for each model and dataset, using the different state representations described above, with the `sum` operator used for history aggregation.[3] Across tasks, we find that the best-performing interpretable model performs on par with RNN, suggesting that interpretable policy modeling is feasible. Accounting for historical information, not just the current context X_t, is critical to achieving a good model fit. The best-performing sequence model, which takes the entire history as input, generally outperforms non-sequential models that rely on hand-crafted states. However, the difference in average AUROC is small when providing non-sequential models with a state that captures more than the current context X_t and the previous action A_{t-1}. In general, aggregating history through summarization yields the best results. This is particularly true for ADNI, where the difference between aggregation operators is significant. With a well-constructed state, LR can perform significantly better than what is suggested in prior work by, e.g., Pace et al. (2022).

In all tasks, models using the state $S_t = A_{t-1}$ perform surprisingly well in terms of average test AUROC. It is reasonable to ask: Is this metric alone a reliable measure of performance? Surely, actions depend on more than the previous choice? To understand this better, we stratify the Sepsis results over treatment stage and by patient groups, defined by the rate of change of the NEWS2 score as in Luo et al. (2024). We identify six subgroups, enumerated 1–6, where subgroups 1 and 6 correspond to patients that have a large negative and a large positive rate of change of the NEWS2 score, respectively.

In Figure 2, we show the distribution of AUROC for subgroups 1, 3, 4, and 6 using all states except

3. Confidence intervals and calibration errors are included in Table 16 and 15, in Appendix C. Results for other aggregation operators are shown in Table 13 and 14.

Table 3: Average test AUROC, expressed as a percentage, in all four tasks: ADNI, RA, Sepsis, and COPD. The upper section contains models with different hand-crafted states; the lower section contains representation learning methods. MLP and RNN are included as benchmarks. History aggregation $\bar{H}$ is performed using the `sum` operator. Confidence intervals are included in Table 16 in Appendix C.

	ADNI				**RA**			**Sepsis**			**COPD**		
State	RS	LR	DT	MLP	LR	DT	MLP	LR	DT	MLP	LR	DT	MLP
X_t	54.2	56.2	53.9	55.6	61.7	58.8	61.1	82.1	78.2	84.1	77.9	74.7	78.8
A_{t-1}	52.0	53.9	53.8	53.7	94.7	94.7	94.7	88.0	90.6	91.1	92.9	95.0	95.0
$H_{(0)}$	53.4	56.8	54.3	56.8	95.6	95.7	96.1	91.3	92.1	94.7	94.0	96.0	95.4
$\bar{H}_t$	63.0	64.4	64.9	64.1	90.5	92.0	94.0	84.6	85.2	89.1	91.1	89.3	93.5
$H_{(0)}, \bar{H}_t$	63.7	65.3	65.0	65.8	96.1	96.5	96.9	91.9	92.3	95.3	94.7	96.7	96.3
$H_{(1)}, \bar{H}_t$	63.4	65.6	65.4	66.0	96.0	96.4	96.9	92.2	92.5	95.5	94.7	96.8	96.4
$H_{(2)}, \bar{H}_t$	62.9	65.4	65.3	66.8	96.0	96.4	96.7	92.3	92.6	95.5	94.7	96.8	96.3
	ADNI				**RA**			**Sepsis**			**COPD**		
State	CPR	PSN	RDT	RNN	PSN	RDT	RNN	PSN	RDT	RNN	PSN	RDT	RNN
H_t	68.7	66.7	62.8	68.0	96.2	90.0	96.8	94.9	77.0	95.7	96.2	81.9	96.5

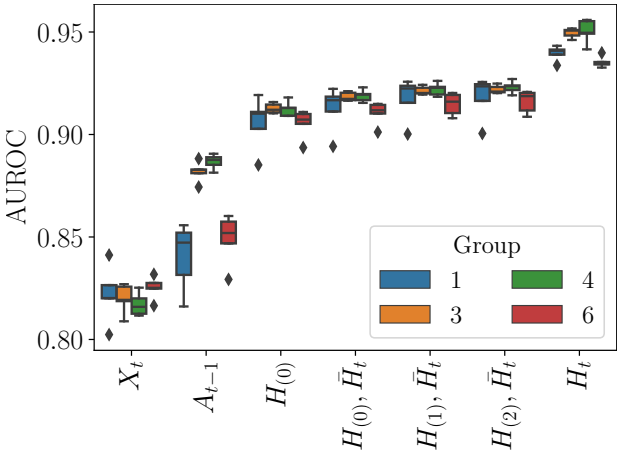

Figure 2: AUROC across states and patient groups, identified based on the rate of change of the NEWS2 score, in Sepsis. PSN is used with $S_t = H_t$, LR with the others.

$\bar{H}_t$. LR is fit to hand-crafted states, whereas PSN is used with $S_t = H_t$. We clearly see that the previous action is an insufficient state for subgroups 1 and 6, i.e., patient groups that are likely to have higher variation in their treatment compared to other groups. Table 4 clarifies the relative difference in AUROC, expressed as a percentage, between the models.

Table 4: Percentage difference in AUROC between PSN and LR, fit to the Sepsis data using each state representation in the column "State". G1, G3, G4, and G6 denote different patient groups based on the rate of change of the NEWS2 score; G1 and G6 have higher variability than G3 and G4.

State	G1	G3	G4	G6
X_t	−12.5	−13.7	−14.0	−11.8
A_{t-1}	−10.5	−7.1	−6.7	−9.2
$H_{(0)}$	−3.6	−3.9	−4.0	−3.2
$H_{(0)}, \bar{H}_t$	−2.9	−3.3	−3.3	−2.6
$H_{(1)}, \bar{H}_t$	−2.4	−3.0	−3.0	−2.2
$H_{(2)}, \bar{H}_t$	−2.3	−2.9	−2.9	−2.1

State representations that take historical events into account through truncation and aggregation enable LR to approach the performance of PSN.

Figure 3 shows AUROC with respect to the stage of treatment for DT, fit to the Sepsis data using varying state representations, and PSN with $S_t = H_t$. In early stages, we clearly see the shortcomings of the state based on the previous action only. However, in later stages, A_{t-1} is a good predictor for the doctor's decision. It may be the case that patients' conditions stabilize and that clinicians reuse the same treatment in subsequent stages. The significance of history is

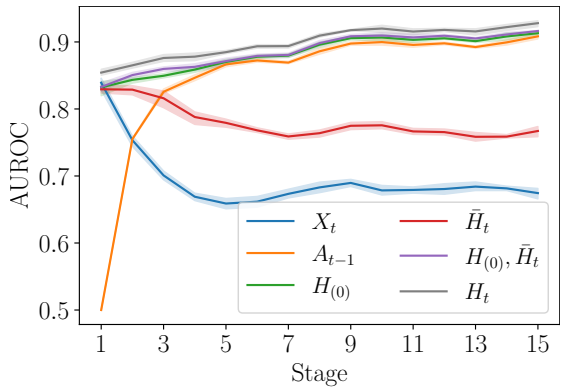

Figure 3: AUROC for different states across stages of treatment in Sepsis. PSN is used with $S_t = H_t$, DT with the others.

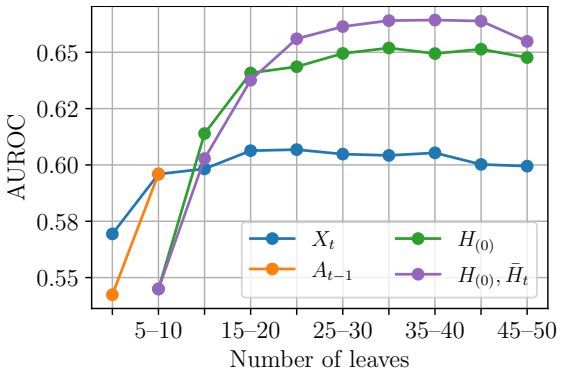

Figure 4: Therapy switch prediction for RA. AUROC against the number of leaves for DT fit using different states. With $S_t = A_{t-1}$, the tree can have at most 5–10 leaves.

illustrated by the difference between X_t and $\bar{H}_t$, with the latter consistently providing a better model fit.

3.5. Modeling Policies for Explanation, Implementation and Evaluation

Learning an interpretable model of the behavior policy is required to explain decision-making and can help verify assumptions in off-policy evaluation. But is it generally possible to learn an interpretable model that performs well? And is there a cost associated with it? In Figure 4, we plot AUROC against the number of leaves in decision trees fit to the RA data using four different state representations. We measure AUROC in critical states where a switch of treatment was made. A near-optimal model, obtained with the state $\{H_{(0)}, \bar{H}_t\}$, requires around 30 leaves and may be difficult for humans to comprehend in its entirety, making it less suitable for explanation. However, it could still provide insights into specific decisions, if implemented in practice, as its decision paths (depth) are fairly short.

For implementation, it is logical to ask: How would the actions suggested by a simple model such as LR differ from those suggested by the best-in-class model? In Figure 5, we investigate this by comparing LR, fit to the RA data using $S_t = H_{(0)}$, to RNN, focusing on the states where a change of treatment was made. In most cases, the predicted actions align, but LR confuses, for example, the less frequent non-TNF combination therapy with the more frequent TNF combination and csDMARD therapies. When

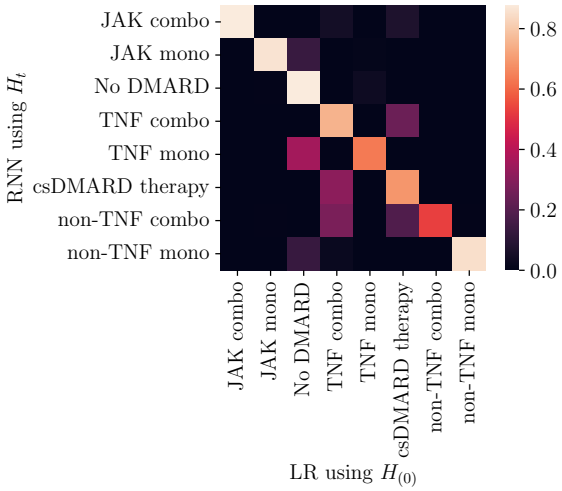

Figure 5: Errors in therapy switch prediction for RA. Confusion matrix between RNN using $S_t = H_t$ and LR using $S_t = H_{(0)}$. See Appendix A for therapy definitions.

the state representation and/or model is overly simplified, we risk losing precision in rare cases.

Off-policy evaluation of new policies is often performed using importance weighing, see Section 2. A crucial step is to re-weight observed outcomes by the product of inverse probabilities $p_{\hat{\mu}}(a_t \mid s_t)^{-1}$, where a_t is the action taken in state s_t under the behavior policy. In Figure 6, we inspect the median of inverse probability products obtained with LR and RNN when considering different state representations

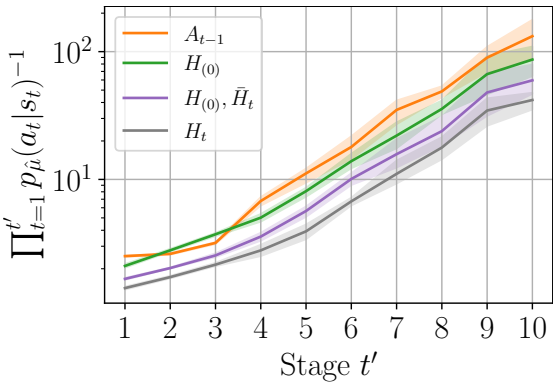

Figure 6: Off-policy evaluation for RA. Median of inverse probability products at stage $t' = 1, 2, ..., 10$ of treatment using LR (A_{t-1}, $H_{(0)}$, and $\{H_{(0)}, \bar{H}_t\}$) and RNN (H_t).

for the former. This product is inversely proportional to the likelihood of observed data; the more individual probabilities deviate from 1, the larger the product. The figure shows that using coarse representations of history, such as A_{t-1} or $H_{(0)}$, leads to inverse probability products that grow much faster (note the logarithmic scale) compared to using a more comprehensive representation of history. This accelerated growth negatively impacts the variance in off-policy evaluation and poses one of the greatest challenges to its practical feasibility (Gottesman et al., 2018).

4. Discussion

In this work, we compared two common approaches to representing patient history for interpretable modeling of clinical policies: hand-crafted summary features and learned sequence representations. In particular, we studied how the quality of the model fit depends on the representation method and the level of detail in history summaries. Across four decision-making tasks, we found it possible to achieve competitive results using simple, manually crafted representations. Combining current patient observations, the most recent treatment, and historical aggregates of prior observations and treatments explained most of the variance in treatment selection. Notably, incorporating recent treatments was critical to model performance. These findings are consistent with clinical guidelines. For example, current recommendations for the management of rheumatoid arthritis focus on broad indicators such as poor prognostic factors rather than details of the patient's medical history (Smolen et al., 2020).

We investigated factors that explain variations in performance across different representation methods. For instance, by breaking down the results by patient subgroups and stages of treatment, we were able to identify shortcomings in simplified representations. Additionally, focusing on therapy selection in rheumatoid arthritis, we highlighted challenges associated with common use cases of interpretable policy modeling. For example, using a coarse history may increase variance in off-policy evaluation. In all experiments, interpretable models using learned representations performed comparably to black-box models, suggesting that interpretable policy learning is generally viable.

We assumed that all direct causes of treatment selections were captured within the observed patient histories. In practice, we were limited to the variables that were actually measured, which means there could be unmeasured confounders. In ADNI, we observed an average AUROC of 0.6–0.7, suggesting that variance in MRI scan ordering is not fully explained by the available variables. However, the published diagnostic policy for mild cognitive impairment, which can be modeled with the variables used here, shows no clear evidence of omitted variables (Pace et al., 2022). The variance may instead stem from differences across institutions and practitioners.

Another limitation is that we examined only a limited set of manually created history summaries. For example, it would be possible to combine different aggregation methods or derive additional features from historical data, such as changes in critical observations over time. We leave this extension for future research. The primary goal of this work was to understand the overarching impact of historical information in policy modeling, rather than to fine-tune representations for individual tasks.

An interesting direction for future work is to construct policy models that explicitly depend on the stage of treatment. As shown in Figure 2, current patient observations are important for accurately predicting the initial treatments of sepsis. Later in the process, the treatment is often repeated, suggesting that the patients' conditions stabilize. However, although a simple model may be sufficient to explain overall patterns, it risks introducing severe bias in specific use cases such as policy evaluation.

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.

Hideki Asoh, Masanori Shiro1 Shotaro Akaho, Toshihiro Kamishima, Koiti Hasida, Eiji Aramaki, and Takahide Kohro. An application of inverse reinforcement learning to medical records of diabetes treatment. In *ECML PKDD 2013 workshop on reinforcement learning with generalized feedback*, 2013.

Mousumi Banerjee, Evan Reynolds, Hedvig B Andersson, and Brahmajee K Nallamothu. Tree-based analysis: a practical approach to create clinical decision-making tools. *Circulation: Cardiovascular Quality and Outcomes*, 12(5):e004879, 2019.

Dimitris Bertsimas, Predrag Klasnja, Susan Murphy, and Liangyuan Na. Data-driven interpretable policy construction for personalized mobile health. In *Proceedings of the 2022 IEEE International Conference on Digital Health*, pages 13–22. IEEE, 2022.

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

Dillon Chrimes et al. Using decision trees as an expert system for clinical decision support for covid-19. *Interactive Journal of Medical Research*, 12(1): e42540, 2023.

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*, 2024.

Pablo Escandell-Montero, Milena Chermisi, Jose M Martinez-Martinez, Juan Gomez-Sanchis, Carlo Barbieri, Emilio Soria-Olivas, Flavio Mari, Joan Vila-Francés, Andrea Stopper, Emanuele Gatti, et al. Optimization of anemia treatment in hemodialysis patients via reinforcement learning. *Artificial Intelligence in Medicine*, 62(1):47–60, 2014.

Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.

Ping Feng, Xiao-Hua Zhou, Qing-Ming Zou, Ming-Yu Fan, and Xiao-Song Li. Generalized propensity score for estimating the average treatment effect of multiple treatments. *Statistics in Medicine*, 31(7): 681–697, 2012.

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.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature Medicine*, 25(1):16–18, 2019.

Arthur Guez, Robert D Vincent, Massimo Avoli, and Joelle Pineau. Adaptive treatment of epilepsy via batch-mode reinforcement learning. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, volume 8, pages 1671–1678, 2018.

Mehak Gupta, Brennan Gallamoza, Nicolas Cutrona, Pranjal Dhakal, Raphael Poulain, and Rahmatollah Beheshti. An extensive data processing pipeline for MIMIC-IV. In *Proceedings of the 2nd Machine Learning for Health symposium*, volume 193, pages 311–325. PMLR, 2022.

Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25 (1):65–69, 2019.

Alihan Hüyük, Daniel Jarrett, and Mihaela van der Schaar. Explaining by imitating: Understanding decisions by interpretable policy learning. In *Proceedings of the 9th International Conference on Learning Representations*, 2021.

Matt Inada-Kim and Emmanuel Nsutebu. NEWS 2: an opportunity to standardise the management of deterioration and sepsis. *BMJ*, 360:k1260, 2018.

Mahsa Oroojeni Mohammad Javad, Stephen Olusegun Agboola, Kamal Jethwani, Abe Zeid, Sagar Kamarthi, et al. A reinforcement learning–based method for management of type 1 diabetes: exploratory study. *JMIR Diabetes*, 4(3):e12905, 2019.

Alistair E.W. 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.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023.

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.

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

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 32nd AAAI Conference on Artificial Intelligence*, 2018.

Rongmei Lin, Matthew D Stanley, Mohammad M Ghassemi, and Shamim Nemati. A deep deterministic policy gradient approach to medication dosing and surveillance in the ICU. In *40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4927–4931. IEEE, 2018.

Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

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*, 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*, pages 1339–1349. PMLR, 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, 2024.

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 & Data Mining*, pages 903–913, 2019.

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 33*, pages 18819–18831, 2020.

Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring calibration in deep learning. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca

Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, 2019.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Doina Precup. Eligibility traces for off-policy policy evaluation. In *Proceedings of the 17th International Conference on Machine Learning*, 2000.

Aniruddh Raghu, Matthieu Komorowski, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach. In *Proceedings of the 2nd Machine Learning for Healthcare Conference*, pages 147–163. PMLR, 2017.

Josef S. Smolen, Robert B.M. Landewé, Johannes W.J. Bijlsma, Gerd R. Burmester, Maxime Dougados, Andreas Kerschbaumer, Iain B. McInnes, Alexandre Sepriano, Ronald F. Van Vollenhoven, Maarten De Wit, et al. EULAR recommendations for the management of rheumatoid arthritis with synthetic and biological disease-modifying antirheumatic drugs: 2019 update. *Annals of the Rheumatic Diseases*, 79(6):685–699, 2020.

Marieke Dingena Spreeuwenberg, Anna Bartak, Marcel A. Croon, Jacques A. Hagenaars, Jan J.V. Busschbach, Helene Andrea, Jos Twisk, and Theo Stijnen. The multiple propensity score as control for bias in the comparison of more than two treatment arms: an introduction from a case study in mental health. *Medical Care*, 48(2):166–174, 2010.

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

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

Marian Tietz, Thomas J. Fan, Daniel Nouri, Benjamin Bossan, and skorch Developers. *skorch: A scikit-learn compatible neural network library that wraps PyTorch*, 2017.

Faraz Torabi, Garrett Warnell, and Peter Stone. Behavioral cloning from observation. *arXiv preprint arXiv:1805.01954*, 2018.

Berk Ustun and Cynthia Rudin. Learning optimized risk scores. *Journal of Machine Learning Research*, 20(150):1–75, 2019.

Chandra Prasetyo Utomo, Xue Li, and Weitong Chen. Treatment recommendation in critical care: A scalable and interpretable approach in partially observable health states. In *Proceedings of the 39th International Conference on Information Systems*, 2018.

Lu Wang, Wei Zhang, Xiaofeng He, and Hongyuan Zha. Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2447–2456, 2018.

Wei-Hung Weng, Mingwu Gao, Ze He, Susu Yan, and Peter Szolovits. Representation and reinforcement learning for personalized glycemic control in septic patients. *arXiv preprint arXiv:1712.00654*, 2017.

Appendix A. Dataset Descriptions

In this section, we describe the datasets used in our experiments (ADNI, RA, Sepsis, and COPD) in more detail. When using a state representation based on history truncation, we replaced missing context observations for $t \leq k$ with the corresponding observation at the first time step. Missing information about the previous action(s) was replaced with "csDMARD therapy" (RA), "no MRI scan" (ADNI), and 0 (Sepsis and COPD).

A.1. Alzheimer's Disease

Following Hüyük et al. (2021) and Pace et al. (2022), we compiled a dataset of 1,605 patients from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`https://adni.loni.usc.edu/`). The raw data was loaded using the function `adnimerge` provided in the R package "ADNIMERGE", see `https:`

//adni.bitbucket.io. Specifically, we filtered out visits lacking a CDR-SB measurement, visits with a separation of more than six months, and patients with fewer than two visits.[4] We estimated the behavior policy for determining whether patients with suspected cognitive impairment should undergo a magnetic resonance imaging (MRI) scan. Patient observations included the CDR-SB score as well as factors such as age, gender, marital status, education level, and possession of the APOE4 allele. Following Hüyük et al. (2021) and Pace et al. (2022), we categorized the CDR-SB score as normal (0–0.5), questionable (0.5–4.5) or severe (4.5–18.0). The previous action was encoded in the outcome of any MRI scan ordered at the previous visit (no MRI scan ordered; hippocampal volume below average, average, or above average). Hippocampus volumes that deviated by ± 0.5 standard deviations of the average hippocampus volume were classified as below and above average, respectively. We describe the context variables in Table 5.

A.2. Rheumatoid Arthritis

We used data from the CorEvitas RA registry (Kremer, 2005), an ongoing longitudinal clinical registry in the US, to model the behavior policy for choosing disease-modifying antirheumatic drug (DMARD) therapy for patients with RA. The standard procedure in treating these patients involves initiating a conventional synthetic DMARD (csDMARD) and incorporating a biologic or targeted synthetic DMARD (b/tsDMARD) if the initial therapy fails, see for example Smolen et al. (2020). Biologic DMARDs are commonly divided into Tumor necrosis factor (TNF) biologics and non-TNF biologics. Janus kinase inhibitors (JAKi) are the most frequently used tsDMARDs.

The raw dataset contained data from 42,068 patients enrolled in the registry between January 2012 and December 2021. We removed 1,323 patients for whom information on therapy changes were missing or apparently incorrect. Since patients may have joined the registry at different times in their disease course, we focused on subsequences starting with the initiation of the first b/tsDMARD. We excluded patients with a history of b/tsDMARD treatment at registry enrollment and patients who did not start any b/tsDMARD within the registry. Additionally, we excluded all visits, as well as any subsequent visits, where multiple b/tsDMARDs were prescribed, as this is not clinically recommended. This left us with 4,391 patients in the final cohort.

There were no constraints placed on the follow-up visits in terms of regularity, although the registry protocol suggests visits every six months in line with clinical practice. Each patient was monitored up until their final documented visit or the data cut date of December 31, 2021, whichever came earlier. Consequently, the total number of registry visits and the length of follow-up differed among patients. We included 33 variables, see Table 6 and 7, in the context vector X_t.

Following previous work by Matsson et al. (2024), we studied changes between classes of DMARDs rather than changes between individual DMARDs. The following classes of drugs were studied: csDMARD therapy ("csDMARD"), TNF biologic monotherapy ("TNF (m)"), TNF biologic combination therapy ("TNF (c)"), non-TNF biologic monotherapy ("non-TNF (m)"), non-TNF biologic combination therapy ("non-TNF (c)"), JAKi monotherapy ("JAK (m)"), JAKi combination therapy ("JAK (c)"), and no DMARD therapy ("No DMARD").

A.3. Sepsis

Utilizing data from the MIMIC-III database (Johnson et al., 2016), we modeled the behavior policy for administering vasopressors and intravenous (IV) fluids to patients diagnosed with sepsis in the ICU. The data, comprising 20,932 patients, was structured as multivariate time series with a discrete time step of four hours using code provided by Komorowski et al. (2018). The patients were followed for up to 72 hours, but some individual sequences were shorter as a result of discharge or death of the patient. The doses of vasopressors and IV fluids were discretized into 5 distinct levels and then combined to form a 25-dimensional action space. We included a subset of the available features, see Table 8 for details. We represented the previous action in terms of the actual doses of vasopressors and IV fluids in the previous 4-h period.

To highlight differences between models and state representations, we stratified the Sepsis results by patient subgroups according to the average rate of change of the National Early Warning Score 2 (NEWS2) score (Inada-Kim and Nsutebu, 2018). Following Luo et al. (2024), we used the following inter-

4. The sum of boxes of the clinical dementia rating (CDR-SB) is a measure of dementia.

Table 5: A summary of the variables included in the context vector X_t in the ADNI experiments and their baseline statistics. We applied history aggregation and history truncation only to the variables in the lower section of the table. For each variable, N represents the number of patients with non-missing baseline information.

Variable	N	Statistics
Age in years, median (IQR)	1605	73.9 (69.3, 78.8)
Education level, median (IQR)	1605	16 (14, 18)
Female, n (%)	1605	715 (44.5)
Marital status, n (%)	1605	
Married		1217 (75.8)
Divorced		134 (8.3)
Widowed		191 (11.9)
Never married		56 (3.5)
Unknown		7 (0.4)
APOE4 alleles, n (%)	1605	
0		840 (52.3)
1		600 (37.4)
2		165 (10.3)
CDR-SB, n (%)	1605	
Normal		453 (28.2)
Questionable		985 (61.4)
Severe		167 (10.4)

Table 6: A summary of variables included in the context vector X_t in the RA experiment and their baseline statistics. We applied neither history aggregation nor history truncation to these variables. The other context variables, for which we applied these operations, are listed in Table 7. Baseline refers to the second visit, i.e., the first visit for which we could determine the switch label. For each variable, N represents the number of patients with non-missing baseline information.

Variable	N	Statistics
Age in years, median (IQR)	4379	59 (50, 67)
Calendar year, median (IQR)	4391	2016 (2013, 2019)
RA duration in years, median (IQR)	4338	3 (1, 9)
Female, n (%)	4383	3355 (76.5)
College completed, n (%)	3997	1507 (37.7)
Work status, n (%)	4291	
Full time		1784 (41.6)
Part time		359 (8.4)
Work at home		363 (8.5)
Student		68 (1.6)
Disabled		450 (10.5)
Retired		1267 (29.5)
Private insurance, n (%)	4391	3128 (71.2)
Medicare insurance, n (%)	4391	1412 (32.2)
Medicaid insurance, n (%)	4391	278 (6.3)
No insurance, n (%)	4391	95 (2.2)

vals for the average rate of change of the NEWS2 score (x): $x < -0.4$; $-0.4 \leq x < -0.15$; $-0.15 \leq x < 0$; $0 \leq x < 0.15$, $0.15 \leq x < 0.4$, and $x > 0.4$. In contrast to Luo et al. (2024), we did not consider the

Table 7: A summary of variables included in the context vector X_t in the RA experiment and their baseline statistics. We applied both history aggregation and history truncation to these variables. The other context variables, for which we did not apply these operations, are listed in Table 6. Baseline refers to the second visit, i.e., the first visit for which we could determine the switch label. For each variable, N represents the number of patients with non-missing baseline information.

Variable	N	Statistics
BMI, n (%)	4304	
Underweight		49 (1.1)
Healthy		965 (22.4)
Overweight		1313 (30.5)
Obesity		1977 (45.9)
Blood pressure, n (%)	4321	
Normal		1073 (24.8)
Elevated		660 (15.3)
Hypertension stage 1		1407 (32.6)
Hypertension stage 2		1154 (26.7)
Hypertension stage 3		27 (0.6)
CDAI, n (%)	4308	
Remission		489 (11.4)
Low		1025 (23.8)
Moderate		1405 (32.6)
High		1389 (32.2)
Smoker, n (%)	3757	644 (17.1)
Drinker, n (%)	4281	1934 (45.2)
Currently pregnant, n (%)	2982	4 (0.1)
Pregnant since last visit, n (%)	2418	13 (0.5)
CCP positive, n (%)	721	407 (56.4)
RF positive, n (%)	791	499 (63.1)
PPD positive, n (%)	780	45 (5.8)
Erosive disease, n (%)	3403	253 (7.4)
Joint space narrowing, n (%)	985	541 (54.9)
Joint deformity , n (%)	965	154 (16.0)
Comorbidities, n (%)		
Severe infections	4391	66 (1.5)
Metabolic diseases	4390	312 (7.1)
Cardiovascular diseases	4390	486 (11.1)
Respiratory diseases	4353	112 (2.6)
Cancer	4390	125 (2.8)
GI and liver diseases	4390	60 (1.4)
Musculoskeletal disorders	4340	1296 (29.9)
Other diseases	4390	516 (11.8)

variance in x to further separate the patients within each group.

A.4. Chronic Obstructive Pulmonary Disease

To extract the COPD dataset, we executed the `main` script in the MIMIC-IV pipeline provided by Gupta et al. (2022), focusing on ICU patients diagnosed with COPD as the chronic disease of interest, totaling 8,535 patients. Initially, we incorporated data on diagnoses, procedures, medications, outputs, and chart events. Subsequently, we conducted a clinical grouping of diagnoses based on their medical codes to streamline the feature space. In addition, ICD-9 codes were mapped to ICD-10 codes in cases where both coding systems were used. The preprocessing of the COPD dataset further entailed the cleaning of lab and chart events through outlier removal and

Table 8: A summary of the variables included in the context vector X_t in the sepsis experiment and their baseline statistics. We applied history aggregation and history truncation only to the variables in the lower section of the table. For each variable, N represents the number of patients with non-missing baseline information.

Variable	N	Statistics
Age in years, median (IQR)	20932	66.1 (53.7, 77.9)
Female, n (%)	20932	9250 (44.2)
Heart rate, median (IQR)	20932	87.2 (75.7, 99.8)
SysBP, median (IQR)	20932	118.2 (105.4, 133.8)
DiaBP, median (IQR)	20932	56.8 (48.2, 66.0)
MeanBP, median (IQR)	20932	77.2 (69.0, 87.2)
Shock index, median (IQR)	20932	0.7 (0.6, 0.9)
Hemoglobin, median (IQR)	20932	10.5 (9.3, 12.1)
Blood urea nitrogen, median (IQR)	20932	22.9 (15.0, 37.3)
Creatinine, median (IQR)	20932	1.0 (0.8, 1.6)
Total urine output, median (IQR)	20932	0.0 (0.0, 250.0)
Base excess, median (IQR)	20932	0.0 (−2.2, 3.0)
Lactate, median (IQR)	20932	1.7 (1.2, 2.5)
pH, median (IQR)	20932	7.4 (7.3, 7.4)
HCO3, median (IQR)	20932	24.3 (21.0, 27.4)
PaO_2/FiO_2 ratio, median (IQR)	20932	265.0 (173.3, 417.5)
Elixhauser, median (IQR)	20932	4.0 (2.0, 5.0)
SOFA, median (IQR)	20932	7.0 (5.0, 9.0)

unit conversion, resulting in a final dataset of 7,977 patients. Outlier removal aimed to eliminate values that exceeded the 0.75 percentile threshold and those that fell below the 0.25 percentile threshold across all values for each time point.

We formatted the data collected over 72 hours as multidimensional time series with a discrete time step of 4 hours, resulting in 18 observations for each patient. We investigated the behavior policy for managing IV fluids (mainly Dextrose 5% and NaCl 0.9%) and sedative drugs, combined into 25 discrete actions. Sedative drugs for induction and maintenance of general anesthesia included propofol and fentanyl concentrate. The previous action was represented through the actual doses of IV fluids and sedatives in the previous 4-h period. Missing values were zero-imputed. In cases where values were recorded at different times, forward and backward imputation were applied to minimize bias. We removed columns that had more than 80 % missingness. The final cohort comprised demographic data and chart events, representing records on vital signs such as blood pressure, heart rate, respiratory rate, and body temperature. We included 40 variables in the context vector X_t, as detailed in Table 9 and 10.

Appendix B. Experimental Details

The results presented in Table 3 were obtained by, for each dataset, state representation, and model class, averaging the AUROC of the *best performing* model across five different splits of the dataset. Specifically, for each such setting, five candidate models were trained using hyperparameters randomly sampled from predefined distributions, and the best performing model was selected as the one with the highest AUROC (ADNI and RA) or accuracy (Sepsis and COPD) on a held-out validation set comprising 20 % of the training data. The hyperparameter distributions for each model are shown in Table 11 and 12. Models based on the contextualized policy recovery framework are not included in the table as we trained them using the code provided by Deuschel et al. (2024), using their hyperparameter ranges. We refer to their work for details.

The neural networks, including the prototype-based models and the recurrent decision tree, were implemented using PyTorch (Paszke et al., 2019) in combination with skorch (Tietz et al., 2017) and trained on Nvidia Tesla T4 GPUs. Model parameters were optimized using the cross-entropy loss and

Table 9: A summary of the variables included in the context vector X_t in the COPD experiments and their baseline statistics. We applied neither history aggregation nor history truncation to these variables. The other context variables, for which we applied these operations, are listed in Table 10. For each variable, N represents the number of patients with non-missing baseline information.

Variable	N	Statistics
Age in years, median (IQR)	7977	67.0 (56.0, 77.0)
Female, n (%)	7977	3472 (43.52)
Ethnicity (self-reported), n (%)	7977	
White		5417 (67.91)
Black/African American		747 (9.36)
Hispanic/Latino		242 (3.03)
Asian		203 (2.54)
American Indian/Alaska Native		16 (0.20)
Other		375 (4.70)
Unknown/Unable to obtain		977 (12.24)
Health insurance, n (%)	7977	
Medicare		4066 (50.97)
Medicaid		522 (6.54)
Other		3389 (42.48)

the Adam optimizer with default parameters. ReLU and hyperbolic tangent were used as activation function in feedforward and recurrent neural networks, respectively. Early stopping was applied to the training if there was no improvement in performance for 5 (ADNI and RA) or 25 (Sepsis and COPD) consecutive epochs. The logistic regression and the decision tree classifier were implemented using the scikit-learn library (Pedregosa et al., 2011). For the scoring system, we used the implementation provided in Ustun and Rudin (2019). The computational time required to produce the results presented in this paper was approximately 2000 core-hours.

For the prototype-based models, three regularization terms, encouraging diversity, clustering, and evidence, were added to the loss function, see Ming et al. (2019) for details. We set the parameters for clustering and evidence regularization to 0.001 and sampled parameter values for diversity regularization, see Table 11. Following Matsson and Johansson (2022), each learned prototype i was a subsequence of length $t_i \leq T$ of a patient sequence in the training data. Prototype projections, see Equation (6) in Matsson and Johansson (2022), were performed every fifth epoch. The recurrent decision tree was implemented as in Pace et al. (2022), using the predictive distribution from the leaf with the greatest path probability. We found that the post-processing steps in Pace et al. (2022) drastically reduced the performance of the model; hence, we evaluated the recurrent decision trees without applying any post-processing steps.

Figure 4 shows how the performance of decision trees varies with their complexity, measured by the number of leaves, for different state representations in RA. Since we could not control the number of leaves directly, we trained 500 different models for each state representation, using randomly selected hyperparameters. We then binned the models based on their complexity and selected the best-performing model in each "complexity bucket" (e.g., 10–20 leaves) to present in the figure. We only performed this experiment for a single split of the data.

Appendix C. Supplementary Results

In Table 15, we show the average test calibration error for each model and dataset, using the different state representations described in Section 3.3. We consider the expected calibration error (ECE) and the static calibration error (SCE) (Nixon et al., 2019), a multiclass extension of ECE. Specifically, we report ECE for ADNI and SCE for RA, Sepsis, and COPD. In Table 16, we report 95 % confidence intervals, based on the bootstrap distribution of the average AUROC, for the results in Table 3. Table 13 and Table 14 show the average test AUROC with history aggregation performed using the `max` and `mean` operator, respectively.

Table 10: A summary of the variables included in the context vector X_t in the COPD experiment and their baseline statistics. We applied both history aggregation and history truncation to these variables. The other context variables, for which we did not apply these operations, are listed in Table 9. For each variable, N represents the number of patients with non-missing baseline information. We present the original naming of the features from the MIMIC database for reproducibility.

Variable	N	Statistics
CHART 220045, median (IQR)	7958	84.5 (72.6, 97.3)
CHART 220046, median (IQR)	7954	120.0 (30.0, 130.0)
CHART 220047, median (IQR)	7956	50.0 (12.5, 30.0)
CHART 220179, median (IQR)	7629	108.5 (91.3,125.0)
CHART 220180, median (IQR)	7627	58.0 (46.0, 68.5)
CHART 220181, median (IQR)	7631	71.0 (59.0, 82.0)
CHART 220210, median (IQR)	7957	19.3 (16.3, 22.5)
CHART 220228, median (IQR)	7899	10.5 (0.0, 10.5)
CHART 220277, median (IQR)	7959	96.8 (94.8, 98.5)
CHART 220545, median (IQR)	7899	25.7 (0.0, 31.6)
CHART 220546, median (IQR)	7949	7.5 (0.0, 12.3)
CHART 220602, median (IQR)	7955	99.0 (21.0, 106.0)
CHART 220615, median (IQR)	7930	0.7 (0.0, 1.2)
CHART 220621, median (IQR)	7971	101.0 (0.0, 136.0)
CHART 220635, median (IQR)	7946	1.8 (0.0, 2.1)
CHART 220645, median (IQR)	7964	135.0 (31.5, 140.0)
CHART 223751, median (IQR)	7558	160.0 (0.0, 140.0)
CHART 223752, median (IQR)	7552	90.0 (0.0, 90.0)
CHART 223769, median (IQR)	7956	100.0 (25.0, 100.0)
CHART 223770, median (IQR)	7956	90.0 (92.0, 22.5)
CHART 224161, median (IQR)	7957	30.0 (7.5, 35.0)
CHART 224162, median (IQR)	7955	8.0 (2.0, 8.0)
CHART 225624, median (IQR)	7935	13.0 (0.0, 26.0)
CHART 225625, median (IQR)	7939	7.8 (0.0, 8.5)
CHART 225677, median (IQR)	7941	2.6 (0.0, 3.7)
CHART 226253, median (IQR)	7947	85.0 (21.3, 86.5)
CHART 227073, median (IQR)	7969	11.0 (0.0, 15.0)
CHART 227442, median (IQR)	7971	3.7 (0.8, 4.2)
CHART 227443, median (IQR)	7941	20.0 (0.0, 24.0)
CHART 227457, median (IQR)	7942	110.0 (0.0, 204.0)
CHART 227465, median (IQR)	7573	12.0 (0.0, 14.6)
CHART 227466, median (IQR)	7563	26.3 (0.0, 33.2)
CHART 223761, median (IQR)	7722	97.7 (24.6, 101.5)

Table 11: Experiment-independent hyperparameters and their respective search space for all models.

Model	Hyperparameter	Search space
RS	max coefficient value	$\{3, 4, 5, 6, 7, 8\}$
	max model size	$\{3, 4, 5, 6, 7\}$
	positive class weight	$\{1, 2, 3, 4, 5\}$
LR	penalty	{L2}
	C	$\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$
	max iterations	{2000}
DT	criterion	{gini, entropy}
	min samples to split	$\{2, 4, 8, 16, 32, 64, 128\}$
MLP	hidden dimensions (encoder)	$\{(16,), (32,), (64,), (16, 16), (32, 32), (64, 64)\}$
	output dimensions (encoder)	$\{16, 32, 64\}$
PSN	prototype threshold d_{min}	$\{1, 2, 3, 4, 5\}$
	diversity regularization λ_d	$\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}\}$
	output dimensions (encoder)	$\{16, 32, 64\}$
	number of layers (encoder)	$\{1, 2\}$
RDT	initial depth	$\{1, 2\}$
	splitting penalty λ	$\{-3, -2, -1\}$
	max depth	$\{3, 4, 5\}$
	evolution prediction δ_1	$\{10^{-3}, 10^{-2}, 10^{-1}\}$
	evolution prediction δ_2	$\{10^{-3}, 10^{-2}, 10^{-1}\}$
	history dimension	$\{5, 10, 15, 20\}$
RNN	output dimensions (encoder)	$\{16, 32, 64\}$
	number of layers (encoder)	$\{1, 2\}$

Table 12: Experiment-dependent hyperparameters of DT and the neural network-based models, along with their search space on different datasets.

Model	Hyperparameter	ADNI	RA	Sepsis/COPD
DT	max depth	$\{3, 5, 7, 9, 11, 13, 15\}$	$\{2, 3, 4, 5, 6, 7, 8\}$	$\{3, 5, 7, 9, 11, 13, 15\}$
MLP	learning rate	$\{10^{-3}, 10^{-2}\}$	$\{10^{-3}, 10^{-2}\}$	$\{10^{-4}, 10^{-3}, 10^{-2}\}$
	max epochs	{20}	{50}	{500}
	batch size	$\{16, 32, 64\}$	$\{128, 256\}$	$\{256, 512, 1024\}$
PSN	learning rate	$\{10^{-3}, 10^{-2}\}$	$\{10^{-3}, 10^{-2}\}$	$\{10^{-4}, 10^{-3}, 10^{-2}\}$
	max epochs	{20}	{50}	{500}
	batch size	$\{16, 32, 64\}$	$\{32, 64\}$	$\{16, 32, 64\}$
	number of prototypes	$\{2, 4, 6, 8, 10\}$	$\{2, 4, 6, 8, 10\}$	$\{5, 10, 15, 20, 25, 30\}$
RDT	learning rate	$\{10^{-3}, 10^{-2}\}$	$\{10^{-3}, 10^{-2}\}$	$\{10^{-4}, 10^{-3}, 10^{-2}\}$
	max epochs	{20}	{50}	{500}
	batch size	$\{16, 32, 64\}$	$\{32, 64\}$	$\{16, 32, 64\}$
RNN	learning rate	$\{10^{-3}, 10^{-2}\}$	$\{10^{-3}, 10^{-2}\}$	$\{10^{-4}, 10^{-3}, 10^{-2}\}$
	max epochs	{20}	{50}	{500}
	batch size	$\{16, 32, 64\}$	$\{32, 64\}$	$\{16, 32, 64\}$

Table 13: Average test AUROC, expressed as a percentage, for different state representations and behavior policy models in ADNI, RA, Sepsis, and COPD. History aggregation is performed using the `mean` operator.

	ADNI				**RA**			**Sepsis**			**COPD**		
State	RS	LR	DT	MLP	LR	DT	MLP	LR	DT	MLP	LR	DT	MLP
$\bar{H}_t$	54.5	56.2	57.9	56.5	91.2	91.2	92.7	86.5	84.2	89.3	91.1	93.2	93.2
$H_{(0)}, \bar{H}_t$	53.9	56.5	58.2	56.9	95.7	96.0	96.3	91.9	92.2	95.3	94.2	96.3	96.1
$H_{(1)}, \bar{H}_t$	55.6	57.3	57.6	59.4	95.7	96.0	96.3	92.1	92.5	95.5	94.2	96.4	96.1
$H_{(2)}, \bar{H}_t$	54.0	57.5	57.8	59.9	95.8	96.0	96.2	92.2	92.2	95.5	94.1	96.4	96.0

Table 14: Average test AUROC, expressed as a percentage, for different state representations and behavior policy models in ADNI, RA, Sepsis, and COPD. History aggregation is performed using the `max` operator.

	ADNI				**RA**			**Sepsis**			**COPD**		
State	RS	LR	DT	MLP	LR	DT	MLP	LR	DT	MLP	LR	DT	MLP
$\bar{H}_t$	54.1	55.6	54.6	55.6	90.8	89.8	91.9	84.3	83.5	88.7	90.3	92.6	92.9
$H_{(0)}, \bar{H}_t$	54.9	57.2	54.4	57.2	95.9	96.0	96.6	91.9	92.8	95.2	94.8	96.3	96.5
$H_{(1)}, \bar{H}_t$	55.1	58.3	58.2	59.5	95.9	95.9	96.6	92.3	92.1	95.5	94.8	96.4	96.6
$H_{(2)}, \bar{H}_t$	55.8	58.1	58.0	60.4	95.9	96.0	96.6	92.4	92.3	95.6	94.7	96.5	96.5

Table 15: Average test calibration error, expressed as a percentage, for different state representations and behavior policy models in ADNI, RA, Sepsis, and COPD. History aggregation is performed using the `sum` operator. We report ECE for ADNI and SCE for RA, Sepsis, and COPD.

	ADNI				**RA**			**Sepsis**			**COPD**		
State	RS	LR	DT	MLP	LR	DT	MLP	LR	DT	MLP	LR	DT	MLP
X_t	8.6	1.9	3.5	2.2	1.1	1.3	1.6	0.2	0.3	0.2	0.4	0.4	0.5
A_{t-1}	8.8	1.8	2.0	3.3	0.4	0.4	0.7	1.0	0.1	0.2	0.9	0.2	0.3
$H_{(0)}$	8.2	2.3	4.0	2.8	0.6	0.4	0.8	0.6	0.1	0.2	0.6	0.2	0.3
$\bar{H}_t$	6.5	2.8	2.7	3.9	3.5	1.3	1.1	0.4	0.2	0.2	0.6	0.5	0.4
$H_{(0)}, \bar{H}_t$	10.1	2.6	2.6	3.6	0.9	0.5	0.9	0.6	0.1	0.2	0.5	0.2	0.4
$H_{(1)}, \bar{H}_t$	7.6	2.7	3.3	2.9	1.0	0.5	0.8	0.5	0.1	0.2	0.5	0.2	0.4
$H_{(2)}, \bar{H}_t$	7.9	2.7	3.1	2.4	1.0	0.5	1.0	0.5	0.1	0.2	0.5	0.2	0.4

	ADNI				**RA**			**Sepsis**			**COPD**		
State	CPR	PSN	RDT	RNN	PSN	RDT	RNN	PSN	RDT	RNN	PSN	RDT	RNN
H_t	2.2	3.8	2.4	4.4	1.1	4.0	0.8	0.5	0.9	0.2	0.6	0.3	0.4

Table 16: Average test AUROC, expressed as a percentage, for the different states and behavior policy models in ADNI, RA, Sepsis, and COPD. History aggregation is performed using the `sum` operator. The 95 % confidence intervals are based on the bootstrap distribution of the average AUROC.

Data	State	RS	LR	DT	MLP	PSN	RDT	RNN
ADNI	X_t	54.2 (51.4, 57.0)	56.2 (54.4, 57.4)	53.9 (52.1, 55.9)	55.6 (54.1, 57.0)	-	-	-
	A_{t-1}	52.0 (51.1, 53.0)	53.9 (52.3, 55.6)	53.8 (52.4, 55.4)	53.7 (52.5, 55.4)	-	-	-
	$H_{(0)}$	53.4 (51.1, 55.7)	56.8 (55.2, 58.1)	54.3 (52.7, 56.1)	56.8 (54.9, 58.8)	-	-	-
	$\bar{H}_t$	63.0 (61.3, 64.7)	64.4 (62.8, 65.9)	64.9 (63.1, 66.7)	64.1 (62.2, 65.3)	-	-	-
	$H_{(0)}, \bar{H}_t$	63.7 (62.9, 64.7)	65.3 (64.1, 66.4)	65.0 (63.1, 66.8)	65.8 (65.1, 66.5)	-	-	-
	$H_{(1)}, \bar{H}_t$	63.4 (62.2, 64.6)	65.6 (64.4, 66.8)	65.4 (63.6, 67.1)	66.0 (65.3, 67.0)	-	-	-
	$H_{(2)}, \bar{H}_t$	62.9 (61.7, 64.2)	65.4 (64.2, 66.7)	65.3 (63.6, 66.9)	66.8 (66.1, 67.6)	-	-	-
	H_t	-	-	-	-	66.7 (65.6, 67.8)	62.8 (61.0, 64.7)	68.0 (67.2, 68.9)
RA	X_t	-	61.7 (61.2, 62.4)	58.8 (58.1, 59.5)	61.1 (60.0, 62.1)	-	-	-
	A_{t-1}	-	94.7 (94.4, 94.9)	94.7 (94.4, 94.9)	94.7 (94.4, 94.9)	-	-	-
	$H_{(0)}$	-	95.6 (95.4, 95.7)	95.7 (95.5, 95.9)	96.1 (95.9, 96.2)	-	-	-
	$\bar{H}_t$	-	90.5 (90.1, 90.9)	92.0 (91.4, 92.9)	94.0 (93.9, 94.2)	-	-	-
	$H_{(0)}, \bar{H}_t$	-	96.1 (95.9, 96.2)	96.5 (96.3, 96.6)	96.9 (96.7, 97.0)	-	-	-
	$H_{(1)}, \bar{H}_t$	-	96.0 (95.9, 96.1)	96.4 (96.2, 96.6)	96.9 (96.7, 97.0)	-	-	-
	$H_{(2)}, \bar{H}_t$	-	96.0 (95.8, 96.1)	96.4 (96.1, 96.6)	96.7 (96.6, 96.9)	-	-	-
	H_t	-	-	-	-	96.2 (96.1, 96.4)	90.0 (85.9, 94.1)	96.8 (96.7, 97.0)
Sepsis	X_t	-	82.1 (81.8, 82.4)	78.2 (77.3, 78.9)	84.1 (83.9, 84.3)	-	-	-
	A_{t-1}	-	88.0 (87.9, 88.1)	90.6 (90.1, 91.0)	91.1 (91.0, 91.2)	-	-	-
	$H_{(0)}$	-	91.3 (91.2, 91.4)	92.1 (90.9, 92.9)	94.7 (94.6, 94.8)	-	-	-
	$\bar{H}_t$	-	84.6 (84.4, 84.8)	85.2 (84.4, 86.0)	89.1 (89.0, 89.2)	-	-	-
	$H_{(0)}, \bar{H}_t$	-	91.9 (91.8, 92.0)	92.3 (91.4, 92.9)	95.3 (95.2, 95.3)	-	-	-
	$H_{(1)}, \bar{H}_t$	-	92.2 (92.1, 92.3)	92.5 (92.2, 92.9)	95.5 (95.4, 95.5)	-	-	-
	$H_{(2)}, \bar{H}_t$	-	92.3 (92.2, 92.4)	92.6 (92.2, 92.9)	95.5 (95.4, 95.6)	-	-	-
	H_t	-	-	-	-	94.9 (94.7, 95.1)	77.0 (69.5, 85.5)	95.7 (95.7, 95.8)
COPD	X_t	-	77.9 (77.4, 78.3)	74.7 (72.9, 75.9)	78.7 (78.1, 79.4)	-	-	-
	A_{t-1}	-	92.9 (92.5, 93.2)	95.0 (94.6, 95.2)	94.9 (94.9, 95.0)	-	-	-
	$H_{(0)}$	-	94.0 (93.8, 94.2)	96.0 (95.7, 96.3)	95.5 (95.2, 95.7)	-	-	-
	$\bar{H}_t$	-	91.1 (90.8, 91.4)	89.3 (86.7, 91.4)	93.6 (93.2, 93.9)	-	-	-
	$H_{(0)}, \bar{H}_t$	-	94.7 (94.6, 94.9)	96.7 (96.2, 97.0)	96.3 (96.1, 96.5)	-	-	-
	$H_{(1)}, \bar{H}_t$	-	94.7 (94.5, 94.9)	96.8 (96.4, 97.0)	96.4 (96.2, 96.6)	-	-	-
	$H_{(2)}, \bar{H}_t$	-	94.7 (94.5, 94.8)	96.8 (96.4, 97.0)	96.3 (96.1, 96.5)	-	-	-
	H_t	-	-	-	-	96.2 (96.0, 96.5)	81.9 (76.1, 86.2)	96.5 (96.3, 96.6)