machine learning & knowledge extraction

MDPI

Article

Learning Optimal Dynamic Treatment Regime from Observational Clinical Data through Reinforcement Learning

Seyum Abebe *, Irene Poli, Roger D. Jones and Debora Slanzi

European Centre for Living Technology, Ca' Foscari University of Venice, 30123 Venice, Italy
* Correspondence: seyumassefa.abebe@unive.it

Abstract: In medicine, dynamic treatment regimes (DTRs) have emerged to guide personalized treatment decisions for patients, accounting for their unique characteristics. However, existing methods for determining optimal DTRs face limitations, often due to reliance on linear models unsuitable for complex disease analysis and a focus on outcome prediction over treatment effect estimation. To overcome these challenges, decision tree-based reinforcement learning approaches have been proposed. Our study aims to evaluate the performance and feasibility of such algorithms: tree-based reinforcement learning (T-RL), DTR-Causal Tree (DTR-CT), DTR-Causal Forest (DTR-CF), stochastic tree-based reinforcement learning (SL-RL), and Q-learning with Random Forest. Using real-world clinical data, we conducted experiments to compare algorithm performances. Evaluation metrics included the proportion of correctly assigned patients to recommended treatments and the empirical mean with standard deviation of expected counterfactual outcomes based on estimated optimal treatment strategies. This research not only highlights the potential of decision tree-based reinforcement learning for dynamic treatment regimes but also contributes to advancing personalized medicine by offering nuanced and effective treatment recommendations.

Keywords: dynamic treatment regime; observational clinical data; reinforcement learning

Citation: Abebe, S.; Poli, I.; Jones, R.D.; Slanzi, D. Learning Optimal Dynamic Treatment Regime from Observational Clinical Data through Reinforcement Learning. *Mach. Learn. Knowl. Extr.* **2024**, *6*, 1798–1817. https://doi.org/10.3390/make6030088

Academic Editor: Krzysztof J Cios

Received: 27 May 2024
Revised: 25 July 2024
Accepted: 25 July 2024
Published: 30 July 2024

1. Introduction

In various clinical practices, it is frequently required to modify treatment over time because of the considerable differences in individual responses to treatment and to accommodate the progressive, often cyclical, nature of many chronic diseases and conditions. For example, in managing diabetic kidney disease, treatment plans must be regularly updated to address how patients respond to medication and the progression of kidney damage that can vary over time. Adjustments might include changing medication types, and dosages, or incorporating new therapies as the disease advances or stabilizes [1]. In this regard, precision medicine [2–4] is becoming the most relevant and innovative approach in healthcare that considers the patient's individual and unique characteristics, such as their genetics, environment, and lifestyle, to tailor treatments to their specific needs. Personalized therapies are based on decision rules that consider the patient's state of health, such as symptoms, test results, and medical history. Hence, the growing interest in developing dynamic treatment regimes (DTRs) can guide clinical decision-making by providing personalized treatment recommendations to individual patients. Unlike traditional *one-size-fits-all* approaches, DTR [5–9] aims to optimize patient disease progression by identifying treatment strategies. This approach holds great promise for improving patient conditions and advancing the field of precision medicine.

Dynamic treatment regimes (DTRs) are predetermined sequences of treatment decision rules. They are strategically devised to provide clinicians with informed guidance on the decision of whether and how to modify, and subsequently re-modify, treatment strategies over time in response to the evolving condition of an individual. The DTR encompasses several treatment stages, with each stage utilizing patient-specific medical history and

current disease status information to formulate individualized treatment recommendations for the subsequent phase [10]. DTRs are particularly useful for managing chronic conditions [11], in which patients may respond differently to different treatments over time.

DTR can be defined as a set of rules that map a patient's individual characteristics and treatment history to a recommended treatment [8]. The rules in a DTR can be based on a variety of factors, including patient demographics, medical history, laboratory test results, and genetic information. DTRs can be implemented in different ways, including through clinical decision support systems, electronic health records, and mobile apps. The identification of optimal DTRs offers an effective vehicle for the personalized management of diseases and helps physicians to identify the best treatment strategies dynamically and individually based on clinical evidence, thus providing a key foundation for better healthcare [11].

Research on DTRs dates back to Robins et al. [7,12]. Most of the developed approaches relied on linear low-dimensional parametric models [8,13–16]. Due to the limitations of standard regression methods in capturing the complexities of DTRs, more advanced statistical methods, such as dynamic conditional models (DCMs) and dynamic marginal structural models (DMSMs), have been proposed to estimate the causal effects of DTRs in observational data [17]. A dynamic conditional model (DCM) is a statistical model that estimates the average effects of treatments on patients, conditional upon their medical history. This indicates that the predicted effects are specifically tailored to individuals with similar medical backgrounds. Dynamic conditional models track a patient's medical history over time to estimate the effect of a treatment on the patient's outcome at a given point in time. There are a number of different methods that can be used to estimate treatment effects in dynamic conditional models, including Q-learning [18,19], the parametric G-formula [5,9,20], and G-estimation [8,12].

Recently, machine learning methods, especially reinforcement learning (RL) have been proposed for addressing the complexities in learning DTRs [21] from observational clinical data. Thus, there has been considerable interest in converting dynamic conditional models (DCMs) into reinforcement learning (RL) problems using observational clinical data. Ref. [22] proposed a method called adaptive contrast weighted learning (ACWL) to estimate the average effects of different treatments on patients. ACWL uses a type of machine learning called decision tree rules to learn how different treatments affect patients with different characteristics, such as their age, sex, and medical history. ACWL also combines two other statistical methods, doubly robust augmented inverse probability weighting (AIPW) estimators and classification algorithms, to improve the accuracy of its estimates. Ref. [23] introduced a tree-based method known as LZ, which aims to directly estimate optimal treatment strategies. LZ adapts the reinforcement learning task to a decision tree framework, incorporating an unsupervised purity measure. Simultaneously, it retains the benefits of decision trees, including their simplicity for comprehension and interpretation. It also maintains its ability to handle various treatment options and outcome types (e.g., continuous or categorical) without making assumptions about data distribution. However, LZ is designed for single-stage decision problems and may be susceptible to model misspecification. More recently, refs. [24–26] applied decision lists to construct interpretable DTRs, which comprise a sequence of "if–then" clauses that map patient covariates to recommended treatments. A decision list can be viewed as a special case of tree-based rules, where the rules are ordered and learned one after another [27]. These list-based approaches prove highly advantageous when the aim is not only to optimize health outcomes but also to reduce the expenses associated with covariate measurements. However, without cost information, a list-based method may be more restrictive than a tree-based method. Then again, in order to achieve simplicity and interpretability, refs. [24,25] limited each rule to encompass a maximum of two covariates, a limitation that could pose challenges when dealing with more intricate treatment strategies.

Tao et al. [28] expanded upon the research conducted by [22,23]. They developed a tree-based reinforcement learning (T-RL) method used to estimate optimal DTRs in a

multi-stage, multi-treatment setting. Another tree-based method that has been proposed for providing explainable treatment recommendations is Stochastic Tree Search for Estimating Optimal Dynamic Treatment Regimes (ST-RL) by [29]. ST-RL builds decision trees at each stage by modeling counterfactual outcomes through nonparametric regression and then uses a stochastic approach with a Markov chain Monte Carlo algorithm to find the best tree-structured decision rule. However, the aforementioned tree-based methods do not explicitly model causal effects. To overcome this aspect [21] proposed a causal tree-based reinforcement learning method that directly estimates treatment effects with a causal interpretation via a specific splitting criterion for the decision trees.

Hybrid algorithms have been also developed using a combination of Q-learning and other regression algorithms such as Random Forest and decision trees. Regression algorithms are used to approximate the Q-function and predict the Q-value for each possible action in the current state [30].

In this study, we conducted several experiments on observational clinical data to evaluate the performance of algorithms in identifying the optimal treatment regime. Observational clinical data provides a rich source of information for developing and evaluating DTRs. These data consist of real-world patient information collected in routine clinical practice, offering a comprehensive view of patients' characteristics, treatments received, and their associated outcomes. Leveraging observational clinical data also allows researchers to analyze large and diverse patient populations, capturing the complexity and heterogeneity of real-world healthcare settings. This heterogeneity can arise from various factors, including differences in patient characteristics, common diseases, treatment approaches, genetic variations, environmental factors, and other variables [31]. Hence, this heterogeneity in the progression of the disease will have several implications in precision medicine; such as treatment effectiveness [32], disease progonosis [33], and risk stratification [34,35].

The algorithms evaluated in this experimental study are tree-based reinforcement learning (T-RL), Stochastic Tree Search for Estimating Optimal Dynamic Treatment Regimes (ST-RL), the causal tree-based reinforcement learning method, and Q-learning with Random Forest. The experiment involves analyzing the observational data of the patient population affected by diabetic kidney disease (DKD) and leveraging their medical history and treatment outcomes to generate personalized treatment recommendations. By comparing the predicted treatment outcomes against the observed outcomes in the clinical data, we assessed the effectiveness of each algorithm in identifying optimal treatment strategies.

The results of this study have implications for the advancement of precision medicine and individualized treatment approaches. By leveraging observational clinical data and advanced algorithms, we aim to contribute to the growing body of knowledge on DTRs and their potential for improving patient outcomes. This research can inform future investigations and guide clinical practice in tailoring treatments to individual patients, ultimately leading to better patient care and outcomes.

2. Dynamic Treatment Regimes

Concept and Notation

Let $i = 1, ..., N$ refer to patients, $t = 1, ..., T$ refer to decision stages, and O_t denote the vector of patient characteristics accumulated during the treatment period t. Let A_t denotes a multi-categorical treatment indicator variable with the observed value $a_t \in A_t = \{1, ..., K_t\}$, where K_t ($K_t \geq 2$) is the number of treatment options at the tth stage. O_{T+1} denotes the entire clinical observation history of a patient up to the end of T. We denote Y as the chosen response to the treatment, with higher values of Y being preferred. Hence, the observed data on patients are $H_T = \{(A_{1i}, ..., A_{Ti}, O_{T+1,i}^{\top})\}_{i=1}^{N}$, which describes the complete patient history through time T.

The concept of compiling patient history over time can be conceptualized as a sequential process. Initially, at baseline, O_1 is collected, such as demographic, clinical, and lifestyle patient characteristics. Subsequently, the first treatment decision A_1 is then determined. Thus, the patient history at stage 1 is simply represented as $H_1 = O_1$. Following this, the

patient's response to the first treatment is recorded as O_2, and the response variable Y_1 is evaluated. Consequently, the history at stage 2 becomes $H_2 = \{O_1, A_1, Y_1, X_2\}$. This continues until the last stage.

A Dynamic Treatment Regimen (DTR) model defines a sequence of individualized decision treatment rules, $r = (r_1, ..., r_T)$, where r_t is a mapping of patient history at stage t to the domain of treatment assignment A_t. One of the many ways to define and identify optimal DTR is to consider a counterfactual framework for causal inference [5] and start from the last treatment stage in reverse sequential order. The counterfactual framework used involves comparing what happened (the observed outcome) to what would have happened if a different treatment had been applied (the counterfactual or unobserved outcome). In this framework, the idea is to construct counterfactual scenarios that represent what would have occurred under different conditions.

To illustrate the concept of counterfactuals, consider a practical example in the context of evaluating a new antihypertensive drug designed to reduce blood pressure in patients with hypertension. In an ideal randomized clinical trial, participants are assigned to either receive the new drug or a placebo (an inactive substance), ensuring that the comparison between treatments is rigorous and unbiased. In the actual scenario, Patient A is assigned to receive the new drug and experiences a significant reduction in blood pressure. This observed outcome provides a direct measure of the drug's effect on the patient's blood pressure under the treatment condition. The counterfactual scenario, however, involves estimating what would have happened to Patient A's blood pressure if they had been assigned to the placebo group instead of receiving the new drug. This hypothetical outcome, known as the counterfactual outcome, represents the blood pressure level that Patient A would have experienced had they received the placebo. By comparing the actual outcome (the reduction in blood pressure observed with the new drug) to this counterfactual outcome (the hypothetical blood pressure level with the placebo), we can estimate the causal effect of the new drug. This comparison provides insights into the drug's efficacy by highlighting the difference between the observed effect and the potential effect had the patient been given an alternative treatment.

Typically, counterfactuals are not directly observable, but they serve as a reference point for assessing the causal impact of a treatment. At the final stage T, let $Y^*(A_1, ..., A_{T-1}, a_T)$, or $Y^*(a_T)$ for brevity, denote the counterfactual outcome of a patient treated with $a_T \in A_T$ conditional on previous treatments $(A_1, ..., A_{T-1})$, and define $Y^*(r_T)$ as a counterfactual outcome under regime rule r_T and history of a patient H_T. That is,

$$Y^*(r_T) = \sum_{a_T=1}^{K_T} Y^*(a_T) I\{r_T(H_T) = a_T\} \tag{1}$$

The performance of the treatment rule r_T is measured by the counterfactual mean outcome $E\{Y^*(r_T)\}$ when all patients followed r_T, and the optimal treatment rule r_T^{opt} satisfies $E\{Y^*(r_T^{opt})\} \geq E\{Y^*(r_T)\}$ for all potential classes of treatment regimes $\mathcal{R}_\mathcal{T}$, where $r_T \in \mathcal{R}_\mathcal{T}$.

To establish a link between counterfactual outcomes and the observed data, we rely on three well-established assumptions as outlined in the literature [36,37]:

- The first assumption, known as consistency, posits that a patient's actual outcome corresponds to what their outcome would have been if the patient had received the treatment they were actually administered. In essence, the treatment received by a patient is the sole factor influencing their outcome.
- The second assumption concerns the concept of stable unit treatment value, asserting that an individual's outcome remains unaffected by the treatments administered to other patients.
- The third assumption pertains to sequential exchangeability, indicating that the treatment assignment at each time point is assumed to be independent of future potential outcomes given past treatment and the covariate history.

These assumptions state that the observed outcomes are a good reflection of the potential outcomes, that the outcomes of any patient are not affected by the treatments received by other patients, that the treatment assignment is not predetermined, and assuming no confounding by unmeasured factors, treatment assignment at the current time point is independent of potential future outcomes, given the patient's complete past treatment history and medical history up to that point. The propensity score method is used to strengthen this assumption (sequential exchangeability) in the longitudinal data analysis.

Under these assumptions, the optimal rule at the final stage T can be written as:

$$r_T^{opt} = \arg\max_{r_T \in \mathcal{R}_T} E\left[\sum_{a_T=1}^{K_T} E(Y|A_T = a_T, H_T) I\{r_T(H_T) = a_T\}\right], \quad (2)$$

where the outer expectation is taken with respect to the joint distribution of the observed data H_T. Similarly, under the above assumptions, the optimal rule r_t^{opt} at stage t can be defined as:

$$r_t^{opt} = \arg\max_{r_t \in \mathcal{R}_t} E\left[\sum_{a_t=1}^{K_t} E(\hat{Y}_t|A_t = a_t, H_t) I\{r_t(H_t) = a_t\}\right], \quad (3)$$

where $\mathcal{R}_t$ is the set of all potential rules at stage t. $\hat{Y}_T = Y$ at stage T, and at t, $\hat{Y}_t$ can be defined recursively using Bellman's optimality:

$$\hat{Y}_t = E\left\{\hat{Y}_{t+1}|A_{t+1} = r_{t+1}^{opt}(H_{t+1}), H_{t+1}\right\}, t = 1, ..., T-1,$$

that is, the expected outcome assuming optimal rules are followed at all future stages.

3. Machine Learning Models for DTRs

The ability to predict how a patient might respond to medication would shift treatment decisions away from trial and error and reduce disease-associated health and financial burdens. In this regard, machine learning approaches applied to clinical observational datasets offer great promise to deliver personalized medicine [38]. In this section, we will discuss tree-based reinforcement learning algorithms that have been chosen for the experimental study.

3.1. Tree-Based Reinforcement Learning

The tree-based reinforcement learning (T-RL) algorithm directly estimates optimal dynamic treatment regimes (DTRs) in a multi-stage multi-treatment setting [28,39] from observational clinical data to determine the optimal treatment. The algorithm uses a decision tree structure, where a node represents a point in the decision tree where a decision is made based on the value of a specific feature, by splitting a parent node into two child nodes repeatedly, starting with the root node which contains the entire learning sample [40].

Estimating the dynamic treatment regime (DTR) presents a significant challenge. This challenge centers around identifying the optimal treatment for a patient using clinical data that are inaccessible through direct observation. The optimal treatment at stage t, r_t^{opt}, can only be inferred indirectly through the observed treatments and outcomes. This can be achieved by estimating the counterfactual mean outcomes given all possible treatments using the causal framework and the three assumptions stated in Section 2. The selected split at each node should increase the counterfactual mean result, which can serve as a metric of purity in DTR trees, with the overall objective of maximizing the counterfactual mean outcome in the whole population of interest. Hence, the T-RL approach constructs decision trees at each stage, managing the optimization process involving multiple treatment comparisons through a purity metric which is constructed using augmented inverse probability weighted estimators (AIPW) as outlined in the work of [22,41]. This process is used recursively using backward induction, which can handle multiple decision stages effectively.

To maximize the overall outcome for the entire population, each node in a DTR tree should split in a way that improves the outcome for the population, and this can be used to measure the purity of the node split as mentioned above. To put it in notation, let $\eta_a(H) = E(\hat{Y}|A = a, H)$ be the counterfactual outcome and $\pi(H)$ be the estimated propensity score [36]. The AIPW estimator for the counterfactual mean outcome under a given treatment a is calculated as:

$$\mathbb{P}\left[\frac{I(A=a)}{\pi_a(H)}Y + \left\{1 - \frac{I(A=a)}{\pi_a(H)}\right\}\eta_a(H)\right], \tag{4}$$

where $\mathbb{P}$ is the empirical expectation operator. For stage T, the estimation for the counterfactual outcome under a treatment rule r_T ($E(Y_T^*(r_T))$) is defined as:

$$\mathbb{P}\left[\frac{I(A_T = r_T(H_T))}{\pi_{T,A_T}(H_T)}Y + \left\{1 - \frac{I(A_T = r_T(H_T))}{\pi_{T,A_T}(H_T)}\right\}\eta_{T,r_T}(H_T)\right], \tag{5}$$

where $\pi_{T,A_T}(H_T)$ is the propensity score, and $\eta_{T,r_T}(H_T)$ is the estimated conditional mean. Thus, T-RL maximizes the counterfactual mean outcome through each of the nodes by optimizing Equation (5).

3.2. Stochastic Tree-Based Reinforcement Learning

In a multi-stage multi-treatment setting, the Stochastic Tree Search for Estimating Optimal Dynamic Treatment Regimes (ST-RL) method as introduced by [29] is applied to data from either randomized trials or observational studies. ST-RL adopts a stochastic tree-based approach to estimate DTRs. This means that it constructs a decision tree at each stage of treatment, where each node in the tree represents a possible treatment decision. To build a decision tree, ST-RL first models the mean of counterfactual outcomes via non-parametric regression models. Then, it considers a Markov chain Monte Carlo algorithm [42,43] to search for the optimal tree-structured decision rule stochastically. This means that it randomly selects a treatment decision at each stage and then evaluates the outcomes of this decision rule. The decision rule with the best outcomes is then selected as the optimal DTR.

To fit $E(Y^*|A_t = a_t, H_t)$, SL-RL uses Bayesian additive regression trees (BART) and predicts the counterfactual outcomes for each patient. In the backward induction implementation of SL-RL at the final stage T, the final response variable is used to estimate DTR. At each intermediate stage $1 < t < T$, the counterfactual outcome depends on the optimal treatment regimes in all future stages and needs to be predicted.

To mitigate the accumulation of bias from the multi-stage backward induction, the intermediate outcome (also called pseudo-outcome) incorporates both the real observed intermediate response variable value at stage t and the projected future loss resulting from sub-optimal treatments. The pseudo-outcome (PO) will be:

$$PO = Y + \sum_{t=t+1}^{T} \{E[Y^*|r_t^{opt}(H_t), H_t] - E[Y|A_t = a_t, H_t]\},$$

where the pseudo-outcome (PO) is used as the outcome for stage $t-1$ in the backward induction process.

3.3. Causal Tree-Based Method

Tree-based methods proposed for deriving treatment recommendations have been constructed using Classification and Regression Trees (CART) or Random Forests, thus without modeling causal effects. However, to understand the impact of various therapeutic interventions on a patient, it is crucial to investigate the causal consequences of these treatments. Hence, ref. [21] introduced a causal tree approach designed to directly assess treatment effects while providing a causal interpretation. This is achieved through the implementation of a customized splitting criterion for decision trees.

Causal tree learning, designed for estimating treatment effects, employs the optimization of distinct criteria compared to decision tree learning, which focuses on prediction and classification. While decision trees utilize splitting criteria focused on accuracy in predicting a target variable, causal tree learning necessitates the incorporation of two essential heuristics. The first heuristic pertains to the achievement of a balance between treated and untreated individuals within each leaf for a given treatment. This balance is crucial to accurately estimate outcome differences and subsequently calculate an unbiased treatment effect estimation. The second heuristic entails the partitioning of leaves into distinct groups with different outcomes. This division ensures that the accuracy of the estimated treatment effects remains uncompromised.

The causal tree learning algorithm addresses these challenges by employing a modified splitting criterion as a heuristic for determining optimal splits. Instead of relying on conventional decision tree splitting measures such as Gini impurity and information gain, causal tree learning employs the Expected Mean Squared Error for Treatment Effects (EMSE). This metric is explicitly designed for the estimation of heterogeneous treatment effects as introduced by [44].

Given the propensity score and patient medical history, at stage t, EMSE can be formulated as:

$$EMSE = \sum_{t=1}^{T} \sum_{i=1}^{K} \pi_t(H_t)(Y(a_t) - Y^*(r_t))^2 + \tau^2, \quad (6)$$

where τ is the causal effect of treatment on outcome. It is the difference in the expected outcome between patients who receive the treatment and patients who do not receive the treatment. The value of τ is added to the squared difference between the true outcome under the optimal treatment allocation rule and the outcome under the actual treatment allocation rule. This means that the EMSE is increased by the amount of the causal effect of treatment on the outcome. This is because the optimal treatment allocation rule is the one that maximizes the difference in the expected outcome between patients who receive the treatment and patients who do not receive the treatment.

To obtain the expected counterfactual mean, we first calculate the expected squared error for each possible treatment rule given a patient history

$$\text{Expected squared error} = \sum_{i=1}^{K} \pi_t(H_t)(Y_{a_t} - Y^*(r_t))^2$$

We choose the treatment rule with the lowest expected squared error and calculate the expected counterfactual mean for the chosen treatment assignment as

$$\text{Expected counterfactual mean } = \sum_{i=1}^{K} \pi_{a_t} Y^*(r_t),$$

where $Y^*(r_t)$ is the counterfactual outcome under the optimal treatment.

The EMSE (Expected Mean Squared Error) formula, applied when splitting causal tree nodes for identifying dynamic treatment regimes, chooses the treatment assignment that minimizes the anticipated squared error. This selection takes into account the current state of the data, the estimated causal effect of treatment on the outcome, and the probabilities of assigning each treatment. The optimal treatment assignment, therefore, minimizes the expected squared error.

Ref. [21] uses the EMSE splitting criterion in the binary treatment option scenario, and we modify it for our experiment as in Equation (6) above to fit it for the multi-treatment environment.

4. The Experimental Study: DTRs for Diabetic Kidney Disease

In this study, we perform an experimental analysis of data specifically centered around patients exhibiting a distinct state of diabetic kidney disease (DKD) to deduce optimal treatment strategies. The dataset under consideration was meticulously gathered from disparate

prospective observational studies, notably, the PROVALID studies, encompassing a diverse spectrum of chronic kidney disease (CKD) states as documented by [45,46].

4.1. The Dataset

The PROVALID (PROspective cohort study in patients with type 2 diabetes mellitus for VALIDation of biomarkers) study prospectively collected data on 4000 individuals with type 2 diabetes mellitus (DM2). The data encompass detailed information on the participants' medical history, physical examination findings, laboratory measurements, and prescribed medications. Data collection was conducted sequentially, including the documentation of treatment regimens [47]. Notably, treatment selection was restricted to a predefined set of four medications. These are as follows:

- Renin-Angiotensin-System-inhibitor (RASi)-only treatment;
- A combination of the Sodium-Glucose Transporter 2 inhibitor (SGLT2i) and the RASi treatment;
- A combination of the Glucagon-Like Peptide 1 receptor agonist (GLP1a) and the RASi treatment.
- A combination of the MineraloCorticoid Receptor Antagonist (MCRa) and the RASi treatment.

Patients receiving RASi monotherapy or RASi in combination with another drug either remained on their current treatment or were transitioned to a different combination therapy.

In our experimental investigation, we consider a subset of patients from the PROVALID dataset, which consists of longitudinal data (covariates of the dataset are listed in Appendix A Table A1). The primary objective is to evaluate the performance of algorithms in determining optimal treatment decisions. By only selecting patients having at least three consecutive visits, the dataset comprises 241 patients, with 125 male and 116 female participants. Patient histories are constructed using 30 selected variables derived from the longitudinal data with the Bayesian Network approach as outlined in the next subsection. The selected variables are displayed in Appendix B Table A2. The outcome for each stage is determined by the estimated Glomerular Filtration Rate (eGFR) value recorded each visit, denoted by t = 0, 1, 2, 3 for the baseline, year one, year two, and year three, respectively.

Treatments for diabetic kidney disease (DKD) are prescribed annually. At each stage, patients are offered one of four treatment options: RASi, a combination of SGLT2i and RASi, a combination of GLP1a and RASi, and a combination of MCRa and RASi.

During the first year, 92% of patients received a RASi-only drug, while in year 2, 81% of patients were treated with RASi, and the rest with other drug combinations. At year 3, patients who received RASi treatment decreased to 79%. The primary outcome of interest is the estimated glomerular filtration rate (eGFR) observed after each treatment year. In this context, a higher eGFR value is expected to indicate improvement in the disease.

4.2. Variable Selection

To address the high dimensionality of the PROVALID dataset and elucidate the network of relationships underlying diabetic kidney disease (DKD) pathophysiology, we adopt the approach of Bayesian Networks (BNs) [2,48] considering the properties of the Markov Blanket (MB) of the target variable Y. BNs are a well-established method in the medical field for representing and reasoning under uncertainty [49–51]. They are structured as directed acyclic graphs (DAGs), where nodes represent variables within the system, and directed arcs depict probabilistic dependencies between them.

BN estimation entails two key steps: structure learning, which identifies the network's topological structure, and parameter learning, which determines the probability distributions. Notably, various data-driven approaches exist for BN estimation [52], many of which allow the incorporation of prior knowledge from the literature and clinical practice. This integration enhances model informativeness and mitigates the effects of data noise and variability.

Within a BN framework, the Markov Blanket (MB) of a node (variable) encompasses directly dependent nodes. This approach to variable selection identifies the minimal subset of variables containing all information necessary for predicting the target variable, rendering additional variables redundant. Importantly, this methodology is theoretically optimal, guaranteeing the selection of the most effective variable subset. This approach not only achieves dimensionality reduction but also enhances model interpretability and computational efficiency.

4.3. Experimental Setup

We perform our experimental evaluation by splitting the PROVALID dataset into a 70:20:10 train–test–validation split through stratified sampling using the drug option as a subgroup, ensuring samples include patients from every subgroup. Since there is no straightforward way to assess purity in tree-based reinforcement learning methods, we include the concept of maximizing the counterfactual average result by utilizing a 10-fold cross-validation estimator to determine the counterfactual mean outcome. Specifically, we use nine sub-samples as training data and the remaining sub-sample as test data. We repeat the process 10 times, with each sub-sample being the test data once.

Due to the absence of verifiable ground-truth data containing unequivocally accurate information regarding true treatments, this study operates under the assumption that the validation set serves as a surrogate ground-truth dataset. Consequently, the treatments delineated at each stage of this set are postulated to represent medically validated optimal interventions. This assumption is made in recognition of the inherent challenge posed by the unavailability of an absolute reference dataset, and it serves as a pragmatic approach to approximating the most efficacious treatments within the scope of this investigation. It is acknowledged that this reliance on the validation set introduces an element of uncertainty, and efforts have been undertaken to mitigate biases and ensure robustness in the analysis and interpretation of the results.

In addition to the previously explained algorithms, we add the Q-learning with Random Forest approach to compare these different tree-growing strategies with a classic well-known tree-growing strategy.

Measurement Metrics

Diabetic kidney disease (DKD), a complication of diabetes, significantly impacts kidney function. eGFR (estimated Glomerular Filtration Rate) serves as a critical indicator of kidney health in DKD patients. Identifying optimal treatment regimes is essential for managing eGFR decline and slowing disease progression. However, due to the complex and dynamic nature of DKD, pinpointing the most effective treatment pathway for individual patients can be challenging. To assess the effectiveness of machine learning models discussed in Section 3 in the context of DKD treatment, we need robust evaluation metrics. Here, we introduce two key metrics for evaluating their performance:

1. **Expected Mean and Standard Deviation value of eGFR ($E\{Y^*(r^{opt})\}$)**—calculated from the counterfactual eGFR under the estimated optimal treatment regime selected at a given stage. This is the average eGFR value that we would expect if the patient were to receive the estimated optimal treatment regime. A higher expected mean eGFR suggests that the optimal treatment regime is likely to be more effective in preserving kidney function. If the mean eGFR under the optimal treatment is significantly higher compared to the expected mean eGFR under the current or other regimes, it indicates that the optimal treatment is better at maintaining eGFR levels.
2. **Optimal Classification Rate (Optimality Percentage)**—a comprehensive comparison with the corresponding ground-truth treatments, which reveals the percentage rate of subjects accurately classified (assigned) into their respective optimal treatment categories, providing a quantitative measure of the model's accuracy and predictive validity. In other words, this is the percentage of subjects correctly assigned to their

optimal treatment categories based on the model's predictions, compared to the ground truth.

These two metrics are interconnected. A high optimality percentage indicates the model is accurately assigning patients to the optimal treatment categories, which, ideally, translates to better health outcomes reflected in a higher expected mean eGFR and lower standard deviation. This signifies the effectiveness of the model in identifying treatment regimes that maintain or improve kidney function.

4.4. Experimental Results

In this section, we present a comprehensive evaluation of algorithmic performance via a series of experiments conducted to assess both single-stage and two-stage reinforcement learning methodologies in determining optimal treatment decisions for individual patients. In each distinct scenario, the training datasets are employed to ascertain the optimal regime. Subsequently, the model is executed on the test datasets to validate the impartiality of the estimated models. The validation set is then utilized to compute the optimality and counterfactual mean, leveraging the known underlying truth.

4.4.1. Single Stage

In this scenario, we consider a single stage with $T = 1$ and four treatment options with $K = 4$. The treatment A is denoted as $1, 2, 3, 4$ to represent RASi, RASi + SGLT2i, RASi + GLP1a, and RASi + MCRA treatment options, respectively.

Table 1 presents the performance measurements of all algorithms considered in a single-stage scenario with baseline covariates. For the tree-based reinforcement learning (T-RL), DTR-Causal Tree (DTR-CT), and DTR-Causal Forest (DTR-CF) algorithms that rely on treatment assignment probabilities, we employ multinomial logistic regression with the observed treatment as the dependent variable and all baseline covariates as explanatory variables. Additionally, T-RL also requires the specification of an outcome regression model for $E(Y|X)$, for which we choose a Random Forest regression model. The experimental results presented in the table compare the performance of the methods in terms of the optimality percentage (Opt%) and the expected mean eGFR $E\{Y^*(r^{opt})\}$ with their respective standard deviations.

Table 1. Evaluation results for one-stage scenario. Opt% presents the mean of the percentage of patients correctly classified to their optimal treatments on the validation set. $E\{Y^*(r^{opt})\}$ represents the empirical mean and empirical standard error estimates of the expected counterfactual outcome under the estimated optimal regime.

Method	**Optimality (Opt)%**	$E\{Y^*(r^{opt})\}$
Q-learning with Random Forest (Q-RF)	51	60.6 (18.5)
DTR-Causal Tree (DTR-CT)	76.5	64 (16.4)
DTR-Causal Forest (DTR-CF)	78	65.8 (16.04)
Stochastic tree-based reinforcement learning (SL-RL)	73	63.8 (16.1)
Tree-based reinforcement learning (T-RL)	61	61.3 (17.7)

Among these methods, the DTR-Causal Forest (DTR-CF) demonstrates the highest optimality percentage at 78%, indicating its superior capability in assigning optimal treatments compared to other methods. This method also shows a high expected mean outcome of 65.8 with a relatively low standard deviation of 16.04, further signifying its consistency and reliability of how the optimal treatment from this algorithm slows down eGFR decline. DTR-Causal Tree (DTR-CT) follows closely with an optimality of 76.5% and an expected mean outcome of 64, suggesting that while slightly less optimal than DTR-CF, it remains a strong contender in terms of performance and stability (standard deviation of 16.4).

Stochastic tree-based reinforcement learning (SL-RL) and tree-based reinforcement learning (T-RL) present moderate performance with optimality percentages of 73% and

61%, respectively. The expected mean outcomes for these methods are 63.8 (standard deviation 16.1) for SL-RL and 61.3 (standard deviation 17.7) for T-RL, indicating that while they perform reasonably well, their variability is slightly higher compared to DTR-based methods, further indicating the unreliable effect of the treatments assigned by the algorithms.

Q-learning with Random Forest (Q-RF) shows the lowest optimality percentage at 51% and an expected mean outcome of 60.6, with a higher standard deviation of 18.5. This suggests that Q-RF is less effective in assigning optimal treatments and exhibits greater variability in its outcomes compared to the other methods evaluated.

In summary, the DTR-Causal Forest (DTR-CF) and DTR-Causal Tree (DTR-CT) methods outperform the other methods in terms of both optimality and expected mean outcomes, with DTR-CF being the most stable and reliable method. The SL-RL and T-RL methods provide moderate performance, while Q-RF demonstrates the least optimality and highest variability, indicating its relative inefficiency in this context.

4.4.2. Two Stage

We extend our experiment to encompass four treatment options in a two-stage scenario. To establish a baseline, we utilize the initial two years of clinical observations, incorporating sixty covariate values. Subsequently, we delineate the disease progression from the second year to the third year as the first stage, characterized by the utilization of thirty covariate values. Similarly, the transition from year three to year four constitutes the second stage, with an analogous configuration of thirty covariates. This staged approach allows for a nuanced examination of the evolving dynamics in the clinical data over the specified temporal intervals, facilitating a comprehensive understanding of the treatment landscape for patients within the designated study period.

The results described in Table 2 confirm that DTR-CF has better performance in assigning optimal treatments to patients. The DTR-Causal Forest (DTR-CF) algorithm demonstrates the highest optimality percentage at 85%, indicating its superior efficacy in identifying optimal treatments across the two stages. Correspondingly, it achieves the highest expected mean outcome of 78.03 with a standard deviation of 13.9, showcasing both its stability and consistency. The DTR-Causal Tree (DTR-CT) algorithm follows closely with an optimality percentage of 82.3% and an expected mean outcome of 74, along with a standard deviation of 14.2, confirming its strong performance and stability in this multi-stage context.

Table 2. Evaluation results for the two-stage scenario. Opt% presents the mean of the percentage of patients correctly classified to their optimal treatments on the validation set. $E\{Y^*(r^{opt})\}$ represents the empirical mean and empirical standard error estimates of the expected counterfactual outcome under the estimated optimal regime.

Algorithm	**Optimality (Opt)%**	$E\{Y^*(r^{opt})\}$
Q-learning with Random Forest (Q-RF)	57	68.34 (16.5)
DTR-Causal Tree (DTR-CT)	82.3	74 (14.2)
DTR-Causal Forest (DTR-CF)	85	78.03 (13.9)
Stochastic tree-based reinforcement learning (SL-RL)	73	72.43 (15.81)
Tree-based reinforcement learning (T-RL)	71.5	69.7 (16.2)

Stochastic tree-based reinforcement learning (SL-RL) and tree-based reinforcement learning (T-RL) show moderate performance levels with optimality percentages of 73% and 71.5%, respectively. SL-RL achieves an expected mean outcome of 72.43 with a standard deviation of 15.81, while T-RL achieves an expected mean outcome of 69.7 with a standard deviation of 16.2. These results indicate that while these algorithms perform reasonably well in a two-stage reinforcement learning setting, their variability is slightly higher compared to the DTR-based methods.

The Q-learning with Random Forest (Q-RF) algorithm shows the lowest optimality percentage at 57% and an expected mean outcome of 68.34, with a standard deviation of 16.5. This suggests that Q-RF is less effective in identifying optimal treatments and exhibits greater variability in its outcomes compared to the other algorithms evaluated, particularly in a multi-stage context.

The two-stage scenario generally shows improved performance metrics across all methods. The optimality percentages and expected mean outcomes are higher in the two-stage setting compared to the one-stage setting, indicating that the additional decision stage contributes to better optimization.

5. Conclusions

This study evaluated the effectiveness of various tree-based reinforcement learning algorithms in identifying optimal, multi-stage multi-treatment regimes for patients with diabetic kidney disease (DKD). We applied a multi-stage, multi-treatment framework on a clinical observational dataset. To assess the algorithms' ability to learn optimal treatment sequences and predict their impact on patient outcomes, we utilized two key metrics: the optimality percentage and expected counterfactual outcome with its standard deviation. The optimality percentage metric reflects the model's accuracy in assigning patients to the treatment regime that would have resulted in the best possible outcome, compared to a predefined ground truth (e.g., expert-recommended treatment plan or data from clinical trials). A high optimality percentage indicates the model's ability to replicate established best practices or identify even more effective treatment pathways. The Expected Counterfactual Outcome with Standard Deviation metric delves into the counterfactual scenario, which represents the predicted outcome (e.g., eGFR) that a patient would experience if they receive the treatment regime identified by the model as optimal.

The experimental results in Tables 1 and 2 indicate that both DTR-CF and DTR-CT algorithms hold significant promise for optimizing treatment regimes in diabetic kidney disease (DKD). These algorithms achieved the highest optimality percentages, indicating their superior accuracy in assigning patients to the most beneficial multi-stage treatment sequences compared to other methods. Additionally, the low standard deviations in the expected eGFR for both DTR-CF and DTR-CT suggest consistent treatment recommendations across the patient population. This consistency implies that these algorithms are likely to lead to similar improvements in kidney function for a broader range of DKD patients.

Furthermore, incorporating the Bayesian Network methodology significantly enhanced the analysis by enabling us to identify the most relevant variables influencing treatment decisions for DKD patients. Unlike traditional methods that might analyze all available variables, Bayesian Networks leverage the concept of Markov Blankets. Relevant variable clusters represent sets of variables that shield a specific variable from the influence of all other variables in the network. By focusing on the variables within a patient's Markov blanket for treatment decisions, the Bayesian Network approach streamlines the analysis and reduces the risk of incorporating irrelevant or redundant factors.

This focus on truly relevant variables strengthens the study's robustness in two key ways. First, it minimizes the potential for overfitting, which can occur when models become too reliant on specific details within the dataset and struggle to generalize to new data. Second, it clarifies the causal relationships between variables, providing a more transparent understanding of how different factors interact and influence treatment outcomes.

Looking forward, exploring these algorithms on larger datasets and including additional clinical factors could further refine treatment recommendations for DKD patients. Additionally, employing advanced model selection techniques can pinpoint the most impactful features, leading to even more precise and robust findings.

Overall, this study contributes to the development of personalized treatment strategies for diabetic kidney disease by showing the potential of tree-based reinforcement learning algorithms.

Author Contributions: S.A.: Conceptualization, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing—original draft, Writing—review and editing. I.P.: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Resources, Supervision, Writing—original draft, Writing—review and editing. R.D.J.: Writing—review and editing. D.S.: Data curation, Writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This work is funded by the European Union's Horizon 2020 research and innovation program under grant agreement No. 848011. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Institutional Review Board Statement: The studies involving human participants were reviewed and approved by the Ethics Committee of the Medical University Innsbruck; DC-ren approval number: EK Nr:1188/2020, date 18 June 2020. The DC-ren cohort consists of patients from PROVALID and written informed consent to participate in this study was obtained from all patients. The PROVALID dataset used in this study was approved by the local Institutional Review Board (IRB) in each participating country, and are listed below. Signing an informed consent was a prerequisite for study participation in all countries. Austria: Ethical approval from the Ethics Committee of the Medical University Innsbruck AN4959 322/4.5370/5.9 (4012a); 29 January 2013 and approval of the Ethics Committee of Upper Austria, Study Nr. I-1-11; 30 December 2010. Hungary: Approval from Semmelweis University, Regional and Institutional Committee Of Science And Research Ethics; No.12656-0/2011-EKU (421/PV11.); 17 June 2011. United Kingdom: Approval from WoSRES, NHS; Rec. Reference:12/WS/0005 (13 January 2012). Netherlands: Approval of the Medical Ethical Committee of the University Medical Center Groningen, ABRnr. NL35350.042.11. Poland: Approval from Ethics Committee of the Medical University of Silesia, KNW/022/KB1/78/11/, date 7 June 2011.

Data Availability Statement: The data analyzed in this study is subject to the following licenses/restrictions: data owned by the European Union DC-ren project. Requests to access these datasets should be directed to GM, gert.mayer@i-med.ac.at.

Acknowledgments: The authors would like to acknowledge many useful reviews and conversations with all the members of the DC-ren consortium (https://dc-ren.eu/), accessed on 27 May 2024.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Covariates in the PROVALID Study

Table A1. Covariates in the PROVALID study.

Variable ACRONYM	Variable DESCRIPTION
GE	Gender
HEIGHT	Height
ADMD	Age at DM2 diagnosis
AHDT	Age at HT diagnosis
SDMAV	Severity of DM2 at first visit in PROVALID
DDMAV	DM2 duration at the first visit in PROVALID (first for the patient sequence)
DDMT	Duration of DM2 pharmacological treatment at first visit in PROVALID (first for the patient sequence)
HTDAV	HT duration at first visit in PROVALID (first for the patient sequence)
SHTAV	Severity of HT at visit in PROVALID (+1 for each HT drug)
DHTT	Duration of HT pharmacological treatment at firt visit in PROVALID (first for the patient sequence)
PHDRB	Personal history of diabetic retinopathy at baseline
PHRDB	Personal history of renal disease at baseline
PHHFB	Personal history of heart failure stage III or IV at baseline

Table A1. *Cont.*

Variable ACRONYM	Variable DESCRIPTION
PHCADB	Personal history of coronary artery disease (any angina, myocardial infarction, coronary intervention) at baseline
PHPADB	Personal history of peripheral artery disease (Claudicatio, amputation, etc) at baseline
PHCVDB	Personal history of cerebrovascular disease (stroke, TIA, PRIND)
SMOK	Smoking
FHRD	Family history of renal disease
FHHT	Family history of hypertension
FHDM	Family history of type 2 diabetes
FHCVD	Family history of cardiovascular disease
FHM	Family history of malignancy
BW	Body weight [kg]
SBP	Systolic BP
DBP	Diastolic BP
AGEV	Age at visit
BMI	Body Mass Index
MABP	Mean arterial blood pressure
PP	Pulse pressure
BG	Blood glucose
HBA1C	HbA1C
SCR	Serum creatinine
TOTCHOL	Serum cholesterol (total)
LDLCHOL	Serum cholesterol (LDL)
HDLCHOL	Serum cholesterol (HDL)
STRIG	Serum triglycerides
SPOT	Serum potassium
HB	Hemoglobin
SALB	Serum albumin
CRP	CRP
EGFR	eGFR
UACR	mean UACR
LDLHDLR	LDL/HDL cholesterol ratio
EVLDLCHOL	(new) estimated VLDL based on the Friedewald equation—only for STRIG < 400
ELDLCHOL	(new) estimated LDL based on the Friedewald equation—only for STRIG < 400
ELDLHDLR	(new) LDL/HDL cholesterol ratio based on data—when available—or estimation
UCREA	Urinary creatinine
CA_CL_num	Calcium concentration in serum—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$
PHOS_CL_num	Phosphate concentration in serum—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$

Table A1. *Cont.*

Variable ACRONYM	Variable DESCRIPTION
CST3_num	Cystatin C concentration in serum—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$
CPEP_CL_num	C-peptide concentration in serum—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$
FFA_CL_num	Free Fatty Acids concentration in serum—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$
UA_CL_num	Uric Acid concentration in serum—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$
SO_CL_num	Sodium concentration in urine—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$
POT_CL_num	Potassium concentration in urine—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$
CHL_CL_num	Chloride concentration in urine—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$
UNA24H	24 h urinary sodium excretion
NIDR	New incidence of diabetic retinopathy (DR)
NIMI	New incidence of non fatal myocardial infarction (NFMI)
NIS	New incidence of non fatal stroke (NFS)
NIHF	New incidence of heart failure (stage III/IV)
NICAD	New incidence of coronary artery disease (CAD)
NICVD	New incidence of cerebrovascular disease (CD)
NIPAD	New incidence of peripheral artery disease (PAD)
AHBB	Beta-receptor blockers
AHCA	Calcium antagonists
AHCAAH	Centrally acting antihypertensives
AHARB	Alpha-receptor blockers
AHDV	Direct vasodilators
ADSU	Sulfonylureas
ADPPI	Meglitinides (glinides)
ADGL	DPPIV inhibitors or GLP1 analogs
ADGLIT	Thiazolinediones (glitazones)
ADMET	Biguanides (metformin)
ADAGI	Alpha-Glucosidase inhibitors
ADI	Insulins
LLCFA	Clofibric acid derivative
LLSTAT	Statins
LLOTHER	Other lipid-lowering drugs (ezetimibe, omega 3 acid)
APASA	ASA
APTPD	Thienopyridine derivatives
APDIP	Dipyridamole
APGPI	GPIIb/IIIa inhibitors
APOTHER	Other platelet aggregation inhibitors (ticagrelor)
VDAC	Alfacalcidol

Table A1. *Cont.*

Variable ACRONYM	Variable DESCRIPTION
VDCCF	Colecalciferol
EPODA	DarbEpoetin alfa
EPOEA	Epoetin alfa
EPOEB	poetin beta
IO	Oral iron
PBCB	Calcium-based
DLOOP	Loop diuretics
DTH	Thiazides
DPS	Potassium-saving diuretics
AC	Analgesics combinations
ASC	Single-component analgesics
TAH	Group "TAH"
TAD	Group "TAD"
TADI	TADI
TLL	Group "TLL"
TEPO	Group "TEPO"
TDIU	Group "TDIU"
MMP7_LUM_num	Matrix Metallopeptidase 7 concentration in serum—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$
VEGFA_LUM_num	Vascular Endothelial Growth Factor A concentration in serum—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$
AGER_LUM_num	Advanced Glycosylation End-Product Specific Receptor concentration in serum—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$
LEP_LUM_num	Leptin concentration in serum—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$
ICAM1_LUM_num	Intercellular Adhesion Molecule 1 concentration in serum NA below min and above max replaced by $0.5 * min$ and $1.5 * max$
TNFRSF1A_LUM_num	TNF Receptor Superfamily Member 1A concentration in serum—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$
IL18_LUM_num	Interleukin 18 concentration in serum—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$
DPP4_LUM_num	Dipeptidyl Peptidase 4 concentration in serum—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$
LGALS3_LUM_num	Galectin 3 concentration in serum—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$
SERPINE1_LUM_num	Serpin Family E Member 1 concentration in serum—NA below min and above max replaced by 0.5*min and 1.5*max
ADIPOQ_LUM_num	Adiponectin, C1Q And Collagen Domain Containing concentration in serum—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$
EGF_MESO_num_norm	epidermal growth factor concentration in urine normalized by UCREA—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$

Table A1. *Cont.*

Variable ACRONYM	Variable DESCRIPTION
FGF21_MESO_num_norm	fibroblast growth factor 21 concentration in urine normalized by UCREA—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$
IL6_MESO_num_norm	Interleukin 6 concentration in urine normalized by UCREA—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$
HAVCR1_MESO_num_norm	hepatitis A virus cellular receptor 1 concentration in urine normalized by UCREA—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$
CCL2_MESO_num_norm	C-C motif chemokine ligand 2 concentration in urine normalized by UCREA—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$
MMP2_MESO_num_norm	matrix metallopeptidase 2 concentration in urine normalized by UCREA—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$
MMP9_MESO_num_norm	matrix metallopeptidase 9 concentration in urine normalized by UCREA—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$
LCN2_MESO_num_norm	lipocalin-2 concentration in urine normalized by UCREA—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$
NPHS1_MESO_num_norm	NPHS1 adhesion molecule, nephrin concentration in urine normalized by UCREA—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$
THBS1_MESO_num_norm	thrombospondin 1 concentration in urine normalized by UCREA—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$

Appendix B. Covariates Selected by Bayesian Network

Table A2. Covariates after variable selection using Bayesian network in the PROVALID study.

Variable ACRONYM	Variable DESCRIPTION
GE	Gender
DLOOP	Loop diuretics
SCR	Serum creatinine
CST3_num	Cystatin C concentration in serum—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$
PHRDB	Personal history of renal disease at baseline
EGF_MESO_num_norm	epidermal growth factor concentration in urine normalized by UCREA—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$
FGF21_MESO_num_norm	fibroblast growth factor 21 concentration in urine normalized by UCREA—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$
HB	Hemoglobin
HDLCHOL	Serum cholesterol (HDL)
ICAM1_LUM_num	Intercellular Adhesion Molecule 1 concentration in serum NA below min and above max replaced by $0.5 * min$ and $1.5 * max$
LEP_LUM_num	Leptin concentration in serum—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$

Table A2. *Cont.*

Variable ACRONYM	Variable DESCRIPTION
MMP7_LUM_num	Matrix Metallopeptidase 7 concentration in serum—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$
SPOT	Serum potassium
TNFRSF1A_LUM_num	TNF Receptor Superfamily Member 1A concentration in serum—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$
UACR	mean UACR
CRP	CRP
LDLCHOL	Serum Cholesterol (LDL)
HBA1C	HbA1C
BG	Blood glucose
ADMET	Biguanides (metformin)
GE	Gender
UCREA	Urinary creatinine
DBP	Diastolic BP
LGALS3_LUM	Galectin 3 concentration in serum—NA below min and above max replaced by $0.5 * min$ and $1.5 * max$
ADMD	Age at DM2 diagnosis
STRIG	Serum triglycerides
TOTCHOL	Serum cholesterol (total)
ADIPOQ_LUM_num	Adiponectin, C1Q, And Collagen Domain Containing concentration in serum—NA below min and above max replaced by 0.5*min and 1.5*max
AGEV	Age at visit
BMI	Body Mass Index

References

1. Pugliese, G.; Penno, G.; Natali, A.; Barutta, F.; Di Paolo, S.; Reboldi, G.; Gesualdo, L.; De Nicola, L. Diabetic kidney disease: New clinical and therapeutic issues. Joint position statement of the Italian Diabetes Society and the Italian Society of Nephrology on "The natural history of diabetic kidney disease and treatment of hyperglycemia in patients with type 2 diabetes and impaired renal function". *Nutr. Metab. Cardiovasc. Dis.* **2019**, *29*, 1127–1150. [PubMed]
2. Koller, D.; Friedman, N. *Probabilistic Graphical Models: Principles and Techniques*; MIT Press: Cambridge, MA, USA, 2009.
3. König, I.R.; Fuchs, O.; Hansen, G.; von Mutius, E.; Kopp, M.V. What is precision medicine? *Eur. Respir. J.* **2017**, *50*, 1700391. [PubMed]
4. Ginsburg, G.S.; Phillips, K.A. Precision medicine: From science to value. *Health Aff.* **2018**, *37*, 694–701.
5. Robins, J. A new approach to causal inference in mortality studies with a sustained exposure period—Application to control of the healthy worker survivor effect. *Math. Model.* **1986**, *7*, 1393–1512.
6. Robins, J.M. Correcting for non-compliance in randomized trials using structural nested mean models. *Commun. Stat. Theory Methods* **1994**, *23*, 2379–2412. [CrossRef]
7. Robins, J.M. Causal inference from complex longitudinal data. In *Proceedings of the Latent Variable Modeling and Applications to Causality*; Lecture Notes in Statistics; Springer: New York, NY, USA; Berlin/Heidelberg, Germany, 1997; pp. 69–117.
8. Murphy, S.A. Optimal dynamic treatment regimes. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2003**, *65*, 331–355.
9. Chakraborty, B.; Moodie, E.E. *Statistical Methods for Dynamic Treatment Regimes*; Springer: Berlin/Heidelberg, Germany, 2013; Volume 10, pp. 978–981.
10. Chakraborty, B.; Murphy, S.A. Dynamic treatment regimes. *Annu. Rev. Stat. Its Appl.* **2014**, *1*, 447–464. [CrossRef]
11. Wagner, E.H.; Austin, B.T.; Davis, C.; Hindmarsh, M.; Schaefer, J.; Bonomi, A. Improving chronic illness care: Translating evidence into action. *Health Aff.* **2001**, *20*, 64–78. [CrossRef] [PubMed]
12. Robins, J.M. Optimal structural nested models for optimal sequential decisions. In Proceedings of the Second Seattle Symposium in Biostatistics: Analysis of Correlated Data; Springer: Berlin/Heidelberg, Germany; New York, NY, USA, 2004; pp. 189–326.

13. Murphy, S.A.; van der Laan, M.J.; Robins, J.M.; Conduct Problems Prevention Research Group. Marginal mean models for dynamic regimes. *J. Am. Stat. Assoc.* **2001**, *96*, 1410–1423. [CrossRef]
14. Moodie, E.E.; Chakraborty, B.; Kramer, M.S. Q-learning for estimating optimal dynamic treatment rules from observational data. *Can. J. Stat.* **2012**, *40*, 629–645.
15. Wallace, M.P.; Moodie, E.E.; Stephens, D.A. Dynamic treatment regimen estimation via regression-based techniques: Introducing r package dtrreg. *J. Stat. Softw.* **2017**, *80*, 1–20.
16. Tsiatis, A.A.; Davidian, M.; Holloway, S.T.; Laber, E.B. *Dynamic Treatment Regimes: Statistical Methods for Precision Medicine*; CRC press: Boca Raton, FL, USA, 2019.
17. van der Laan, M.J.; Petersen, M.L.; Joffe, M.M. History-adjusted marginal structural models and statically-optimal dynamic treatment regimens. *Int. J. Biostat.* **2005**, *1*. [CrossRef]
18. Watkins, C.J.; Dayan, P. Q-learning. *Mach. Learn.* **1992**, *8*, 279–292. [CrossRef]
19. Murphy, S.A. A Generalization Error for Q-Learning. 2005. Available online: https://www.jmlr.org/papers/volume6/murphy05a/murphy05a.pdf (accessed on 27 May 2024).
20. Mahar, R.K.; McGuinness, M.B.; Chakraborty, B.; Carlin, J.B.; IJzerman, M.J.; Simpson, J.A. A scoping review of studies using observational data to optimise dynamic treatment regimens. *BMC Med. Res. Methodol.* **2021**, *21*, 39.
21. Blumlein, T.; Persson, J.; Feuerriegel, S. Learning optimal dynamic treatment regimes using causal tree methods in medicine. In Proceedings of the Machine Learning for Healthcare Conference. PMLR, Durham, NC, USA, 5–6 August 2022; pp. 146–171.
22. Tao, Y.; Wang, L. Adaptive contrast weighted learning for multi-stage multi-treatment decision-making. *Biometrics* **2017**, *73*, 145–155. [CrossRef] [PubMed]
23. Laber, E.B.; Zhao, Y.Q. Tree-based methods for individualized treatment regimes. *Biometrika* **2015**, *102*, 501–514.
24. Zhang, Y.; Laber, E.B.; Tsiatis, A.; Davidian, M. Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics* **2015**, *71*, 895–904. [PubMed]
25. Zhang, Y.; Laber, E.B.; Davidian, M.; Tsiatis, A.A. Interpretable dynamic treatment regimes. *J. Am. Stat. Assoc.* **2018**, *113*, 1541–1549. [CrossRef]
26. Lakkaraju, H.; Rudin, C. Learning cost-effective and interpretable treatment regimes. In Proceedings of the Artificial Intelligence and Statistics. PMLR, Fort Lauderdale, FL, USA, 20–22 April 2017; pp. 166–175.
27. Rivest, R.L. Learning decision lists. *Mach. Learn.* **1987**, *2*, 229–246. [CrossRef]
28. Tao, Y.; Wang, L.; Almirall, D. Tree-based reinforcement learning for estimating optimal dynamic treatment regimes. *Ann. Appl. Stat.* **2018**, *12*, 1914. [CrossRef]
29. Sun, Y.; Wang, L. Stochastic tree search for estimating optimal dynamic treatment regimes. *J. Am. Stat. Assoc.* **2021**, *116*, 421–432. [CrossRef]
30. Min, J.; Elliott, L.T. Q-learning with online random forests. *arXiv* **2022**, arXiv:2204.03771.
31. Alyass, A.; Turcotte, M.; Meyre, D. From big data analysis to personalized medicine for all: Challenges and opportunities. *BMC Med. Genom.* **2015**, *8*, 33. [CrossRef]
32. Mathur, S.; Sutton, J. Personalized medicine could transform healthcare. *Biomed. Rep.* **2017**, *7*, 3–5. [PubMed]
33. Denson, L.A.; Curran, M.; McGovern, D.P.; Koltun, W.A.; Duerr, R.H.; Kim, S.C.; Sartor, R.B.; Sylvester, F.A.; Abraham, C.; de Zoeten, E.F.; et al. Challenges in IBD research: Precision medicine. *Inflamm. Bowel Dis.* **2019**, *25*, S31–S39.
34. Martin, T.P.; Hanusa, B.H.; Kapoor, W.N. Risk stratification of patients with syncope. *Ann. Emerg. Med.* **1997**, *29*, 459–466. [PubMed]
35. Roberts, M.C. Implementation challenges for risk-stratified screening in the era of precision medicine. *JAMA Oncol.* **2018**, *4*, 1484–1485. [CrossRef]
36. Rosenbaum, P.R.; Rubin, D.B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **1983**, *70*, 41–55.
37. Robins, J.M.; Hernán, M.A. Estimation of the causal effects of time-varying exposures. *Longitud. Data Anal.* **2009**, *553*, 599.
38. Plant, D.; Barton, A. Machine learning in precision medicine: Lessons to learn. *Nat. Rev. Rheumatol.* **2021**, *17*, 5–6. [CrossRef]
39. Zhou, N.; Brook, R.D.; Dinov, I.D.; Wang, L. Optimal dynamic treatment regime estimation using information extraction from unstructured clinical text. *Biom. J.* **2022**, *64*, 805–817. [CrossRef] [PubMed]
40. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Routledge: Wadsworth, OH, USA; Belmont, MA, USA, 1984; ISBN 978-0412048418.
41. Robins, J.M.; Rotnitzky, A.; Zhao, L.P. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Am. Stat. Assoc.* **1995**, *90*, 106–121. [CrossRef]
42. Chipman, H.A.; George, E.I.; McCulloch, R.E. Bayesian CART model search. *J. Am. Stat. Assoc.* **1998**, *93*, 935–948. [CrossRef]
43. Wu, Y.; Tjelmeland, H.; West, M. Bayesian CART: Prior specification and posterior simulation. *J. Comput. Graph. Stats.* **2007**, *16*, 44–66.
44. Athey, S.; Imbens, G. Recursive partitioning for heterogeneous causal effects. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 7353–7360. [CrossRef] [PubMed]
45. Mayer, G.; Eder, S.; Rosivall, L.; Voros, P.; Heerspink, H.L.; de Zeeuw, D.; Czerwienska, B.; Wiecek, A.; Hillyard, D.; Mark, P.; et al. Baseline Data from the Multinational Prospective Cohort Study for Validation of Biomarkers (Provalid). *Nephrol. Dial. Transplant.* **2016**, *31*, 1482. [CrossRef]

46. Eder, S.; Leierer, J.; Kerschbaum, J.; Rosivall, L.; Wiecek, A.; de Zeeuw, D.; Mark, P.B.; Heinze, G.; Rossing, P.; Heerspink, H.L.; et al. A prospective cohort study in patients with type 2 diabetes mellitus for validation of biomarkers (PROVALID)—Study design and baseline characteristics. *Kidney Blood Press. Res.* **2018**, *43*, 181–190. [CrossRef] [PubMed]
47. Gregorich, M.; Heinzel, A.; Kammer, M.; Meiselbach, H.; Böger, C.; Eckardt, K.U.; Mayer, G.; Heinze, G.; Oberbauer, R. A prediction model for the decline in renal function in people with type 2 diabetes mellitus: Study protocol. *Diagn. Progn. Res.* **2021**, *5*, 19.
48. Scutari, M.; Denis, J.B. *Bayesian Networks: With Examples in R*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2021.
49. Scutari, M.; Auconi, P.; Caldarelli, G.; Franchi, L. Bayesian networks analysis of malocclusion data. *Sci. Rep.* **2017**, *7*, 15236. [CrossRef]
50. Arora, P.; Boyne, D.; Slater, J.J.; Gupta, A.; Brenner, D.R.; Druzdzel, M.J. Bayesian networks for risk prediction using real-world data: A tool for precision medicine. *Value Health* **2019**, *22*, 439–445.
51. Shen, J.; Liu, F.; Xu, M.; Fu, L.; Dong, Z.; Wu, J. Decision support analysis for risk identification and control of patients affected by COVID-19 based on Bayesian Networks. *Expert Syst. Appl.* **2022**, *196*, 116547. [CrossRef]
52. Kitson, N.K.; Constantinou, A.C.; Guo, Z.; Liu, Y.; Chobtham, K. A survey of Bayesian Network structure learning. *Artif. Intell. Rev.* **2023**, *56*, 8721–8814.