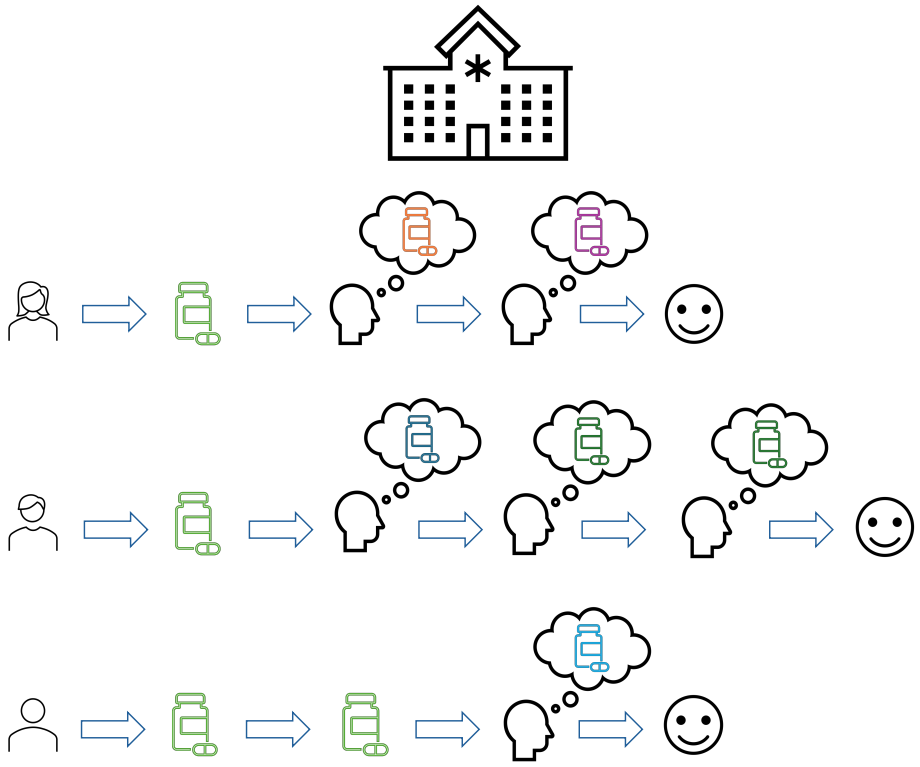
CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG



Policy Learning and Off-Policy Evaluation with Application to the Sequential Treatment of Rheumatoid Arthritis

Master's Thesis in Computer Science and Engineering

Yaochen Rao, Jinming Wei

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2024

Master's thesis 2024

Policy Learning and Off-Policy Evaluation with Application to the Sequential Treatment of Rheumatoid Arthritis

Yaochen Rao, Jinming Wei

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2024

Policy Learning and Off-Policy Evaluation with Application to the Sequential Treatment of Rheumatoid Arthritis
Yaochen Rao, Jinming Wei

Supervisor: Anton Matsson, Computer Science and Engineering
Advisor: Fredrik Johansson, Computer Science and Engineering
Examiner: Marina Axelson-Fisk, Mathematical Sciences

Master's Thesis 2024
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Description of the picture on the cover page (if applicable)

Typeset in LaTeX
Gothenburg, Sweden 2024

Policy Learning and Off-Policy Evaluation with Application to the Sequential Treatment of Rheumatoid Arthritis

Yaochen Rao and Jinming Wei
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

This thesis addresses the challenge of optimizing treatment strategies for patients with Rheumatoid Arthritis (RA) after switch to second-line biologic or targeted synthetic disease-modifying antirheumatic drugs (b/tsDMARDs) therapy. Given the complexity of RA treatment and the limitations of randomized controlled trials (RCTs) in real-world settings, we leverage historical data from the CorEvitas RA registry to model and evaluate alternative treatment strategies. We employ interpretable machine learning models to capture existing treatment decision patterns and propose a target policy tailored to individual patient needs. Our approach integrates features derived from patient history, treatment patterns, and clinical characteristics to enhance prediction accuracy. We introduce a novel two-stage combined model that first predicts whether a patient will switch therapies and then determines the most appropriate subsequent therapy. The proposed target policy is compared against current guidelines, such as those provided by the European Alliance of Associations for Rheumatology (EULAR), to identify potential improvements. Finally, we validate the proposed policy through off-policy evaluation, utilizing techniques such as importance sampling and weighted importance sampling to assess its effectiveness in real-world scenarios. The findings suggest that the proposed approach can lead to more personalized and effective treatment plans, potentially improving patient outcomes in RA management.

Keywords: machine learning, interpretable machine learning, policy learning, off-policy evaluation, sequential decision making, rheumatoid arthritis, treatment strategy optimization.

Acknowledgements

Throughout the process of completing this masters thesis, we have received a great deal of help and support from many people, and we would like to take this opportunity to express our most sincere gratitude to everyone who has assisted and guided us along the way.

First and foremost, we would like to express our sincere appreciation to our supervisors, Anton Matsson and Fredrik Johansson. Throughout the research process, they have provided us with endless patience and guidance. From the initial conception of the research idea to the final completion of the thesis, they always encouraged us to explore new ideas with an open mind and provided valuable advice and feedback. Without their help, we would not have been able to complete our thesis as smoothly as we did. We are deeply thankful for their willingness to tirelessly answer our questions and guide us out of confusion when we felt lost.

We would also like to thank CorEvitas for generously providing the valuable data used in this research. Without this data, our study would not have been possible.

Finally, we would like to thank our families and friends for their endless support and encouragement throughout our pursuit of a master's degree. We would also like to thank our fellow students for their many valuable ideas and suggestions throughout this journey.

Once again, we would like to thank everyone who has helped and supported us. Your encouragement and guidance have been so valuable.

Yaochen Rao and Jinming Wei, 2024-08-19

Contents

List of Figures

List of Tables

1

Introduction

Rheumatoid Arthritis (RA) is a chronic disease that causes inflammation around the body, with common symptoms including joint pain, stiffness, tenderness, fever, and swelling. These symptoms can significantly impact a patient's mobility and quality of daily life. Data from 1980 to 2018 indicate that the global prevalence of rheumatoid arthritis is approximately 0.46%, meaning that 460 out of every 100,000 people suffer from this disease [1]. The treatment process for RA is often complex and prolonged, generally requiring lifelong therapy [2].

For chronic diseases like RA, finding the best treatment for a patient can take several rounds of trial and error. This is because the number of possible therapies is large, and what predicts a good response to these therapies is poorly understood [3], [4]. Although most people with RA eventually find a treatment that works for them, not everyone does. Many patients spend several months trying medications that do not work [4]. Given the incurable nature of RA, treatment goals have shifted from curing the disease to alleviating it and reducing symptoms [5]. Therefore, it is critical to find an effective therapy that can control the disease as early as possible.

In the treatment of RA, disease-modifying anti-rheumatic drugs (DMARDs) are used to slow the disease's progression by reducing inflammation. Conventional synthetic DMARDs (csDMARDs) form the basis of RA treatment. With a better understanding of the pathophysiology of RA, new therapeutic approaches such as biologic DMARDs (bDMARDs) and targeted synthetic DMARDs (tsDMARDs) are emerging to provide precision medicine for individuals [3]. Although various clinical guidelines recommend the initial prescription of csDMARDs as a first-line treatment strategy [6], [7], there is still no consensus on how to choose between bDMARDs and tsDMARDs. This uncertainty is largely due to variations in patient response to treatment, differences in drug mechanisms, and considerations of potential side effects and cost [8], [9]. As the pharmaceutical market continues to expand, so does the complexity of choosing the right treatment for RA patients. An increasing number of available medications provides more options, but also requires more careful decision-making to effectively tailor treatment plans.

Physicians devise treatment plans for RA patients based on factors such as current state of the disease, medical history, preferences, insurance coverage, and clinical guidelines. This sequential decision-making process is dynamic and patient-centered, and each decision can influence the outcome of future treatments. Physicians must continually adjust the treatment plan based on the patient's response to treatment,

and the effectiveness of the entire process depends on making the proper decisions. The resulting treatment patterns, which represent the sequences of therapies and adjustments made over time in response to the patient's evolving condition, are known as the *behavior policy* [10].

Although we can estimate the effectiveness of current behavior policies, estimating alternative treatment plans poses significant challenges. Common evaluation methods often rely on randomized controlled trials (RCTs) [11]. An RCT is a study design that can be used to assess the effects of a single treatment choice or multiple treatment choices. In an RCT that focuses on a single treatment choice, participants are randomly assigned to either the treatment group or the control group. When conducting RCTs that involve comprehensive randomization of a series of treatment choices, multiple treatment options and sequences result in a vast variety of treatment combinations. In our work, with a large number of RA patients and highly diverse treatment combinations, the likelihood that any two patients will receive the same treatment sequence is very low. Therefore, RCTs is not feasible in our situation. This brings us to the importance of observational studies, which allows us to use existing treatment patterns to estimate the outcomes of a different treatment patterns.

Given the pressing need for effective treatment strategies in RA patients, coupled with the limitations of using RCTs in real-world settings, our work uses historical data from CorEvitas RA registry [12] to evaluate alternative treatment strategy, that is, *target policy*. We use interpretable machine learning models to model the behavior policy, including both established interpretable machine learning models and our developed interpretable two-stage combined model. Based on the modeled behavior policy, we propose the target policy, ensuring that the new policy is supported by the data. Interpretable machine learning models can support domain experts in explaining clinical decisions and help us validate the results [13]. Moreover, interpretability allows us to check the quality of the model fit and identify potentially missing confounding variables—variables that affect both treatment decisions and outcomes [14]. We formulate the target policy by choosing the therapy with the highest probability according to the modeled behavior policy, which represent proposed new clinical guidelines. This target policy is then evaluated and compared with the existing guideline of the European Alliance of Associations for Rheumatology (EULAR, 2022) [6]. Lastly, we employ off-policy evaluation methods to assess the effectiveness of this targeted policy, thereby validating the efficacy and feasibility of our proposed treatment strategy in practice.

1.1 Objective

The overall objective of this project is to provide suggestions for subsequent treatment paths for RA patients after first-line therapy with csDMARDs. Based on this overall objective, this project can be divided into three sub-objectives: modeling behavioral policy, proposing target policy, and performing off-policy evaluation.

To model the behavior policy, we use interpretable machine learning models. In

this part, we consider two approaches to predicting the treatment chosen by the physician given the patient's data. This treatment may be the same as before, since patients do not always switch therapies. The first approach is to directly use a machine learning model to predict which therapy a patient will switch to. The second approach adopts a "divide and conquer" strategy, breaking down the complex task of predicting therapy into two subproblems: predicting whether a patient will switch therapy, and if so, predicting which therapy the patient will switch to. We will answer the following questions:

- *Q1: If the patient switches therapy, what is the reason for the switch?*
- *Q2: If the patient switches therapy, to which therapy does the patient switch?*
- *Q3: Which of the two methods is better for predicting therapy in RA patients?*

We propose a target policy based on the behavior policy by selecting the most common treatment. This target policy address areas of ambiguity or lack of clarity in existing RA treatment guidelines, particularly focusing on how to select b/tsDMARDs for different patients after the first-line therapy csDMARDs. By leveraging insights from our interpretable machine learning models, we can provide more tailored suggestions to improve the comprehensiveness and applicability of the guidelines in clinical practice. To achieve this, we focus on the following questions:

- *Q4: How would individual patients be treated using the proposed policy?*
- *Q5: How does our proposed policy differ from existing guidelines?*

To validate the effectiveness of the proposed target policy, we performed off-policy evaluation using the behavior policy derived from observational data to evaluate a different target policy. In this part, we use both importance sampling (IS) and weighted importance sampling (WIS) methods for estimation [15]. We will address the following questions:

- *Q6: How does the proposed target policy impact patient outcomes compared to the current behavior policy?*

1.2 Related Work

Interpretable Machine Learning Interpretable machine learning models are designed to make the reasoning process understandable to humans, which is crucial for high-stakes decisions and troubleshooting. This transparency allows users to see and understand how decisions are made, which is essential for trust and reliability. Interpretable models can be applied to tabular data, providing clear and understandable rules or equations to explain their predictions [16]. In this project, we consider two classes of interpretable models: rule-based models and linear models. Rule-based models, including decision trees, decision lists, and decision sets, use "if-the" statements and logical operations (such as "or" and "and") to make predictions, providing understandable reasons for each decision [17]–[19]. These models are robust in handling missing data and outliers. Linear models include linear regression and scoring systems; the former predicts continuous variables through a linear

combination of input features, while the latter uses simple arithmetic operations for classification [20]. Both types of models are widely used for their simplicity and interpretability.

Interpretable Policy Learning In the field of interpretable policy learning, Silva et al. proposed a method using Differentiable Decision Trees (DDTs) in reinforcement learning, which allows optimization by gradient descent. This method overcomes the limitation of traditional decision trees' difficulty in updating within reinforcement learning, integrating policy gradient methods while maintaining model interpretability [21]. However, this approach assumes Markovianity. In recent advances in interpretable policy learning, there has been an increased focus on explaining individual patient trajectories. Hüyük et al. introduced INTERPOLE, a Bayesian method for interpretable policy learning. This approach parameterizes a latent belief space related to decisions, addresses the policy explanation problem, and provides a method to imitate agents' possibly suboptimal policies without assuming unbiased beliefs and objective reasoning [22]. However, this method assumes low-dimensional state spaces. Pace et al. (2022) used recurrent decision trees in their POETREE method to dynamically model behavior policies, adapting over time with aggregated patient data [13]. Although POETREE provides an interpretable model, it requires extensive post-processing to remove uninterpretable components, sacrificing performance, especially in high-dimensional observation spaces.

Machine Learning in RA Treatment Norgeot et al. used historical data with longitudinal deep learning models to predict the disease activity of RA patients at their next rheumatology clinic visit [23]. In addition, the study evaluated the performance differences between hospitals and the interpretability strategies of the models. The results indicate that it is feasible to build models using electronic health record data to predict outcomes of complex diseases and that these models can be shared across different patient populations in different hospitals.

Treatment Guidelines of RA Although EULAR and ACR provide guidelines for the treatment of RA, in some cases these recommendations remain vague. For example, in the EULAR guidelines proposed in 2022, it is suggested that bDMARDs should be added in the second phase of treatment if the patient has poor prognostic factors [6]. However, the guidelines do not provide detailed guidance on the selection of different bDMARDs. This lack of specificity results in physicians often having to rely on their personal experience to choose the appropriate bDMARDs [24].

Propose new treatment strategies in RA Research to propose new treatment strategies for RA typically focuses on different drugs within a specific treatment class. For example, Tao et al. studied the treatment response of RA patients to TNFi therapies (ADA or ETN) before starting treatment, with a focus on TNFi therapy [25]. Similarly, Plant et al. identified and adjusted treatment for patients who did not respond to methotrexate (MTX) at an earlier stage, focusing on csDMARDs therapy [26]. Gosselt et al. compared the performance of machine learning algorithms with traditional logistic regression in predicting insufficient response to

MTX treatment, another study focused on csDMARDs therapy; however, many of these models are not interpretable [27]. In the context of different treatment classes, Morid et al. predicted which patients would move from first-line csDMARDs to second-line b/tsDMARDs therapy, but this study did not distinguish between bDMARDs and tsDMARDs [28].

Off-Policy Evaluation in Healthcare Gottesman et al. introduced the significance of Importance Sampling methods in the evaluation of reinforcement learning algorithms, highlighting the challenges these methods face in handling sparse data and high variance estimates. By analyzing a specific case of sepsis management in the ICU, they demonstrated the performance of methods such as per-decision importance sampling (PDIS), weighted per-decision importance sampling (WPDIS), doubly-robust (DR), and weighted doubly-robust (WDR) in practical applications. They also proposed methods to improve the robustness of evaluations by examining the distribution of weights and using non-deterministic policies. These insights provide important references for the evaluation of reinforcement learning algorithms in healthcare decision-making [29]. Matsson et al. studied off-policy evaluation using Importance Sampling in situations where the behavior policy is unknown and must be estimated from data. They propose a novel approach using prototype learning to estimate the behavior policy, which better explains patterns in policy decisions and value estimates. Through experiments on sepsis management using data from the MIMIC-III database, they demonstrate that learned prototypes can describe differences between policies, assess support for evaluating a given target policy, and provide insights into policy evaluation despite introducing some approximation error [14].

1.3 Contributions

One of the contributions of our project is the creation of a unified framework for interpretable policy learning and off-policy evaluation. This framework includes the use of interpretable machine learning models to model behavior policy, propose target policy based on the modeled behavior policy, and evaluate the proposed target policy using off-policy evaluation.

In the area of interpretable policy learning, we compared two major classes of interpretable machine learning models, evaluating their performance in both binary (whether the RA patients will switch) and multi-class classification problems (which therapy the RA patients will switch to), as well as their interpretability. In addition, we proposed a novel modeling approach, i.e., using a two-stage combined model based on the "divide and conquer" principle, to address the problem of which therapy the RA patients will switch to. Our results show that this new modeling approach performs as well as the traditional approach, which directly uses machine learning models to predict the physician's choice of therapy. Our proposed combined model provides clearer, more useful clinical insights that help improve physician decision making.

In the field of RA treatment, our project provides a comprehensive description of treatment patterns for RA patients, including whether patients switch therapies and, if so, which therapy they switch to. We also explain the reasons behind therapy switches and the specific factors influencing the choice of next therapy. The proposed target policy offers recommendations for areas in the current EULAR guidelines that are ambiguous or unclear, and provides specific treatment plans tailored to RA patients.

1.4 Limitations

The primary limitation of our study come from the singularity of the dataset we used. Although the CorEvitas RA registry dataset is the largest real-world prospective RA study in the world, it is limited to patients from North America. This geographic limitation restricts our rheumatoid arthritis research to a North American population, thereby excluding diverse patient profiles from other parts of the world. Thus, it is uncertain whether the patient population enrolled in the study is representative of the broader RA patient population.

Although RA patients typically receive glucocorticoids as an adjunct to DMARDs, we did not to study this aspect. Furthermore, we are limited to classes of DMARDs rather than specific drugs. It is also possible that there are unidentified confounders in the dataset that could affect the values in the off-policy evaluation. Additionally, our dataset contains a certain degree of missing values, although we have employed imputation methods to fill in these missing value, this might still introduce some bias into our results. Finally, in our project, we did not consult domain experts for suggestions. We made clinical assessments based on the knowledge of RA provided by our supervisors and our own understanding of the disease.

1.5 Thesis Outline

The thesis is structured as follows.

Chapter 2 provides the theoretical background of the project. It begins with an introduction to the basics of rheumatoid arthritis, followed by principles of sequential decision making, particularly in the context of healthcare. Next, it covers interpretable policy learning, providing an overview of interpretable machine learning and ensemble learning models. Finally, it discusses off-policy evaluation, including the estimators used and their limitations.

Chapter 3 focuses on data aspects. It provides a detailed description of the dataset, including statistical analysis and missing values of selected features. It then introduces the therapy class used in this project. It also discusses the wash-out period and follow-up window criteria and how they are handled in our analysis.

Chapter 4 provides a detailed overview of the methods used in this thesis. First, we introduce two approaches we used to model behavior policy—one approach is to use our proposed two-stage combined model, another approach is to directly use

machine learning models to predict the physician's therapy choice. Next, we describe the models, data processing techniques, and experimental setup used in the project. Finally, we explain two methods for off-policy evaluation.

Chapter 5 presents the experimental results and analysis for two approaches to modeling behavior policy. We interpret and validate the results of both methods, perform sanity checks, and evaluate their clinical relevance.

Chapter 6 demonstrates how to propose a target policy based on the modeled behavior policy. We explain the method for proposing the target policy and discuss its implications for the treatment of RA patients from a clinical perspective. In addition, we provide recommendations to clarify ambiguous areas in the EULAR guidelines based on the proposed target policy.

Chapter 7 discusses the validation of our proposed target policy using off-policy evaluation. We compare the performance of different models, discuss the limitations of our evaluation approach, and explore methods to improve the reliability of our results.

Chapter 8 provides a comprehensive discussion of our work, addressing the questions raised in Section 1.1. We discuss our findings from the experimental results and discuss our proposed two-stage combined model.

Chapter 9 concludes the thesis by summarizing the overall results and suggesting possible directions for future work.

2

Background

In this chapter, we introduce the background for our thesis. We first discuss the basics of rheumatoid arthritis, including its symptoms, disease measurement indexes, treatment phases, and therapies. We then delve into the principles of sequential decision making, particularly within the context of healthcare. Next, we introduce interpretable policy learning, providing a brief overview of interpretable machine learning models and ensemble learning models. Finally, we discuss off-policy evaluation, including commonly used estimators and their limitations.

2.1 Rheumatoid Arthritis

Rheumatoid arthritis is a chronic systemic autoimmune disease primarily characterized by painful joint inflammation. The incidence of this disease is significantly higher in women than in men, and is most common in middle-aged people over the age of 55 [30]. RA shows regional and ethnic variability and is influenced by genetic, socioeconomic, environmental and lifestyle factors [31]. Rheumatoid arthritis occurs when the immune system mistakenly attacks the body's own tissues, particularly the lining of the synovial joints, leading to progressive disability, premature death and socioeconomic burden [3]. Common symptoms include joint pain, swelling and stiffness, and even limited range of movement. These symptoms are typically symmetrical, affecting the same joints on both sides of the body [32].

Despite the availability of several medications for the treatment of RA, the disease remains incurable [4]. The main treatment goal is to alleviate symptoms through early intervention to achieve low levels of disease activity [3], [5]. To quantify different aspects of disease activity and evaluate the effectiveness of treatments, several assessment tools have been developed, including Disease Activity Score-28 (DAS28), Clinical Disease Activity Index (CDAI), and Health Assessment Questionnaire Disability Index (HAQ-DI). DAS28 is a quantitative method that assesses the swelling and tenderness of 28 specific joints, along with erythrocyte sedimentation rate (ESR) or C-reactive protein (CRP) and the patient's self-report of health status [33]. In contrast, CDAI offers a simpler measurement approach that does not rely on acute-phase reactants like ESR or CRP. The CDAI score ranges from 0 to 76 and is categorized as follows: CDAI 2.8 indicates remission; CDAI $>$ 2.8 and 10 signifies low disease activity; CDAI $>$ 10 and 22 represents moderate disease activity; and CDAI $>$ 22 indicates high disease activity. The CDAI score is determined by calcu-

lating the number of swollen and tender joints out of 28, combined with the overall assessments of disease activity by both the patient and the physician [34]. Unlike DAS28 and CDAI, which focus on disease activity, HAQ-DI is used to assess the degree of physical functional disability in RA patients. It evaluates the patient's ability to perform daily activities such as dressing, walking, and grooming through a questionnaire [35].

Disease-modifying antirheumatic drugs (DMARDs) have been widely used to reduce pain and inflammation in RA patients. DMARDs include conventional synthetic DMARDs (csDMARDs), biological DMARDs (bDMARDs), and targeted synthetic DMARDs (tsDMARDs). Figure 2.1 shows the EULAR algorithm for RA management [6], [7], csDMARDs therapy is typically recommended as the first-line treatment (Phase I). If treatment with csDMARDs does not lead to sufficient improvement of symptoms and reduction of inflammation after 3 months and if patients do not achieve the target (such as low disease activity) after 6 months, the guidelines suggest adding a bDMARD or a JAK inhibitor (tsDMARD) following a risk assessment (Phase II). If Phase II failed, the guidelines recommend switching to another bDMARD or tsDMARD (Phase III). Should the patient still not improve after 3 months and not achieve the target after 6 months, the guidelines recommend continuing to switch to another bDMARD or tsDMARD. Currently, there is no consensus on choosing between bDMARDs and tsDMARDs in Phase II and Phase III [6], [7]. The choice in b/tsDMARDs can be influenced by several factors, including patient characteristics, previous treatment responses, comorbidities, insurance, and potential side effects [3], [4], [8], [9].

2.2 Sequential Decision Making

Sequential decision making can be viewed as a series of adaptive interactions between an agent and its environment, where the agent's decision is made based on the current state of the environment to achieve specific long-term goals. These long-term goals could be maximizing final investment returns in stock trading, minimizing transportation costs in the logistics sector, or reducing mortality rates from certain diseases in the medical field. Sequential decision making processes are often described by trajectories of state variables S_t, action variables A_t, and reward variables R_t that evolve over discrete time steps $t = 1, 2, 3, \ldots, T$, where T is the end time. At each time step t, the agent receives some representation of the state of the environment, represented as $S_t \in \mathcal{S}$, and based on that chooses an action $A_t \in \mathcal{A}$. One time step later, the environment provides feedback in the form of a numerical reward $R_t \in \mathcal{R} \subset \mathbb{R}$, which evaluates the effectiveness of the action. The outcome of this action then transitions the environment to a new state, S_{t+1}, setting the stage for the next decision. After passing the time steps t, the resulting trajectory is a sequence of $S_1, A_1, R_1, S_2, A_2, R_2, \ldots, S_t, A_t, R_t$.

In sequential decision making, a policy tells the agent which action should be taken given a particular state. There are deterministic policies and stochastic policies. A deterministic policy π is a mapping from a state S to an action A, that is, $\pi : S \to A$. A stochastic policy maps a state S to probability of selecting each possible action,

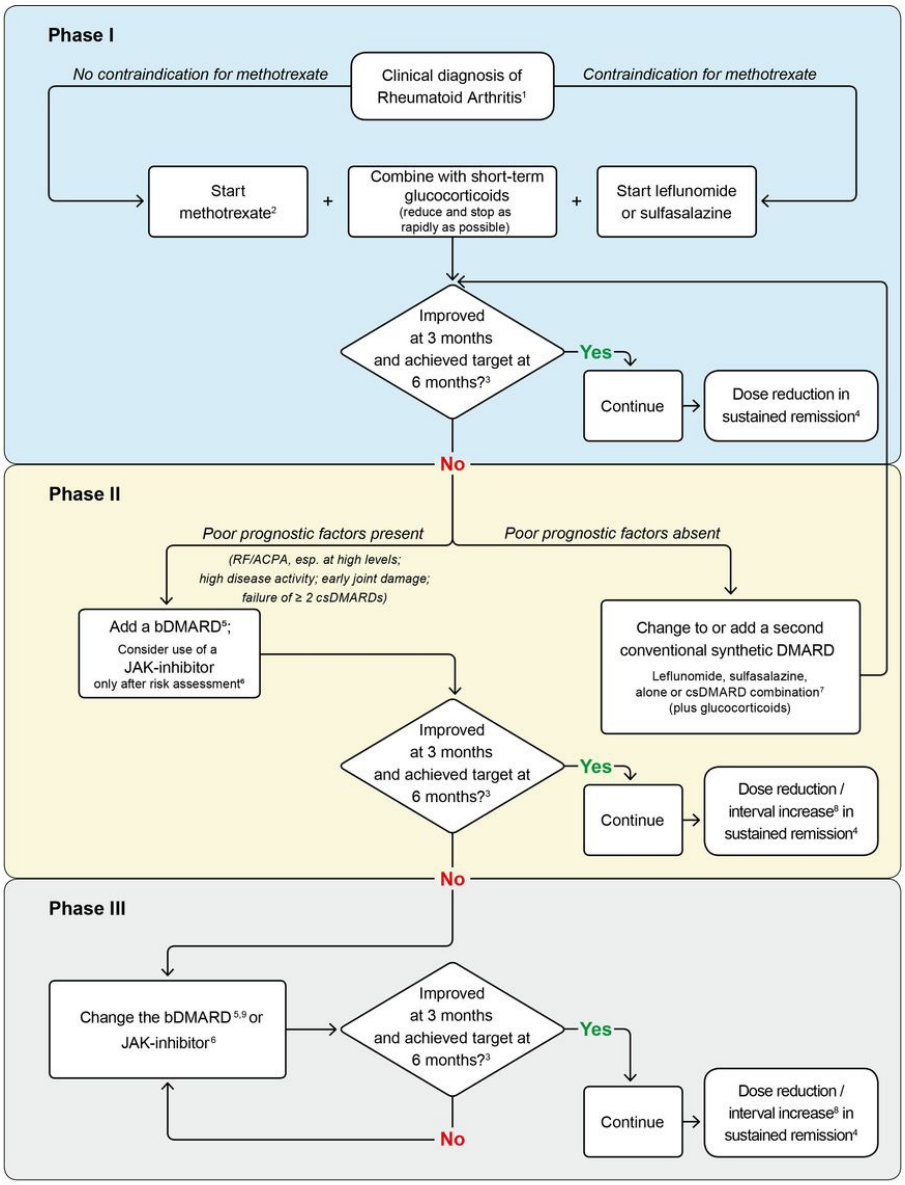


Figure 2.1: The flowchart outlining the latest treatment guidelines for RA is taken directly from the updated EULAR recommendations. This figure provides a visual representation of the treatment algorithm, including the three-phase therapy approach.

that is, $\pi : S \to \Delta_{\mathcal{A}}$. We use $p_\pi(A_t = a|S_t = s)$ to denote the probability that action $A_t = a$ is taken at time t, given state $S_t = s$, under policy π.

A Markov decision process (MDP) is defined as a tuple $(\mathcal{S}, \mathcal{A}, p, \mathcal{R})$, consisting of a state space $\mathcal{S}$, an action space $\mathcal{A}$, a reward space $\mathcal{R}$, and a joint probability distribution $p(S_{t+1}, R_t \mid S_t, A_t)$. This probability distribution specifies the likelihood of transitioning to the next state S_{t+1} and receiving a reward R_t, given current state S_t and current action A_t, thus defining the dynamics of the environment. The MDP has the Markov property, which means that the future state depends only on the current state and action, and not on any previous states or actions. Given this

property, we can form the following state transition and reward probability:

$$p(S_{t+1} = s+1, R_t = r \mid S_1, A_1, R_1, \ldots, S_t, A_t) = p(S_{t+1} = s+1, R_t = r \mid S_t, A_t) \tag{2.1}$$

Also, the probabilities for action, and reward can be obtained based on the current state and action, disregarding the past.

$$p(A_t = a \mid S_1, A_1, R_1, \ldots, S_{t-1}, A_{t-1}, R_{t-1}, S_t) = p(A_t = a \mid S_t) \tag{2.2}$$

$$p(R_t = r \mid S_1, A_1, R_1, \ldots, S_{t-1}, A_{t-1}, R_{t-1}, S_t, A_t) = p(R_t = r \mid S_t, A_t) \tag{2.3}$$

In a finite horizon setting, the value of a policy $V(\pi)$ can be defined as the expected sum of rewards from time step $t = 1$ to time step T. This can be written as:

$$V(\pi) = E\left[\sum_{t=1}^{T} R_t \mid \{A_t \leftarrow \pi(S_t)\}_{t=1}^{T}\right] \tag{2.4}$$

where R_t is the reward received at time t, $\{A_t \leftarrow \pi(S_t)\}_{t=1}^{T}$ specifies that the actions A_t at each time step t are determined by policy π, based on the corresponding state S_t.

2.3 Sequential Decision Making in Healthcare

Our project models sequential decision making in healthcare, focusing on the sequential treatment of patients with RA. In our work, each decision stage corresponds to a clinical visit for the patient, with each stage represented by different time steps t. In this context, the agent corresponds to a rheumatologist. At a given time step t, the state S_t represents the basis for the doctor's decision-making. The action A_t is the therapy selected by the rheumatologist at the current time step t. The reward R_t represents the treatment outcome from the previous time step $t-1$ to the current time step t.

As we discussed in section 2.2, sequential decision making problems in many domains are typically viewed as MDPs. However, this assumption does not hold for our focus on the treatment of RA. Because our state is not fully observable, at each stage t, we cannot completely know the current state S_t, we can only know the observations O_t. Therefore, we are working with a partially observable Markov decision process (POMDP). In practice, when considering the next steps in a patient's treatment, physicians make decisions based not only on the patient's current status and therapy received, but also on the patient's medical history [36]. By using the history to construct an extended state, we can capture more information, which helps to better infer the current true state.

Considering the influence of this historical factor, we introduce the concept of *history* in the sequential decision making for RA treatment. The history includes past actions, observations, and rewards. At stage t, the history H_t can be defined as $H_t := (O_1, A_1, R_1, \ldots, O_{t-1}, A_{t-1}, R_{t-1})$. With the introduction of history, the extended state S_t can be constructed as $S_t := (O_t, H_t)$.

2.4 Policy Learning

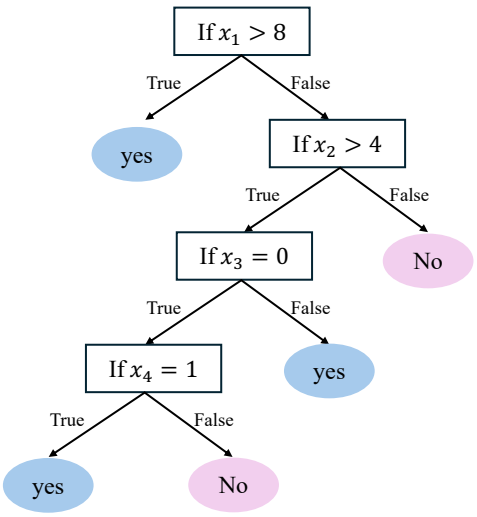
Using interpretable machine learning to model physicians' decision patterns (i.e., behavior policies) in healthcare is critical because it provides a transparent and understandable basis for decision making. This transparency enhances clinical trust and helps to describe current medical practices.

Although black-box models typically exhibit high performance, they are not interpretable. To model behavior policy, we use six interpretable machine learning models, two ensemble learning models as our black-box models, and a dummy model as the baseline. The interpretable machine learning models we used in this project can be categorized into two main classes: rule-based models and linear models.

2.4.1 Interpretable ML: Rule-Based Models

Rule-based models are a type of machine learning model that use a series of "if-then" statements to make predictions. There are three types of rule-based models: decision sets, decision lists, and decision trees.

Decision sets are an unordered collection of if-then rules, where the satisfaction of any rule leads to the corresponding prediction. Decision lists are an ordered sequence of if-then-else rules, evaluated in order until one rule applies, providing the prediction. Decision trees are hierarchical structures where each node represents a condition on a feature, and the leaf nodes represent the predictions [16]. Figure 2.2 shows examples of binary prediction problems for these three types of models, illustrating the differences among them.



(a)

IF $x_1 < 6$ **and** $x_2 < 8$ **THEN** predict yes
ELSE IF $x_2 < 4$ **THEN** predict no
ELSE IF $x_3 = 1$ **and** $x_4 = 0$ **THEN** predict no
ELSE predict yes

(b)

IF $x_1 > 10$ **THEN** predict yes
IF $x_2 < 5$ **and** $x_1 < 8$ **THEN** predict yes
IF $x_3 = 0$ **THEN** predict no
ELSE predict no

(c)

Figure 2.2: Examples of binary prediction problems include (a) a decision tree, (b) a decision list, and (c) a decision set. These examples are inspired by those provided by Rudin et al. in their work on interpretable machine learning [16].

2.4.2 Interpretable ML: Linear Models

Linear models are a class of statistical models that predict an outcome based on a linear combination of input features, including logistic regression and risk score models.

Logistic regression predicts probabilities by modeling the relationship between predictor variables and the outcome as a linear combination, which is then transformed by a logistic function. Risk score models are linear classification models that provide quick, interpretable predictions by assigning point values (weights) to various predictor variables, which are then summed to assess risk [16]. Both models are interpretable because examining the weights of the predictor variables allows us to understand their influence on the predicted outcome.

2.4.3 Non-Interpretable ML: Ensemble Learning Models

Ensemble learning models combine the predictions of multiple individual models to improve overall performance and robustness. The key idea is that by aggregating multiple models, the ensemble can achieve better accuracy and generalization than any single model alone. Bagging, boosting, and stacking are three main types of ensemble learning methods. Bagging trains multiple models on different subsets of the data and averages their predictions to reduce variance. Boosting sequentially corrects errors of prior models to reduce bias. Stacking combines predictions from multiple models using a metamodel for the final prediction.

Random Forest is a bagging ensemble learning method that builds multiple decision trees using random subsets of data and features, and combines their predictions through averaging or majority voting for improved accuracy and robustness [37]. XGBoost is a powerful boosting algorithm that sequentially builds models to correct the errors of previous ones, using gradient boosting techniques to optimize performance and achieve high predictive accuracy [38].

2.5 Off-Policy Evaluation

Off-policy evaluation (OPE) is to estimate the value $V(\pi)$ of a target policy π, using data collected under a different behavior policy μ. This is particularly useful in our settings where experimenting with the target policy directly is impractical and unethical. OPE addresses the question: What would happen if we implemented the unobserved actions suggested by the target policy π? This question is a matter of counterfactual reasoning, as we can only observe actions implemented under the behavior policy μ.

A classic method in OPE is importance sampling (IS), which estimates the expected value of samples by weighting them according to the ratio of their probabilities under the target distribution and the proposal distribution [39]. The IS estimator is consistent and unbiased [15]. However, when the behavior and target policies are very different, this method can suffer from high variance [40]. To mitigate this issue, weighted importance sampling (WIS) is widely used. This approach normalizes the

importance weights to sum to one, reducing the impact of extreme weights and thereby providing a more stable estimation. Though WIS reduces variance, it is a consistent but biased estimator [15].

3

Data

In this chapter, we provide a detailed description of the data used in this project. We start with an overview of the dataset, including the statistical and missing value descriptions of the selected features. Next, we introduce the exposures within the dataset, which refer to the treatments patients received. We also discuss the wash-out period, the time required for the complete metabolism and elimination of previously used therapy, and the follow-up window, the period during which physicians evaluate a patient's condition and therapy outcomes after the previous therapy. Finally, we discuss how we handle these two criteria in our analysis.

3.1 Dataset Description

We use the dataset from CorEvitas, a large RA registry based in the United States, covering the period from January 2012 to December 2021 [12]. The original dataset includes data on 42,068 patients. We exclude 905 patients with missing or inconsistent information on therapy changes. Finally, we obtain data on 41,163 patients. Figure 3.1 shows the flowchart of data examination process.

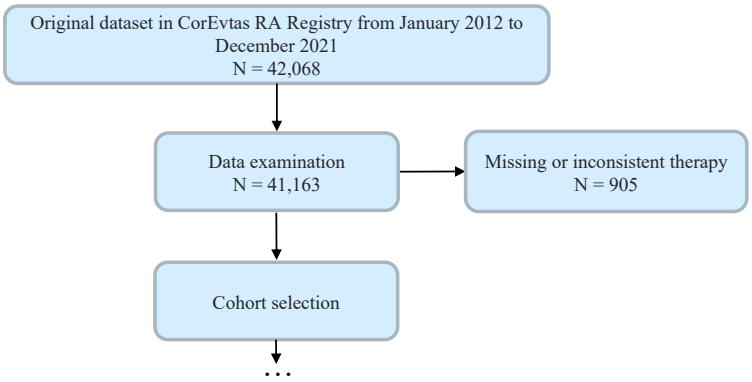


Figure 3.1: The flowchart of data examination process.

During each follow-up, every patient is recorded until the last registry visit or the data cut-off date of December 31, 2021. Therefore, the total number of recorded visits and the duration of follow-up vary between patients. We address this issue by using a fixed follow-up duration in OPE, such as limiting the follow-up window to 2 years from the first b/tsdmard.

This dataset includes a wide range of variables, such as patient demographic information, clinical characteristics, infections, comorbidities, and adverse events. To align with our research objectives, we consider the first use of b/tsDMARDs after cohort selection (detailed in Section 4.5) as the baseline. Table 3.1 and Table 3.2 present the baseline statistics and missing percentages of the features selected for this project. Section 4.5 details our cohort selection approach. We address all missing values, with Section 4.7 outlining our imputation strategy.

3.2 Exposures

Given the large number of available DMARDs and their potential combinations, to simplify the complexity associated with the multitude of drug options, Matsson et al. [41] study changes between classes of DMARDs rather than changes between individual DMARDs. Following their work, we study the following drug classes, both have monotherapy and combination therapy except csDMARDs:

- Conventional synthetic DMARD (csDMARD)
- Tumor necrosis factor inhibitor (TNFi)
- Janus kinase inhibitor (JAKi)
- Interleukin-6 receptor inhibitor (IL-6Ri)
- Abatacept
- Rituximab
- No DMARD

Interleukin-1 receptor inhibitors are excluded due to small sample size. Rare combinations of b/tsDMARDs are categorized as "other" because they are not clinically recommended.

Biological DMARDs (bDMARDs) include TNFi, IL-6Ri, abatacept, and rituximab, while targeted synthetic DMARDs (tsDMARDs) include JAKi. Monotherapy refers to taking a bDMARD or tsDMARD only, while combination therapy means taking a csDMARD in combination with a bDMARD or tsDMARD, see Figure 3.2.

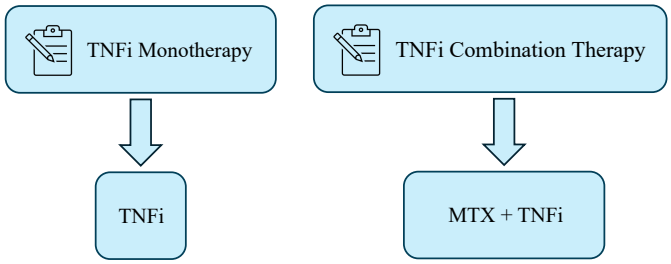


Figure 3.2: Example of TNFi monotherapy and TNFi combination therapy, with methotrexate (MTX) being the most commonly used csDMARD.

3.3 Criteria

Based on our understanding of rheumatoid arthritis, we have established several criteria, including the wash-out period and the follow-up window. This section discusses the definitions of these two criteria and how we handle them in our analysis.

Wash-Out Period In the treatment of RA, the wash-out period refers to the time required for the complete metabolism and elimination of previously used therapy before switching therapies or stopping a particular therapy. The purpose of this period is to avoid interactions between the old and new therapies and to ensure the safety and efficacy of the new therapy. According to previous studies [42], [43], in our project we define that: if the period of not taking any therapy (i.e. No DMARD therapy) is less than two months, we refer to this time as the wash-out period, as can be seen in Figure 3.3. In our data, the average No DMARD period is 150 days, with a median of 112 days. Additionally, approximately 13.4% of patients in our dataset experience a wash-out period.

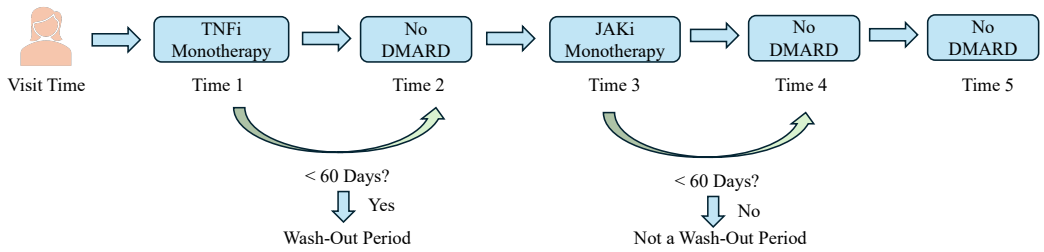


Figure 3.3: An example of defining a wash-out period.

Follow-Up Window The follow-up window is a period during which physicians evaluate a patient's condition and therapy outcomes following the previous therapy. By limiting this window, we can more accurately assess the effectiveness of the previous therapy. In our project, we limit the follow-up window to 3 to 12 months to ensure regularity of therapy and follow-up, thereby laying the foundation for subsequent off-policy evaluation. Figure 3.4 shows an example of defining a follow-up window after removing the a wash-out period illustrated in Figure 3.3.

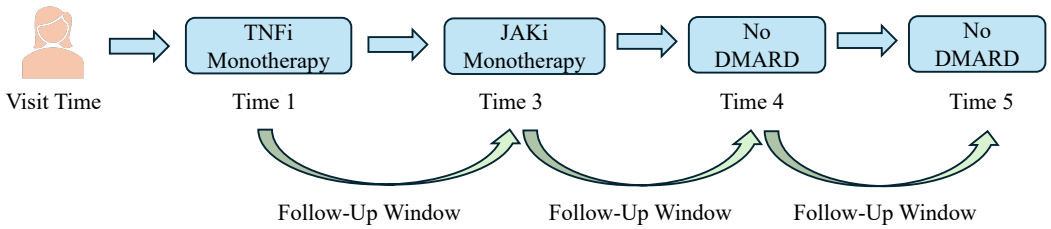


Figure 3.4: An example of defining a follow-up window after remove a wash-out period.

Table 3.1: Variables included in S_t that do not contain history—i.e., features that represent the patient's current status—along with their baseline statistics and missing percentages. Baseline refers to the RA patient's first use of b/tsDMARDs in our data. For each variable, N represents the number of patients with non-missing baseline information, calculated after data examination and cohort selection. Variables that include history are listed in Table 3.2.

Variable	N	Statistics	Missing Percentage
Age in years, median (IQR)	6016	59 (50, 68)	0.3
Calendar year, median (IQR)	6037	2016 (2013, 2019)	0.0
RA duration in years, median (IQR)	5956	3 (1, 9)	1.3
bmi, median (IQR)	5902	29 (25, 34)	2.2
Systolic Blood Pressure, median (IQR)	5924	128 (118, 139)	1.9
Diastolic Blood Pressure, median (IQR)	5925	78 (70, 84)	1.9
Pain (self-reported), median (IQR)	6006	45 (20, 70)	0.5
Fatigue (self-reported), median (IQR)	5979	48 (17, 70)	1.0
CDAI, median (IQR)	5899	14 (7, 25)	2.3
DAS, median (IQR)	3366	3 (2, 4)	44.2
HAQ, median (IQR)	6000	0 (0, 1)	0.6
Smoker, n (%)	5232	884 (16.9)	13.3
Drinker, n (%)	5867	2602 (44.3)	2.8
Pregnant, n (%)	4158	6 (0.1)	31.1
CCP Positive, n (%)	884	483 (54.6)	85.4
RF Positive, n (%)	959	586 (61.1)	84.1
Female, n (%)	6022	4629 (76.9)	0.2
Ethnicity (self-reported), n (%)	5955		1.4
White		4721 (79.3)	
Hispanic		542 (9.1)	
Black		476 (8.0)	
Asian		118 (2.0)	
Other		98 (1.6)	
Final education, n (%)	5824		3.5
Primary school		153 (2.6)	
High school		2289 (39.3)	
College		3339 (57.3)	
Do not remember		43 (0.7)	
Work status, n (%)	5875		2.7
Full time		2431 (41.4)	
Part time		482 (8.2)	
Work at home		505 (8.6)	
Student		78 (1.3)	
Disabled		649 (11.0)	
Retired		1730 (29.4)	
Private insurance, n (%)	6037	4276 (70.8)	0.0
Medicare insurance, n (%)	6037	1960 (32.5)	0.0
Medicaid insurance, n (%)	6037	397 (6.6)	0.0
No insurance, n (%)	6037	118 (2.0)	0.0

Table 3.2: Variables included in S_t that contain history—Features that include the patient's previous visit records—features that include the patient's previous visit records—along with their baseline statistics. Baseline refers to the RA patient's first use of b/tsDMARDs in our data. For each variable, N represents the number of patients with non-missing baseline information, calculated after data examination and cohort selection. Other variables without history are listed in Table 3.1.

Variable	N	Statistics
Comorbidities, n (%)		
Metabolic diseases	6037	376 (6.2)
Cardiovascular diseases	6037	608 (10.1)
Respiratory diseases	6037	141 (2.3)
Drug-induced lupus	6037	1 (0.0)
Cancer	6037	150 (2.5)
GI and liver diseases	6037	83 (1.4)
Other diseases	6037	1402 (4.2)
Severe Infections, n (%)	6037	101 (1.7)

4

Methodology

In this chapter, we first introduce two methods for modeling behavior policy, around which all our work is centered. Next, we detail the three main types of models used in modeling behavior policy—linear models, rule-based models, and black-box models. We also discuss the evaluation metrics for these models.

We then move on to data processing methods. We firstly provide a general overview of the data processing steps and how these steps fit into our two prediction tasks. We then delve into the construction of data preprocessing pipelines and their functionalities. Following this, we discuss how to handle specific criteria, including the follow-up window and wash-out period. Lastly, we explore data transformation pipelines for handling missing values and feature encoding.

Finally, we cover the experimental setup, including data splitting methods and model training techniques. We also introduce off-policy evaluation methods, specifically importance sampling and weighted importance sampling estimators.

4.1 Two Approaches for Modeling Behavior Policy

To predict which therapy a RA patient might switch to, the common approach is to directly use a machine learning model to fit the data and make predictions. Although this method is straightforward and easy to implement, it may lack interpretability when dealing with complex situation. To address this issue, we propose another approach based on the "divide and conque" idea, which involves breaking down a complex problem into multiple sub-problems and solving these sub-problems to answer the complex question. Figure 4.1 illustrates these two approaches.

The first approach directly using a machine learning model to fit the data. Specifically, we use historical data of RA patients to train a multi-class machine learning model, which can predict which therapy each patient might switch to. Although this method is simple and direct, it may lack interpretability or do not give valuable information when dealing with complex situation, making it difficult for physicians to understand the decision-making process of the model.

The second approach adopts the "divide and conquer" principle, breaking down the problem into two sub-questions: "Will the RA patient switch therapy?" and

"If the RA patient switches therapy, which therapy will they switch to?" This method involves a combined model with two stages. The first stage uses a binary classification model to determine the probability of a patient switching therapy. For patients predicted to switch, the second stage employs a multi-class model to predict the specific therapy they will switch to. By combining the outputs of these two sub-models, we can calculate the overall probability of a patient switching therapy. This approach aims to enhance interpretability, making the decision-making process clearer at each step.

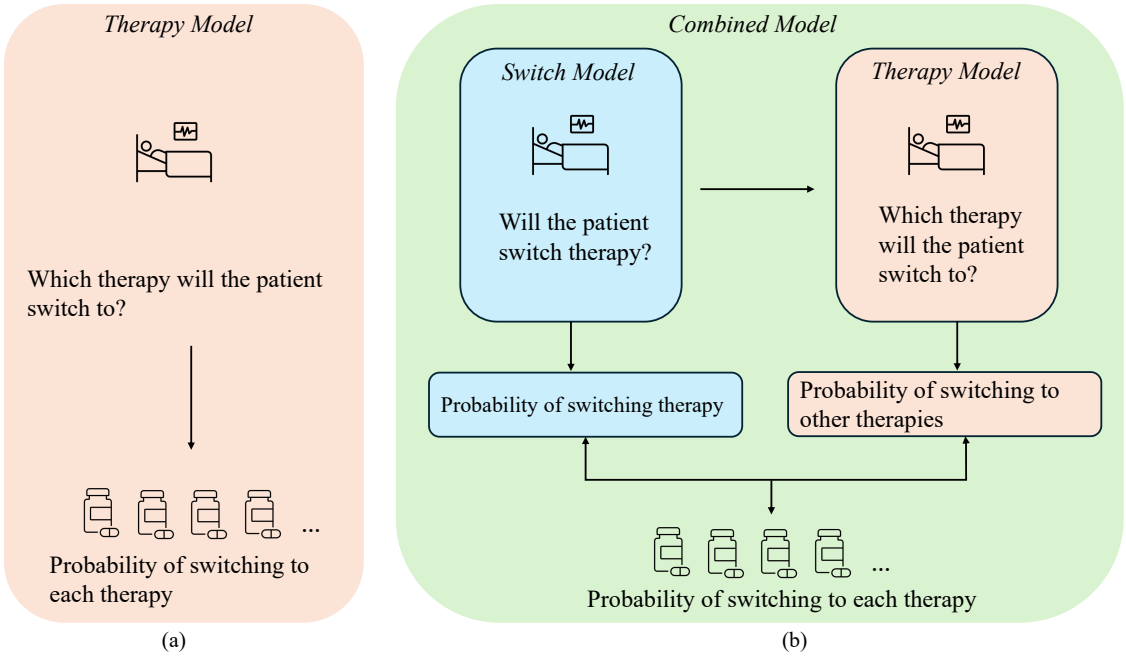


Figure 4.1: Two approaches for modeling behavior policy. (a) shows the first approach that directly uses a single therapy model, while (b) illustrates the second approach that employs a combined model based on the "divide and conquer" principle.

4.2 Models

In this section, we provide a detailed introduction to the models used in this project. We categorize these models into three main types: linear models, rule-based models, and ensemble learning models. Table 4.1 shows all the models used in our project.

4.2.1 Linear Models

Linear models are widely used for their simplicity and interpretability, with feature weights intuitively explaining how inputs influence outputs. Additionally, when sparsity is enforced through regularization or constraints, the model relies on fewer key features, further enhancing clarity and ease of understanding [20].

Table 4.1: The models used in our project.

Model	Interpretable by design	Accepts Multi-Class	Model Type	Complexity measure
Boosted rule set (BRS)	✓	✓	Rule set	# rules
Greedy rule list (RL)	✓	✗	Rule list	# rules
Hierarchical shrinkage wrapper (RT)	✓	✓	Rule tree	# leaves
Decision tree (DT)	✓	✓	Rule tree	# leaves
Risk scores (RS)	✓	✗	Linear model	# rules
Logistic regression (LR)	✓	✓	Linear model	# coefficients
Rule ensembles (RE)	✓	✗	Linear model	# rules
eXtreme gradient boosting (XGB)	✗	✓	Ensemble learning	N/A
Random forest (RF)	✗	✓	Ensemble learning	N/A

Logistic Regression. Logistic regression calculates the probability of an instance belonging to a particular category by passing the results of linear regression through a logistic function, such as the Sigmoid function for binary classification or the Softmax function for multi-class classification. This transformation ensures the output is a probability value between 0 and 1. The model then uses these probabilities to make predictions, with the goal of minimizing the difference between the predicted and actual class labels using a loss function, typically the cross-entropy loss.

Risk Scores. Risk scores are simple classification models that let users make quick risk predictions by adding and subtracting a few small integer numbers. In this project, we use RISKSLIM, a classic risk score model. RISKSLIM is a simple and interpretable binary classification model that uses small integer coefficients to make predictions. It aims to minimize prediction error while keeping the model simple by penalizing the number of non-zero coefficients [44]. This balance is achieved by solving an optimization problem that combines logistic loss with a sparsity-inducing penalty, making RISKSLIM particularly suitable for applications requiring transparency and ease of interpretation.

Rule Ensemble. Rule ensemble (RE) learns sparse linear models that incorporate automatically detected interaction effects in the form of decision rules, combining a decision tree model with a linear model to enhance interpretability [19]. Rule ensembles generates a set of rules from a decision tree and then uses these rules, along with the original features, as inputs to a linear model, tailored for either regression or classification tasks.

4.2.2 Rule-Based Models

Rule-based models are logical models that consist of statements that include "if-then", "or", and "and" clauses. These statements provide human-understandable reasons for each prediction, such as "IF age >= 65 AND has comorbidities, THEN recommend switching therapy". As we introduced in Section 2.4, rule-based models can be categorized into three types: decision sets, decision lists, and decision tree, as shown in Table 5.5, 5.6, and Figure 5.2, respectively.

Decision trees. We use the hierarchical shrinkage wrapper (RT) and the decision tree (DT) to represent the decision tree category. The decision tree is a classic machine learning model widely used for its simplicity and interpretability and the hierarchical shrinkage wrapper is an advanced decision tree model that enhances performance and robustness. RT is a novel post-hoc algorithm designed to avoid structural modifications to trees while addressing the overfitting problem [45]. RT focuses on regularization by shrinking predictions within each node towards the sample means of its ancestral nodes. Instead of modifying the trees, RT enhances regularization by nudging predictions within each node towards the average of its ancestors. DT makes decisions by recursively splitting the data set into smaller subsets. At each node, the model splits the data into two subsets based on a specific feature threshold. Each data point is assigned to either the left or right subtree based on the threshold decision without ambiguity. The process starts at the root node and continues to the leaf nodes that satisfy the stopping condition, which represent the final decision or prediction.

Decision lists. A decision list is a series of if-then statements. During the prediction process, it tests each rule sequentially until it finds the first one that satisfies the conditions. We use greedy rule lists (RL) as a representative of the decision lists category. Greedy rule lists is a binary classification algorithm that iteratively splits on one feature at a time along a single path to maximize the probability of class 1. It constructs rules by selecting the most significant feature at each step, creating a simple decision path.

Decision sets. A decision set acts like a democracy among rules, where each rule can have different levels of influence or voting power [46]. The rules within a decision set are unordered, and each rule is composed of a conjunction of conditions, allowing multiple rules to be evaluated independently. Within the category of decision set models, we utilize Boosted Rule Sets (BRS). BRS employs the Adaboost algorithm to sequentially fit a set of rules [47]. This algorithm iteratively applies a weak classifier to modified versions of the data, increasing the weights of misclassified observations so that each subsequent learner focuses on correcting the errors made by its predecessor. The final predictions are made by aggregating the individual rule predictions through a weighted majority vote.

4.2.3 Ensemble Learning Models

Although interpretable models provide transparency and explainability, they may not capture complex patterns in the data, which can lead to reduced accuracy. Therefore, we introduce two ensemble learning models as complementary black-box models to improve performance: random forest (RF) and extreme gradient boosting (XGB).

Random forest. A random forest is constructed by a multitude of decision trees during training [37]. These decision trees fit on various sub-samples of the dataset and use averaging to improve predictive accuracy and control overfitting [48]. For

classification tasks, the output of the random forest is the class selected by the majority of the trees. Each tree in the forest is built from a random subset of features, which helps to de-correlate the trees and reduce variance. By aggregating the predictions of several different trees, the random forest reduces the risk of overfitting to noisy data and improves generalization to unseen data.

Extreme gradient boosting. Extreme gradient boosting is an implementation of gradient-boosting decision trees [38]. XGB iteratively refines an ensemble of weak learners (such as decision trees) that optimize a predefined objective function through gradient descent. Through this iterative process, XGB strategically assigns higher weights to misclassified instances, progressively enhancing the model's predictive accuracy. Additionally, XGB includes regularization and parallel processing, which help prevent overfitting. The regularization terms control the complexity of the model, reducing the likelihood of overfitting, while the parallel processing capabilities significantly speed up computation. Furthermore, XGB uses a second-order Taylor approximation to more accurately capture the gradients, leading to better optimization and faster convergence.

4.3 Evaluation Metrics

We use Accuracy, Area Under the Receiver Operating Characteristic Curve (AUC-ROC), and Area Under the Precision-Recall Curve (AUC-PR) to evaluate our model's performance on classification tasks because each provides a unique perspective: Accuracy measures overall correctness but can be misleading on imbalanced datasets; AUC-ROC evaluates the ability to distinguish between classes but may overestimate performance in highly imbalanced scenarios; AUC-PR focuses on positive class performance, making it useful for imbalanced datasets, though less intuitive than ROC. Using these metrics together gives a comprehensive view of the model's performance, especially important for our imbalanced dataset, where the number of patients who stay far exceeds those who switch, and our focus is on accurately identifying switch patients.

Additionally, we use Expected Calibration Error (ECE) and Brier Score to assess the model's calibration [49]. Calibration refers to how well the predicted probabilities reflect the true likelihood of events. Together, these metrics help us ensure that the model's predicted probabilities are reliable and well-calibrated, which is crucial for making decisions based on those probabilities.

AUC-PR. AUC-PR is the area under the precision-recall curve. The precision-recall (PR) curve represents the trade-off between precision (the proportion of true positive predictions out of all positive predictions) and recall (the proportion of true positive predictions out of all true positive instances in the data set). AUC-PR summarizes this trade-off by calculating the area under the PR curve. A higher AUC-PR value indicates better model performance, especially in accurately identifying the positive class. This metric is particularly useful in situations with unbalanced data sets where the positive class is rare.

Expected calibration error. Expected calibration error (ECE) measures the the calibration of probabilistic classifiers, assessing the agreement between predicted probabilities and empirical frequencies. Mathematically, ECE Score can be expressed as:

$$\text{ECE} = \sum_{m=1}^{M} \frac{N_m}{N} \left|\text{Acc}_m - \text{Conf}_m\right| \tag{4.1}$$

where M is the number of bins, which refers to ranges of predicted probability values into which predictions are grouped. N_m is the number of instances in bin m, N is the total number of instances, Acc_m is the accuracy of the predictions in bin m and $Conf_m$ is the confidence of the predictions in bin, where confidence indicates how sure the model is about its prediction.

Brier score. The Brier score is to evaluate the accuracy of binary probabilistic predictions. It measures the mean squared difference between predicted probabilities and the actual outcomes. It can be calculated as:

$$\text{Brier score} = \frac{1}{N} \sum_{i=1}^{N} (f_i - o_i)^2 \tag{4.2}$$

where N is the total number of predictions, f_i denotes the forecast probability of event i and o_i represents the outcome (0 or 1) for event i.

4.4 Data Processing Steps

To model behavior policy using the two approaches we mentioned in Section 4.1, we need our data to accommodate two scenarios: binary classification (whether to switch therapy) and multi-class classification (which therapy to switch to). Therefore, we divide our data processing into two main parts. The first part handles the multi-class classification problem of determining which therapy to switch to. And the second part focuses on predicting whether to switch therapy.

4.4.1 Multi-Class Classification Task

The first part of our data processing is for the multi-class classification task. The main steps are:

1. Data loading: Loading initial data for further preprocessing.
2. Patient cohort selection and defining index visit: Performing cohort selection to ensure that the selected patients are those who started b/tsDMARDs therapy. Additionally, we define each patient's index visit as their first switch to b/tsDMARDs.
3. Adding features: This includes applying a one-step history truncation, which means considering only the most recent visit in the patient's record. We also add previous actions, indicating whether the patient switched therapy at their last visit.

4. Handling specific criteria: Processing specific criteria, such as the wash-out period and follow-up window mentioned in section 3.3.

5. Data transformation: Transforming different types of data, ensuring the data is ready for model training.

6. Data splitting: Splitting the data into training, validation, and test sets while maintaining group consistency (i.e., patients) to prevent data leakage.

4.4.2 Binary Classification Task

This part focuses on predicting whether a patient will switch therapy. It builds on the multi-class classification task with the following additional steps:

1. Using existing functionality: Applying all the data processing steps from the multi-class classification task.

2. Generate switch labels: Processing the therapy label to represent a switch label (whether the patient has switched therapy).

4.5 Data Preprocessing Pipelines

In our project's data preprocessing, the first step is the selection of the patient cohort. After selecting the patient cohort, we further process the data. This include employing one-step history truncation, incorporating features of previous actions and defining the index visit.

Cohort selection. The selected cohort should be consistent with our overall objective, that is, to investigate the subsequent treatment patterns in RA patients. Based on our overall objective, we made the following assumptions for the patient cohort:

- At the first visit in the data, the patient is older than 18 years of age.
- At the first visit in the data, the patient has no history of b/tsDMARDs.
- In the patient's entire visit record, there is at least one visit where b/tsDMARD therapy is initiated.

One-step history truncation. One-step history truncation refers to truncating the history to the last time step, meaning we look back one time step. This involves using the therapy and the action from the last visit to represent the most recent history. For example, as shown in the Figure 4.2, if the patient's therapy at Time 3 is TNFi monotherapy, this patient's previous therapy would be the therapy at Time 2 — csDMARD. For patients with no previous therapy, impute the missing previous therapy with 'csDMARD' since we start from the first b/tsDMARD.

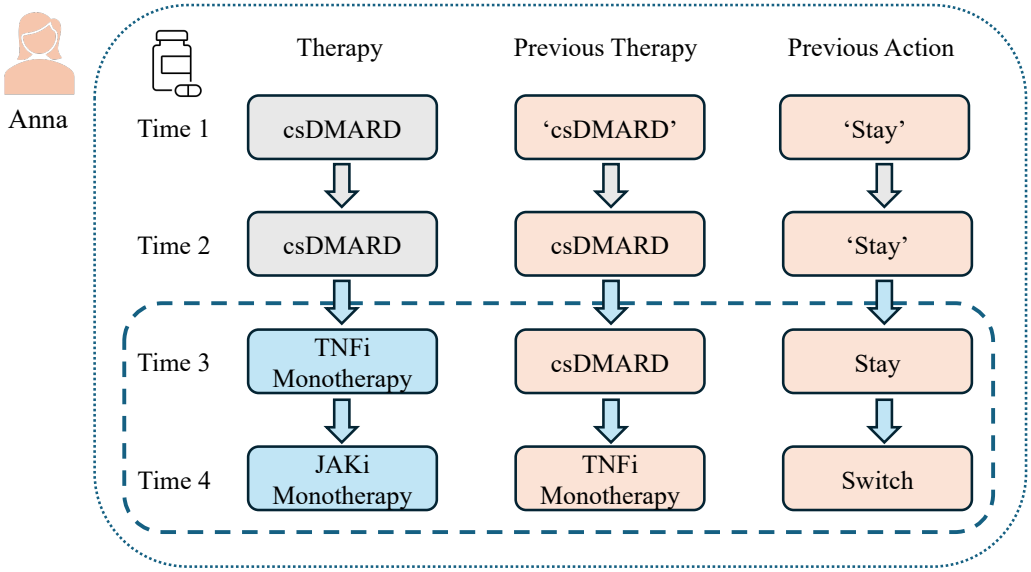


Figure 4.2: An example of patient data after passing through the data preprocessing pipeline. The short dotted line box represents the complete patient's record, while the long dotted line box represents the data after defining the index visit. Quotation marks indicate data that we have filled in. This example is provided solely to demonstrate the data preprocessing pipeline and does not reflect actual patient data.

Previous action. Previous action refers to whether the patient switched therapy at their last visit. For each patient, we calculate their previous action at the current time step. As shown in the Figure 4.2, the previous action for the patient at Time 3 to be 'stay' because the patient did not switch therapy from csDMARD at Time 1 to csDMARD at Time 2. For unknown previous actions, such as those at Time 1 and Time 2, we impute these unknown previous actions as 'stay'. This is because the unknown previous actions only occur in the first and second time step of the patient's sequence in our data. After selecting the patient cohort, the patients used only csDMARDs and never tried b/tsDMARDs before their first visit, so we impute the previous actions in the first and second time step in our data as 'stay'.

Index visit. To align the sequences, we define the index visit as the patient's first change to b/tsDMARD, meaning the first row in each patient's sequence data is their first switch to b/tsDMARD therapy. As shown in Figure 4.2, after defining index visit, we truncate the patient's entire sequence data to the data within the blue long-dashed box, with Time 3 becoming the first row of data for this patient.

4.6 Handling Specific Criteria

As introduced in Section 3.3, our project defines the criteria for both the wash-out period and the follow-up window. The criteria for the wash-out period is no therapy (denoted as no DMARD in our data) within 2 months after the last therapy, while

the criteria for the follow-up window is that the time between two visits is limited to 3 to 12 months.

For the wash-out period, we first group the data by each patient, so the following processing steps apply to each patient's sequences. Next, as shown in Figure 3.3, we calculate the period of no DMARD for each patient. If the duration of no DMARD is less than 2 months, we consider it a wash-out period and delete the visit records where the therapy is no DMARD. Specifically, we remove the Time 2 record in Figure 3.3.

Figure 3.4 shows how to calculate follow-up window for each patient after removing wash-out period for that patient. For the follow-up window, we also group the data by each patient and process the sequence data for each patient. Next, we calculate the time difference between the current visit record and the previous visit record, which is our follow-up window. In the context of OPE, to ensure accuracy, following our assumptions, if the follow-up window for a visit record is less than 3 months or greater than 12 months. This approach allows us to retain more data while ensuring the reliability of our analysis.

Figure 4.5 shows the flow chat of complete data preprocessing pipelines and the process of handling specific criteria, starting from the examined data, including the number of patients remaining after each step.

4.7 Data Transformation Pipelines

We input the data that has been processed through the data preprocessing pipelines and specific criteria handling into the data transformation pipelines for handling missing values and encoding features. For numerical data, categorical data, and boolean data, we transform them in different ways accordingly in our pipeline, as shown in Figure 4.3.

For numerical data, we impute missing values using the `SimpleImputer` from scikit-learn with the mean value. Then, depending on the model, we use `StandardScaler`, `KBinsDiscretizer`, or leave the features unencoded. `StandardScaler` standardizes features by removing the mean and scaling them to unit variance. `KBinsDiscretizer` is a scikit-learn transformer that discretizes continuous features into k bins using the quantile strategy.

For categorical data, we impute the missing values using `SimpleImputer` with the most frequent value. Then, we use `OneHotEncoder` from scikit-learn to encode the features. For boolean data, we also impute the missing values using `SimpleImputer` with the most frequent value, then we use binary encoding to encode features.

4.8 Experiment Setup

We split the dataset into 80% training set and 20% test set. Then, the 80% training set is further divided into 80% training data and 20% validation data. This gives

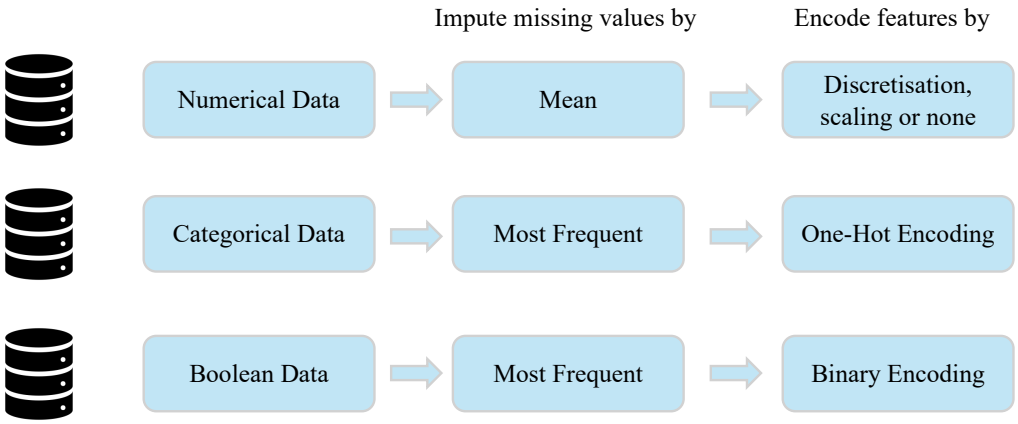


Figure 4.3: Data transformation pipelines for different data types.

us 64% of the total data for training the model, 16% for evaluating the model performance, and 20% for off-policy evaluation, as shown in Figure 4.4. All splits are based on patient ID, ensuring that the records of a patient only appear in either the training set, test set, or validation set. This approach prevents data leakage, which occurs when information from outside the training dataset is inadvertently used to train the model, leading to overly optimistic performance estimates during training or validation.

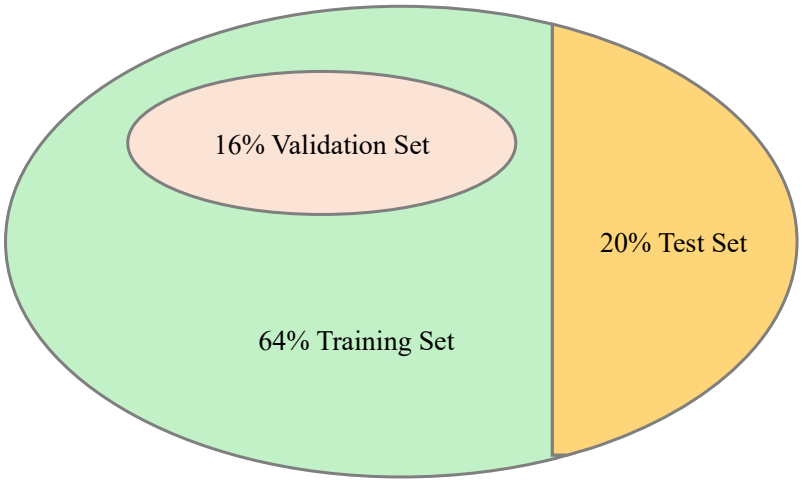


Figure 4.4: The dataset split graph shows how the dataset is split, where the percentage of each section indicates its proportion in the original dataset.

For each model training, we can split the data multiple times using different random seeds and specify the number of splits. By default, each model undergoes 10 data splits for training. For each model, during each data split and model training, we use `RandomizedSearchCV` to tune the model parameters on the training set. The model with the best parameters is selected based on the ROCAUC score from 10-fold cross-validation. After finding the model with the best parameters, we evaluate this model using different evaluation metrics, which are detailed in Section 4.3. For different

models, all models are trained based on the same data splits. After all models have been trained with 10 data splits, each model selects its own final model.

4.9 Off-Policy Evaluation Methods

Our thesis addresses the problem of learning a behavior policy μ from a dataset $\mathcal{D}$ and propose a target policy π based on the behavior policy μ. We evaluate the effectiveness of the proposed target policy π using two off-policy evaluation methods— importance sampling and weighted importance sampling. We define the notations used in this context as follows:

- The behavior policy μ is the probability distribution of taking an action A_t given the state S_t at time t. This is denoted as $p_\mu(A_t \mid S_t)$.
- The target policy π, which we aim to evaluate, dictates the probability distribution of taking an action A_t given the state S_t at time t. This is denoted as $p_\pi(A_t \mid S_t)$.

4.9.1 Importance Sampling

Importance sampling is a method for estimating the expected value of a random variable under one distribution using samples drawn from another distribution. This technique allows us to re-weight the samples obtained from the behavior policy μ to reflect the target policy π. Through importance sampling, the estimator for the expected reward under the target policy π is given by:

$$\hat{V}_{\text{IS}}(\pi) = \frac{1}{N} \sum_{i=1}^{N} R_t^{(i)} w_i \tag{4.3}$$

where N is the total number of patients, $R_t^{(i)}$ is the reward at time step t for the i^{th} patient, and w_i is the importance weight of i^{th} patient. The importance weight w_i can be represented as:

$$w_i = \prod_{t=0}^{T} \frac{p_\pi(A_t^{(i)}|H_t^{(i)})}{p_\mu(A_t^{(i)}|H_t^{(i)})}$$

This weighting approach places more emphasis on samples that are rare under the sampling distribution p_μ but frequent under the target distribution p_π. If p_μ and p_π are the same, then the weights for all samples are 1. The importance sampling estimator gives consistent and unbiased estimates of the samples [15].

4.9.2 Weighted Importance Sampling

When the behavior policy μ and the target policy π are significantly different, the importance sampling method can suffer from high variance. To mitigate this issue, we use weighted importance sampling (WIS). WIS reduces variance by normalizing the importance weights, resulting in a more stable estimator. The weighted importance

sampling estimator is calculated by performing a weighted average of the samples with weight w_i. The estimator for the expected return under the target policy π is given by:

$$\hat{V}_{\text{WIS}}(\pi) = \frac{\sum_{i=1}^{N} R_t^{(i)} w_i}{\sum_{i=1}^{N} w_i} \tag{4.4}$$

This estimator reduces the high variance problem of importance sampling because it normalizes the weights. When an unlikely event occurs, the corresponding weight can be very large, causing significant variation in the importance sampling estimator. By normalizing the weights, WIS smooths out these large variations, resulting in a more stable and reliable estimate of the expected return under the target policy π.

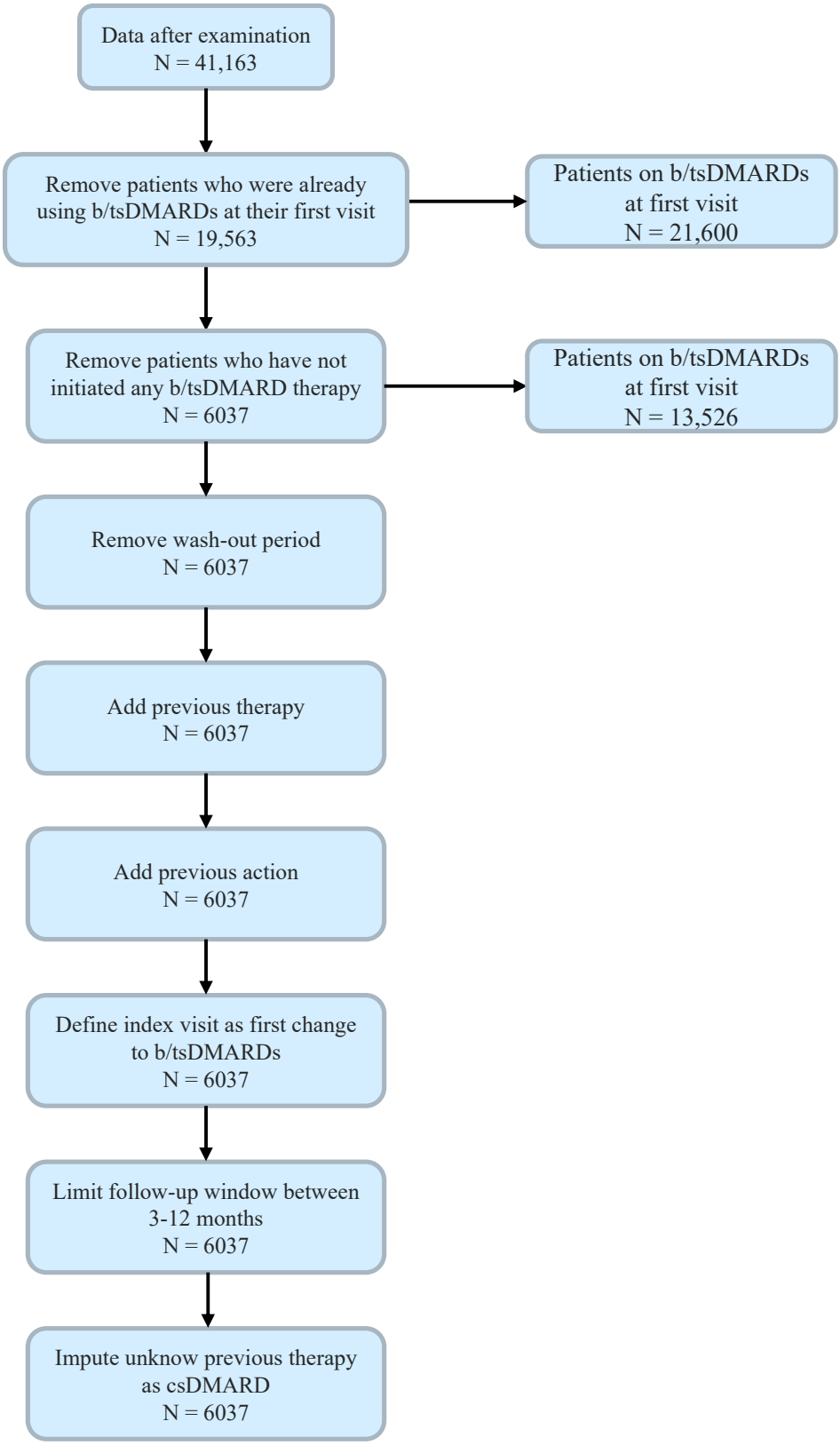


Figure 4.5: The flowchart shows the data preprocessing pipelines and the number of patients remaining after each data processing step. The notation N remains unchanged for several steps, as the number of patients is still 6037 after these steps.

5

Model Behavior Policy

In this chapter, we present the experimental results and analysis for two approaches to modeling the behavior policy. The first approach directly uses machine learning models to make predictions. The second approach, based on the divide and conquer idea, uses our proposed two-stage combined model for predictions. For each approach, we provide examples of interpretable machine learning models used in that approach, with a particular focus on decision trees. Decision trees are highlighted because they effectively divide patients into groups and describe the decision paths for those groups. We also explain the decision patterns described by the decision trees, perform sanity checks, and evaluate their clinical relevance.

5.1 Direct Multi-Class Prediction

The first approach for modeling behavior policy is directly using machine learning models that support multi-class classification to predict which therapy a RA patient switch to based on historical data. In this method, we use all the machine learning models listed in the Table 4.1 that support multi-class classification, including interpretable models such as BRS, RT, DT, and LR, the baseline dummy model, and black-box models such as XGB and RF. All models were trained using the same data splits, and the parameters of each model were optimized through 10-fold cross-validation and random search. We use accuracy, AUCROC and SCE as the evaluation metrics. The performance of the models on the validation set is shown in Table 5.1.

Table 5.1: Model performance for predicting RA patient therapy, average over 10 different data splits.

	Black-Box Models		Interpretable ML Models					
Metric	**XGB**	**RF**	**BRS**	**RT**	**DT**	**LR**	**Dummy**	**Combined**
Accuracy	0.709	0.587	0.639	0.651	0.611	0.710	0.421	0.645
AUC	0.853	0.786	0.776	0.867	0.845	0.843	0.500	0.837
SCE Score	0.013	0.076	0.082	0.049	0.048	0.007	0.005	0.032

Although linear models (LR), decision sets (BRS) and decision trees (RT and DT) all provide interpretable predictions, linear models only show the importance weights of features, as illustrated in Table 5.4 and Figure 5.5. Decision sets categorize patients into multiple groups based on rules, but do not offer complete decision

paths, as shown in Table 5.5. However, decision trees present full decision paths, which naturally places patients into specific bucket, making the decision-making process easier to understand and interpret. Here, we focus use decision trees to explain the decision-making process of physicians for different RA patients.

In decision trees, the nodes represent different criteria for making decisions, while the paths represent different decision-making processes, and the leaf nodes correspond to patient groups that follow these decisions. By using DT to predict patient therapy, we can generate decision trees shown in Figure 5.3. We focus on explaining the decision paths for DT because it is a standard and classic decision tree, while RT is more advanced and state-of-the-art.

5.1.1 Explain Patterns in Decision Tree

Figure 5.3 shows the complete decision tree of the DT model used to predict therapy for RA patients. The paths from the root node to the leaf nodes illustrate the decision making process. Each node shows the probabilities of switching to different therapies under different conditions, along with the therapy corresponding to the highest probability. Figure 5.4 shows a subtree of the DT model focusing on patients who have switched therapies. This subtree collapses the nodes predicting patients who did not switch therapies and keeps the leaf nodes for those who did. The decision process is detailed as follows:

The overall interpretation of this DT is that patient's most often stay on the same treatment, which is a drawback with using a single tree to predict therapy, because we want the DT to show the switch patterns. The first decision path begins by assessing whether the patient's previous therapy was Abatacept monotherapy. If it was, the model predicts that the patient will continue on Abatacept monotherapy, indicating no change in therapy. If the patient's previous therapy was not Abatacept monotherapy, the model then checks whether the previous therapy was IL-6Ri combination therapy. For patients who were previously on IL-6Ri combination therapy, the prediction is that they will continue on IL-6Ri combination therapy, again indicating no switch. The model then considers patients who were previously on Abatacept combination therapy, JAKi monotherapy, or JAKi combination therapy. Similar to the previous decisions, the model predicts that these patients will continue on their respective therapies, highlighting a pattern of non-switching for these specific therapies.

For patients whose previous therapy was TNFi monotherapy, the decision tree evaluating the patients CDAI. If the CDAI is 9.35 or lower, the probability of continuing TNFi monotherapy is relatively high at 0.66. However, if the CDAI exceeds 9.35, the probability of continuing with TNFi monotherapy significantly drops to 0.32. This indicates that higher disease activity makes continuing TNFi monotherapy less likely. In cases where the previous therapy was TNFi combination therapy, the model first assesses the CDAI. If the CDAI is 4.95 or lower, the decision further depends on whether the visit occurred before or after 2015. Visits before 2015 show a 0.43 probability of continuing TNFi combination therapy, whereas visits after 2015 increase this probability to 0.55. If the CDAI is greater than 4.95, the decision tree

then examines the patient's history with b/tsDMARDs. Patients with a history of b/tsDMARDs have a 0.27 probability of continuing TNFi combination therapy, compared to a lower 0.14 probability for those without such a history.

If the patients previous therapy was csDMARD, the decision path again considers the history with b/tsDMARDs. Patients with a b/tsDMARDs history are predicted to continue with csDMARD. Conversely, those without such a history are predicted to switch to JAKi combination therapy. For patients whose previous therapy does not match any of the mentioned therapies, the model predicts switching to Rituximab monotherapy. If the patient had no previous therapy, the model suggests that no treatment should be continued.

5.1.2 Sanity Checks and Clinical Relevance

It is reasonable to make decisions based on these three features. Previous therapy is the most important evaluation criterion because physicians need to understand the patients treatment history and adjust the current treatment plan based on the patients response to previous treatments. Understanding a patients previous therapy helps physicians assess which treatments have been tried, which ones were effective, and which ones may have failed, thereby formulating a more effective treatment plan.

CDAI, as a key indicator of disease activity, is also a reasonable evaluation feature. Different CDAI values represent varying levels of disease severity, so it is important and reasonable to consider the CDAI when selecting a treatment plan to ensure it effectively controls the disease. For example, a higher CDAI may require stronger or more aggressive treatment strategies, while a lower CDAI might allow for maintaining the current therapy. In the decision paths we previously mentioned, when CDAI is greater than 9.35, the probability of a patient continuing with TNFi combination therapy is lower. This is because a CDAI between 2.8 and 10 indicates a low disease activity, while a CDAI between 10 and 22 indicates a moderate disease activity, meaning the patients condition is more severe. This implies that the patient may need stronger or more aggressive treatment strategies, thus reducing the likelihood of continuing with the current therapy.

Having a history of b/tsDMARDs treatment is also an important evaluation feature because these drugs are usually used for patients who do not respond well to conventional treatments, and this is also an indicator for the first time step. Knowing whether a patient has used b/tsDMARDs and their effects can help physicians choose the appropriate treatment when formulating a new plan. For example, as explained in the decision tree, this might involve using TNFi combination therapy or JAKi combination therapy.

Clinical relevance. The interpretability of the DT is crucial for clinical practice. By identifying key factors influencing treatment decisions, such as previous therapy, CDAI, and history of b/tsDMARDs, the model can help clinicians make data-driven choices. However, the model's decision paths lack clinical value. Specifically, as shown in Figure 5.3, out of the 15 paths in the decision tree, only two predict

therapy switches, while the remaining 13 predict no change in therapy. This is because directly using the model to handle complex problems struggles to solve two problems simultaneously: predicting whether patients will switch therapies and predicting which therapy they will switch to. The model's performance indicates that its decision paths favor continuing current therapies and fail to capture more complex switching patterns in the data within small tree. We use a pruned decision tree to maintain simplicity and interpretability. While an unpruned tree increases complexity and provides more leaf nodes predicting switches, its clinical value is limited. We do not want clinicians to have to search through a large, complex tree for switch paths. Instead, we want a decision tree that is straightforward and easy to interpret. As most patients do not switch therapies, the model performs well on evaluation metrics. Although the DT provides reasonable explanations and good performance, predicting therapy switches within a small tree is more important for our project. Therefore, despite its strong interpretability, the model lacks sufficient clinical value in practical applications.

5.2 Divide and Conquer Two-Stage Prediction

The second approach for modeling behavior policy adopts the idea of "divide and conquer", we propose a two-stage combined model to predict which therapy a RA patient switch to. Unlike directly using a therapy model to predict therapies, our proposed combined model predicts therapies by combining the outputs of two separate models, as shown in Figure 5.1. This model composes two stages:

- **Switch Decision Stage:** This stage determines whether a patient will switch therapies using a switch model that predicts the probability of switching.
- **Therapy Selection Stage:** For patients predicted to switch, this stage identifies the specific therapy they will switch to, using a therapy model that predicts the probability of switching to each alternative therapy (excluding the current one).

By combining these stages, we calculate the probability of switching to each therapy as follows:

$$\Pr(\text{therapy}) = \Pr(\text{stay}) + \Pr(\text{switch}) \times \Pr(\text{switch to each therapy}) \tag{5.1}$$

5.2.1 Switch Prediction

In the first stage of the combined model, we want to answer the question of whether the patient will switch their therapy. We use all the models in Table 4.1 to model the patient's therapy switching patterns, including interpretable machine learning models, linear models and black-box models. In the therapy switching task, the models perform as shown in Table 5.2.

To evaluate model performance, we focus on AUCROC and AUC-PR. AUC represents the overall performance of the model, taking into account its ability to identify

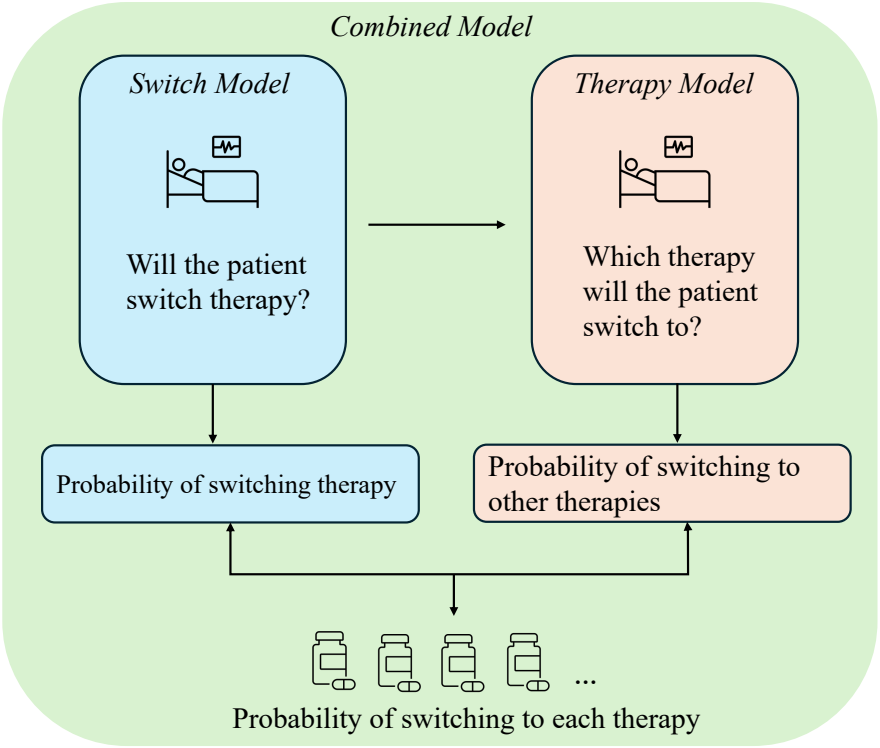


Figure 5.1: The structure of the combined model.

Table 5.2: Model performance for predicting whether RA patients will switch their therapy.

	Black-Box Models		Interpretable Machine Learning Models							
Metric	**XGB**	**RF**	**RE**	**BRS**	**RL**	**RT**	**DT**	**RS**	**LR**	**Dummy**
Accuracy	0.791	0.763	0.778	0.788	0.768	0.786	0.777	0.774	0.790	0.571
AUC	0.853	0.837	0.850	0.848	0.827	0.846	0.840	0.821	0.849	0.500
AUC-PR	0.859	0.842	0.856	0.856	0.847	0.855	0.848	0.845	0.856	0.714
ECE Score	0.017	0.158	0.149	0.209	0.010	0.044	0.045	0.035	0.013	0.009
Brier Score	0.142	0.180	0.174	0.199	0.151	0.146	0.150	0.152	0.144	0.245

both positive cases (switch therapy) and negative cases (stay on current therapy). The closer the AUC value is to 1, the better the classification ability of the model across different thresholds. We are particularly interested in the accuracy of the model in predicting therapy switching, i.e. label 1. Switching a patient's therapy requires careful decision making by the physician, making AUC-PR a value metric to examine. A high AUC-PR indicates that the model is good at identifying true positives among the positive predictions. In this context, a high AUC-PR means that the model effectively balances sensitivity (recall) and precision, providing reliable predictions that can support clinical decisions.

Based on the results of Table 5.2, we can see that both black-box models and interpretable machine learning models perform better than the baseline model (Dummy). Although the decision sets models (RE and BRS) perform the best among the interpretable models, they have higher calibration errors. Overall, XGB stands out

across all metrics, particularly in terms of overall model performance and calibration error.

5.2.2 Examples of Switch Models

To predict whether a patient will switch therapy, Figures 5.7, 5.2, and Table 5.6 show the use of decision sets (BRS), decision trees (DT) and decision lists (RL), respectively. Tables 5.4 and 5.5, and Figure 5.5 demonstrate the use of linear models LR, RE, and RS, respectively. Through this binary classification example, we can see that while linear models RE, LR, RS demonstrate the importance of features in predicting outcomes, they fail to classify patients into different groups based on decision paths. BRS and RS use rule sets and rule lists to classify patients into different groups, but these groups may overlap. Decision trees, on the other hand, can categorize patients into non-overlapping groups through clear decision paths, making decision trees an ideal model for describing switching patterns in RA patients. Next, the decision paths provided by the decision tree are explained in more detail.

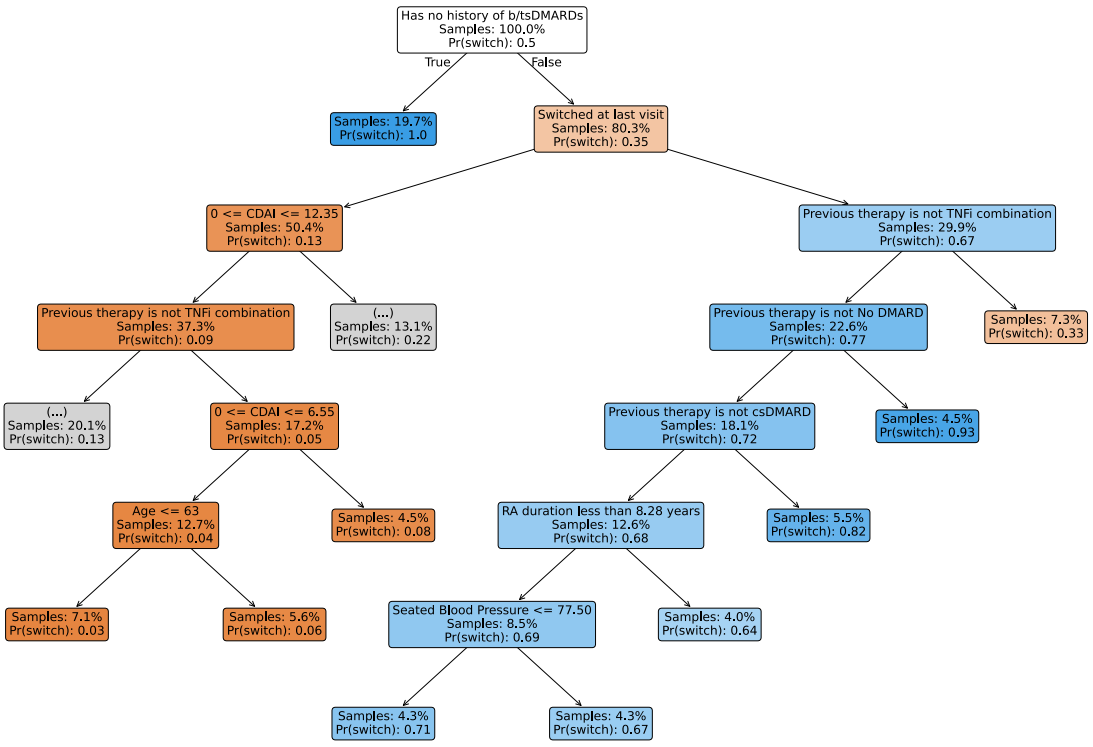


Figure 5.2: An example of using a decision tree to predict whether patients will switch therapy.

Figure 5.2 shows a subtree of the decision tree, where the full decision tree is shown in Appendix, Figure A.1. This decision tree indicates that if a patient has no history of b/tsDMARD treatment, they will most likely switch therapy. This is consistent with the patient cohort selection criteria and the index visit definition, where all selected patients have no history of b/tsDMARD treatment and the index visit marks their first switch to b/tsDMARD.

In the decision tree, orange nodes represent patient groups with a low likelihood of switching. Specifically, these patients have a history of b/tsDMARD treatment, did not switch therapy at their previous visit, are currently on TNFi combination therapy, and are in remission (as determined by CDAI values). For these patients, the probability of switching therapies is 0.04. In addition, the likelihood of switching is also influenced by age. Older patients may be more sensitive to drug side effects or may have other health problems that could lead to a higher tendency to switch therapies.

Blue nodes represent patient groups with a high probability of switching. Specifically, patients with a history of b/tsDMARD treatment who switched therapies at their previous visit and were either on no DMARD or csDMARDs therapy at their previous visit have a probability of switching of 0.93 and 0.82, respectively. These patients were likely in a wash-out period, designed to eliminate the effects of previous therapies to help switch to next therapy. For patients with a history of b/tsDMARD therapy who switched at their previous visit, were not in a wash-out period, and had RA less than 8 years, the probability of switching is 0.69. This reflects the current challenge for RA patients to try multiple therapies early in their disease to determine which is most effective for them.

5.2.3 Therapy Prediction

In the combined model, we can use any of the switch models from Table 5.2 to predict switching. As previously mentioned, the decision tree can naturally group patients based on decision paths, making it an ideal model for describing the treatment patterns of RA patients. Therefore, we use a decision tree as the estimator for predicting therapy.

To evaluate the impact of different switch estimators on the overall performance and interpretability of the model, we conducted two sets of experiments. Using the same data split, the therapy estimator was consistently a decision tree. We compared the best performing black box model, XGB, as the switch estimator with a more interpretable decision tree as the switch estimator. The overall AUC and accuracy of the propensity model, along with the AUC of the two estimators in the experiments, are shown in Table 5.3.

Table 5.3: Evaluation metrics of two different combined models under the same data split.

Metric	**Combined Model 1 (XGB+DT)**	**Combined Model 2(DT+DT)**
Accuracy	0.646	0.652
AUC	0.836	0.821
Switch Model AUC	0.813	0.831
Therapy Model AUC	0.685	0.667
SCE	0.032	0.032

In the two sets of experiments, Figure 5.10 shows a subtree of the therapy model from the first combined model experiment, where the black box XGB serves as the switch model, while a decision tree is used as the therapy model. Figure 5.9 and

Figure 5.8 show the full decision tree and its subtree, respectively, from the second combined model experiment, where decision trees are used for both the switch and therapy models.

5.2.4 Explain Patterns in Combined Model

The sub decision tree in Figure 5.8 can be explained based on whether the patient has previously used b/tsDMARDs:

For patients who have never used b/tsDMARDs:

- Generally, most patients' first use of b/tsDMARDs is TNFi combination therapy if treatment starts before 2018. If treatment starts after 2018, most patients' first use of b/tsDMARDs is JAKi combination therapy.

TNFi combination therapy is widely applied between 2012-2014. For patients treated outside 2012-2014:

- If the previous therapy is a csDMARD and the patient is older (71-94 years), Abatacept combination therapy is used. If the patient is younger than 71 years and treated in recent years (2018-2022, our data through 2022), JAKi combination therapy is used. If treated before 2018, TNFi combination therapy is used.

- If the previous therapy is not csDMARD and the patient has high disease activity, TNFi combination therapy is used. If the patient does not have high disease activity, JAKi monotherapy is used.

For patients who have previously used b/tsDMARDs:

- If the previous therapy is TNFi combination or monotherapy, the patient switches to csDMARD. If the previous treatment is csDMARD, the patient stops the therapy process.

- If the previous therapy does not belong to any of the above and the therapy is processed between 2021-2022, JAKi monotherapy is used.

- If the previous treatment is not one of the above and the therapy is not performed between 2021-2022, IL-6Ri monotherapy is used if the patient has no history of other diseases. If the patient has a history of other diseases, Abatacept monotherapy is used.

The sub-decision tree in Figure 5.10 shares similar decision patterns with the decision tree in Figure 5.8. Both decision trees primarily base their decisions on previous therapy and CDAI. Additionally, the therapy decision tree in the first combined model considers the patient's pain and fatigue. Given that combined model 2 performs similarly to combined model 1 and is composed of two decision trees, making it more interpretable, we will focus on explaining combined model 2.

5.2.5 Sanity Checks and Clinical Relevance

The therapy model in the first combined model (DT+DT) shown in Figure 5.8, which bases its decisions on the patient's previous therapy, therapy year, CDAI, and comorbidity history, is reasonable. As mentioned in Section 5.1.2, understanding a patient's previous therapy is a sound and essential factor in determining therapy changes. Clinically, knowing a patient's treatment history helps clinicians assess the patient's response to previous therapies and allows for more appropriate treatment decisions.

Because JAKi therapy was developed in 2017, the decision tree accurately captures the increased use of JAKi drugs since 2018. Before 2018, TNFi treatments were widely used, but with the advent of JAKi, their clinical use has increased significantly.

The patient's CDAI level reflects the current severity of the patient's RA. The decision tree in Figure 5.8 appropriately determines therapy based on whether the patient has high disease activity, which indicates severe RA. Severe disease requires more robust treatments, and TNFi are highly effective in controlling high disease activity in RA due to their ability to inhibit multiple inflammatory pathways. For patients with lower CDAI levels, indicating less severe disease, JAKi monotherapy is a good therapy to try, as it can reduce unnecessary medication burden and potential side effects.

It also makes sense to base therapy decisions on the patient's comorbidity history. Clinically, physicians need to consider comorbidities to better individualize treatment plans and select the most appropriate therapies for each patient's situation. By understanding a patient's comorbidity history, clinicians can avoid therapies that may exacerbate existing conditions and select those with the best efficacy for the individual, thereby improving safety and overall treatment outcomes. The decision tree in Figure 5.8 takes into account whether the patient has a history of other conditions, including psoriasis, which refers to the patient's history of psoriasis; depression, which indicates a history of depressive symptoms; fibromyalgia, referring to a history of widespread muscle pain and fatigue; other neurological conditions, indicating a history of various neurological disorders; hospitalized bleeding events, which refer to a history of severe bleeding requiring hospitalization; non-hospitalized bleeding events, indicating a history of milder bleeding events not requiring hospitalization; and other conditions, which refer to the patient's history of other unspecified health issues.

These comorbidities are relevant to RA treatment because they affect the choice and effectiveness of therapy. For example, certain RA therapies may affect psoriasis. Fibromyalgia often coexists with RA, and recognizing this helps to optimize treatment. Neurological conditions may be exacerbated by certain RA therapies, requiring careful selection. Patients with a history of bleeding require special attention due to the increased risk posed by some therapies. For patients with a history of these comorbidities, the decision tree shows that IL-6Ri monotherapy is often preferred. IL-6 plays a role in various inflammatory and immune responses and has been implicated in conditions such as depression, fibromyalgia and psoriasis. IL-6R

inhibitors (IL-6Ri) can effectively control inflammation and are well tolerated in various patient populations. For patients without a history of these comorbidities, abatacept monotherapy is typically chosen due to its reduced side effects and stable efficacy, making it more suitable for these patients.

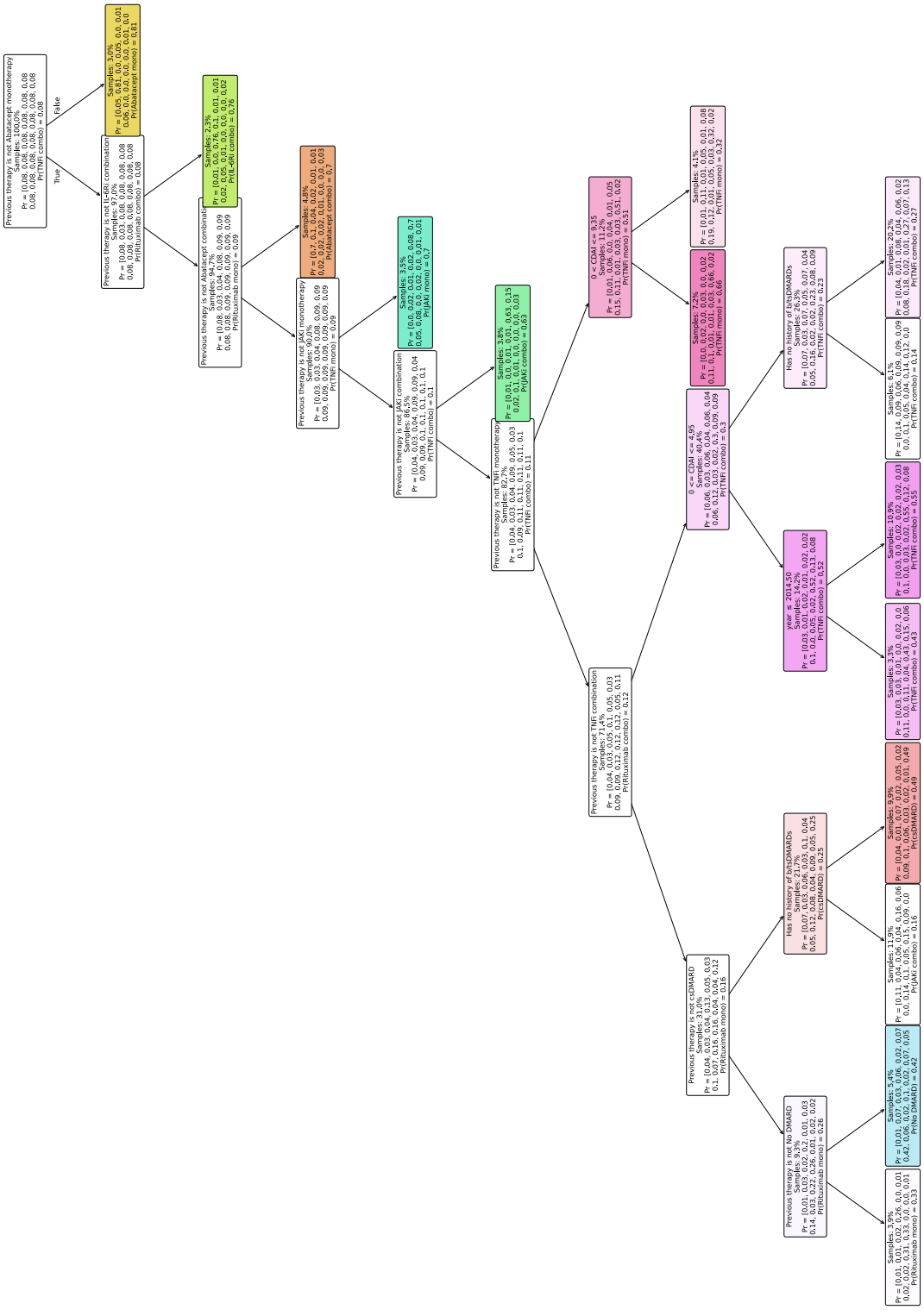


Figure 5.3: A complete decision tree to directly predict therapy for RA patients. The probability list corresponds to the probability of using the following therapies, respectively: Abatacept combo, Abatacept mono, IL-6Ri combo, IL-6Ri mono, JAKi combo, JAKi mono, No DMARD, Other, Rituximab combo, Rituximab mono, TNFi combo, TNFi mono, and csDMARD.

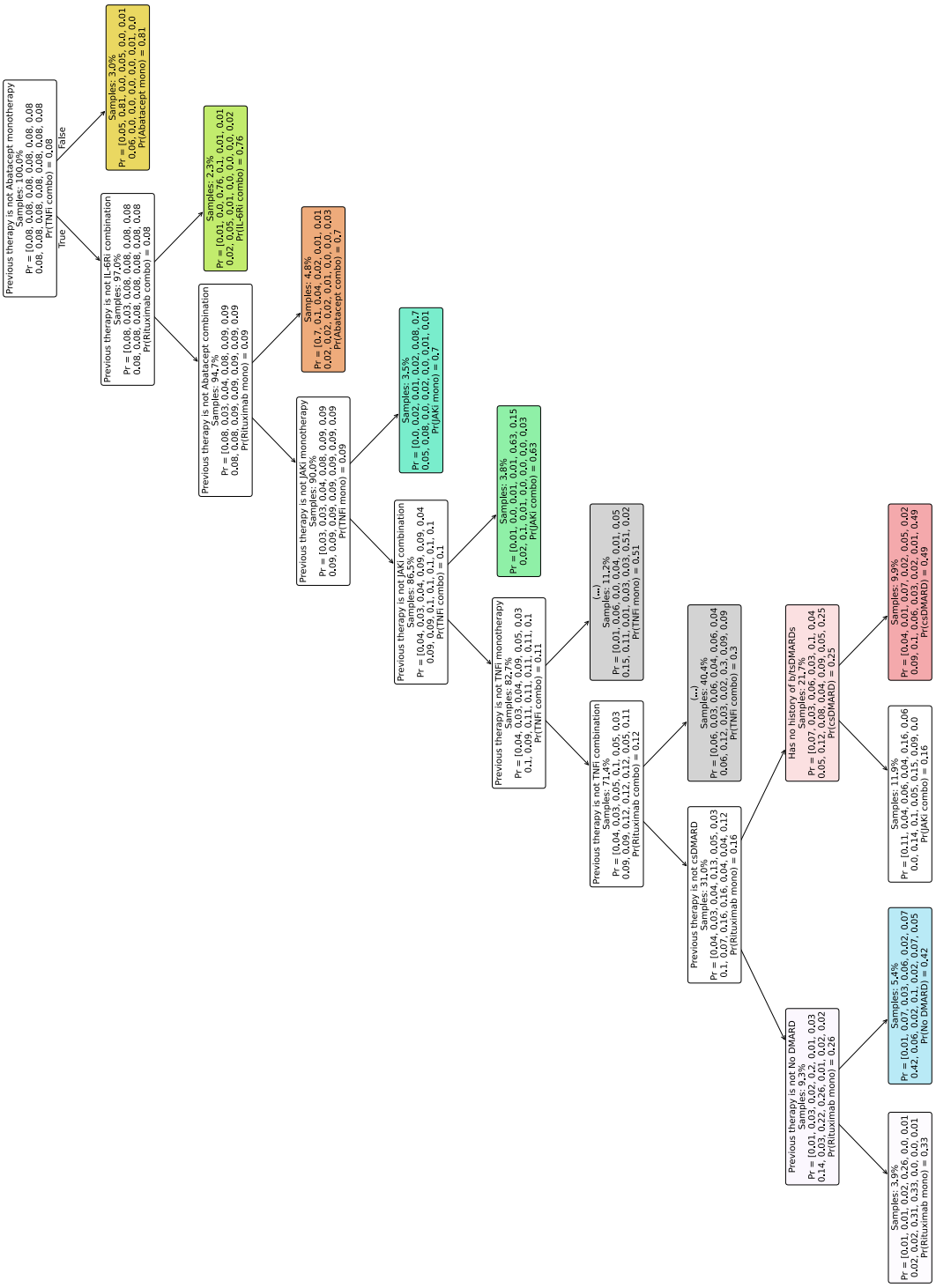


Figure 5.4: A sub-decision tree to directly predict therapy for RA patients. This sub-tree collapses all leaf nodes representing patients who did not switch therapies. The probability list corresponds to the probability of using the following therapies, respectively: Abatacept combo, Abatacept mono, IL-6Ri combo, IL-6Ri mono, JAKi combo, JAKi mono, No DMARD, Other, Rituximab combo, Rituximab mono, TNFi combo, TNFi mono, and csDMARD.

Table 5.4: The LR model lists the top five features with the highest positive coefficients and the top five features with the largest absolute negative coefficients, highlighting the most critical positive and negative contributors in the model.

Model	Rule	Coefficient
LR	Previous therapy is other	1.633
	Switched at last visit	0.984
	Comorbidity of GI or liver	0.906
	Comorbidity of drug-induced reactions	0.800
	Comorbidity of metabolic	0.747
	...	...
	History of b/tsDMARDs	−6.871
	Previous therapy is TNFi combination therapy	−0.533
	No insurance	−0.475
	Previous therapy is IL-6Ri monotherapy	−0.367
	$0 \leq$ CDAI ≤ 2.5	−0.351

Table 5.5: The RE model presents a set of rules ranked by importance, where importance indicates the contribution of linear items and rule items to the prediction of whether to switch treatment.

Rule	**Coeff.**	**Impor.**
History of b/tsDMARDs	2.860	1.141
History of b/tsDMARDs AND $20.7 \leq$ CDAI ≤ 76 AND Previous therapy is csDMARDs	−0.416	0.205
History of b/tsDMARDs AND $20.7 \leq$ CDAI ≤ 76 AND Switched at last visit	−0.335	0.164
History of b/tsDMARDs AND CCP Positive AND Switched at last visit	−0.258	0.129
History of b/tsDMARDs AND $70 \leq$ Pain ≤ 100 AND $4.47 \leq$ DAS ≤ 8.8 AND Switched at last visit AND CCP Positive AND Infections	−0.240	0.114
History of b/tsDMARDs AND $44 \leq$ Pain ≤ 70 AND Didn't use therapy at last visit	−0.169	0.082
History of b/tsDMARDs AND Previous therapy is TNFi combination therapy AND Switch at last visit	0.202	0.082
History of b/tsDMARDs AND Previous therapy is TNFi combination therapy	−0.168	0.080
History of b/tsDMARDs AND $70 \leq$ Fatigue ≤ 100 AND Switched at last visit	−0.114	0.055
$70 \leq$ Pain ≤ 100	0.059	0.024
Previous therapy is csDMARDs	0.051	0.021
History of b/tsDMARDs AND $0 \leq$ CDAI ≤ 2.5	−0.023	0.009
$11.5 \leq$ CDAI ≤ 20.800	0.018	0.007
History of b/tsDMARDs AND $70 \leq$ Pain ≤ 100 AND Switched at last visit	−0.004	0.002

	Conditions		Probability	Support
IF	History of b/tsDMARDs	THEN switch prob. is:	100%	3241
IF	Switch at last visit	THEN switch prob. is:	42.4%	5209
IF	$20.7 \leq$ CDAI ≤ 76	THEN switch prob. is:	34.4%	1071
IF	$11.5 \leq$ CDAI ≤ 20.8	THEN switch prob. is:	26.9%	1407
IF	Previous therapy is csDMARDs	THEN switch prob. is:	24.0%	533
IF	Have comorbidity of other disease	THEN switch prob. is:	26.8%	157
IF	Didn't use therapy at last visit	THEN switch prob. is:	25.5%	184
IF	Have comorbidity of cancer	THEN switch prob. is:	36.6%	41
IF	$2.5 \leq$ CDAI ≤ 6	THEN switch prob. is:	16.3%	1378
IF	Pregnant	THEN switch prob. is:	62.5%	8
ELSE		THEN switch prob. is:	11.49%	3063

Table 5.6: The RL model presents a list of rules ranked by probability of switching therapy.

1. $20.8 \leq$ CDAI ≤ 76	1 point	+	...
2. Switched therapy at last visit	1 point	+	...
3. Previous therapy is TNFi combination therapy	-1 point	+	...
4. History of b/tsDMARDs	-5 points	+	...
	SCORE	=	...

SCORE	-6	-5	-4	-3	-1	0	1	2
RISK	11.9%	26.9%	50%	73.1%	95.3%	98.2%	99.3%	99.8%

Figure 5.5: An example of using RS to predict whether patients will switch therapy.

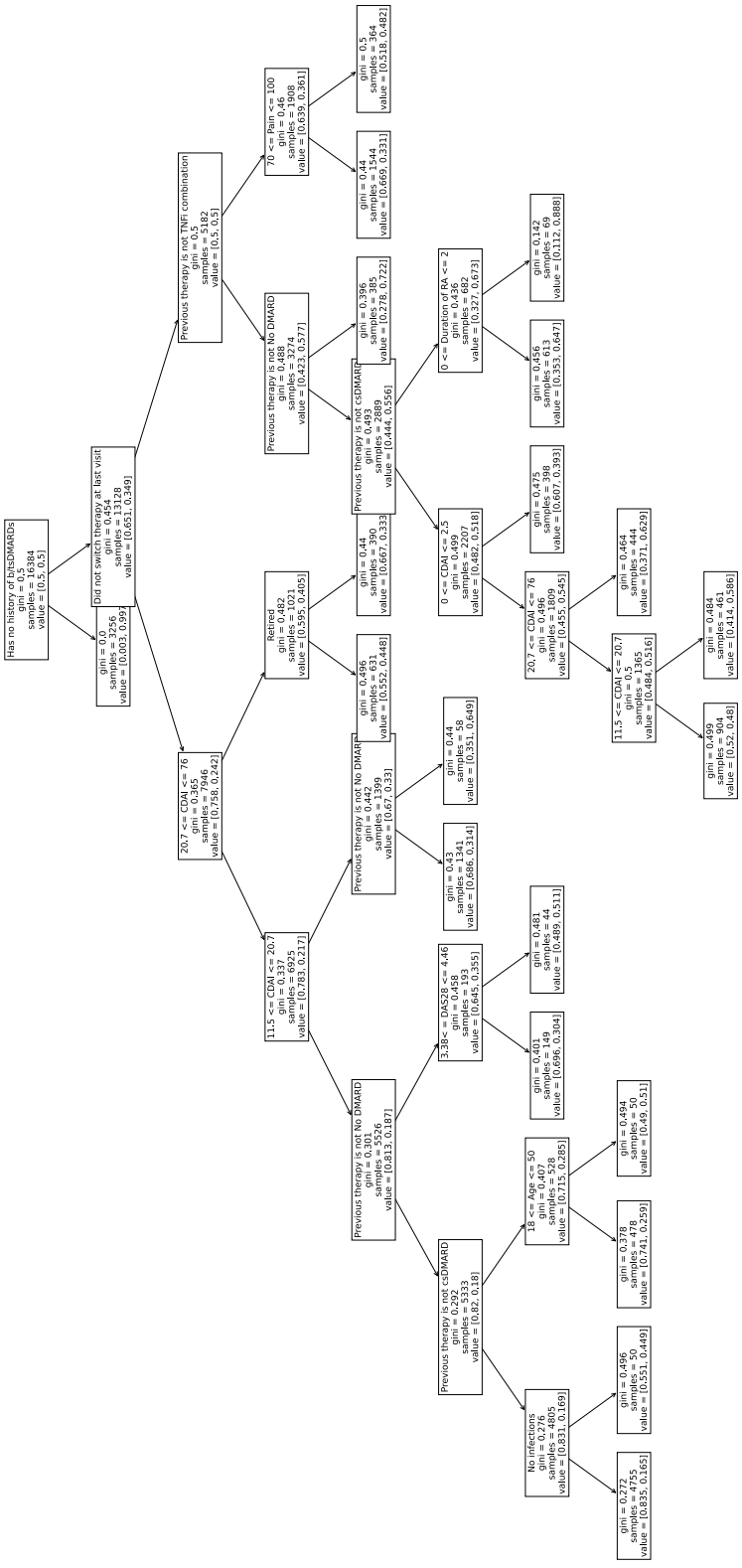


Figure 5.6: The decision tree from RT is used to predict whether patients will switch therapy.

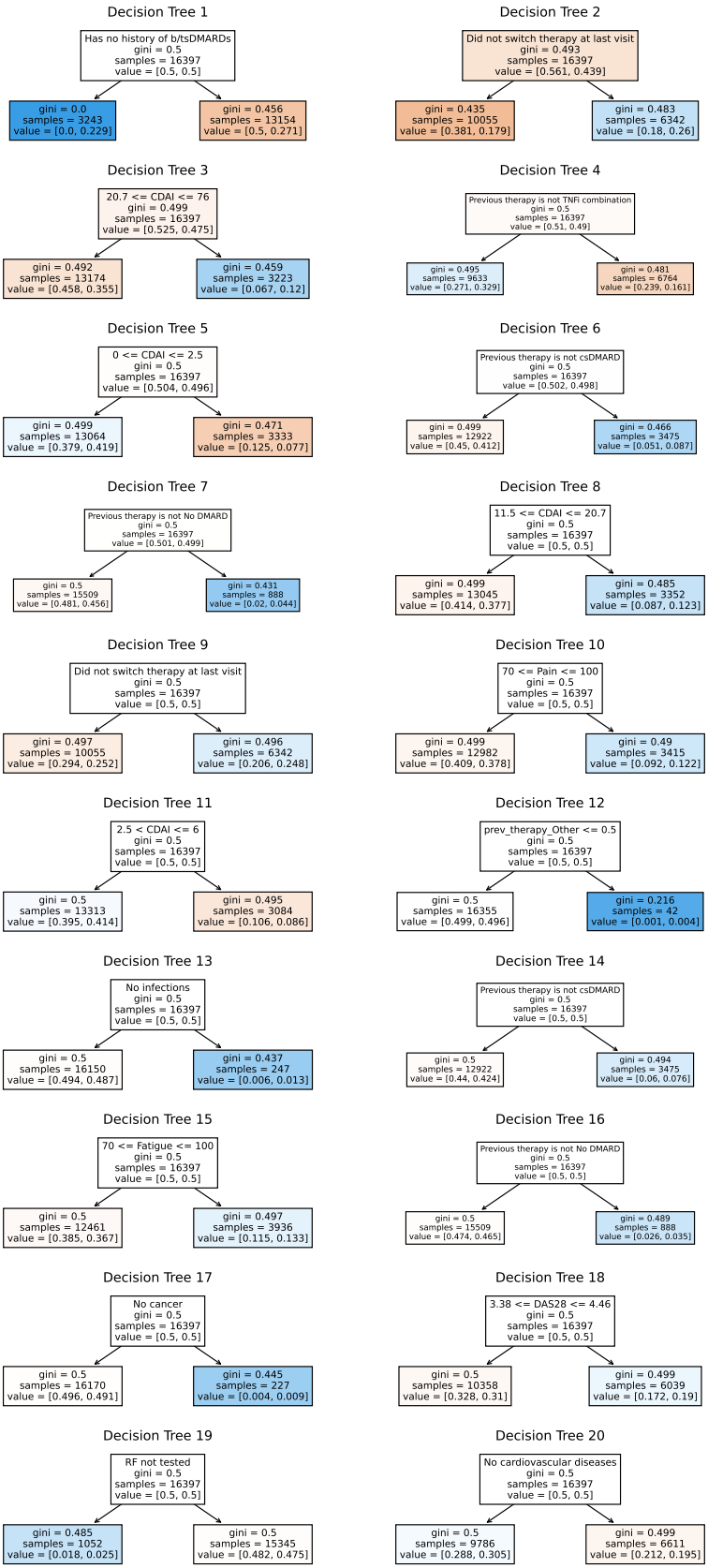


Figure 5.7: Decision trees from the BRS, where each sub-tree represents a rule.

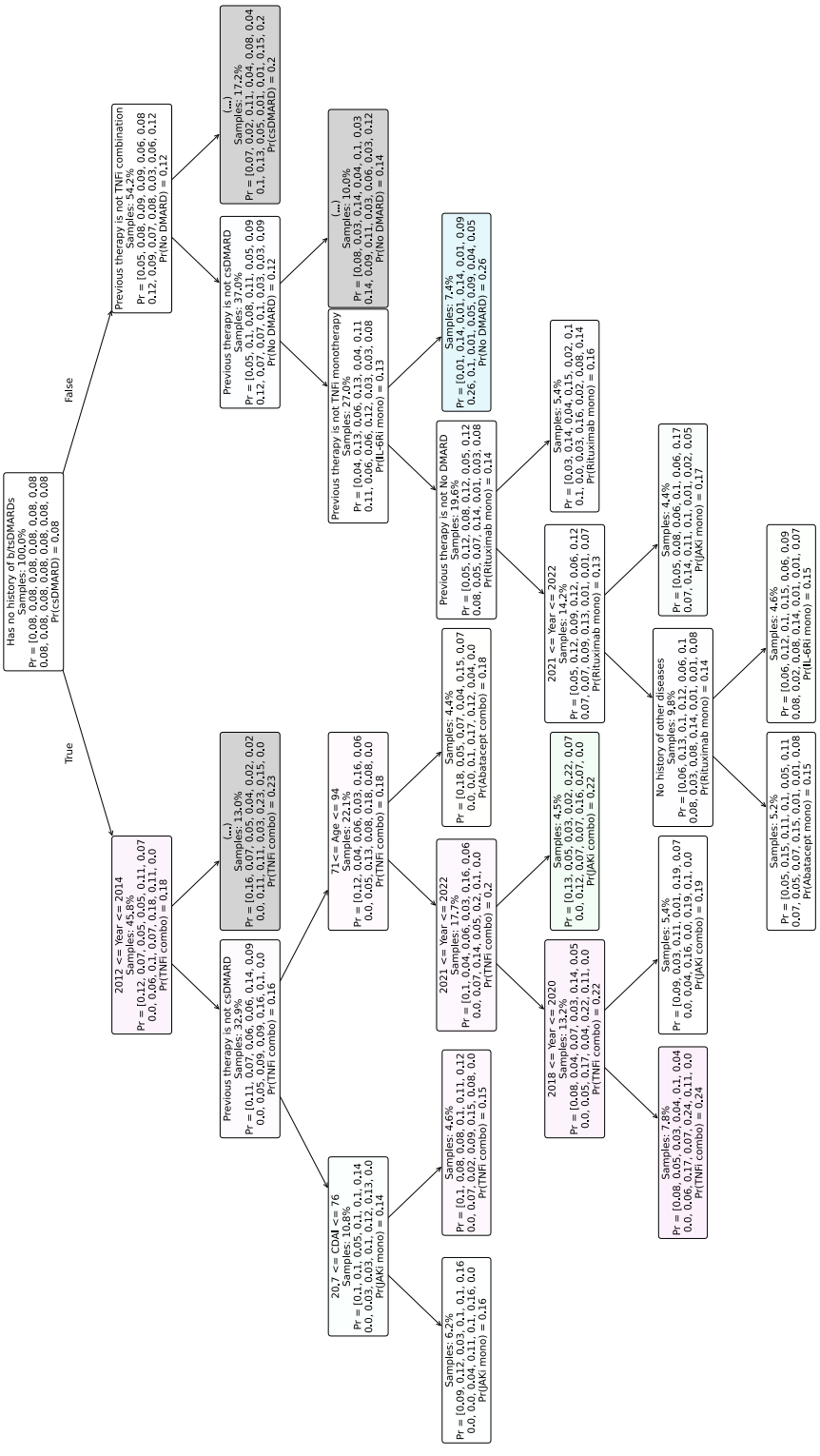


Figure 5.8: The sub-decision tree of the therapy model in the combined model 2 (DT + DT) for predicting therapy for RA patients. The probability list corresponds to the probability of using each of the following therapies: Abatacept combo, Abatacept mono, IL-6Ri combo, IL-6Ri mono, JAKi combo, JAKi mono, no DMARD, other, Rituximab combo, Rituximab mono, TNFi combo, TNFi mono, and csDMARD.

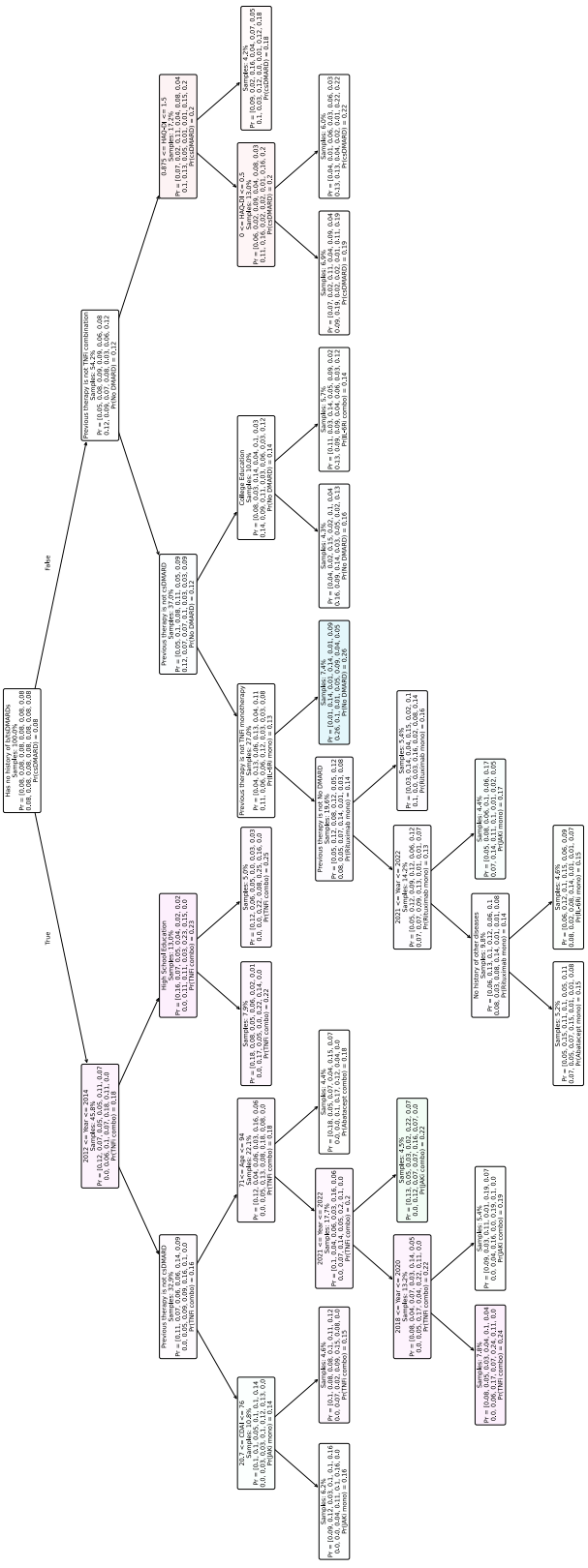


Figure 5.9: The complete decision tree of the therapy model in the combined model 2 (DT + DT) for predicting therapy in RA patients.

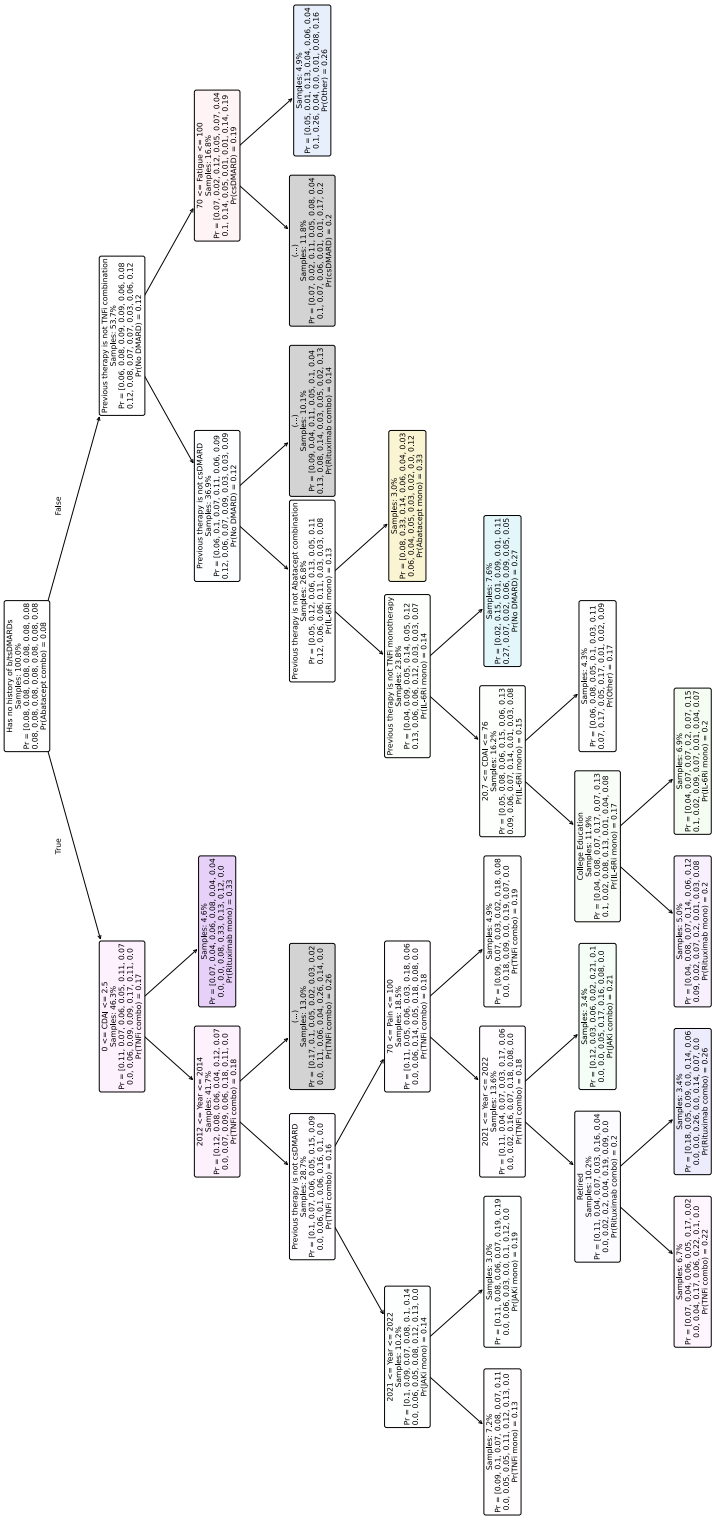


Figure 5.10: The sub-decision tree of the therapy model in the combined model 1 (XGB + DT) for predicting therapy for RA patients. The probability list corresponds to the probability of using each of the following therapies: Abatacept combo, Abatacept mono, IL-6Ri combo, IL-6Ri mono, JAKi combo, JAKi mono, no DMARD, other, Rituximab combo, Rituximab mono, TNFi combo, TNFi mono, and csDMARD.

6

Propose Target Policy

In this chapter, we discuss how to propose target policy based on the modeled behavior policy from Chapter 5. We also provide explanations of our proposed target policy and offer suggestions for the guidelines based on this target policy. When proposing the target policy, we use the most likely policy derived from the behavior policy as the target policy and compare it to the random policy as the baseline. We then explain the clinical significance of the proposed target policy for patients. Finally, we make recommendations for clarifying ambiguous areas in the EULAR guidelines based on the insights gained from our target policy.

6.1 Proposed Target Policy

We use the validated behavior policy obtained from the combined model 2 (DT+DT) to propose the target policy. Specifically, we consider the most likely policy in the behavior policy as the target policy and use a random policy as the baseline for comparison.

Figure 5.8 shows the behavior policy we obtained from combined model 2, with the probability list indicating the likelihood of choosing different therapies. By selecting the therapy with the highest probability (i.e., the most likely therapy), we derive the target policy, as shown in Figure 6.1. Figure A.2 presents the complete proposed target policy. The random policy serves as a baseline for comparison and is generated by randomly selecting therapies based on random seeds, without reference to the behavior policy. Since our proposed target policy is derived from the behavior policy, this means that the decisions we make have been implemented in real clinical settings. This not only provides clinical support, but also satisfies the overlap assumption of OPE. This is crucial because we do not want to implement a policy that has never been tested in clinical practice. Such a policy would lack a real-world basis and may not perform well in actual application.

6.2 Explain Target Policy

Based on Figure 6.1, our proposed target policy for predicting therapy changes in RA patients can be divided into two scenarios: patients using b/tsDMARDs for the first time and patients who have previously used b/tsDMARDs.

For patients who are new to b/tsDMARDs, our target policy recommends that they try TNFi combination therapy. If the patient's CDAI is not in the high disease activity level or the patient is younger than 71 years of age, our target policy recommends JAKi combination therapy. If the patient is older than 71 years, our target policy recommends Abatacept combination therapy.

For patients who have previously used b/tsDMARDs, if the patient is not in a wash-out period and the last treatment was not a TNFi combination therapy or monotherapy, our target policy recommends that they try Rituximab combination therapy or JAKi monotherapy. If the patient has no other medical history, as mentioned in Section 5.2.5, our target policy recommends that patients try Abatacept monotherapy. If the patient has a history of other diseases, our target policy recommends that patients try IL-6Ri monotherapy.

Suggestions to the guidelines Since our project focuses only on therapy classes, we assume that physicians select therapies from each class in the same way they normally do. Based on our proposed target policy and in accordance with the EULAR guidelines, we offer the following recommendations for the current EULAR guidelines: In phase 2, if poor prognostic factors are present, it is recommended to add a b/tsDMARD. We recommend adding TNFi therapy, while also considering the patient's CDAI level and age to assess whether JAKi therapy should be used. In Phase 3, if patients need to switch b/tsDMARDs, we recommend selecting the replacement therapy based on the patient's previous therapy and history of comorbidities.

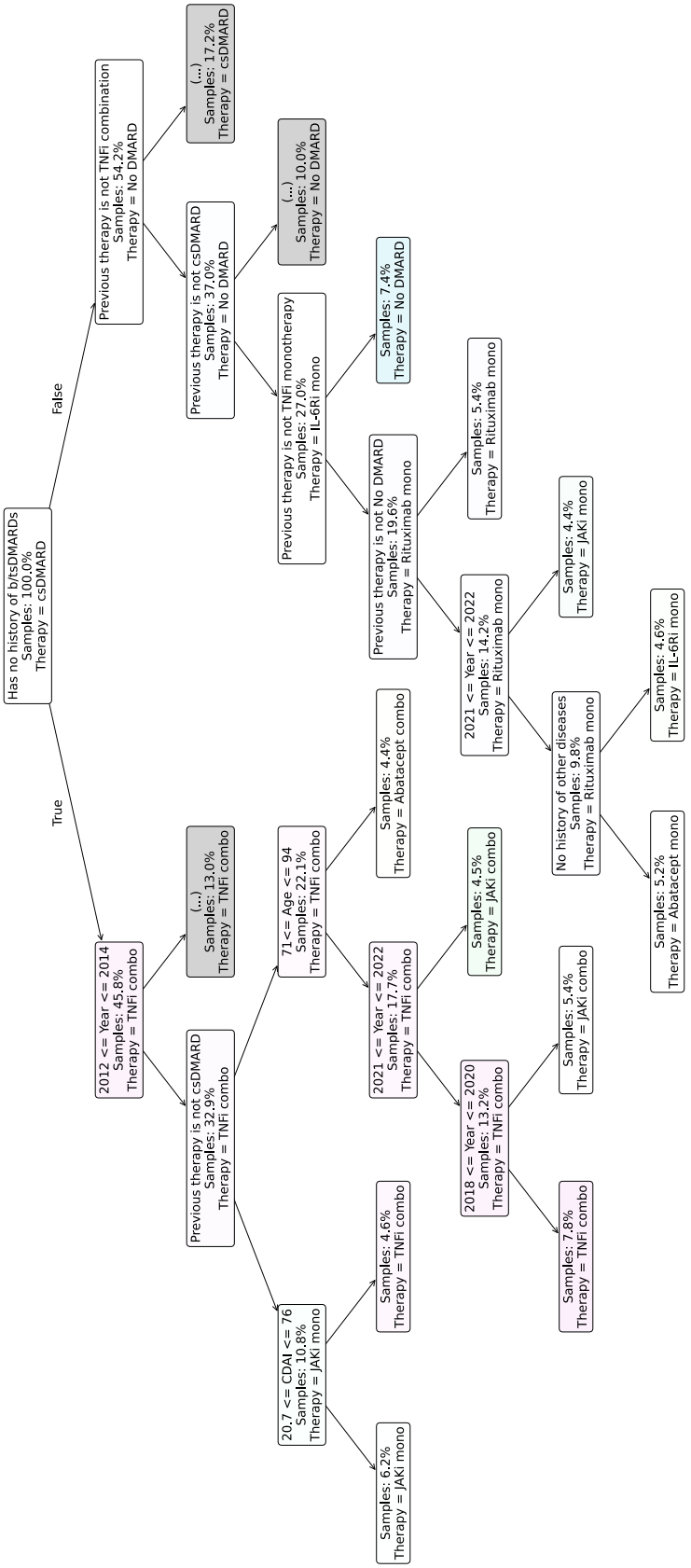


Figure 6.1: Proposed target policy based on combined model 2 (XGB+XGB).

7

Off-Policy Evaluation

While the effectiveness of the proposed target policy on the combined model is uncertain, evaluating it across different models helps us understand its robustness and performance consistency. This chapter presents the results of the WIS estimator for the **Random Policy**, which refers to a baseline policy where actions are selected randomly without any specific strategy, and the models trained in Chapter 5. The effective sample size for off-policy evaluation is 5149.

It is important to note that the probability of the Random Policy in our experiment is genuinely random, which may fail to meet the overlap assumption. The WIS estimator is a crucial tool in policy evaluation, providing a way to estimate the expected return of a target policy using data generated from a different policy. Additionally, the dummy model, which uses a "prior" strategy that it predicts the class label based on the class distribution in the training set, is not expected to produce accurate off-policy evaluation (OPE) estimates.

The WIS estimator can be significantly affected by extreme weights. Therefore, this chapter first presents the raw WIS estimator results, followed by the WIS estimator results after excluding extreme weights (with a threshold of 1000).

7.1 Raw WIS estimator

We set the rewards as $10-CDAI$, out of two reasons. One is the fact that the higher $CDAI$ is, the worse condition the patient is at while the rewards is on the contrary. So we should apply $-CDAI$ here. As for the 10 in the rewards, it's because 10 is the threshold to define low disease activity and moderate disease activity. Consequently, we use rewards as $10 - CDAI$ in the off-policy evaluation part. And the value of behavior policy is -3.4.

Figure 7.1 presents a comparative analysis of the weighted importance sampling (WIS) estimates $\tilde{V}_{WIS}(pi)$ under a Random policy across several machine learning models: combined model 2 (DT + DT), LR, dummy, RT, DT, RF, XGB and BRS. The y-axis ranges from -400 to over 100, representing the $\tilde{V}_{WIS}(\pi)$ values, while the x-axis lists the models. By splitting the sample differently for 100 times, we get the distribution of WIS values, illustrated by each box. With the boxes indicating the interquartile range (IQR), the median represented by a line within the box, and whiskers extending to 1.5 times the IQR; outliers are marked as individual points.

A red dashed line denotes the behavior policy value $\tilde{V}(\mu)$, serving as a benchmark for comparison.

The plot reveals substantial variability in performance among the models. For instance, the combined model exhibits narrow variability, indicating consistency. In contrast, the dummy model shows a wide range and high variance in its WIS estimates. Notably, some models, such as LR and RT, have medians above the behavior policy value. Conversely, models like DT and XGB have medians close to the behavior policy value.

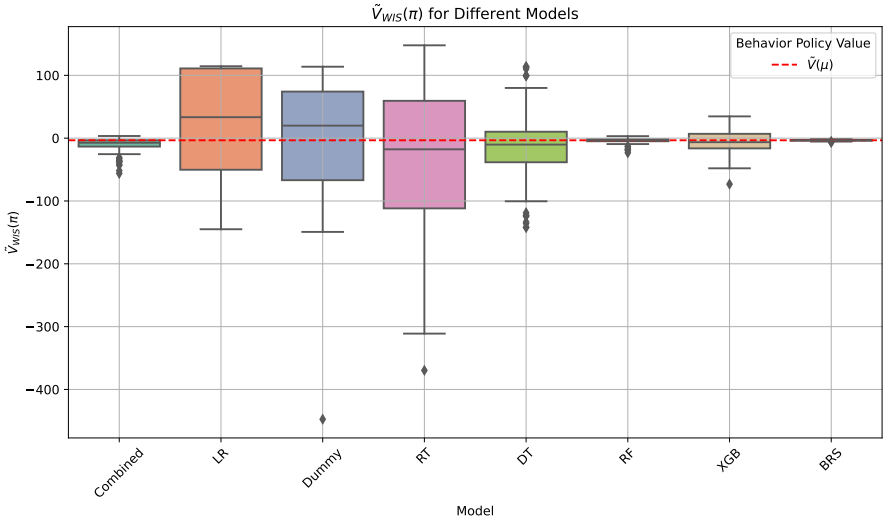


Figure 7.1: Raw WIS Estimates for Different Models under Random Policy

Figure 7.2 illustrates the comparison of weighted importance sampling (WIS) estimates $\tilde{V}_{WIS}(pi)$ under the Most Likely policy for various machine learning models.

The plot reveals variability in model performance. Models such as LR, dummy, and RT exhibit larger medians and variability, suggesting higher potential performance but with greater inconsistency. In contrast, models like combined model and XGB show narrower ranges and medians closer to the behavior policy value. Notably, BRS demonstrates the highest variance among all models, highlighting significant performance fluctuations.

7.2 Limitations with raw WIS estimator

There are limitations with the raw WIS estimator that we need to address. The first issue is the presence of extreme weights. As depicted in Figure 7.3, although most weights are relatively small (less than 100), there are three instances of extremely large weights. These extreme weights can disproportionately affect the final results. To mitigate this, we need to establish a threshold and filter out these excessive weights[50].

Another concern is censoring. Censoring occurs when information on the time to an outcome event is not available for all study participants [51]. To address this,

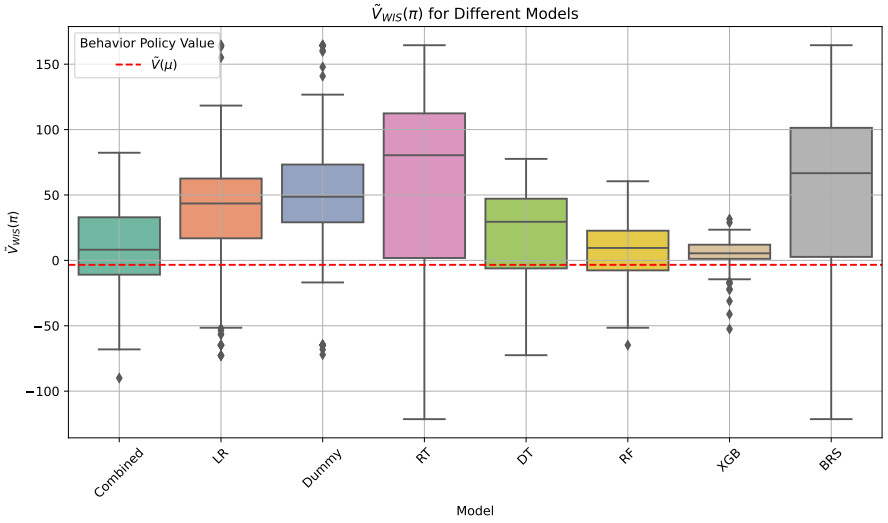


Figure 7.2: Raw WIS Estimates for Different Models under The Most Likely Policy

we should limit the period for off-policy evaluation to a specific time-frame. In our project, this period is set to three years.

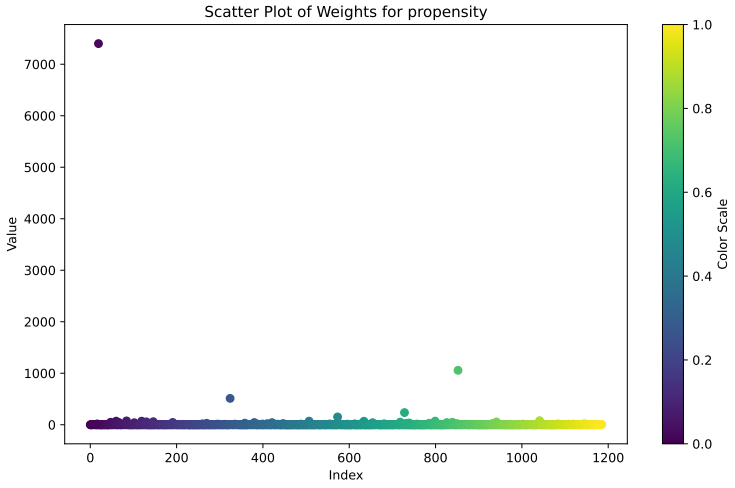


Figure 7.3: Scatter plot for propensity model under the most likely policy

7.3 Filtered WIS estimator

To address the issues of extreme weights and censoring in the raw WIS estimator, we removed weights over 1000. The results of the filtered WIS estimates are presented in Figure 7.4.

Figure 7.4 is a box plot comparing the weighted importance sampling (WIS) estimates $\tilde{V}_{WIS}(pi)$ under the Random policy for various machine learning models as we mentioned above.

The plot reveals notable differences compared to the raw WIS estimates. Models like combined model exhibit narrow variability, indicating improved consistency, while models like Dummy still display wide ranges and high variance. Some models, such as LR and RT, have medians above the behavior policy value, suggesting better performance. Others, like DT and XGB, have medians close to the behavior policy value, indicating similar performance.

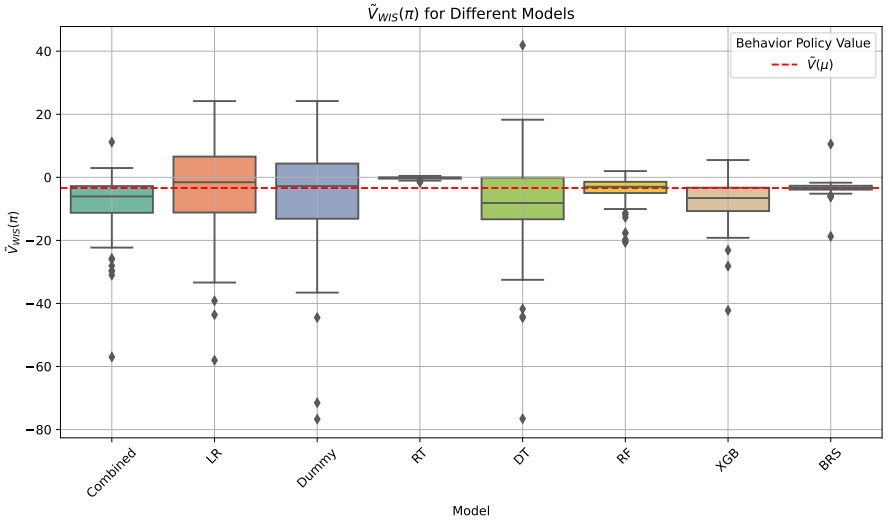


Figure 7.4: Filtered WIS Estimates for Different Under Random Policy

Figure 7.5 compares the weighted importance sampling (WIS) estimates $\tilde{V}_{WIS}(pi)$ under the Most Likely Policy for various machine learning models.

Overall, we can tell from the plot that the Most Likely Policy performs better than the Random Policy.

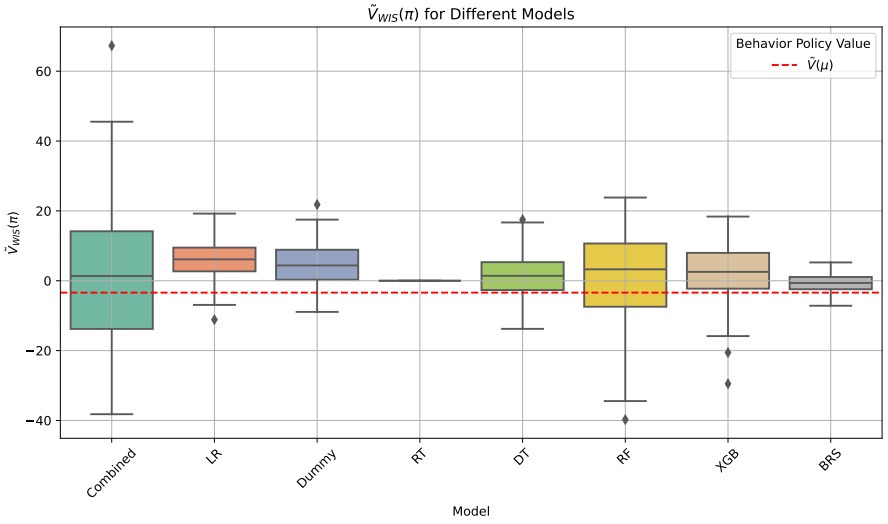


Figure 7.5: Filtered WIS Estimates for Different Under the Most Likely Policy

8

Discussion

In this chapter, we first address the six questions we brought up in Section 1.1. Next, we compare the results of using interpretable models versus black-box models in the two-stage combined model. We then discuss the observed increase in the frequency of use of JAKi over time in modeled behavior policy. Finally, we provide a detailed discussion of our proposed two-stage combined model.

Q1: If the patient switches therapy, what is the reason for the switch? Based on the switch decision tree in Figure 5.2, we can see that previous therapy is an important factor in determining whether a patient will switch therapy. If the patient previously used TNFi combination therapy, the probability of switching is only 0.33. However, if the patient is currently in a wash-out period, the probability of switching is high. Additionally, the duration of RA, seated blood pressure, CDAI level, and age are important factors in determining whether a patient continues with the current therapy.

Q2: If the patient switches therapy, to which therapy does the patient switch? Based on the decision trees shown in Figures 5.3 and 5.8, we can see that the prediction of switching therapy depends on whether the patient is using b/tsDMARDs for the first time or has used them previously. For patients using b/tsDMARDs for the first time, the most common switch is to TNFi combination therapy, followed by JAKi combination therapy. Patients of a certain age tend to switch to Abatacept combination therapy. For patients who have previously used b/tsDMARDs, the switch is to Rituximab combination therapy, JAKi monotherapy, Abatacept monotherapy, or IL-6Ri monotherapy. Prediction depends on their previous therapy, whether they are in a wash-out period, and their history of comorbidities, as detailed in Section 5.2.

Q3: Which of the two approaches is better for predicting therapy in RA patients? In this question, we compare the two approaches mentioned in Chapter 5 for predicting the physician's treatment choice given the patient's data. The first approach uses a machine learning model directly for prediction, while the second approach uses a two-stage combined model based on the divide-and-conquer principle. As shown in Table 5.1, both methods outperform our baseline dummy model in terms of both performance metrics and calibration error, with the first method slightly outperforming the second. However, in terms of model

interpretability and clinical value, the model provided by the first method, as shown in Figure 5.3, is significantly less valuable than the model provided by the second method, as shown in Figure 5.8. The decision tree model from the first method mainly predicts no switch in therapy, which is correct based on the data set, but not as clinically insightful as we would like. In contrast, the second method addresses this issue by using a two-stage prediction process. The first stage determines whether a patient needs to switch therapy, and the second stage focuses on predicting which therapy they should switch to. This approach provides more clinically valuable insights because we are interested in the actual switching patterns.

Q4: How would individual patients be treated using the proposed policy? Based on our proposed target policy, for patients who switched therapy at their last visit, are in CDAI remission or low disease activity, or were previously on TNFi combination therapy, our target policy recommends that they stay on their current therapy. For patients who are switching therapies, our target policy first determines whether they are using b/tsDMARDs for the first time or have used them before. If they are new to b/tsDMARDs, our target policy recommends that they try combination TNFi therapy, combination JAKi therapy, or combination Abatacept therapy. For patients who have previously used b/tsDMARDs, our target policy suggests that they try Rituximab combination therapy, JAKi monotherapy, Abatacept monotherapy, or IL-6Ri monotherapy based on their previous therapy, whether they are in a wash-out period, and their history of comorbidities. We provide a detailed explanation of the target policy's treatment recommendations for these different types of patients in Section 6.2.

Q5: How does our proposed policy differ from existing guidelines? Our proposed target policy provides suggestions for clarifying ambiguous areas in the EULAR guidelines, particularly in phase 2 and phase 3. In phase 2, if poor prognostic factors are present, the guidelines recommend adding b/tsDMARDs therapy. Based on our target policy, we further suggest adding TNFi therapy while also considering the patients CDAI level and age to assess whether JAKi therapy should be used. In phase 3, the EULAR guidelines suggest changing the patient's b/tsDMARDs. If patients need to switch b/tsDMARDs, we recommend selecting the replacement therapy based on the patients previous therapy and history of comorbidities.

Q6: How does the proposed target policy impact patient outcomes compared to the current behavior policy? Based on Figures 7.5 and Figure 7.4, we can see that our proposed target policy outperforms the baseline random policy. From Figure and Figure, we can see that the average value of the target policy is higher than that of the behavior policy, indicating that the target policy is better than the existing behavior policy in terms of evaluating the severity of the patient's RA. In other words, using the target policy can achieve better results than the behavior policy. The higher performance of the target policy suggests that using the target policy can potentially improve the treatment effectiveness in practical applications. The better performance of the goal policy over the behavior policy in OPE

also shows that our combined model can successfully identify decision paths that are better than the current strategy.

Different interpretable machine learning models in modeling behavior policy. As shown in Table 5.2, the two decision set models have high performance metrics but significant calibration errors. In contrast, the RL model in the decision list and logistic regression in the linear model both show good performance and very low calibration errors, while the RS model in the linear model performs worse than these two. Among all the interpretable machine learning models, the decision trees models, DT and RT, show good balanced performance in terms of both overall performance and calibration errors. Considering the applicability of these models to our project, decision sets and decision lists can only group patients based on rules, often resulting in overlapping groups and failing to provide clear decision paths. While linear models can indicate the importance of features, they cannot group patients based on this feature importance and still fail to describe the decision paths. On the other hand, the decision tree model can effectively illustrate decision paths and group patients into distinct, non-overlapping categories based on these paths. Therefore, the decision tree is the most ideal interpretable machine learning model for our project.

Interpretable vs. black-box models in two-stage combined model. From Figure 5.8 and Figure 5.10, we can see that the therapy decision tree in combined model 1 (XGB + DT) and the therapy model in combined model 2 (DT + DT) exhibit similar decision patterns when predicting the treatment chosen by the physician. In the combined models, using a black-box model (such as XGB) as the switch model is comparable in interpretability to using interpretable machine learning models (such as DT), with only minor performance differences. This suggests that in this context, the added complexity of black-box models does not provide a significant advantage over interpretable models. This finding highlights the value of interpretable models, as they offer similar predictive capabilities while clearly illustrating the decision-making process, which is crucial for clinical applications.

Distribution shift across time. As mentioned in Section 5.2, the decision tree illustrates the increased use of JAKi since 2018. Tofacitinib was first approved by the U.S. Food and Drug Administration in November 2012 for the treatment of patients with moderate to severe RA who had an inadequate response to, or were intolerant of, methotrexate [52]. Subsequently, baricitinib was approved in 2018 and upadacitinib in 2019 [53]. The approval of more JAKi drugs has driven the growth trend in JAKi use. Our decision tree captures the 2018 time point well, describing the shift in the distribution of JAKi usage over time. Before 2018, the decision tree recommended TNFi combination therapy; after 2018, it recommended JAKi combination therapy.

Two-stage combined model. The two-step combined model uses a "divide and conquer" approach that closely mimics human decision making. By breaking the complex problem into two separate tasks - predicting whether a patient will switch

therapy and then predicting which therapy they will switch to - this model improves interpretability and allows for more refined and understandable predictions. In contrast, using a single machine learning model to perform both prediction tasks simultaneously can yield good evaluation metrics. However, this approach can be problematic for highly complex problems. Our observations suggest that such models often end up predicting the most common patterns in the data (patients staying on previous therapy), while we are more interested in the less common patterns (therapy switching). This approach can lead to predictions that are statistically sound, but lack clinical value and cannot provide meaningful insight into the decision-making process. By using the two-stage combined model, we can ensure that each aspect of the prediction task is addressed with specific focus, leading to more interpretable results. This is especially important in the medical field, where understanding the rationale behind a prediction is as important as the prediction itself. The divide and conquer strategy of the combined model not only mirrors the logical steps a human expert might take, but also allows us to dissect and understand the intricacies of each prediction, making it a more reliable and insightful tool in complex scenarios.

9

Conclusion

In this thesis, we investigated the application of policy learning and off-policy evaluation to the sequential treatment of rheumatoid arthritis. Our work focuses on using interpretable machine learning models to better understand and improve clinical decision-making processes for rheumatoid arthritis treatment.

To model the behavior policy of RA patients using historical patient data, we proposed a two-stage combined model based on the divide-and-conquer principle. We compare this approach to a common approach that directly uses machine learning models to predict the physician's choice of therapy. In both approaches, we use a mixture of interpretable machine learning models and black-box models. For interpretable machine learning models, we use rule-based models, including decision sets, decision trees, and decision lists, as well as linear models such as risk scores and logistic regression. For black-box models, we use two ensemble learning methods: random forest and XGBoost. By using interpretable machine learning models, we are able to capture the decision-making patterns of physicians when selecting therapies for RA patients.

We also compare these two approaches. Our results show that our proposed two-stage combined model performs comparably to the direct approach using decision trees but provides more clinically valuable decision paths. This demonstrates that the combined model not only maintains performance, but also improves the practical applicability of the decision process in clinical settings. The two-stage model divides the complex problem into two stages, mimicking the physician's decision-making process. This structured approach results in clearer, more useful insights that help improve clinical decision making.

To propose a target policy, we use the most likely policy, which selects the therapy with the highest probability in the modeled behavior policy. We also use a randomly generated policy as our baseline. We explain what this proposed target policy means for the treatment of RA patients and provide suggestions for clarifying ambiguous areas in the EULAR guideline.

Our off-policy evaluation validates the effectiveness of our proposed target policy. Using importance sampling and weighted importance sampling methods, we validate that the target policy can potentially improve treatment effectiveness compared to the existing behavior policy. This highlights the importance of using data-driven approaches to refine treatment guidelines.

In conclusion, our thesis highlights the potential of interpretable machine learning models in optimizing RA treatment strategies and suggests that the use of a two-step model may provide better clinical value compared to the direct use of machine learning models to predict the physician's choice of therapy. By providing transparent and useful insights, these interpretable machine learning models can assist clinicians in making informed decisions, ultimately improving patient care and outcomes.

9.1 Future Work

Future work can extend this research in several ways to further improve the modeling and evaluation of RA treatment strategies. First, sequential models such as recurrent neural networks and long-term memory networks can be used to model behavior policy, as these models help to more accurately and comprehensively understand decision patterns that change over time. Second, instead of relying solely on the single strategy provided by the most likely policy, multiple strategies can be explored when proposing the target policy. For example, the two therapies with the highest probabilities in the behavior policy can be selected and their probabilities averaged to 0.5 to form our proposed target policy. This approach can provide more diverse and flexible treatment recommendations. Finally, in the off-policy evaluation, more estimators can be used to improve the accuracy of the evaluation. For example, using methods such as the doubly robust estimator and the direct method estimator can provide more reliable and comprehensive policy evaluation results. Additionally, while modeling behavior policy, we observed that csDMARDs may also indicate a wash-out period. Therefore, the definition of the wash-out period needs further refinement.

Bibliography

[1] K. Almutairi, J. Nossent, D. Preen, H. Keen, and C. Inderjeeth, "The global prevalence of rheumatoid arthritis: A meta-analysis based on a systematic review," *Rheumatology international*, vol. 41, no. 5, pp. 863–877, 2021.

[2] A. Finckh, B. Gilbert, B. Hodkinson, *et al.*, "Global epidemiology of rheumatoid arthritis," *Nature Reviews Rheumatology*, vol. 18, no. 10, pp. 591–602, 2022.

[3] Q. Guo, Y. Wang, D. Xu, J. Nossent, N. J. Pavlos, and J. Xu, "Rheumatoid arthritis: Pathological mechanisms and modern pharmacologic therapies," *Bone research*, vol. 6, no. 1, p. 15, 2018.

[4] J. Alam, I. Jantan, and S. N. A. Bukhari, "Rheumatoid arthritis: Recent advances on its etiology, role of cytokines and pharmacotherapy," *Biomedicine & Pharmacotherapy*, vol. 92, pp. 615–633, 2017.

[5] P. Emery and M. Salmon, "Early rheumatoid arthritis: Time to aim for remission?" *Annals of the Rheumatic Diseases*, vol. 54, no. 12, p. 944, 1995.

[6] J. S. Smolen, R. B. Landewé, S. A. Bergstra, *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*, vol. 82, no. 1, pp. 3–18, 2023.

[7] L. Fraenkel, J. M. Bathon, B. R. England, *et al.*, "2021 american college of rheumatology guideline for the treatment of rheumatoid arthritis," *Arthritis & Rheumatology*, vol. 73, no. 7, pp. 1108–1123, 2021.

[8] B. Chastek, C.-I. Chen, C. Proudfoot, S. Shinde, A. Kuznik, and W. Wei, "Treatment persistence and healthcare costs among patients with rheumatoid arthritis changing biologics in the usa," *Advances in Therapy*, vol. 34, pp. 2422–2435, 2017.

[9] E. Sullivan, J. Kershaw, S. Blackburn, J. Choi, J. R. Curtis, and S. Boklage, "Biologic disease-modifying antirheumatic drug prescription patterns for rheumatoid arthritis among united states physicians," *Rheumatology and Therapy*, vol. 7, pp. 383–400, 2020.

[10] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[11] K. F. Schulz, D. G. Altman, and D. Moher, "Consort 2010 statement: Updated guidelines for reporting parallel group randomised trials," *Journal of Pharmacology and pharmacotherapeutics*, vol. 1, no. 2, pp. 100–107, 2010.

[12] J. Kremer, "The corrona database," *Annals of the Rheumatic Diseases*, vol. 64, no. suppl 4, pp. iv37–iv41, 2005.

[13] A. Pace, A. J. Chan, and M. van der Schaar, "Poetree: Interpretable policy learning with adaptive decision trees," *arXiv preprint*, 2022. eprint: 2203.08057.

[14] A. Matsson and F. D. Johansson, "Case-based off-policy evaluation using prototype learning," in *Uncertainty in Artificial Intelligence*, PMLR, 2022, pp. 1339–1349.

[15] R. Y. Rubinstein and D. P. Kroese, *Simulation and the Monte Carlo method*. John Wiley & Sons, 2016.

[16] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, "Interpretable machine learning: Fundamental principles and 10 grand challenges," *Statistic Surveys*, vol. 16, pp. 1–85, 2022.

[17] L. Breiman, *Classification and regression trees*. Routledge, 2017.

[18] R. L. Rivest, "Learning decision lists," *Machine learning*, vol. 2, pp. 229–246, 1987.

[19] J. H. Friedman and B. E. Popescu, "Predictive learning via rule ensembles," 2008.

[20] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009, vol. 2.

[21] A. Silva, T. Killian, I. D. J. Rodriguez, S.-H. Son, and M. Gombolay, "Optimization methods for interpretable differentiable decision trees in reinforcement learning," *arXiv preprint arXiv:1903.09338*, 2019.

[22] A. Hüyük, D. Jarrett, and M. van der Schaar, "Explaining by imitating: Understanding decisions by interpretable policy learning," *arXiv preprint arXiv:2310.19831*, 2023.

[23] B. Norgeot, B. S. Glicksberg, L. Trupin, *et al.*, "Assessment of a deep learning model based on electronic health record data to forecast clinical outcomes in patients with rheumatoid arthritis," *JAMA network open*, vol. 2, no. 3, e190606–e190606, 2019.

[24] J. S. Smolen, R. Landewé, F. C. Breedveld, *et al.*, "Eular recommendations for the management of rheumatoid arthritis with synthetic and biological disease-modifying antirheumatic drugs," *Annals of the rheumatic diseases*, vol. 69, no. 6, pp. 964–975, 2010.

[25] W. Tao, A. N. Concepcion, M. Vianen, *et al.*, "Multiomics and machine learning accurately predict clinical response to adalimumab and etanercept therapy in patients with rheumatoid arthritis," *Arthritis & Rheumatology*, vol. 73, no. 2, pp. 212–222, 2021.

[26] D. Plant, M. Maciejewski, S. Smith, *et al.*, "Profiling of gene expression biomarkers as a classifier of methotrexate nonresponse in patients with rheumatoid arthritis," *Arthritis & Rheumatology*, vol. 71, no. 5, pp. 678–684, 2019.

[27] H. R. Gosselt, M. M. Verhoeven, M. Bulatovi-alasan, *et al.*, "Complex machine-learning algorithms and multivariable logistic regression on par in the prediction of insufficient clinical response to methotrexate in rheumatoid arthritis," *Journal of personalized medicine*, vol. 11, no. 1, p. 44, 2021.

[28] M. A. Morid, M. Lau, and G. Del Fiol, "Predictive analytics for step-up therapy: Supervised or semi-supervised learning?" *Journal of Biomedical Informatics*, vol. 119, p. 103 842, 2021.

[29] O. Gottesman, F. Johansson, J. Meier, *et al.*, "Evaluating reinforcement learning algorithms in observational health settings," *arXiv preprint arXiv:1805.12298*, 2018.

[30] T. Vos, S. S. Lim, C. Abbafati, *et al.*, "Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: A systematic analysis for the global burden of disease study 2019," *The lancet*, vol. 396, no. 10258, pp. 1204–1222, 2020.

[31] Y. Shapira, N. Agmon-Levin, and Y. Shoenfeld, "Geoepidemiology of autoimmune rheumatic diseases," *Nature Reviews Rheumatology*, vol. 6, no. 8, pp. 468–476, 2010.

[32] G. S. Firestein, "Evolving concepts of rheumatoid arthritis," *Nature*, vol. 423, no. 6937, pp. 356–361, 2003.

[33] M. Prevoo, M. Van'T Hof, H. Kuper, M. Van Leeuwen, L. Van De Putte, and P. Van Riel, "Modified disease activity scores that include twenty-eight-joint counts development and validation in a prospective longitudinal study of patients with rheumatoid arthritis," *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, vol. 38, no. 1, pp. 44–48, 1995.

[34] D. Aletaha and J. Smolen, "The simplified disease activity index (sdai) and the clinical disease activity index (cdai): A review of their usefulness and validity in rheumatoid arthritis," *Clinical and experimental rheumatology*, vol. 23, no. 5, S100, 2005.

[35] J. F. Fries, P. Spitz, R. G. Kraines, and H. R. Holman, "Measurement of patient outcome in arthritis," *Arthritis & Rheumatism*, vol. 23, no. 2, pp. 137–145, 1980.

[36] S. Liu, K. C. See, K. Y. Ngiam, L. A. Celi, X. Sun, and M. Feng, "Reinforcement learning for clinical decision support in critical care: Comprehensive review," *Journal of medical Internet research*, vol. 22, no. 7, e18477, 2020.

[37] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[38] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[39] C. P. Robert, G. Casella, and G. Casella, *Monte Carlo statistical methods*. Springer, 1999, vol. 2.

[40] A. R. Mahmood, H. P. Van Hasselt, and R. S. Sutton, "Weighted importance sampling for off-policy learning with linear function approximation," in *Advances in neural information processing systems*, vol. 27, 2014.

[41] A. Matsson, D. Solomon, M. Crabtree, R. Harrison, H. Litman, and F. Johansson, "Patterns in the sequential treatment of patients with rheumatoid arthritis starting a biologic or targeted synthetic diseasemodifying antirheumatic drug: 10year experience from a usbased registry," *ACR Open Rheumatology*, 2023.

[42] M. Schiff, C. Pritchard, J. E. Huffstutter, *et al.*, "The 6-month safety and efficacy of abatacept in patients with rheumatoid arthritis who underwent a

washout after anti-tumour necrosis factor therapy or were directly switched to abatacept: The arrive trial," *Annals of the rheumatic diseases*, vol. 68, no. 11, pp. 1708–1714, 2009.

[43] V. P. Bykerk, A. J. Östör, J. Alvaro-Gracia, *et al.*, "Tocilizumab in patients with active rheumatoid arthritis and inadequate responses to dmards and/or tnf inhibitors: A large, open-label study close to clinical practice," *Annals of the rheumatic diseases*, vol. 71, no. 12, pp. 1950–1954, 2012.

[44] B. Ustun and C. Rudin, "Learning optimized risk scores," *Journal of Machine Learning Research*, vol. 20, no. 150, pp. 1–75, 2019.

[45] A. Agarwal, Y. S. Tan, O. Ronen, C. Singh, and B. Yu, "Hierarchical shrinkage: Improving the accuracy and interpretability of tree-based methods," *ArXiv preprint arXiv:220200858*, Feb. 2022.

[46] C. Molnar, *Interpretable machine learning*. Lulu. com, 2020.

[47] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[48] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed. Springer, 2008, ISBN: 0-387-95284-5.

[49] J. Nixon, M. W. Dusenberry, L. Zhang, G. Jerfel, and D. Tran, "Measuring calibration in deep learning.," in *CVPR workshops*, vol. 2, 2019.

[50] W. J. Dixon and K. K. Yuen, "Trimming and winsorization: A review," *Statistische Hefte*, vol. 15, no. 2, pp. 157–170, 1974.

[51] S. Prinja, N. Gupta, and R. Verma, "Censoring in clinical trials: Review of survival analysis techniques," *Indian Journal of Community Medicine*, vol. 35, no. 2, pp. 217–221, 2010.

[52] K. Traynor, "Fda approves tofacitinib for rheumatoid arthritis.," *American Journal of Health-System Pharmacy*, vol. 69, no. 24, 2012.

[53] R. Harrington, S. A. Al Nokhatha, and R. Conway, "Jak inhibitors in rheumatoid arthritis: An evidence-based review on the emerging clinical data," *Journal of inflammation research*, pp. 519–531, 2020.

A

Appendix 1

A.1 Experiment Details

Table A.1: Moddel hyperparameters and their respective search space for all models.

Model	Hyperparameter	Search space
RE	tree size	$\{2, 3, 4, 5\}$
	max rules	$\{51015\}$
LR	C	$\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$
	max iterations	$\{1000, 1500, 2000, 2500\}$
DT	max depth	{None, 3, 5, 7, 9}
	min samples to split	$\{0.02, 0.04, 0.06, 0.08\}$
	min samples of the leaf	$\{0.02, 0.04, 0.06, 0.08\}$
BSR	number of estimators	{5, 10, 15, 20, 30}
	learning rate	$\{0.6, 0.8, 1.0\}$
	min samples of the leaf	$\{0.02, 0.04, 0.06, 0.08\}$
RL	max depth	{3, 5, 8, 10}
RT	max leaf nodes	{None, 10, 20, 30, 40}
RF	number of estimators	{100, 200, 300, 400, 500}
	max depth	{None, 5, 10, 20}
	min samples to split	$\{0.02, 0.04, 0.06, 0.08\}$
	min samples of the leaf	$\{0.02, 0.04, 0.06, 0.08\}$
XGB	learning rate	{0.01, 0.05, 0.1, 0.2}
	max depth	{3, 5, 7, 9, 12}

A.2 Supplementary Figures

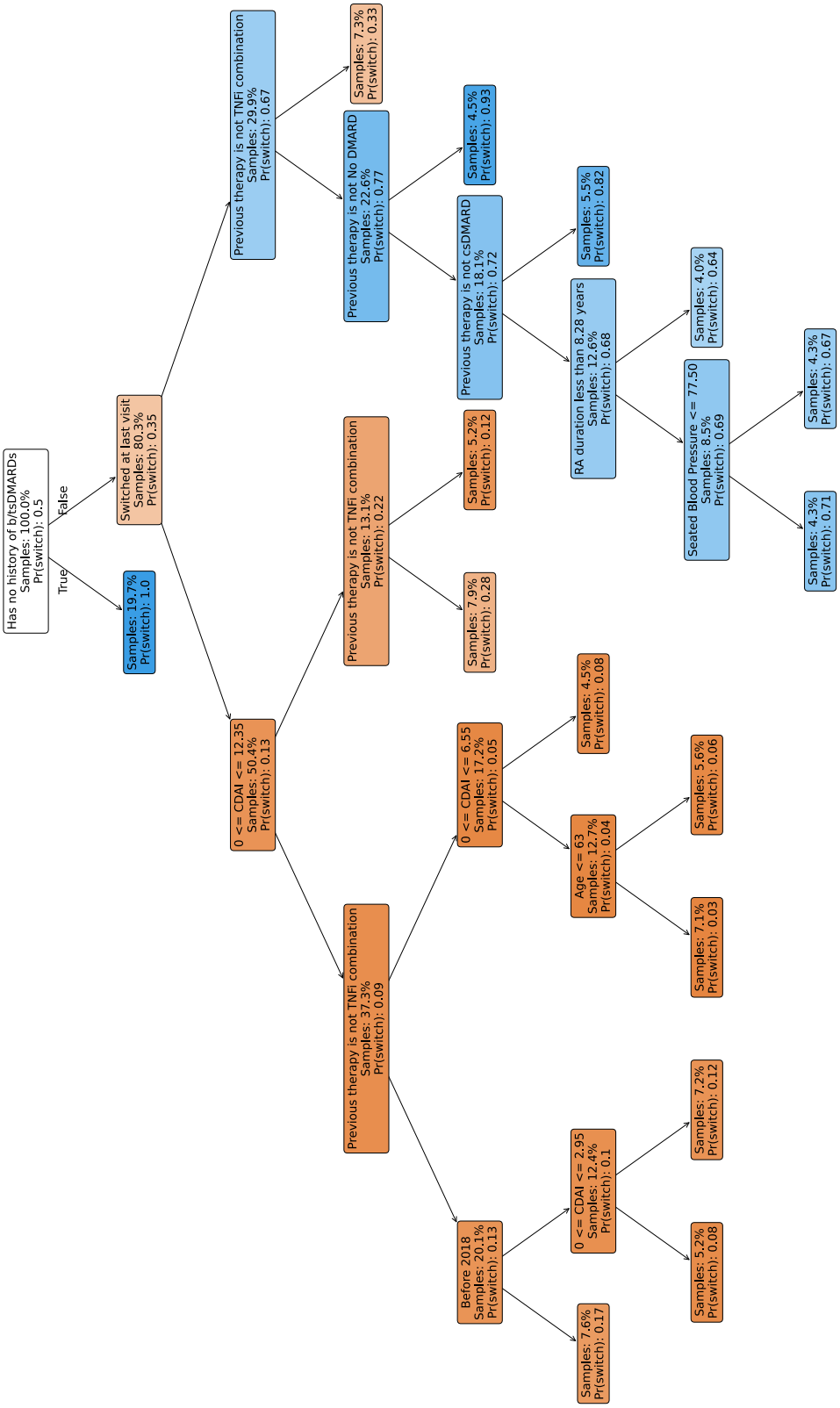


Figure A.1: Complete decision tree to predict whether patients will switch therapy.

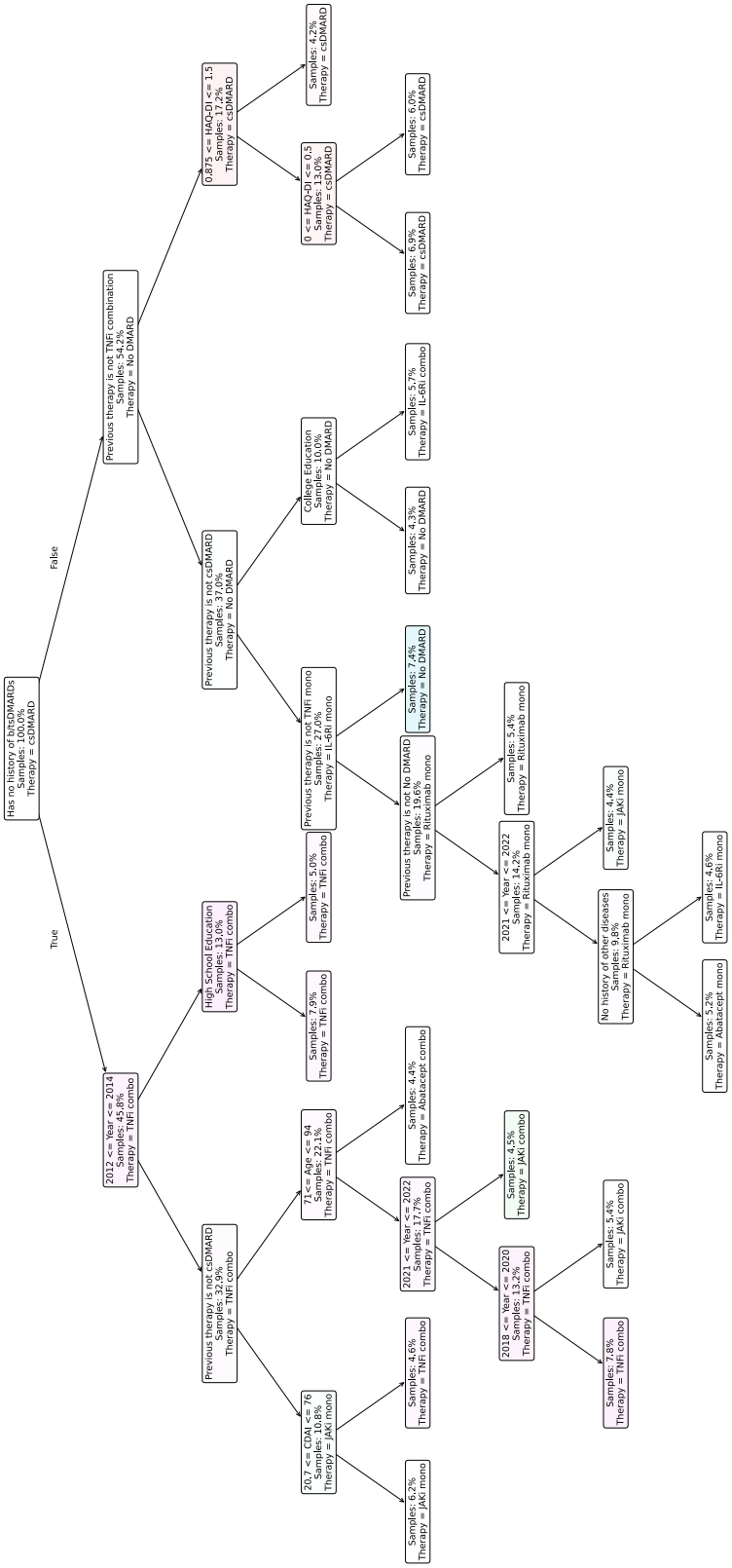


Figure A.2: Complete proposed target policy based on combined model 2 (XGB+XGB).