Bachelor’s Programme in Science and Technology

Machine learning for personalized treatment decision-making

Linh Ngo

Aalto University

BACHELOR’S THESIS

Aalto University
BACHELOR'S THESIS 2022

Machine learning for personalized treatment decision-making

Linh Ngo

Thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Science in Technology.
Otaniemi, 09 Sep 2022

Supervisor: associate professor Maarit Käpylä
Advisor: doctoral candidate Sophie Wharrie

Aalto University
School of Science
Bachelor's Programme in Science and Technology

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Linh Ngo

Title
Machine learning for personalized treatment decision-making

School School of Science

Degree programme Bachelor's Programme in Science and Technology

Major Data Science **Code** SCI3095

Supervisor associate professor Maarit Käpylä

Advisor doctoral candidate Sophie Wharrie

Level Bachelor's thesis **Date** 09 Sep 2022 **Pages** 30 + 6 **Language** English

Abstract

Treatments are rarely effective for all patients, and under all circumstances. A treatment that works for one patient may not be applied to others, and a treatment that works for one earlier may not be effective later (e.g. due to drug resistance). Since there is no one-size-fits-all way to treatment, personalized treatment by incorporating patient characteristics and tailoring the patient needs into treatment decisions appears as a transformative approach to healthcare.

In general, clinical experts usually have to make decisions about whether to prescribe a treatment to a patient or not. To determine the best treatment to administer to a patient, information about the treatment effects of each possible treatment actions is required. The problem of estimating treatment effects can be formulated as a causal inference problem, and machine learning-based methods for causal inference can be utilized to estimate individualized treatment effects.

Moreover, healthcare decisions also include the search for treatment regimes that yield the best outcome in all states of a patient's disease. Those regimes work in a dynamic fashion as the patient's state has to be updated over time with regards to the progression of the disease. Reinforcement learning with the main mechanism of feedback and improvement has strong potential to learn dynamic treatment regimes, thus reinforcement learning methods are usually applied to derive an optimal treatment strategy in the time-varying and dynamic setting.

This thesis is a literature review which aims to study and compare two different machine learning approaches for personalized treatment decision-making: causal inference for individualized treatment effects estimation and reinforcement learning for optimal dynamic treatment regimes estimation. For each approach, background knowledge, the objective, the overview of data sources, machine learning methods, evaluation methods, and applications in recent years are introduced.

This thesis concludes that both approaches show great potential in providing personalized treatment recommendations and transforming healthcare. However, there are still some limitations that need to be addressed to make more interpretable and trustful decisions. One suggestion would be to take advantage of the capabilities of both approaches and incorporate causal model into the learning process of reinforcement learning.

Keywords machine learning, personalized treatment decision-making, causal inference, reinforcement learning, dynamic treatment regimes, individualized treatment effects

urn https://aaltodoc.aalto.fi

Contents

1. Introduction

Over the last few years, machine learning (ML) has made a significant impact in the healthcare domain by assisting doctors with personalized treatment decision-making for their patients. A personalized treatment decision involves making treatment choices for the patient that aim to optimize their individual health outcomes. The increasing adoption of electronic health records (EHRs) as a means of storing patients medical history facilities the heterogeneity of observational data for various patients and diseases (Bica, Ahmed M Alaa, Lambert, et al., 2021). This source of observational data can also be utilized in ML algorithms to support doctors with the decision-making process.

Treatment decisions that doctors may typically encounter include deciding whether to administer a treatment or not (binary treatment decisions), and long-term monitoring of treatments for patients with chronic diseases or patients in intensive care units (sequential treatment decisions). Causal inference and reinforcement learning approaches have been developed for both these types of problems. Generally, causal inference allows one to quantify the causal effects of every possible treatment conditioned on a patient's health characteristics, hence the most appropriate treatment option can be determined. However, causal inference using observational data usually faces a problem of underestimating or overestimating the treatment effects due to the biases from confounding variables and treatment selection. Reinforcement learning (RL) with the main mechanism of learning from experiences and adjusting the policy to achieve the best long-term outcome is a suitable approach for determining dynamic sequential treatment decisions that need to be adjusted based on the progression of a patient's disease. However, the RL approach towards personalized decision-making poses difficulties in the interpretability of the models, the formulation of reward function that can balance the trade-off among the toxicity, efficacy, and cost of the treatment, and the balance between exploration and exploitation.

Therefore, this thesis is a literature review which aims to critically analyse the feasibility of different ML techniques for personalized treatment decision-making. This thesis specifically focuses on two main approaches for personalized treatment decision-making. The first approach includes using ML methods for causal inference to quantify the individualized treatment effects from observational data. The second approach consists of using RL methods to find an optimal policy for dynamic treatment regimes (DTRs). To address this purpose, this thesis seeks to answer the following research questions:

- How machine learning has been used for personalized treatment decision-making?
- What are the similarities and differences of causal inference and reinforcement learning approaches for personalized decision-making?
- What are the current state-of-the-art approaches for personalized treatment decision-making?
- What medical applications have been utilized from these approaches?

The remainder of this thesis is structured into five sections. The second section introduces some important background knowledge about casual inference, individualized treatment effects, dynamic treatment regimes, and reinforcement learning. The third section discusses the methodology used to conduct literature review and presents literature review results. The fourth section analyzes the literature review results and makes comparison of the findings made in the previous section. The fifth section of this thesis presents the conclusion for the thesis.

2. Background

Personalized treatment decision-making aims to incorporate patient characteristics and patient needs into treatment decisions by selecting the most appropriate treatments to prescribe for a patient. This decision-making process usually includes quantifying the association of an intervention or a series of intervention on a patient's health outcome to determine the efficacy of possible treatments. Individualized treatment effect inference is built based on the concept of causal inference, thus ML methods for causal inference can be utilized to estimate treatment effect. These methods can be applied on observational data such as EHRs to estimate the individualized treatment effect.

Moreover, healthcare decisions also include the search for treatment regimes that yield the best outcome in all states of a patient's disease. Those regimes work in a dynamic fashion, as the patient's state has to be updated over time with regard to the progression of the disease. Reinforcement learning with the main mechanism of feedback and improvement has strong potential to learn DTRs, thus RL methods are usually applied to derive an optimal treatment strategies in the time-varying and dynamic setting.

Background information about causal inference using ML methods and dynamic treatment regimes estimation using RL methods are presented in Section 2.1 and Section 2.2 respectively.

2.1 Causal inference

2.1.1 Individualized treatment effects

Individualized treatment effect (ITE) estimation aims to determine and quantify the effect of an intervention on a patient's health outcomes. ITE can be formulated into a causal model by applying the classical potential framework introduced by Neyman and Rubin (Rubin, 2005). The framework using observational data is described as follows.

Given observational data: X_i, W_i, Y_i

- Each patient i has features: $X_i \in X \subset R^d$
- Treatment assignment: $W_i \in \{0, 1\}$ (1: treated, 0: untreated)
- Two potential outcomes: $Y_i^{(1)}, Y_i^{(0)} \in R$

In the framework, the potential outcomes $Y_i^{(1)}$ and $Y_i^{(0)}$ - also known as factual outcome and counterfactual outcome – respectively correspond to the outcome of the i^{th} subject if the individual received the treatment $W_i = 1$ and if he/she does not receive the treatment $W_i = 0$.

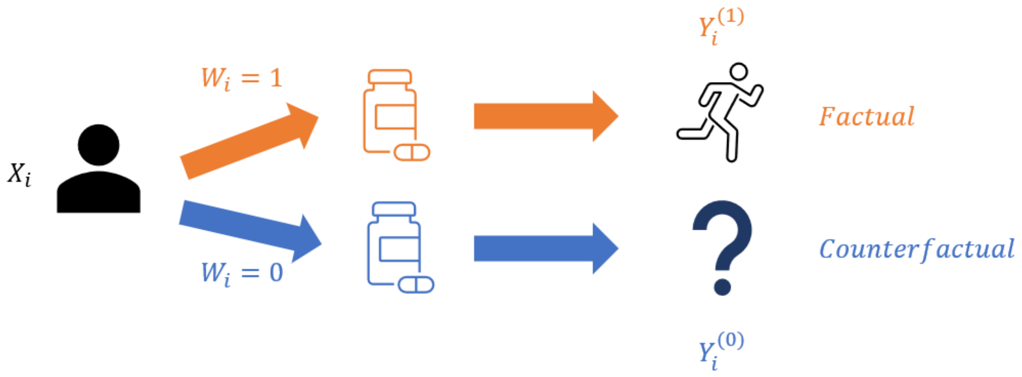

Figure 2.1. Relationship between features, treatment assignment, and potential outcomes

2.1.2 Data for individualized treatment effects estimation

The gold standard designs for measuring treatment effects on outcomes are randomized controlled trials (RCTs), in which subjects are randomly allocated among available treatment groups. However, RCTs are not always practical in the evaluation of the effectiveness of certain treatments, since conducting RCTs is expensive, time-consuming, and sometimes infeasible due to ethical constraints and difficulties in patient recruitment. Furthermore, strict inclusion and exclusion criteria in RCTs leads to the fact that a patient sample is usually not a representative of patient population. Therefore, the causal conclusions from RCTs do not support decision-making choices at the individual level, as the inferences are usually made at the population level. In contrast, observational studies are less expensive and faster alternatives to clinical trials (Ahmed M. Alaa and Schaar, 2017).

Over the last few years, with the advent and wide-scale adoption of EHRs, there is a huge potential for estimating ITE by utilizing these observational data. EHRs store baseline patient features (e.g., age, sex, and genetic information) and longitudinal information about follow-up clinic visits and lab tests (e.g., specific treatments administered and patient's performance status under treatments) (Bica, Ahmed M Alaa, Lambert, et al., 2021). Although data from EHRs is valuable, estimating ITE from observational data still poses the challenges of bias from treatment selection and confounding variables.

In the observational setting, the decision about which treatment to assign to an individual is usually made based on subject characteristics. Particularly, in clinical practices, doctors usually take into consideration patients' current medical characteristics when assigning treatments. For example, cancer patients who have more aggressive tumours may receive more extensive treatments, but also may get worse outcomes. This leads to the treatment selection bias problem and a possibly wrong conclusion: it is dangerous to prescribe extensive treatments to patients with aggressive tumours (Bica, Ahmed M Alaa, Lambert, et al., 2021). Moreover, some factors, such as socioeconomic or environmental factors, cannot be fully measured and become hidden confounders, which affects the actual causal relationship between the treatment and the outcome. The presence of confounding factors in observational data can lead to biased estimates (Bica, Ahmed M Alaa, Lambert, et al., 2021).

2.1.3 Causal inference

The task of estimating ITE can be formulated as a question of quantifying the causal effects of the treatment on a person's health outcome. Using the potential outcome framework in section 2.1.1, the individualized treatment effects (ITE) for an individual with a feature $X_i = x$ is specified as the expected difference between the treated outcome and the control outcome (Bica, Ahmed M Alaa, Lambert, et al., 2021):

$$T(x) = E[Y_i^{(1)} - Y_i^{(0)} | \ X_i = x].$$

It can be observed from the equation above that both potential outcomes must be known to estimate the causal effect sizes. However, the key challenge of individual-based causal inference is that only the treated outcome is available for each particular person. Without knowing the counterfactual outcome, we cannot estimate the causal ground truth effect sizes. To solve this problem, observational patient data from EHRs can be utilized to train causal inference models. Given a training data set consisting of patients who received the treatment and patients who did not receive the treatment, these two groups of patients can be used to estimate the responses of each different treatment options. After being trained, the model then can be utilized to estimate the potential outcomes for new patients.

Performing casual inference using observational data requires two assumptions to be made: the overlap assumption and the unconfoundedness assumption. The overlap assumption holds when each patient always has a chance to receive each treatment option. The unconfoundedness assumption states that the treatment assignment must not depend on the outcomes, given patient's covariates. However, in reality, the unconfoundedness assumption rarely holds because the existence of confounding variables cannot be tested in the absence of counterfactual outcomes. Thus, its plausibility depends on subject-matter knowledge.

Causal inference methods can be utilized to estimate ITE in both the static setting (one-time decision) and in the time-series setting (patient history and treatment timing are considered) (Bica, Ahmed M Alaa, Lambert, et al., 2021). The ITE estimates learned from the causal inference models support experts in choosing the best treatment out of all possible treatments. In the longitudinal setting, causal inference methods are used to model patients' trajectories in order to predict the potential outcomes of future possible treatments. By learning from the treatment responses over time, ITE estimates in the time-varying treatment settings can support experts in, for example, understanding more about the progression of diseases under different treatment plans, the optimal timing for assigning treatments, and the response of individual patients to medication over time (Bica, A. Alaa, and Van Der Schaar, 2020).

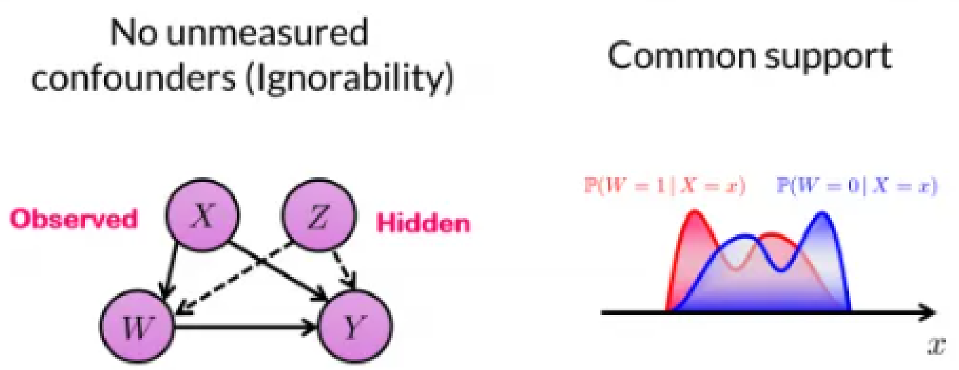

Figure 2.2. Assumptions for causal inference (*Individualized treatment effect inference | | Van der Schaar Lab* 2022)

2.2 Reinforcement learning

Reinforcement learning (RL) is a subfield of ML with the focus on goal-directed learning. The two most important distinguished features of RL are trial-and-error search and delayed reward (Sutton and Barto, 2018). In the RL approach, the agent interacts with the surrounding environment and receives rewards after every action. Each interaction generates new information about the system and improves the current knowledge and future interaction with the system. The aim of an agent is to maximize the numerical rewards in the long run. In Figure 2.3, the interaction between agent and environment in RL is presented. (Sutton and Barto, 2018).

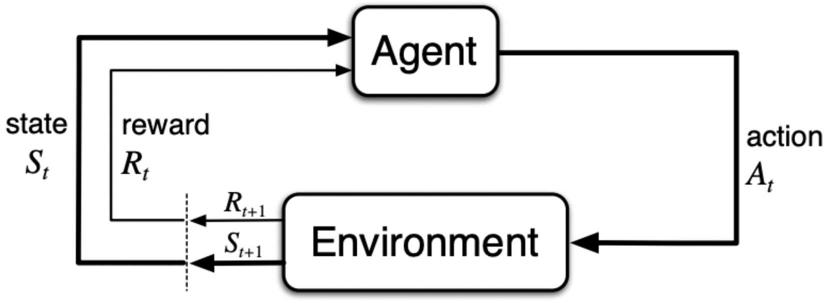

Figure 2.3. The agent–environment interaction in reinforcement learning (Sutton and Barto, 2018)

Along with the agent and the environment, an RL system is composed of four main sub-elements: a policy, a reward signal, a value function, and a model of the environment (Sutton and Barto, 2018).

A policy is a set of stimulus-response association which determines behaviour of an agent at a given time. A policy maps each state to feasible actions to take at each stage. The policy alone is sufficient to determine the functioning of a system (Sutton and Barto, 2018).

A reward signal defines the goal for an agent. The agent receives a reward after every action and aims to maximize long-term rewards. The reward signal thus measures the performance of an agent towards the goal. The reward signal is the principal mechanism for modifying the policy: the policy may be altered if an action followed by the policy yields low reward (Sutton and Barto, 2018).

A value function specifies how good is a state in a long run. In particular, a value of a state is equal to the expected accumulated reward starting from that state. The sole goal of the agent is to choose actions that yield the highest values, not the highest rewards, as those actions maximize long-term rewards (Sutton and Barto, 2018). A model describes the behaviour of the environment. With a state and an action available, the model might forecast the next state and rewards (Sutton and

Barto, 2018).

The exploration-exploitation trade-off is a fundamental challenge presented in the RL approach. At each time step, the agent has two options: either explore new actions in the environment by selecting a suboptimal action, or exploit the current knowledge by choosing the actions that yield the best reward so far. While exploration encourages the agent to gain new information about the state space to make the better actions in the future, exploitation takes advantage of the current knowledge in order to maximize immediate rewards (Sutton and Barto, 2018). The dilemma between these is that we have to balance between exploration or exploitation to let the agent discover various actions and progressively favour the best performance (Sutton and Barto, 2018). By this way, the agent can learn to take optimal actions by taking into account different possibilities in the state space (Sutton and Barto, 2018).

Another challenging task in RL comes from the formulation of reward functions. In some medical contexts, the outcomes of a given treatment can be represented numerically, and the reward function is usually a mapping from several independent parameters into some integers (Yu, J. Liu, et al., 2021). For example, in cancer treatment, a really low negative number is usually used as a threshold to penalize for patient death, and a positive number is considered as a reward for a cured patient (Yu, J. Liu, et al., 2021). This quantification considers an approach of trading-off between efficacy and toxicity, and thus has an impact on the resulting treatment strategies. However, these reward functions are usually defined by clinical experts, whose ways of formulating these numbers may vary greatly (Yu, J. Liu, et al., 2021).

2.2.1 Dynamic Treatment Regimes

2.2.1.1 Definition

Dynamic treatment regimes (DTRs) are alternatively known as dynamic treatment policies, adaptive treatment strategies, or adaptive interventions (Yu, J. Liu, et al., 2021). In the context of healthcare, a DTR consists of a set of tailored decision rules that determines the intervention to prescribe to a patient at each assessment stage, with the consideration of achieving the best long-term treatment outcomes (Yu, J. Liu, et al., 2021).

2.2.1.2 DTR problem

The DTR problem formulation below is adapted from Y. Zhang and Schaar (2020). Given observational data with N individuals and T time steps. Each individual is characterized by

- a baseline covariate vector $Z = H_0$, (e.g., age and gender),
- a covariate vector X_t, a treatment assignment A_t, an outcome variable Y_t at each time step $t \in [T] = \{1, ..., T\}$, and
- the history up to time t: $H_t = (X_{[t]}, A_{[t]}, Z)$ and $\widetilde{H}_t = (X_{[t]}, A_{[t-1]}, Z)$.

Given the history up to time t: $\widetilde{H}_t = (X_{[t]}, A_{[t-1]}, Z)$:

- At time t, the treatment rule d_t is a function that assigns A_t based on $\widetilde{H}_t$.
- A DTR is $\mathbf{d} = \{d_1, ..., d_T\}$

Assume data distribution:

$$P(O) = P(H_0) \prod_{t=1}^{T} P(X_t|H_{t-1}) P(A_t|\widetilde{H}_t) P(Y_t|H_t)$$

The purpose is to find the treatment rules or DTR that maximizes the expected total outcomes over time, or in RL term, maximizes the value function. The value function at time t under the treatment rules $d_t, ..., d_T$ is defined as

$$V_t(d_{t:T}) = E_{P_{d_{t:T}}}[\prod_{r=t}^{T} Y_r].$$

The value function under the DTR $\mathbf{d} = \{d_1, ..., d_T\}$ is denoted as $V(d)$.

The optimal DTR is defined as $\mathbf{d}^* = argmax_d V(d)$

2.2.1.3 Data for DTR estimation

According to Chakraborty and Murphy (2014), the two most common sources for the construction of DTRs are from observational data and sequentially randomized data.

Data from observation studies is the most used source of data to construct DTRs (Chakraborty and Murphy, 2014). In observational studies, treatment assignment is unknown and presumed to be not randomized. In longitudinal studies, individuals are repeatedly examined to detect any changes over a period of time. Longitudinal observation data can be drawn from a variety of sources such as databases from hospitals, EHRs, randomized encouragement trials, and cohort studies (Chakraborty and Murphy, 2014).

Recently, experimental designs known as Sequential Multiple Assignment Randomized Trial (SMART) designs has experienced a rapid growth in practice (Chakraborty and Murphy, 2014). SMART design consists of multiple-stage trials, with each stage corresponding to a critical decision point. At each stage, patients are allocated randomly to one of the available treatment actions. Treatment options at randomization depend on intermediate outcome and/or treatment history, thus ethical constraints are not violated. A SMART is used to inform the construction of optimal DTRs.

For a concrete example, a SMART study for Adaptive Pharmacological and Behavioural Treatments for Children with Attention Deficit/Hyperactivity Disorder (ADHD) is presented in the Figure 2.4. In the trial, each child with ADHD is allocated randomly to one of the possible initial treatments (the circle with an "R" indicates "Randomization"); that is, a behaviour modification (BMOD) or an oral methamphetamine (MEDS). After two months, the response of each child to the first treatment is evaluated by the Impairment Rating Scale (IRS) and the individualized List of Target Behaviours (LTB) measurement. The second treatment depends on the classification of the children to the first treatment as responders or non-responders. Consequently, responders to the first treatment continue with this treatment, while non-responders to the first treatment are assigned randomly to either an augmentation (BMOD+MEDS) or an intensification of the first treatment. The main outcome measured is the score of the school performance at the end of the study.

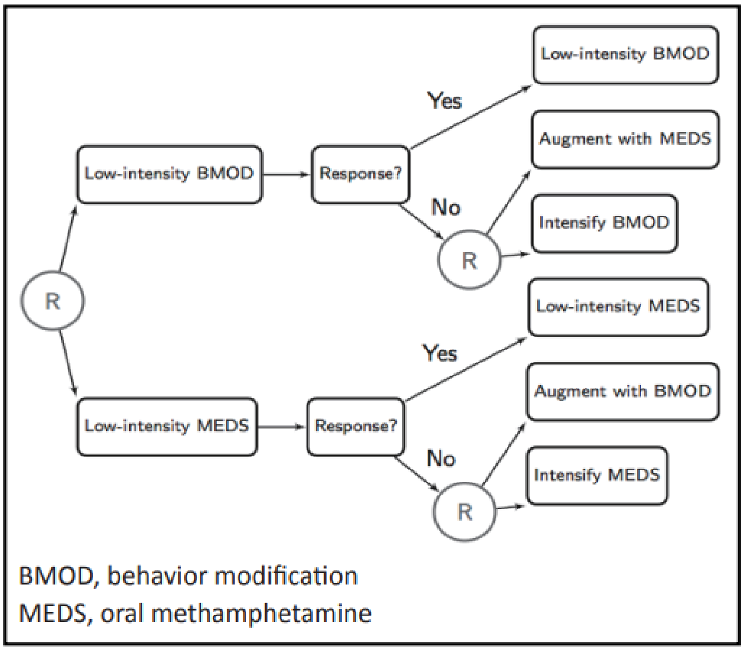

Figure 2.4. An example of a SMART study (Y. Liu, Donglin, and Yuanjia, 2014)

In the above example, there are four adaptive interventions: two interventions associated with the first treatment, and two interventions associated with the second treatments for non-responders. The intervention for the responders is counted only once.

The effectiveness of DTRs can be evaluated based on the notation of potential outcomes. At any single stage, given two possible treatments, only one treatment can be received, and only the treated outcome is observed. Therefore, estimating the causal effect of a DTR requires two assumptions are needed: the Stable Unit Treatment Value Assumption (SUTVA) and the no unmeasured confounders assumption. The SUTVA assumption states that an outcome of a subject is independent of other subjects' treatment assignments. The unconfoundedness assumption states that the newly allocated treatments are conditional on the patient's trajectories, but not associated with potential future outcomes.

2.2.1.4 DTR and RL

RL is a suitable approach for applications in healthcare due to numerous reasons. Firstly, RL allows making time-dependent decisions for each patient at a given time, which supports personalized treatment in DTRs. This precise treatment also holds without any well- represented mathematical model or explicit relationship between treatments and outcomes. Secondly, in medical treatments, the effect of an intervention may not be immediately observed. As RL approach enables to make decisions without information about the effectiveness of the current actions but focuses more on future reward in the long term, RL is well-suited for solving DTR problems considering the time delay.

The design of DTR fits well into the concept of RL. A DTR as a series of decision rule is analogous to a policy in RL, with states corresponding to information about patient's states and actions corresponding to treatment options at each stage. The reward functions define treatment outcomes.

3. Methods and Results

3.1 Methodology

This thesis is a literature review on two different approaches of using ML in personalized treatment decision-making. The first one is individual treatment effects estimation with ML methods for casual inference. The second one is optimal DTRs estimation with RL techniques. For each approach, reviewed studies are chosen based on a database search and backward snowballing of the key sources on the topic. The backward snowballing was based on two initial key sources (Bica, Ahmed M Alaa, Lambert, et al., 2021) and (Coronato et al., 2020) proposed by my advisor. These sources were checked in Scopus, and it was determined that they are relatively recently published and have a good number of citations, 23 and 39, respectively. Moreover, the database search also utilized a pearl-growing method. The key terms for each application are mentioned in the section below. The search is conducted on the Scopus and Google Scholar databases. The search queries in these databases included title, keywords, and abstracts. For each resulting paper, the abstract is read and only those papers that are relevant to the topics are chosen.

This thesis utilizes the following inclusion criteria for the primary sources. Firstly, the sources must focus on machine learning methods in the context of personalized healthcare research. Secondly, this study only includes sources that are published in English and after 2005. After comparing against the inclusion/exclusion criteria, there are 53 of suitable sources in total.

3.2 Estimating individualized treatment effects using ML methods for causal inference

3.2.1 Search terms

This part concentrates on reviewing materials on methods for ITE estimation and applications of ITE estimation for treatment decision-making. The keywords are established by combining search terms with the same categories and joining the categories together with boolean AND operator.

- "individual treatment effects" OR "heterogeneous treatment effects" OR "counterfactual inference"

- "longitudinal" OR "time series" OR "observational data" OR "electronic health records"

The search terms for applications of ITE estimation are established by combining the first bullet point with the name of the disease (such as sepsis, cancer, diabetes, . . .).

3.2.2 Data sources used in previous work

A brief look at the data sources used in experiments of previous work in ITE estimation reveals that three most commonly used database sources are from Infant Health and Development Program (IHDP), TWINS, and JOBS data sets. Most of the work used IHDP data proposed by Hill (2011) .The IDHP is a semi-simulated dataset with synthesized potential outcomes with the purpose of estimating the effects of high-quality child care and specialist home visits on future cognitive test scores. The TWINS is a semisynthetic data set consists of records of twin births in the USA from 1989-1991. The Jobs is a real-world dataset which intends to estimate the effect of the job training on income and employment status after training.

Publications	IHDP	TWINS	JOBS	Other
F. Johansson, Shalit, and Sontag (2016)	✓			News
Shalit, F. D. Johansson, and Sontag (2017)	✓		✓	
Ahmed M. Alaa, Weisz, and Schaar (2017)	✓			
Louizos et al. (2017)	✓		✓	
Ahmed M. Alaa and Schaar (2017)	✓			UNOS
Yoon, Jordon, and Schaar (2018)	✓	✓	✓	
Lee, Mastronarde, and Schaar (2018)		✓		
Yao et al. (2018)	✓	✓	✓	
Chen et al. (2019)	✓			Heart failure

Table 3.1. Data sources used in ITE estimation publications

3.2.3 Methods for handling time-dependent confounders when using longitudinal data

- **Propensity score methods** are commonly utilized to handle the treatment selection bias that plagues non-randomized methods. The propensity score was defined by Rubin (1974) as the conditional probability of treatment assignment given a number of observed characteristics:

  $$e_i = P(Z_i = 1 | \mathbf{X}_i)$$

  Propensity score shows the probability that a patient being appointed to a treatment given observed baseline covariates (Bica, Ahmed M Alaa, Lambert, et al., 2021). In practice, the propensity score can be computed using a logistic regression model, given measured baseline characteristics (Austin, 2011).

- **Inverse probability of treatment weighting (IPTW)** is a propensity score method that is applied to adjust the effects of confounding variables in the treatment effects estimation by creating a weighted synthetic sample in which treatment assignment is not dependent on observed baseline characteristics (Austin, 2011). IPTW assigns to each subject i a weight using the inverse of each subject's propensity score e_i:

  $$w_i = \frac{Z_i}{e_i} + \frac{1 - Z_i}{1 - e_i},$$

  where Z_i is an indicator variable denoting whether the i^{th} subject was treated or not (Austin, 2011).

3.2.4 Methods for ITE estimation

The results from the search terms in section 3.2.1 return 17 methods, and those methods used are presented in the table 3.2. The publications are ordered by the publication year.

Publications	Methods used	ML model
Hill (2011)	Bayesian Additive Regression Trees (BART)	Regression Trees
F. Johansson, Shalit, and Sontag (2016)	Balancing Neural Networks (BNN)	Neural Network
Yanbo Xu, Yanxun Xu, and Saria (2016)	Bayesian Nonparametrics model	Bayesian Nonparametrics model
Shalit, F. D. Johansson, and Sontag (2017)	Counterfactual Regression (CFR)	Multitask Neural Network
Ahmed M. Alaa, Weisz, and Schaar (2017)	Deep Counterfactual Networks with Propensity-Dropout (DCN-PD)	Multitask Neural Network
Louizos et al. (2017)	Causal Effect Variational Autoencoder (CEVAE)	Variational Autoencoders
Ahmed M. Alaa and Schaar (2017)	Causal Multi-task Gaussian Processes (CMGP)	Multitask Gaussian Processes
Yoon, Jordon, and Schaar (2018)	Generative Adversarial Nets for inference of Individualized Treatment Effects (GANITE)	Generative Adversarial Network
Lim (2018)	Recurrent Marginal Structural Networks (R-MSNs)	Sequence-to-sequence RNN model
Lee, Mastronarde, and Schaar (2018)	Causal Effect using a Generative Adversarial Network (CE-GAN)	Generative Adversarial Networks
Wager and Athey (2018)	Causal Forest	Random Forests
Yao et al. (2018)	Similarity Preserved Individual Treatment Effect (SITE)	Multitask Neural Network
Chen et al. (2019)	multitask deep learning—K-nearest neighbours (MTDL-KNN)	Multitask Neural Network
Sugasawa and Noma (2019)	Gradient Boosting Trees (GBT)	Regression Trees
Bica, A. Alaa, and Van Der Schaar (2020)	Time Series Deconfounder	Multitask Neural Network
Bica, Ahmed M Alaa, Jordon, et al. (2020)	Counterfactual Recurrent Network (CRN)	Sequence-to-sequence RNN model
Bica, Jordon, and Schaar (2020)	SCIGAN (eStimating the effects of Continuous Interventions using GANs)	Generative Adversarial Network

Table 3.2. Methods for ITE estimation

Different ML methods have been applied to estimate ITE in both the static setting and the longitudinal setting.

- Bayesian Additive Regression Trees (BART), Causal Forests, and Gradient Boosting Trees (GBTs) use tree-based estimators.

- Causal Multi-task Gaussian Processes (CMGP) is a Gaussian process-based method.

- Counterfactual Regression (CFR), Balancing Neural Network (BNN), and Similarity Preserved Individual Treatment Effect (SITE) are deep representation learning based methods. BNN and CFR methods use neural network model to learn a balanced representation so that the distribution difference between treated and control populations is minimized in the embedding space, while SITE preserves the local similarity information and balances data distributions at the same time.

- MTDL-KNN uses multitask deep learning strategy to train deep neural network to learn the hidden representations of patient features and uses K nearest neighbors (KNN) -a matching-based method-to estimate the counterfactual outcomes.

- GANITE utilizes Generative Adversarial Nets (GAN) framework to model ITE estimation function for inferring potential outcomes under both binary and multiple treatments.

- Deep Counterfactual Networks with Propensity Dropout uses a deep multitask network to model an individual's potential outcomes and apply a propensity-dropout regularization scheme to handle selection bias.

- The mentioned above are causal inference methods that can only be applied in the static setting. Time Series Deconfounder, Recurrent Marginal Structural Networks (R-MSNs), and Counterfactual Recurrent Network are methods for causal inference in the longitudinal setting.

- The Time Series Deconfounder takes advantage of the dependencies of treatment assignments over time to derive substitutes for hidden confounders, hence obtaining unbiased estimates of the treatment effects. The highlight of Time

Series Deconfounder is that this method uses weaker assumptions than existing methods and shows effectiveness in removing the bias even when multi-cause hidden confounders exist.

- The Counterfactual Recurrent Network (CRN) leverages domain adversarial training and representation learning to build treatment invariant representations that can be applied to remove the time-varying bias from confounders. CRN then integrates the domain adversarial training procedures as a part of sequence-to-sequence architecture to predict counterfactual outcomes for treatment plans. The CRN approach can be used to choose optimal treatments, find optimal timing for treatment, and finding the optimal ending time for treatment to assess the best health outcomes.

- SCIGAN is a framework for estimating response curves for continuous and many-level-discrete interventions from observational data. SCIGAN uses a modified generative adversarial networks (GANs) model to learn the data distribution of the unobserved counterfactual outcomes. Thus, SCIGAN can be used to learn an inference model, enabling estimating these counterfactual for a new sample.

3.2.5 Applications of ITE estimation

Those methods mentioned above can be applied to various designs, some example of the applications of those methods are described below.

Publications	sepsis	breast cancer	lung cancer	cardiovascular disease
Yanbo Xu, Yanxun Xu, and Saria (2016)	✓			
Hu et al. (2021)			✓	
Amsterdam et al. (2022)			✓	
Schrod et al. (2022)		✓		
Bica, A. Alaa, and Van Der Schaar (2020)	✓			
Duan et al. (2019)				✓
Sugasawa and Noma (2019)		✓		
Tabib and Larocque (2020)		✓		
W. Zhang et al. (2017)		✓		
Bica, Ahmed M Alaa, Jordon, et al. (2020)	✓			

Table 3.3. Applications of ITE estimation in treatment decision-making

Ahmed M. Alaa and Schaar (2017) evaluate the ability of CMGP model to estimate the survival benefits of applying Left Ventricular Assistance Devices (LVADs) on patients waiting for a heart transplantation. Chen et al. (2019) evaluates the ability of the MTDL-KNN method using a real-world clinical data set including information of heart failure patients to estimate the effects of eight treatment strategies combined from three major medications recommended for heart failure patients (ACEI/ARB, beta-blockers, and aldosterone antagonist) on the hospital readmission rate in one year. Bica, A. Alaa, and Van Der Schaar (2020) applies the Time Series Decounfounder method to a real-world ICU data set containing information about septic patients to measure the effect of vasopressors, antibiotics, and mechanical ventilator on the patients' white blood cell count, blood pressure, and oxygen saturation. Bica, Ahmed M Alaa, Jordon, et al. (2020) evaluates the ability of their proposed CRN model to estimate the individualized effect of antibiotics on the white blood cell count using data from ICUs. Yanbo Xu, Yanxun Xu, and Saria (2016) applies Bayesian Nonparametrics model on a real-world clinical data sets containing information of septic patients to estimate the treatment responses of intermittent hemodialysis and continuous renal replacement therapy, the main treatment choices for managing blood pressure and kidney functions, on creatine levels, a measure of kidney deterioration. Berrevoets et al. (2020) introduces the OrganITE, an organ-to-patient assignment methodology for modelling ITEs for organ transplants. OrganITE takes into consideration its estimation of the potential outcomes and the scarcity of organs when assigning organs.

3.2.6 Evaluation of causal inference methods for individualized treatment effect estimation

Validation of causal inference models is an important task to translate algorithmic advances into practice. In supervised ML, cross-validation is a popular approach to determine the best model to choose from. However, evaluation of causal inference methods from a real data is a complicated task as the true ground truth causal effects cannot be observed due to the missing of counterfactual data. One of the straightforward approach is to create a synthetic observational data that generate patient outcomes under different treatment actions, which enables one to observe ground truth treatment effects to validate the robustness of causal models on synthesized data. However, this approach is not likely to reflect the same complexity of real-world data (Bica, Ahmed M Alaa, Lambert, et al., 2021). In practice, several simple heuristics for estimating model performance have been used, but such heuristics don't provide any theoretical guarantees and can fail to generalize in some scenarios. Bica, Ahmed M Alaa, Lambert, et al. (2021) also listed out different approaches of performing internal model evaluation and external model evaluation. Internal model evaluation aims to select a model that performs the best on the observational test set, while external model is used to assess the accuracy of the model used once it has been deployed in practice (Bica, Ahmed M Alaa, Lambert, et al., 2021). For internal model evaluation, Bica, Ahmed M Alaa, Lambert, et al. (2021) introduced a state-of-the-art model validation procedure that can approximate the loss of causal inference methods even in the absence of counterfactual data. This procedure uses influence functions, which are essentially the functional derivatives of a loss function, to approximate the loss function of a method on a given data set (A. Alaa and Van Der Schaar, 2019). For external model evaluation, information about the patient's health outcomes can be compared with recommendation of the causal inference model to evaluate whether the assigned treatments improve the patient health outcomes.

3.3 Estimating optimal DTR using reinforcement learning

3.3.1 Search terms

The utilized search terms for collecting materials about optimal DTR estimation using RL include (disjunction indicated by commas): “dynamic treatment” AND “reinforcement learning”, “adaptive treatment” AND “reinforcement learning”, “dynamic treatment regimes” AND “reinforcement leaning”, and “adaptive treatment regimes” AND “reinforcement learning”.

3.3.2 Data sources used in previous work

A brief look at the resources reveals that most publications in the field evaluate and apply their methods to various diseases, ranging from chronic diseases (e.g. cancer, diabetes, anemia) to critical care (e.g. sepsis). Since different diseases require different patterns about which patients and diseases are represented within the data, data sources used in the work of optimal DTR estimation using RL methods vary greatly depends on the research purposes and the application tasks. For example, the work in the optimization of a deep-brain stimulation strategy to treat epilepsy (a severe brain disease) uses recordings of animal brain tissues to evaluate the performance of their model, since the acquisition of recordings of human brain is expensive and requires strict protocols. Meanwhile, a lot of work in the application of RL to treat sepsis utilize the publicly available dataset MIMIC-III that consists of information about patients admitted to an ICU during an 11-year period. In conclusion, the variation of patterns required in the dataset and the availability of public dataset by disease account for the fact that there is no benchmark dataset in the research area of optimal DTR using RL.

3.3.3 Methods for optimal DTR estimation

Publications	**Methods used**
Moodie, Chakraborty, and Kramer (2012)	Q-Learning
Baniya et al. (2017)	Q-Learning
Krakow et al. (2017)	Q-Learning
N. Liu et al. (2018)	Deep Neural Network (DNN) + online deep Q-learning
Tao, L. Wang, and Almirall (2018)	Tree-based reinforcement learning (T-RL)
L. Wang et al. (2018)	Supervised Reinforcement Learning with Recurrent Neural Network (SRL-RNN)
N. Liu et al. (2019)	Deep Q-Network (DQN)
Tang et al. (2021)	step-adjusted tree-based reinforcement learning (SAT-Learning)
Sun and L. Wang (2021)	stochastic tree-based reinforcement learning (ST-RL)
Zhou et al. (2022)	Tree-based reinforcement learning (T-RL)
Li et al. (2022)	EHRs based deep Q network (EHRs-DQN)

Table 3.4. Methods for optimal DTR estimation

Different RL methods have been applied to estimate the optimal DTR.

- Q-Learning is a model-free RL approach that is particularly popular in the DTR literature (Chakraborty and Murphy, 2014). Q-learning employs dynamic programming to construct the optimal DTRs using a backward recursive stagewise estimation to identify sequences of actions that maximize some long-term reward.

- Deep Q-Network (DQN) integrates Q-learning with deep neural networks to approximate the Q-functions. By integrating deep learning with RL, DQN is a promising method to optimize DTRs with high-dimensional treatment options (e.g. high-dimensional action space) and heterogeneous decision stages.

- Supervised Reinforcement Learning with Recurrent Neural Network (SRL-RNN) combines supervised learning and RL to study a policy to suggest personalized treatments. Supervised learning learns a policy by matching the indicator signals which denotes doctor prescription, hence aims to minimize the difference between treatment prescribed by doctors and recommended output. RL learns a policy by maximizing evaluation signals, which indicate cumulative rewards from survival rates. By combining indicator signal and evaluation signal simultaneously, SRL-RNN can learn an integrated policy that guarantees a safe performance and optimal dynamic treatments.

3.3.4 Applications of DTR in treatment decision-making

Due to the inherent delay, RL is usually used to automate the decision-making within treatment regimes. DTRs design has helped for chronic diseases and improved critical care using the data collected in intensive care units (ICUs).

Disease	Publications	Applications	Methods
Anemia	Adam E Gaweda, Muezzinoglu, Aronoff, et al. (2005)	Perform individualized pharmacotherapy in the management of renal anemia	SARSA
	Adam E Gaweda, Muezzinoglu, Jacobs, et al. (2006)	Combine Model Predictive Control (MPC) to simulate patient responses with RL to estimate dosage strategy, which helps in the management of anemia caused by the failure of kidney	MPC + SARSA
	Escandell-Montero, M. Martínez-Martínez, et al. (2011)	Derive an optimal dosing strategy for Erythropoiesis Stimulating Agents (ESAs) in the management of anemia in patients with hemodialysis	Fitted Q-iteration
	Malof and Adam E. Gaweda (2011)		
	Escandell-Montero, Chermisi, et al. (2014)		
Diabetes	Adam E Gaweda, Muezzinoglu, Aronoff, et al. (2005)	Perform individualized pharmacotherapy in the management of renal anemia	SARSA

Table 3.5. Applications of DTR in chronic diseases

Several recent papers in the field have showed the applications of RL to the development of personalized treatments for sepsis, a life-threatening infection.

Disease	Publications	Highlights	Methods
Sepsis	Raghu, Komorowski, Celi, et al. (2017)	Derive treatment rules for patients with sepsis using continuous state-space models and DRL	Dueling Double-Deep Q Networks + Q-learning
	Raghu, Komorowski, Ahmed, et al. (2017)		
	Peng et al. (2018)	Introduce a mixture-of-experts (MoE) approach to learn improved fluid and vasopressor administration strategies for patients with sepsis using ICUs observational data	deep RL (DRL) + kernel RL (KRL)
	Yu, Ren, and J. Liu (2019)	Propose a model to learn the best reward functions from a set of presumably optimal treatment trajectories using retrospective real medical data	deep inverse RL with Mini-Tree model (DIRL-MT)

Table 3.6. Applications of DTR in critical care

3.3.5 Evaluation of RL methods in healthcare

Evaluation of the performance of RL-based methods is critical to ensure that the model runs safely in practice. Essentially, in RL, the quality of a policy is quantified by the expected cumulative reward, and the optimal policy is the one that yields the highest expected cumulative reward. Coronato et al. (2020) proposed two approaches for the evaluation of the performance of RL methods. The first approach is to plot the accumulated sum of rewards as a function of the number of steps. This plot consists of three important elements: the slope of the reward that shows the degree of goodness of the policy after the algorithm is stabilized, the minimum of the curve that shows how much reward-value the agent lost before it starts to improve, and the zero-crossing points that show how many episodes the agent takes until he regains its cost of learning (Coronato et al., 2020). The second approach is to plot the average reward—the cumulative reward per set of steps. This approach shows the value of the policy eventually learned and whether the algorithm has stopped learning, but often suffer from large variations for early times (Coronato et al., 2020).

4. Analysis

4.1 Discussion

This thesis presents two different approaches that aim to personalize the treatment decision-making process. The ITE estimation approach is usually applied to problems where quantification of treatment effect should be known to understand the effectiveness of an intervention to a health outcome variable. Estimation of ITE can be done both in the cross-sectional setting and in the longitudinal setting. In the longitudinal setting, the problem of estimating treatment effects over time can sometimes involve the problem of learning the optimal treatment options that would yield the highest reward. This is similar to the problem of estimating the optimal DTR using reinforcement learning. However, given the same objective, two approaches handle the problem in different ways. With the causal inference model, the main mechanism is to estimate the counterfactual outcomes under all possible treatment options, hence allowing one to find the optimal treatment. Alternatively, with the DTR designs using RL, the best course of treatments is usually defined through the process of evaluative feedback and improvement.

The results from both approaches hence support decision-making process in different aspect. Given all the treatment effects estimations of all possible treatment options from causal inference models, clinical experts and patients can observe all the estimated potential outcomes and decide whether to perform a planned intervention or not. Otherwise, counterfactual outcomes cannot be derived from the RL models. The RL algorithms usually output treatment rules that serve as the suggestion and prediction about the next actions that would optimize patients' health outcomes.

Furthermore, the causal inference approach allows one to draw conclusions about the causal relationships about whether an intervention affects outcome, while credit assignment is still one of the ongoing challenging issues in the applications of RL in healthcare. It is usually uncertain in RL problems which of the performed actions lead to the outcomes due to the time delay effects in health outcomes. Without causal model defined clearly in the learning process, causal conclusions cannot be drawn in RL approach.

A great amount of applications has been found for both approaches. In the clinical management of chronic diseases, DTRs are particularly effective as they provide decision rules that adapt throughout disease progression, allowing the new treatments to adjust based on the history of response to past treatments. Alternatively, ITE estimation is usually applied to understand the heterogeneous effects of an intervention on a health outcome, verify the causes of certain diseases by explaining what variables lead to the outcome. Knowing ITE helps decision-makers to answer counterfactual inference questions, such as “Would the patient benefit more or less from the alternative treatment options ?”. Understanding the causal effects can also help understand which groups or types of patient would benefit more or less from the treatment. Since treatments in healthcare are sometimes costly and patient’s responses can vary greatly given the same treatment, knowing ITE can help alleviate financial burdens and improve decision-making process.

Health decisions usually include the trade-off among cost, efficacy, and toxicity. For example, in the cancer treatment, the objective of chemotherapy treatment is to kill aggressive cancer tumour but also to minimize the damage to other normal tissues in the body. This kind of problem can be formulated as multi-objective optimization to achieve a Pareto optimal solution. Multi-objective RL as an emerging research topic in RL provides an opportunity to achieve these kinds of optimization problem, while causal inference approach still lacks of methods to handle multi-objective problems.

In terms of interpretability, the causal inference approaches usually provides a transparent model that offers a clear framework to express assumptions about the data, while RL methods usually suffer from a black-box interpretation. From a set of input data, the models using RL methods usually directly output a policy without immediate results, which makes it difficult to interpret the reasoning behind the output decisions.

4.2 Research direction

4.2.1 Causal inference for ITE estimation

Causal inference methods can be applied to both data collected from RCTs and real-world data to estimate ITEs. RCTs data provide high robustness of conclusions due to randomized treatment decisions; however, strict eligibility criteria employed in RCTs usually limit the representative of the patient population covered in data. While real-world data contains large number of subjects that cover the entire patient population, the control for confounding factors and non-randomized treatment decisions is still the challenging task of using observational data for causal inference. Due to the complementary benefits of RCTs data and real-world data, leveraging these two types of data would harness a full potential of data available.

Causal inference methods can be applied in the cross-sectional setting and in the longitudinal setting. A brief look at causal inference methods for ITE estimation from observational data reveals that more methods are studied in the static setting compared to the longitudinal setting. However, those methods in static setting can not be applied in the estimation of treatment effects in time-varying treatments as they can not handle the dependencies in patient trajectories. To utilize observational data in estimating ITE in longitudinal setting, which are the cases that resemble more of real-world complex treatment scenarios, further work should concentrate more in developing robust methods that can handle time-dependent confounders, model combinations of treatments over time, and estimate ITE of time-dependent treatments in the continuous settings, where associated dosage is involved (Bica, Ahmed M Alaa, Jordon, et al., 2020).

Most of the causal inference methods mentioned above rely on the no unmeasured confounders and the overlap assumptions to be hold. While the overlap assumption can be verifiable, the no hidden confounders is untestable and rarely holds in practice. The violation of such assumptions can lead to the bias in the conclusions made from the causal models. However, most of the current methods still depend on the assumptions of no hidden confounders. Time Series Confounder assumes hidden confounders, but uses weaker assumption of no hidden single-cause confounders.

Another research direction in ITE estimation is to develop accessible tools to find the most important features that account for treatment responses, especially in high-dimensional data. This could lead to better treatment responses and more transparent and interpretable models. (Bica, Ahmed M Alaa, Lambert, et al., 2021).

4.2.2 RL for optimal DTR estimation

In RL, the reward function usually determines the success of the RL-based models. However, there are still some issues related to the reward formulation in healthcare that requires more attention in future research. Currently, the majority of RL applications in healthcare still encode reward functions quantitatively than qualitatively. However, those numerical and quantitative functions are highly subjective and can vary greatly depending on experts' personal experience. Alternatively, preference-based models are provided as the way to establish qualitative functions using ranking functions that sort states, actions, trajectories, and policies. These ranking functions are easier when comparisons between trajectories or suboptimal actions are needed to specify the performance. Moreover, as the medical decisions can involve many objectives (e.g. benefits versus cost, efficacy versus toxicity), multi-objective RL techniques can be utilized to estimate a policy that balances between multiple objectives to achieve a Pareto optimal solution. Although preference based RL and multi-objective RL is a promising area, current work of these areas in medical settings is still limited and only considered simple scenarios with static preferences or fixed objectives (Yu, J. Liu, et al., 2021). Further work can extend on these areas by considering more complex scenarios where patients' preferences and treatment objectives are dynamic and evolving.

Another core challenge of real-world RL is the sparsity of reward feedback. Since the real evaluation outcomes in healthcare usually come at the end of treatment, long-term rewards at the end of a learning episode are more favourable than short-term rewards at each decision step (Yu, J. Liu, et al., 2021). Most of previous studies trying to address this problem in healthcare only concentrate on DTRs within a short horizon, which is usually not the case in practice. Therefore, future work of RL in healthcare should investigate more in tacking sparse reward learning with a long horizon setting.

One fundamental issue in RL is the exploration and exploitation trade-off. In healthcare domains, exploration without safety measures can lead to bad results, while insufficient exploration may result in suboptimal policies and undesirable treatment decisions. Most of existing RL application in healthcare usually use simple heuristic-based exploration strategies such as -greedy strategy, which is impractical in complicated dynamics medical settings with large or continuous state/action spaces (Yu, J. Liu, et al., 2021). Moreover, one has to consider the true cost of exploration when applying exploration strategies in medical settings. Although penalizing an agent with negative scores to discourage unwanted actions work in most situations, this approach is inappropriate in healthcare when an unfavourable action may cause unrecoverable damage to patients' health. Given the importance of exploration strategies in the success of RL applications, safe exploration strategies are of much interest in the further research.

5. Conclusion

This thesis is a literature review that aims to examine how machine learning can be utilized for personalized treatment decision-making in healthcare. The thesis presents two approaches for this problem: causal inference for ITE estimation and RL for optimal DTR estimation. For each approach, background knowledge, the objective, the overview of data sources, ML methods, evaluation methods, and applications in recent years are also included.

A fair volume of publications have been found for these two approaches. As the number of research papers are still growing, ML shows great potential in providing personalized treatment recommendations and transforming healthcare. However, there are still some limitations that exist in two approaches. For causal inference using observational data, the most challenging one is how to handle the confounding variables that can bias the estimates. In this case, in the near future, it is expected that methods that work in the presence of confounders should be developed. For reinforcement learning, the problem lies in the interpretability of models, the formulation of reward functions, and the balance between exploration and exploitation. Tackling these problems is at the forefront

Both causal inference and reinforcement learning have been widely researched, but the combination of these two in research has been really limited. Due to the complementary objective of both approach, incorporating causal models into different aspects of clinical decision support systems by using advances in RL could bridge the gap between these two fields.

Lastly, most of the ML methods are still required external validation from experts. It is noteworthy that ML models are only envisaged as decision support tools, with the final decisions still made by clinical experts and patients.

Bibliography

Alaa, Ahmed and Mihaela Van Der Schaar (Sept. 2019). “Validating Causal Inference Models via Influence Functions”. In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 191–201. URL: `https://proceedings.mlr.press/v97/alaa19a.html`.

Alaa, Ahmed M. and Mihaela van der Schaar (2017). “Bayesian Inference of Individualized Treatment Effects using Multi-task Gaussian Processes”. In: *CoRR* abs/1704.02801. arXiv: `1704.02801`. URL: `http://arxiv.org/abs/1704.02801`.

Alaa, Ahmed M., Michael Weisz, and Mihaela van der Schaar (2017). “Deep Counterfactual Networks with Propensity-Dropout”. In: *CoRR* abs/1706.05966. arXiv: `1706.05966`. URL: `http://arxiv.org/abs/1706.05966`.

Amsterdam, Wouter AC van et al. (2022). “Individual treatment effect estimation in the presence of unobserved confounding using proxies: a cohort study in stage III non-small cell lung cancer”. In: *Scientific reports* 12.1, pp. 1–11.

Austin, Peter C (2011). “An introduction to propensity score methods for reducing the effects of confounding in observational studies”. In: *Multivariate behavioral research* 46.3, pp. 399–424.

Baniya, Abiral et al. (2017). “Adaptive interventions treatment modelling and regimen optimization using sequential multiple assignment randomized trials (SMART) and Q-learning”. In: *IIE Annual Conference. Proceedings*. Institute of Industrial and Systems Engineers (IISE), pp. 1187–1192.

Berrevoets, Jeroen et al. (2020). “OrganITE: Optimal transplant donor organ offering using an individual treatment effect”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 20037–20050. URL: `https://proceedings.neurips.cc/paper/2020/file/e7c573c14a09b84f6b7782ce3965f335-Paper.pdf`.

Bica, Ioana, Ahmed Alaa, and Mihaela Van Der Schaar (2020). “Time series deconfounder: Estimating treatment effects over time in the presence of hidden confounders”. In: *International Conference on Machine Learning*. PMLR, pp. 884–895.

Bica, Ioana, Ahmed M Alaa, James Jordon, et al. (2020). “Estimating counterfactual treatment outcomes over time through adversarially balanced representations”. In: *arXiv preprint arXiv:2002.04083*.

Bica, Ioana, Ahmed M Alaa, Craig Lambert, et al. (2021). “From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges”. In: *Clinical Pharmacology & Therapeutics* 109.1, pp. 87–100.

Bica, Ioana, James Jordon, and Mihaela van der Schaar (2020). “Estimating the Effects of Continuous-valued Interventions using Generative Adversarial Networks”. In: *CoRR* abs/2002.12326. arXiv: 2002.12326. URL: https://arxiv.org/abs/2002.12326.

Chakraborty, Bibhas and Susan A Murphy (2014). “Dynamic treatment regimes”. In: *Annual review of statistics and its application* 1, p. 447.

Chen, Peipei et al. (2019). “Deep representation learning for individualized treatment effect estimation using electronic health records”. In: *Journal of biomedical informatics* 100, p. 103303. DOI: https://doi.org/10.1016/j.jbi.2019.103303. URL: https://reader.elsevier.com/reader/sd/pii/S1532046419302229?token=606970F5A207624BAC6B074E9090F7F0F403A0D26EF7F9E04F720D9E1D304523803E558114E219F92942EAED41BC991originRegion=eu-west-1&originCreation=20220709222611.

Coronato, Antonio et al. (2020). “Reinforcement learning for intelligent healthcare applications: A survey”. In: *Artificial Intelligence in Medicine* 109, p. 101964.

Duan, Tony et al. (2019). “Clinical value of predicting individual treatment effects for intensive blood pressure therapy: a machine learning experiment to estimate treatment effects from randomized trial data”. In: *Circulation: Cardiovascular Quality and Outcomes* 12.3, e005010.

Escandell-Montero, Pablo, Milena Chermisi, et al. (2014). “Optimization of anemia treatment in hemodialysis patients via reinforcement learning”. In: *Artificial intelligence in medicine* 62.1, pp. 47–60.

Escandell-Montero, Pablo, José M. Martínez-Martínez, et al. (2011). “Adaptive treatment of anemia on hemodialysis patients: A reinforcement learning approach”. In: *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pp. 44–49. DOI: 10.1109/CIDM.2011.5949442.

Gaweda, Adam E, Mehmet K Muezzinoglu, George R Aronoff, et al. (2005). “Reinforcement learning approach to individualization of chronic pharmacotherapy”. In: *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.* Vol. 5. IEEE, pp. 3290–3295.

Gaweda, Adam E, Mehmet K Muezzinoglu, Alfred A Jacobs, et al. (2006). “Model predictive control with reinforcement learning for drug delivery in renal anemia management”. In: *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, pp. 5177–5180.

Hill, Jennifer L. (2011). “Bayesian Nonparametric Modeling for Causal Inference”. In: *Journal of Computational and Graphical Statistics* 20.1, pp. 217–240. DOI: 10.1198/jcgs.2010.08162. eprint: https://doi.org/10.1198/jcgs.2010.08162. URL: https://doi.org/10.1198/jcgs.2010.08162.

Hu, Liangyuan et al. (2021). “Estimating heterogeneous survival treatment effects of lung cancer screening approaches: A causal machine learning analysis”. In: *Annals of Epidemiology* 62, pp. 36–42.

Individualized treatment effect inference // Van der Schaar Lab (Aug. 2022). URL: https://www.vanderschaar-lab.com/individualized-treatment-effect-inference/.

Johansson, Fredrik, Uri Shalit, and David Sontag (2016). “Learning representations for counterfactual inference”. In: *International conference on machine learning*. PMLR, pp. 3020–3029.

Krakow, Elizabeth F et al. (2017). “Tools for the precision medicine era: how to develop highly personalized treatment recommendations from cohort and registry data using Q-learning”. In: *American journal of epidemiology* 186.2, pp. 160–172. DOI: 10.1093/aje/kwx027.

Lee, Changhee, Nicholas Mastronarde, and Mihaela van der Schaar (2018). “Estimation of Individual Treatment Effect in Latent Confounder Models via Adversarial Learning”. In: *CoRR* abs/1811.08943. arXiv: 1811.08943. URL: http://arxiv.org/abs/1811.08943.

Li, Tianhao et al. (2022). “Electronic health records based reinforcement learning for treatment optimizing”. In: *Information Systems* 104, p. 101878.

Lim, Bryan (2018). “Forecasting Treatment Responses Over Time Using Recurrent Marginal Structural Networks”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2018/file/56e6a93212e4482d99c84a639d254b67-Paper.pdf.

Liu, Ning et al. (2018). “Deep Reinforcement Learning for Dynamic Treatment Regimes on Medical Registry Data”. In: *CoRR* abs/1801.09271. arXiv: 1801.09271. URL: http://arxiv.org/abs/1801.09271.

– (2019). “Learning the dynamic treatment regimes from medical registry data through deep Q-network”. In: *Scientific reports* 9.1, pp. 1–10.

Liu, Ying, ZENG Donglin, and WANG Yuanjia (2014). “Use of personalized dynamic treatment regimes (DTRs) and sequential multiple assignment randomized trials (SMARTs) in mental health studies”. In: *Shanghai archives of psychiatry* 26.6, p. 376.

Louizos, Christos et al. (2017). “Causal effect inference with deep latent-variable models”. In: *Advances in neural information processing systems* 30.

Malof, Jordan M. and Adam E. Gaweda (2011). “Optimizing drug therapy with Reinforcement Learning: The case of Anemia Management”. In: *The 2011 International Joint Conference on Neural Networks*, pp. 2088–2092. DOI: 10.1109/IJCNN.2011.6033485.

Moodie, Erica EM, Bibhas Chakraborty, and Michael S Kramer (2012). “Q-learning for estimating optimal dynamic treatment rules from observational data”. In: *Canadian Journal of Statistics* 40.4, pp. 629–645.

Peng, Xuefeng et al. (2018). “Improving sepsis treatment strategies by combining deep and kernel-based reinforcement learning”. In: *AMIA Annual Symposium Proceedings*. Vol. 2018. American Medical Informatics Association, p. 887.

Raghu, Aniruddh, Matthieu Komorowski, Imran Ahmed, et al. (2017). “Deep reinforcement learning for sepsis treatment”. In: *arXiv preprint arXiv:1711.09602*.

Raghu, Aniruddh, Matthieu Komorowski, Leo Anthony Celi, et al. (2017). “Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach”. In: *Machine Learning for Healthcare Conference*. PMLR, pp. 147–163.

Rubin, Donald B (1974). “Estimating causal effects of treatments in randomized and nonrandomized studies.” In: *Journal of educational Psychology* 66.5, p. 688.

– (2005). “Causal Inference Using Potential Outcomes”. In: *Journal of the American Statistical Association* 100.469, pp. 322–331. DOI: 10.1198/016214504000001880. eprint: https://doi.org/10.1198/016214504000001880. URL: https://doi.org/10.1198/016214504000001880.

Schrod, Stefan et al. (2022). “BITES: Balanced Individual Treatment Effect for Survival data”. In: *arXiv preprint arXiv:2201.03448*.

Shalit, Uri, Fredrik D Johansson, and David Sontag (2017). “Estimating individual treatment effect: generalization bounds and algorithms”. In: *International Conference on Machine Learning*. PMLR, pp. 3076–3085.

Sugasawa, Shonosuke and Hisashi Noma (2019). “Estimating individual treatment effects by gradient boosting trees”. In: *Statistics in medicine* 38.26, pp. 5146–5159.

Sun, Yilun and Lu Wang (2021). “Stochastic Tree Search for Estimating Optimal Dynamic Treatment Regimes”. In: *Journal of the American Statistical Association* 116.533, pp. 421–432. DOI: 10.1080/01621459.2020.1819294. eprint:

https://doi.org/10.1080/01621459.2020.1819294. URL: https://doi.org/10.1080/01621459.2020.1819294.

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

Tabib, Sami and Denis Larocque (2020). "Non-parametric individual treatment effect estimation for survival data with random forests". In: *Bioinformatics* 36.2, pp. 629–636.

Tang, Ming et al. (2021). "Step-adjusted tree-based reinforcement learning for evaluating nested dynamic treatment regimes using test-and-treat observational data". In: *Statistics in Medicine* 40.27, pp. 6164–6177.

Tao, Yebin, Lu Wang, and Daniel Almirall (2018). "Tree-based reinforcement learning for estimating optimal dynamic treatment regimes". In: *The annals of applied statistics* 12.3, p. 1914.

Wager, Stefan and Susan Athey (2018). "Estimation and inference of heterogeneous treatment effects using random forests". In: *Journal of the American Statistical Association* 113.523, pp. 1228–1242.

Wang, Lu et al. (2018). "Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation". In: *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2447–2456.

Xu, Yanbo, Yanxun Xu, and Suchi Saria (18–19 Aug 2016). "A Non-parametric Bayesian Approach for Estimating Treatment-Response Curves from Sparse Time Series". In: *Proceedings of the 1st Machine Learning for Healthcare Conference*. Ed. by Finale Doshi-Velez et al. Vol. 56. Proceedings of Machine Learning Research. Northeastern University, Boston, MA, USA: PMLR, pp. 282–300. URL: https://proceedings.mlr.press/v56/Xu16.html.

Yao, Liuyi et al. (2018). "Representation Learning for Treatment Effect Estimation from Observational Data". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2018/file/a50abba8132a77191791390c3eb19fe7-Paper.pdf.

Yoon, Jinsung, James Jordon, and Mihaela van der Schaar (2018). "GANITE: Estimation of Individualized Treatment Effects using Generative Adversarial Nets". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=ByKWUeWA-.

Yu, Chao, Jiming Liu, et al. (2021). "Reinforcement learning in healthcare: A survey". In: *ACM Computing Surveys (CSUR)* 55.1, pp. 1–36.

Yu, Chao, Guoqi Ren, and Jiming Liu (2019). “Deep inverse reinforcement learning for sepsis treatment”. In: *2019 IEEE international conference on healthcare informatics (ICHI)*. IEEE, pp. 1–3.

Zhang, Weijia et al. (2017). “Mining heterogeneous causal effects for personalized cancer treatment”. In: *Bioinformatics* 33.15, pp. 2372–2378.

Zhang, Yao and Mihaela van der Schaar (2020). “Gradient Regularized V-Learning for Dynamic Treatment Regimes”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., pp. 2245–2256. URL: `https://proceedings.neurips.cc/paper/2020/file/17b3c7061788dbe82de5abe9f6fe22b3-Paper.pdf`.

Zhou, Nina et al. (2022). “Optimal dynamic treatment regime estimation using information extraction from unstructured clinical text”. In: *Biometrical Journal* 64.4, pp. 805–817.