Analytics for Medical Decision Making

Applications to the Management of Treatment-Resistant Depression

Martin COUSINEAU

Doctor of Philosophy

Desautels Faculty of Management

McGill University
Montréal, Québec
October 2019

A thesis submitted to McGill University
in partial fulfillment of the requirements of the degree of
Doctor of Philosophy in Management

© Martin Cousineau 2019

Je dédie cette thèse à ma conjointe
Agathe, mon fils Edouard, mon
futur garçon Albert et mes parents.

Contents

List of Figures

List of Tables

List of Algorithms

Abstract

The objective of this thesis is the use and design of analytics methods (i.e., methods from operations research and management science, artificial intelligence, and statistics) for medical decision making. In particular, this work focuses on methods to assist physicians towards achieving remission in patients suffering from treatment-resistant depression, a severe form of the major depressive disorder. Following a literature review of medical decision making methods relevant for treating depression, this thesis proposes medical decision making methods for (1) finding the best initial treatment modification for incoming patients, (2) characterizing the current timing decisions between appointments and (3) recommending potential successful treatments. All of these tasks are addressed using observational longitudinal data from the Depressive and Suicide Disorders Program of the Douglas Mental Health University Institute in Montreal.

In particular, the first method focuses on the task of using observational data to determine which of five treatment modification strategies is best at the initial visit. To do so, the proposed method balances the five strategy groups using an improved approach for causal inference. The chapter associated with this method is also used as a tutorial to causal inference for the operations research and management science community.

The second method identifies the relevant variables among the patient's, physician's and clinic's characteristics for the timing decisions between appointments. This decision is of importance due to the trade-off between high-frequency appointments that lead to a waste of resources and low-frequency appointments that lead to the degradation of patients. Using imitation learning on data, this method infers these variables and their weights. This knowledge can then be used by the physicians to refine and standardize their practice with respect to this decision.

The third method recommends potential successful treatments using similarities between past patients and treatments. This recommender system consists somewhat of an extension of the first method where causal inference is again used. However, the treatment drugs are now considered instead of the five treatment modification strategies. For this method, we assume that the sequence of treatments that have been administered to the patient does not affect the efficacy of the current treatment.

Abrégé

L'objectif de cette thèse est l'utilisation et la conception de méthodes analytiques (c.-à-d. des méthodes issues de la recherche opérationnelle, de l'intelligence artificielle et des statistiques) pour la prise de décisions médicales. En particulier, ces travaux portent sur les méthodes visant à aider les médecins à obtenir une rémission chez les patients souffrant de dépression résistante au traitement, une forme grave du trouble dépressif majeur. À la suite d'une revue de la littérature sur les méthodes de prise de décision médicale pertinentes pour le traitement de la dépression, cette thèse propose des méthodes de prise de décision médicale pour (1) trouver la meilleure modification initiale du traitement pour les patients entrants, (2) caractériser les décisions actuelles concernant le délai entre les rendez-vous et (3) recommander des traitements potentiellement efficaces. Toutes ces tâches sont abordées à l'aide de données longitudinales d'observation du programme des troubles dépressifs et suicidaires de l'Institut universitaire en santé mentale Douglas à Montréal.

En particulier, la première méthode est axée sur l'utilisation de données d'observation pour déterminer laquelle des cinq stratégies de modification du traitement est la meilleure lors de la visite initiale. Pour ce faire, la méthode proposée équilibre les cinq groupes de stratégies en utilisant une approche améliorée pour l'inférence causale. Le chapitre associé à cette méthode est également utilisé comme un tutoriel d'inférence causale pour la communauté de la recherche opérationnelle.

La deuxième méthode identifie les variables pertinentes parmi les caractéristiques du patient, du médecin et de la clinique pour la prise de décision du délai entre les rendez-vous. Cette décision est importante en raison de l'arbitrage entre les rendez-vous à haute fréquence qui entraînent un gaspillage de ressources et les rendez-vous à basse fréquence qui entraînent la dégradation des patients. En utilisant l'apprentissage par imitation sur les données, cette méthode infère ces variables et leur pondération. Ces connaissances peuvent ensuite être utilisées par les médecins pour affiner et normaliser leur pratique en ce qui a trait à cette décision.

La troisième méthode recommande des traitements potentiellement efficaces en utilisant les similitudes entre les anciens patients et les traitements. Ce système de recommandation consiste en quelque sorte en une extension de la première méthode où l'inférence

causale est de nouveau utilisée. Cependant, les médicaments des traitements sont maintenant considérés au lieu des cinq stratégies de modification du traitement. Pour cette méthode, nous supposons que la séquence des traitements qui a été administrée au patient n'affecte pas l'efficacité du traitement actuel.

Acknowledgements

I would first like to thank my supervisor Professor Vedat Verter for his guidance and great personality during this journey, and for all his editorial feedback on this writing.

Continuing on the academic side, I would as well like to thank Susan A. Murphy who collaborated on Chapter 3 and gave me a warm welcome at the University of Michigan, Ann Arbor. I also extend my gratitude to all my advisory committee members, Professors Joelle Pineau, Erick Delage and Gustavo Turecki, for their comments on my proposal. The same gratitude also goes towards my internal and external examiners, respectively Professors Joelle Pineau and Finale Doshi-Velez, and my oral defence committee members, Professors Wei Qi, Joelle Pineau, Emine Sarigollu and Daniel Frank, for their comments on a previous version of this thesis. All these people helped me to improve this thesis in many ways.

This applied research wouldn't have been possible without the data collected at the Douglas Mental Health University Institute. Therefore, I am grateful to David Guan who helped me with the data collection, and to everyone at this institute who helped me during the data collection phase and the later research phase.

Finally, my acknowledgments would be incomplete without thanking my family. I want to thank my father and mother for always believing in me and pushing me towards higher education. I also thank my partner Agathe for supporting me during this long journey. This has not always been easy; with her, I discovered the joy (and challenges) of having our own family. I hope Edouard and my future son Albert will someday find their own paths.

List of Abbreviations

ADP approximate dynamic programming

AI artificial intelligence

AIDS acquired immune deficiency syndrome

A-learning advantage learning

ALS alternating least squares

ANOVA analysis of variance

ATE average treatment effect

ATT average treatment effect among the treated

BC behavioral cloning

BOSS balance optimization subset selection

CANMAT Canadian Network for Mood and Anxiety Treatments

CATE conditional average treatment effect

CI confidence interval

DCG discounted cumulative gain

DFS disease-free survival

DP dynamic programming

DSDP depressive and suicide disorders program

DSM Diagnostic and Statistical Manual of Mental Disorders

DTR dynamic treatment regime

EBal entropy balancing

FCP fraction of concordant pairs

FM factorization machine

GAM generalized additive model

GAN generative adversarial net

HCI human-computer interaction

HIV human immunodeficiency virus

IDCG ideal discounted cumulative gain

IDPL interactive direct policy learning

IHDP Infant Health and Development Program

IL imitation learning

INR international normalized ratio

IPTW inverse probability of treatment weighting

IRL inverse reinforcement learning

ITE individualized treatment effect

KG knowledge gradient

KL divergence Kullback-Leibler divergence

KMM kernel mean matching

LASSO least absolute shrinkage and selection operator

MAE mean absolute error

MAOI monoamine oxidase inhibitor

MAP-BIRL maximum a posteriori inference for Bayesian inverse reinforcement learning

MDD major depressive disorder

MDM medical decision making

MDP Markov decision process

ML machine learning

MLR multiclass logistic regression

MMP maximum margin planning

MSE mean squared error

MSM marginal structural mean

MWAL multiplicative weights for apprenticeship learning

NDCG normalized discounted cumulative gain

NMSE normalized mean squared error

OR/MS operations research and management science

PE pure exploitation

PF Poisson factorization

POMDP partially observable Markov decision process

PWL piecewise linear

QALY quality-adjusted life year

QIDS Quick Inventory of Depressive Symptomatology

Q-learning quality learning

RBF radial basis function

RCT randomized controlled trial

RKHS reproducing kernel Hilbert space

RL reinforcement learning

RMSE root mean squared error

RS recommender system

RV random variable

SBW stable balancing weights

SCID-I Structured Clinical Interview for DSM Axis I Disorders

SCID-II Structured Clinical Interview for DSM Axis II Personality Disorders

SMART sequential multiple assignment randomized trial

SMDP semi-Markov decision process

SMLIRL semi maximum likelihood inverse reinforcement learning

SNMM structural nested mean model

SNRI serotonin-norepinephrine reuptake inhibitor

SSRI selective serotonin reuptake inhibitor

STAR*D Sequenced Treatment Alternatives to Relieve Depression

SUTVA stable unit treatment value assumption

SVMMatch support vector machine matching

TCA tricyclic antidepressant

TPM transition probability matrix

TRD treatment-resistant depression

Chapter 1

Introduction

Operations research and management science (OR/MS) methods have been applied successfully to a wide range of healthcare settings in the past (i.e., system design and planning, management of operations and medical management) (Pierskalla and Brailer, 1994). In particular, more recently, there has been a surge of interest in the application of OR/MS methods to medical decision making (MDM) problems due to the increasing healthcare costs, the increasing access to better data, the high level of preventable medical errors and the trend towards the uniformization of medical practice (Tunc, Alagoz, and Burnside, 2014). In parallel, while OR/MS was traditionally a practice-oriented field, it has become associated with a defined set of methods and problems; a potentially harmful association in this time of renewed interest in analytics, a field unconstrained with respect to the methods used and problems addressed (Liberatore and Luo, 2010).

This thesis expands over the traditional OR/MS methods by using methods from other fields (e.g., artificial intelligence, statistics, causal inference) that are pillars of analytics to address innovative MDM problems. In particular, this work consists in the use of analytics for MDM to manage treatment-resistant depression (TRD). To do so, this work uses observational longitudinal data from the depressive and suicide disorders program (DSDP) of the Douglas Mental Health University Institute in Montreal.

The rest of this chapter defines and discusses the topics of analytics, medical decision making, treatment-resistant depression and observational data, four fundamental topics to this thesis. Then, the thesis structure and contributions are presented.

1.1 Analytics

There exists multiple definitions of analytics. For example, INFORMS (2017) defines analytics as "the scientific process of transforming data into insights for the purpose of making better decisions" while Gass and Fu (2013) defines it as "data-driven modeling

and analysis for decision making". Both of these definitions contain two important aspects. First, the analytics process needs to be data-driven, i.e., based on data and not on intuition or personal experience. Second, the analytics process is used for decision making, i.e., some action needs to be taken in the end.

The analytics process consists of four steps (Liberatore and Luo, 2010). The first step encompasses the collection, extraction and manipulation of data. The second step consists in the analysis generally categorized as descriptive (What has happened?), predictive (What might happen?) and prescriptive (How to improve the predicted future?). It is important to note that prescriptive analytics needs to understand both *what* might happen and *why* it will happen in order to suggest how to act.[1] The third step consists in the extraction of the insights from the previous questions. Finally, the last step consists in the implementation of the action at the strategic, tactical or operational level. This process does not have to be a linear sequence; there can be several back and forths between the different steps.

The use of the word *analytics* in this thesis assumes both aspects of the previous definitions, i.e., data-driven and decision making. While the latter aspect cannot be enforced, the proposed applications at least try to provide some minimal insights to encourage decision making. With respect to the analytics process, the three original works within this thesis can be classified as descriptive for the work in Chapter 4, and as prescriptive for the works in Chapters 3 and 5.

1.2 Medical Decision Making

The definition of medical decision making given by the Society for Medical Decision Making (*Definition of Medical Decision Making*, n.d.) consists of the following quotation from Schwartz and Bergus (2008):

> Medical decision science is a field that encompasses several related pursuits. As a normative endeavor, it proposes standards for ideal decision making. As a descriptive endeavor, it seeks to explain how physicians and patients routinely make decisions, and has identified both barriers to, and facilitators of, effective decision making. As a prescriptive endeavor, it seeks to develop tools that can guide physicians, their patients, and health care policymakers to make good decisions in practice.

[1] Refer to Spirtes (2010) for a formal definition of predictive and prescriptive modeling.

This definition highlights that the field of MDM is interested in (1) proposing standards/guidelines, (2) explaining decisions and (3) developing tools to make better decisions. In this thesis, we address the latter two objectives. In particular, the work in Chapter 4 consists in the study of the timing decision between appointments while the works in Chapters 3 and 5 propose tools to make better decisions.

Due to the continuous progress in medicine, several open opportunities for analytics lie ahead in the field of MDM. In particular, the OR/MS community identified the following opportunities (Zhang et al., 2013b; Denton et al., 2011):

- *personalized medicine*: the customization of interventions to individual patients
- *patient behavior*: the integration within the models of the patient behavior (e.g., the compliance to treatment);
- *natural history of disease*: the development of new models of the progression of diseases from observational data;
- *future medical interventions*: the inclusion of the possibility of new future treatments within the model;
- *burden of treatment*: new methods and studies to evaluate the impact of treatments (e.g., determination of quality-adjusted life years (QALYs));
- *decision aids*: the development of easy-to-use tools for the physicians and patients with these models;
- *real time decision making*: real time optimization to be used in a variety of context like mobile health;
- *integration of prevention, detection and treatment*: the design of models that encompass the full spectrum of the clinical pathway.

With respect to these opportunities, this thesis involves personalized medicine with Chapter 5, decision aids[2] with Chapters 3 and 5, and partially the integration of prevention, detection and treatment by addressing the timing decision between appointments, in Chapter 4, and the treatments in Chapters 4 and 5.

It is interesting to note that the main challenge in most MDM studies remains the availability, cost and complexity of the data used to construct these models (Zhang et al., 2013b). This challenge supports the need to use observational data, a topic discussed in Section 1.4.

[2]Note that the decision aids tools developed in this thesis are not an operationalization of the current knowledge; they are data-driven.

1.3 Treatment Resistant-Depression

Major depressive disorder (MDD), is amongst the top ten causes of the global burden of disease and is predicted to become the leading cause by 2030 (World Health Organization, 2008). This translates to an estimated economic burden of US$210.5 billion in 2010 for MDD in the United States alone (Lam et al., 2016b).[3] In addition to these aspects, people suffering from MDD are also more prone to commit suicide; their past-year prevalence of suicide attempts in 2012 was twenty times higher than people seeking healthcare for other mental illnesses (Patten et al., 2015).

Unfortunately, up to 15% of the population affected by MDD[4] remains significantly depressed despite the aggressive use of multiple pharmacological and psychotherapeutical approaches. These patients are generally referred to as suffering from treatment-resistant depression (TRD). Although there is no consensus regarding the definition of TRD, a patient suffering from MDD is usually considered treatment-resistant (or refractory) when at least two trials with antidepressants from different pharmacologic classes (adequate in dose, duration, and compliance) fail to produce a significant clinical improvement (Berlim and Turecki, 2007).

TRD patients are quite hard to treat by definition and necessitate a referral to a specialized mental health clinic where pharmacotherapy, psychotherapy and neurostimulation therapy are all possible treatment options. In fact, it is important to note that each of these treatment options consists of a wide array of treatment options, which leads to a combinatorial search for the best treatment.

Adding to this complexity, the current medical literature regarding this best treatment is unclear. While some of the literature covers the treatment of TRD, few of the studies compare more than two treatments, and it is hard to reconcile the prevailing studies because of their different inclusion and exclusion criteria and the multiple definitions of TRD (Berlim, Fleck, and Turecki, 2008). In addition, the guidelines (Kennedy et al., 2016) are primarily designed to treat MDD, i.e., they are mostly concerned in identifying a good initial treatment. Thus, they are of limited use to treat patients suffering from TRD who followed non-effective treatments for some time.

It is also important to note that mental illnesses are generally quite different than physical diseases and complications (e.g., acquired immune deficiency syndrome (AIDS),

[3] There exist no estimates for the economic burden of MDD in Canada, but the economic burden of mental illness was estimated to C$51 billion for 2003 (Lam et al., 2016b).

[4] In Canada, the past-year prevalence of MDD in 2012 was 3.9% with higher prevalence for women and younger age groups (Patten et al., 2015).

diabetes, cancer). For example, in contrast with most physicial diseases and complications, the pathophysiology of MDD is currently unknown (Hasler, 2010) as is also the case for other mental illnesses. Hence, MDM approaches for these mental illnesses might differ substantially from the ones for physical diseases and complications. For example, due to the unavailability of a biological model, these approaches might require to be based entirely on data without any *a priori* model.

The complexity of treatment, and incomplete knowledge with respect to treatment efficiency and pathophysiology makes the area of TRD interesting for this thesis. In particular, this area seems opportune for the use of analytics methods and appears to be untouched by the OR/MS literature as shown in Chapter 2.

1.4 Observational Data

Observational data consists of data that has been collected without any interference in the assignment of treatment (Rosenbaum, 2005). This data can be primarily collected for an observational study, or can be collected for other reasons (e.g., health record) and then analyzed with an observational study. This latter setting is defined as secondary analysis of observational data and is usually more challenging because of an increased missingness of data for example.

While a randomized controlled trial (RCT) (also known as a randomized experiment) is considered the gold standard, there exists several circumstances in which an observational study is a viable option. For example, an observational study can be used to formulate hypotheses to be tested in further RCTs and it can be used for ethical reasons when a treatment is harmful or unwanted.

In addition, of importance to this thesis, observational studies can be used to reduce the gap between the evidence needed and the evidence produced in healthcare. Currently, it appears that only 10-20% of the medical decisions and 50-60% of the medical guidelines recommendations are based on formal evidence (Institute of Medicine, 2013). While it is financially inconceivable to perform RCTs for all the possible decisions faced by physicians, it is more reasonable to address these questions by performing secondary analyses of data.

It is important to note however that observational studies doing prescriptive analytics can be biased because of the uncontrolled treatment assignment. Hence, careful design of observational studies, a topic of the causal inference literature, is required to reduce this bias and approach the reliability of the RCTs.

The MDM methods developed in this thesis are based on observational data, in particular health record data. Thus, this thesis explicitly addresses the limitations of using observational data for prescriptive analytics in Chapters 3 and 5, in contrast to most studies in the healthcare OR/MS literature that use observational data for prescriptive tasks with no explicit discussions of the related issues (see Chapter 2).

1.5 Thesis Structure and Contributions

Chapter 2 reviews methodologies that could assist psychiatrists for MDM at the pharmacotherapy level in order to achieve remission for TRD patients. In particular, this review focuses on analytics (i.e., methodologies from operations research and management science (OR/MS), artificial intelligence (AI) and statistics) and identifies how the different methodologies could be applied to our problem. This chapter also discusses the possible synergies and complementarities among the different methodologies from OR/MS, AI and statistics. In this chapter, it is found that no OR/MS studies are applied to MDD while several such studies exist in the AI and statistics literature. In addition, it is found that the mathematical models from both domains differ in their focus and characteristics.

Chapter 3 discusses some of the challenges of working with observational data (e.g., electronic health records, administrative and claims data), one of the most prominent source of data for the past healthcare OR/MS studies. In particular, it exposes the fundamentals of causal inference (the field addressing inference when manipulating decisions), and improves an existing method to determine the causal effects of the initial treatment modifications for treatment-resistant depression. In this chapter, the improved method is shown to obtain similar results and in many cases the best results when compared to other similar causal inference methods on different simulation models. Then, for the TRD case study where the goal is to determine which of five treatment modification strategies is best at the initial visit, the treatment effects obtained with this improved method are unfortunately not statistically significant with respect to the 95% confidence intervals. Still, some findings are consistent with the medical literature and guidelines.

Chapter 4 studies an often neglected decision within the medical literature, i.e., the time between appointments. This is an important trade-off decision given that more frequent appointments can decrease the total number of different patients seen while less frequent appointments can decrease the patients' well-being. In this chapter, we characterize how this decision is taken at the DSDP using a two-stage framework. First, with the use of semi-structured interviews, potential features used to determine the time between consecutive

appointments are elicited from the four psychiatrists at the DSDP. Unsurprisingly, it appears that similar features are used by each of these psychiatrists; yet, the importance of these features to each psychiatrist cannot be captured by these interviews, the reason for the existence of the framework's second stage. Still, these interviews also capture the variable experience and variable typical times between appointments for each of these psychiatrists. Then, the second stage of the framework consists in the use of different imitation learning (IL) methods. After a brief review of the existing IL methods, three methods are selected for the case study, with one method being a proposed extension. The results of the case study show that methods using discretization do not perform well in this setting and that the importance of each feature appears to differ across physicians when making this timing decision.

Chapter 5 consists in the personalized recommendations of TRD treatments using recommender systems. This chapter consists somewhat of an extension of Chapter 3 where the treatments are now recommended in a personalized way and the treatments now consist of drugs instead of treatment modification strategies. For this chapter, we assume that the sequence of treatments that have been administered to the patient does not affect the efficacy of the current treatment. In this chapter, two different treatment definitions, which make sense from a medical point of view, are used. On these two treatment definitions, we then fit models that use different features available in the data set such as features describing the patient, the treatment and the outcomes resulting from other treatments. According to different metrics, it appears that the models using the most features from the data provide the best results. Thus, the limited number of features describing the patient and the treatment does contain some relevant information. Yet these models are not performing well enough to be used as decision aids. In this work, it is also found that no treatment consistently leads to remission for all patients, but some patients' subgroups are found to respond to the same treatments. These particular treatments could then be assumed as being better than the other treatments which appear to only work for one or two patients.

Finally, Chapter 6 provides concluding remarks and future directions for research.

Chapter 2

A Survey of Medical Decision Making Methods Relevant for Treating Depression

The motivation of this chapter is to review methodologies that could assist psychiatrists for medical decision making (MDM) at the pharmacotherapy level in order to achieve remission for treatment-resistant depression (TRD) patients. In particular, this review focuses on analytics (i.e., methodologies from operations research and management science (OR/MS), artificial intelligence (AI) and statistics) and identifies how the different methodologies could be applied to our problem. Finally, this chapter discusses the possible synergies and complementarities among the different methodologies from OR/MS, AI and statistics.

2.1 Review Process

We define some inclusion/exclusion criteria in order to find the relevant papers to our issue. In order to be considered by this review, a paper needs to respect all of the following criteria:

1. focus on the treatment's part of the clinical pathway (see Figure 2.1), in particular on *prescriptive models for the pharmacotherapy* (e.g., optimization of the selection, timing or dosage of the drugs);

2. focus on the *patient and/or physician perspective(s)* (see Table 2.1);

3. incorporate some notion of the wider definition of *personalized medicine*, i.e., not only restricted to biological markers but also considers any other characteristics (e.g., sociodemographic, clinical) and extrinsic factors (e.g., lifestyle and environmental

FIGURE 2.1: Clinical pathway.

TABLE 2.1: Common stakeholder perspectives in medical decision making (Denton et al., 2011).

Perspective	**Description**
Patient	Considers mostly the treatment effect (e.g., control or cure of the disease, quality of life, side effects, disablement, life expectancy) but might also consider the cost of the treatment if the third party insurance doesn't cover the expenditures.
Physician	Often aligned with that of the patient due to the shared patient-physician decision making process but might have some personal incentives to recommend a particular treatment or no treatment (e.g., profitability, experience, time, difficulty).
Third part payer	Trade-off between the immediate cost of medical treatment and long-term potential cost associated with serious health outcomes due to an not-well treated disease.
Societal	Simultaneously considers the patient, physician and third part payer perspectives. Uses the concept of *willingness-to-pay* to transform these different criteria in monetary value or employ the concept of an *efficient frontier*.

exposures) that induces an adaptation of a treatment to a particular patient (Simon and Perlis, 2010; Burke and Psaty, 2007; Liebman, 2007);

4. use *longitudinal data*, preferably observational data (claims or clinical data) (Hunter, 2006; Overhage and Overhage, 2011); and

5. focus on *data-driven* methods (i.e., methods that are based on data and that do not use biological models).

2.2 Excluded Papers

The inclusion and exclusion criteria lead us to focus only on a small part of the MDM literature. Before reviewing the papers that correspond to our criteria in Section 2.3 and Section 2.4, we will acknowledge some of the excluded papers that also seem important to MDM. These papers were found during our literature review but they unfortunately didn't correspond to one or many of our inclusion/exclusion criteria.

In the excluded papers from the OR/MS literature, there seems to be a great interest in MDM research at the policy level. For example, there is research on the timing of the screening decisions (Helm et al., 2015; Erenay, Alagoz, and Said, 2014; Yang, Goldhaber-Fiebert, and Wein, 2013; Ayer, Alagoz, and Stout, 2012; Zhang et al., 2012c; Alagoz, Ayer, and Erenay, 2010; Chhatwal, Alagoz, and Burnside, 2010; Rauner et al., 2010; Ivy, 2009; Liberatore et al., 2009; Tafazzoli et al., 2009; Maillart et al., 2008; Harper and Jones, 2005; Leshno, Halpern, and Arber, 2003) and on the cost-effectiveness of several drug treatments (Mason et al., 2014; Mason et al., 2012; Denton et al., 2009; Cooper et al., 2006; Paltiel et al., 2004). There is also some research on non-pharmacological treatments such as radiation therapy (Chan et al., 2014; De Boeck, Beliën, and Egyed, 2014; Lavieri et al., 2012; Taskin et al., 2010; Simon, 2009; Bortfeld et al., 2008; Romeijn et al., 2006), dialysis therapy (Lee, Chertow, and Zenios, 2008), organ transplantation (Bertsimas, Farias, and Trichakis, 2013; Alagoz et al., 2007a; Alagoz et al., 2007b; Alagoz et al., 2004; Zenios, 2002) and hip replacement (Keren and Pliskin, 2011; Hazen, 2004). Finally, there is some limited work on the diagnostic of diseases (Lee and Wu, 2009; Rubin, Burnside, and Shachter, 2004).

In the excluded papers from the AI and statistics literature, we found many papers from the areas of data mining (Yoo et al., 2012; Chaovalitwongse, 2009; Bellazzi and Zupan, 2008; Cios and William Moore, 2002), expert systems (Wagholikar, Sundararajan, and Deshpande, 2012; Pandey and Mishra, 2009; Shu-Hsien Liao, 2005), and artificial intelligence and machine learning (Amato et al., 2013; Kononenko, 2001). These papers mostly focus on the diagnostic part of the clinical pathway and are more concerned with the predictive task.

There are also some excluded papers on MDM that wouldn't fall in either of the two previous categories. In particular, there is some work done on the treatment of cancer (Shi et al., 2011) and diabetes (Parker, Doyle, and Peppas, 2001): two diseases with known pathophysiologies that can be represented by biological models. These papers are not truly data-driven because they only use data to parametrize a mathematical model of the disease. Unfortunately, the pathophysiology of major depressive disorder (MDD) is currently unknown (Hasler, 2010) and hence these approaches are not relevant to our case.

Finally, there are some excluded papers that are covering MDD. Some of them study the predictors of quality of life (Ay-Woan et al., 2006), outcome (Berman and Hegel, 2014), response to treatment (Simon and Perlis, 2010) or remission (Gudayol-Ferré et al., 2012; Lin et al., 2011). Others are more concerned with a one-time prediction of the onset (Huang et al., 2014; de Man-van Ginkel et al., 2013; Wong et al., 2012), outcome (Pfeiffer et al., 2015; Berman and Hegel, 2014; Huang et al., 2014), response to treatment (Patel et al., 2015), remission (Liu et al., 2014) or recurrence (Wang et al., 2014) of MDD. Instead of doing a one-time prediction, some papers model the full dynamic of the disease with a Markov chain (Marostica et al., 2015; Bhattacharya, 2014; Oskooyee, 2011), a finite state machine (Demic and Cheng, 2014) or a simulation model (Patten, 2007). There is also some work on the diagnostic of MDD (Wu et al., 2015). Finally, there is some work in expert systems in mental health (Ohayon, 1993; Bronzino, Morelli, and Goethe, 1989) with more powerful frameworks starting to appear (Bennett and Hauser, 2013).

2.3 Operations Research and Management Science

Operations research and management science (OR/MS) methods have been applied successfully to a wide range of healthcare settings in the past (i.e., system design and planning, management of operations and medical management) (Pierskalla and Brailer, 1994). In particular, more recently, there has been a surge of interest in the application of OR/MS methods to MDM because of the increasing healthcare costs, the increasing access to better data, the high level of preventable medical errors and the trend towards the uniformization of medical practice (Tunc, Alagoz, and Burnside, 2014). This section covers applications of OR/MS methods to MDM issues that fit within our previously described inclusion/exclusion criteria. For other examples of OR/MS methods applied to MDM, you can refer to Schaefer et al. (2004) and Zhang et al. (2013b).

2.3.1 Methodologies

Much of the early literature of OR/MS in MDM have focused on Markov models (Alagoz et al., 2010). These models were used to represent disease progressions and compare different policies. They are however limited when optimizing for the best policy. For example, Alagoz et al. (2010) compared a Markov model and a Markov decision process (MDP) model on the problem of the optimal timing of liver transplantation. The solution time was under one second for the MDP and about one minute for the Markov model.

Both gave the same optimal policy and optimal value. However, the Markov model only explored the threshold policies while the MDP model was able to explore all possible policies.

This is why MDP models are becoming increasingly popular for MDM (Alagoz et al., 2010). Schaefer et al. (2004) provided an overview of MDP in the context of MDM. Variants and extensions of this framework such as partially observable Markov decision processes (POMDPs), semi-Markov decision processes (SMDPs) and approximate dynamic programming (ADP) have also been used in this context (Schaefer et al., 2004). For more information regarding these methodologies, refer to Puterman (2005), Bertsekas (2005), Bertsekas (2012), and Powell (2011).

2.3.2 Applications

We now review some applications of OR/MS methods that respect our inclusion/exclusion criteria. These are classified according to whether they address the issue of initiation, switching, sequencing or dosage of treatment.

Treatment Initiation

Shechter et al. (2008) looked at the trade-off between benefits (e.g., avoid side effects and development of resistance) and risks (e.g., irreversible damage to immune system, complications and death) of delaying human immunodeficiency virus (HIV) therapy. They developed an infinite-horizon MDP model to find the optimal time to initiate HIV therapy that maximizes total expected lifetime or quality-adjusted life years (QALYs). They applied their model to data from the Veterans Aging Cohort Study: an observational cohort study of 25,000 HIV+ and 67,000 HIV- individuals. The action space at each state consists in whether to initiate treatment or wait. The state $h \in \mathcal{H} \triangleq \{0, 1, 2, 3, 4\}$ was either death (i.e., $h = 0$) or one of the four CD4 count range (i.e., $h \in \mathcal{H}' \triangleq \{1, 2, 3, 4\}$ where higher is better). Interpolating splines on CD4 count prior to treatment were used in order to determine the transition probability matrix (TPM) P (i.e., natural history model) under constant time intervals. The expected remaining lifetime $R(h)$ upon initiating therapy (i.e., survival model) was computed with Cox proportional hazard ratios between HIV+ and HIV- for each CD4 category that were applied to standard life table data. Their model is

given by:

$$V(h) = \max\left\{ r(h) + \sum_{j=0}^{4} p(j|h)V(j),\ R(h) \right\} \quad \text{for } h \in \mathcal{H}$$
$$V(0) = 0$$

where $r(h)$ is the immediate reward and $p(j|h)$ is given by the TPM P. Their model can work without discounting because of the absorbing state (i.e., death) that is reachable from all states. They provided structural results for the optimal policy and, quality of life and adherence variants of their model. Finally, some sensitivity analysis were done in order to test the sensitivity of the optimal policy and optimal value to changes in the natural history model or survival model.

Shechter, Alagoz, and Roberts (2010) improved over the model proposed by Shechter et al. (2008) by incorporating the possibility of a new treatment development. Their model consists in a finite-horizon model solved by backward induction where the terminal values are given by infinite-horizon models. They apply their model to the same data as in Shechter et al. (2008). An irreversible treatment T_1 is always available to the patient and there exists a possibility that a better treatment T_2, currently under clinical trials, may become available in N months with probability q. They set N to 12 months and they explore different values for q and R^{T_2} (defined later). The state is represented by the health state h (as defined in Shechter et al. (2008)) and by the number of treatments available $m \in \{1, 2\}$. The action space is either $\{W, T_1\}$ when $m = 1$ or $\{W, T_1, T_2\}$ when $m = 2$. However, because T_2 dominates T_1, the latter action space is reduced to $\{W, T_2\}$. The models are the following:

$$V_t(h,1) = \max\left\{ r(h) + \sum_{j=0}^{4} p(j|h)V_{t-1}(j,1),\ R^{T_1}(h) \right\}$$
$$\text{for } h \in \mathcal{H}',\ t = N, N-1, \ldots, 2$$
$$V_1(h,1) = \max\left\{ r(h) + (1-q)\sum_{j=0}^{4} p(j|h)V_0(j,1) + q\sum_{j=0}^{4} p(j|h)V_0(j,2),\ R^{T_1}(h) \right\}$$
$$\text{for } h \in \mathcal{H}'$$
$$V_0(h,m) = \max\left\{ r(h) + \sum_{j=0}^{4} p(j|h)V_0(j,m),\ R^{T_m}(h) \right\} \text{ for } h \in \mathcal{H}',\ m \in \{1,2\}$$
$$V_t(0,m) = 0 \text{ for } m \in \{1,2\},\ t = N, N-1, \ldots, 0$$

where $R^{T_m}(h)$ indicates the expected remaining lifetime upon initiating therapy with

treatment m and health state h. They denoted by $W(h)$ the perceived expected value when considering only T_1 (same model as in Shechter et al. (2008)) and by $W^{\pi}_{act,N}(h)$ the actual expected value when considering only T_1 in the decision process but with an unknown possibility of T_2 becoming available in N months. All these values have the following ordering: $W(h) \leq W^{\pi}_{act,N}(h) \leq V_N(h,1)$. They proved some structural results for their models and then compared the different optimal policies and values of $W(h)$, $W^{\pi}_{act,N}(h)$ and $V_N(h,1)$.

Liu, Brandeau, and Goldhaber-Fiebert (2017) also looked at the issue of technological advances in treatment. They however assumed that the disease progression is deterministic and that the irreversible treatment effectiveness can improve by random amounts over the time horizon. Their model consists in a finite-horizon MDP where the goal is to maximize the total expected QALYs. The available actions at each stage is to treat according to the current treatment effectiveness or to wait for it to improve. The treatment effectiveness (i.e., the state of the model) consists in the probability of success of the treatment. It improves stochastically according to either an incremental innovation model or radical innovation model. After deriving structural properties of their model, they applied it to the treatment of chronic hepatitis C. With three time periods and multiple scenarios, they were able to capture intuitive results on this application.

Kurt et al. (2011) optimized the statin initiation policies for patients with Type 2 diabetes. With a MDP model, they tried to identify the optimal time to initiate the irreversible statin treatment in order to maximize the patient's QALYs. Once the statin treatment is initiated ($m = 1$), the patient's lipid ratio (i.e., the health state), h, is improved by a factor w before progressing according to the natural history model. The transition probability from health state h to health state j at time t under treatment status m is given by:

$$p_t^m(j|h) = \begin{cases} [1 - p_t^m(H+1|h]q(j|h) & \text{if } h, j \in \mathcal{H}', \\ p_t^m(H+1|h) & \text{if } h \in \mathcal{H}', j = H+1 \\ 1 & \text{if } h = j = H+1 \\ 0 & \text{otherwise} \end{cases}$$

where $p_t^m(H+1|h)$ is the probability of an adverse event (i.e., state $H+1$) and $q(j|h)$ is the probability of transitioning to j from h given that the patient does not incur an adverse event. The rewards $r^m(h)$ are defined as the QALYs between each time periods and as 0 if in the absorbing state $H+1$. There are $N+1$ time periods. The last time period is used as an infinite horizon MDP where we assume that the parameters don't change. The

optimality equations are given by:

$$V_t(h) = \begin{cases} \max\left\{r^0(h) + \lambda \sum_{j\in\mathcal{H}} p_t^0(j|h)V_{t+1}(j), R_t(h)\right\} & \text{for } h \in \mathcal{H}, t = 1, \ldots, N \\ \max\left\{r^0(h) + \lambda \sum_{j\in\mathcal{H}} p_{t-1}^0(j|h)V_t(j), R_t(h)\right\} & \text{for } h \in \mathcal{H}, t = N+1 \end{cases}$$

where

$$R_t(h) = \begin{cases} r^1(h) + \lambda \sum_{j\in\mathcal{H}} p_t^1(j|h)R_{t+1}(j) & \text{for } h \in \mathcal{H}, t = 1, \ldots, N \\ r^1(h) + \lambda \sum_{j\in\mathcal{H}} p_{t-1}^1(j|h)R_t(j) & \text{for } h \in \mathcal{H}, t = N+1. \end{cases}$$

$R_t(h)$ corresponds to the total expected rewards of starting statin at time t in health h. λ corresponds to the discount factor. Kurt et al. (2011) derived structural properties for their model and applied it to longitudinal data from Mayo Clinic in Rochester. Through their numerical application, they found that their optimal solutions are both sensitive to the definition of the rewards function and to the treatment effect w. They also found that their model leads to improvements in QALYs over the current treatment guidelines.

Zhang and Denton (2015) provided a robust MDP model for the glycemic control of patients with type 2 diabetes. It is built upon the Markov chain model proposed by Zhang et al. (2014). The goal of the model is to decide which of the multiple available treatments to initiate (or not) at each stage according to the state of the system (i.e., the health state and the remaining treatments) in order to maximize the patient's QALYs. They proposed an interval matrix model with a budget of uncertainty for the uncertainty set of the transition probabilities between health states. The full transition probability matrix consists in the transition probabilities between health states and the probability of entering the absorbing state (i.e., probability of an adverse event). The uncertainty set respects the rectangular uncertainty property to maintain tractability. They provided algorithms in order to solve their robust counterpart and applied this model to the claims data of Zhang et al. (2014) in order to provide the Pareto frontier. They showed that their robust model can be solved in a timely fashion while allowing for the adjustment of conservativeness.

Treatment Switching

Hsih (2010) worked on the glycemic control of type 2 diabetes patients. They developed a MDP model to optimize the selection of medication in order to keep the health measures in the normal ranges. Their objective function is a combination of piecewise linear functions of the four health state's dimensions. The health state is discretized in order to solve it as a MDP model. Their study includes the process of optimal learning to the MDP

model. This model can thus be viewed as a multi-armed bandit model where there is an exploration-exploitation tradeoff. They used a Dirichlet distribution to represent the transition probability matrix. They compared a model called pure exploitation (PE) that uses the current Dirichlet distribution in order to find the optimal action at a stage with a model called knowledge gradient (KG) that also takes into account the learning value of each possible action at that stage. The KG model puts a greater emphasis on the exploration than the PE model. It is however much more expensive to compute. From their experimental results, they found that the KG model outperforms the PE model especially when a prior underestimates the probability of mild effects for a chosen action. They also found intuitive results regarding the use of the different drugs in order to treat type 2 diabetes.

Jiang and Powell (2015) developed a convergent ADP algorithm for monotone value functions. By exploiting the structure of the value functions, it increases the convergence's rate significantly over the other ADP algorithms. It is also faster than backward dynamic programming while providing really good solutions. For the MDP setting without optimal learning of Hsih (2010), they found that their algorithm was able to provide solutions within 1.5% of the optimal solution while significantly less computation.

Treatment Sequencing

Kahruman et al. (2010) optimized the adjuvant endocrine therapy plan for HR+ early stage breast cancer patients who are postmenopausal. They used a mixed integer nonlinear programming model in order to determine the drugs to use for the first and second phases, and the length of these phases. The parameters of their model are based on data from published randomized controlled trial (RCT) results. Their model tries to maximize the disease-free survival (DFS) percentage at the end of the treatment period. It takes into account the different side effects associated with these drugs (i.e., thromboembolic events, cardiovascular disease events, endometrial cancer, bone fractures, hot flushes and vaginal bleeding) through constraints that bound their risk. The equations deriving these risks for each treatment plan are derived from the RCTs data. These equations are the one rendering the model nonlinear. Kahruman et al. (2010) tested different parameters values for their model. They found that the recommended treatment plans depends on the risk bounds for the side effects. They also found that a shorter first phase leads to promising results. Hence, they recommended conducting such RCTs. Finally, they evaluated the DFS percentage and the different side effects risks for treatment plans that were tested in

the published trial results. They found that their model represented well these treatment plans.

Treatment Dosage

He, Zhao, and Powell (2010) worked on the controlled ovarian hyperstimulation cycle of the in vitro fertilization-embryo transfer therapy. They built a discretized finite-horizon stochastic model of the problem in order to minimize the risk of ovarian hyperstimulation syndrome and maximize the probability of pregnancy. Their state consists in three measures frequently used in the medical practice. The decision consists in whether to give two or three ampoules of gonadotropin at each of the time epoch. The time horizon has a maximum length of 20 days but can be stopped before (down to 6 days after initiation of therapy) if the mean diameter of the larger ovary (i.e., one of the state variables) grows larger than 18 mm. The transition probability matrix of the states is splitted into two parts: (1) a deterministic part that depends on the patient class and the dosage and (2) a correlated stochastic part that depends on the patient class and the previous state. The patient class (i.e., normally responsive or highly responsive) is assumed to be known a priori and is assumed to stay the same throughout the full time horizon. The cost function is defined as an additive piecewise linear convex function that depends on two variables of the state. It is incurred when the time horizon finishes or when the stopping condition is met. He, Zhao, and Powell (2010) provided a modified backward dynamic programming (DP) algorithm to solve their problem and did several analysis of their discretized model against a simulation model. They tested the effect of different levels of discretization, the impact of a wrong patient classification and the robustness of their cost function. They were not able to find a structural characterization of the optimal policy.

He, Zhao, and Powell (2012) solved the same problem as He, Zhao, and Powell (2010) but by using an ADP model. They did this in order to improve the solution time of He, Zhao, and Powell (2010) without affecting too much the quality of the solution found. Three different ADP approaches were proposed. The first one, the lookup-table ADP, consists in a lookup table for the value function that is updated according to simulated trajectories. The second one, the separable piecewise linear (PWL) value function approximation, leverages the convexity and separability properties of the value function. It updates the slopes of the linear segments of the value function according to simulated trajectories. It also ensures that the convexity property is maintained. The last one, the indexed piecewise linear function, improves over the second one by indexing these slopes according to intervals of the ovary diameters (i.e., a dimension of the state variable). This indexing allows to take

into account the correlation with the ovary diameters. It, however, increases the number of functions to evaluate. They, thus, use a weighted sum of the indexed functions and PWL functions for this last ADP model. Numerical comparison of these ADP models with the MDP model of He, Zhao, and Powell (2010) shows that the two PWL approximations are superior to the lookup-table approximation because they leverages characteristics of the problem. They are both close to the optimal solution value of the MDP benchmark and they allow to reduce the solution time from 41.2 hours (MDP benchmark) to seconds.

Ibrahim et al. (2016) worked on personalizing the dosage of warfarin; a drug used to prevent hearth attacks, strokes and bloth cloths that accommodates a narrow therapeutic window. They developed a two-stage solution that is in line with medical practice. Their first model corresponds to the initiation stage and learns sequentially about the patient sensitivity to warfarin using POMDP until it converges to an exogenous required accuracy. It is described by the following Bellman's equation:

$$V_t(b) = \min_{d \in \mathcal{A}} \left\{ R(b,d) + \theta \sum_{o \in \mathcal{O}} p(b,d,o) V_{t-1}(\tau(b,d,o)) \right\}$$

where t is the time period, b is the belief state for the sensitivity γ, d is the dosage action, $R(b,d)$ is the immediate risk of having dosage d with a belief for sensitivity b, θ is the discount factor, o is the international normalized ratio (INR) measure, $p(b,d,o)$ is the probability of observing o given b and d and $\tau(b,d,o)$ is the update function for the belief. Assuming a linearly additive Gaussian form for the dose-response model, they provided a closed-form update function for the belief. Their risk function is a quadratic function where the minimum is located at an INR value of 2.5. Their second model corresponds to the maintenance stage and optimize the dosage with respect to the expected risk. It used the final updated distribution for the sensitivity that was learned in the first model. Their second model is the following MDP:

$$V_t(s) = \min_{d \in \mathcal{A}} \left\{ r(s,d) + \theta \sum_{s' \in \mathcal{S}} P_{\mathcal{S}}(s,d,s') V_{t-1}(s') \right\}$$

where s is the INR measure. They proved that the myopic policy is optimal for this MDP model and they are thus able to solve this problem analytically. They also solve this problem for the case where the objective consists in maximizing the time in the therapeutic range. Their two-stage approach was applied to observational data from the Jewish General Hospital in Montreal. They fitted their linearly additive Gaussian model on the

data of the 503 patients in order to find the variance of the error term and the parameters for the relevant fixed effects at the population level. They discretized their states and action space according to the data and literature. It gave them three levels for the sensitivity, five intervals for the INR values and five possible actions. The state-transition probabilities were determined according to the dose-response model. With their application, they found that it suffices to have a short initiation stage in practice and that, if the physician is unsure about the sensitivity, it is better to underestimate it.

2.4 Artificial Intelligence and Statistics

This section covers methodologies and applications at the intersection of artificial intelligence and statistics. In particular, this section will discuss the dynamic treatment regimes (also known as adaptive treatment strategies) literature (Chakraborty and Moodie, 2013), which is an emerging field that specifically address our problem and that is mainly based on the reinforcement learning and causal inference literature.

Reinforcement learning (RL) is a subfield of machine learning that focuses on the search for the best sequence of actions in order to maximize the rewards of a learner while he is interacting with the environment. It includes Markov decision processes (MDPs) seen in Section 2.3 but it also includes settings with unknown system dynamics (i.e., unknown state transition probabilities and reward functions), limited data (i.e., no generative models or expensive data collection) and non-Markovian set-ups (i.e., history dependent). These latter settings are similar to what medical decision making is generally facing. Under these settings, RL is also known as a specific setup of ADP. In this section, we only cover some particuliar methodologies from RL. For a more general overview, refer to Sutton and Barto (1998), Bertsekas and Tsitsiklis (1996), and Kaelbling, Littman, and Moore (1996). For a general framework for designing and testing a RL approach for dynamic treatment regimes (DTRs), refer to Vincent (2014).

Causal inference is a subfield of statistics, epidemiology, economy and social sciences. It is interested in determining the causal effect (i.e., not the association) of a treatment, exposure or intervention on an outcome. There has been a lot of work done over the years by James Robins and colleagues on approaches to model and estimate the joint effects of a sequence of treatments (Vansteelandt and Joffe, 2014). The prior work of Robins described by Vansteelandt and Joffe (2014) is, however, not dynamic (i.e., personalized) with respect to previous observations (i.e., past outcomes and treatments) so it will be skipped in favor of the work done on dynamic treatment regimes described here. Causal inference is also

interested in the statistical estimation of error which will be discussed at the end of this section. For more information about causal inference, refer to Hernán and Robins (2018).

With the dependence of DTRs on subfields of artificial intelligence and statistics, it is easy to see how it can be seen as analytics. There are several good references in the literature that review the work that has been done on DTRs. For a quick introduction, refer to Chakraborty and Murphy (2014), Zhao and Laber (2014), Moodie, Richardson, and Stephens (2007), and Murphy et al. (2007). For a more comprehensive overview, refer to Chakraborty and Moodie (2013).

2.4.1 Methodologies

The DTR literature is generally interested in one of two goals. It either compares two (or more) preconceived DTRs in terms of their outcomes or solves for the optimal DTR (Chakraborty and Moodie, 2013). Many approaches have been developed for these two goals. These approaches are based on the potential-outcome framework that has been developed in the causal inference literature. We first review this framework and some notation before looking at the indirect and direct approaches to accomplish these goals. In this subsection, we also cover some other interesting methodological aspects and the confidence sets that can be constructed over the parameters or the values of the DTRs. Finally, we discuss the sequential multiple assignment randomized trial (SMART) design which is the randomized controlled trial design generally used with DTRs.

Potential-Outcome Framework and Notation

The potential-outcome framework (Robins, 1986) that was developed in the causal inference literature is the basis for the methods that will be discussed next. This framework is necessary to infer the expected potential outcomes that would occur for an individual if he receives treatments different than the ones observed. In order to take the population averages for these expected potential outcomes, the following assumptions are necessary in the context of DTRs (Chakraborty and Moodie, 2013):

1. axiom of consistency (i.e., potential outcome under the observed treatment and the observed outcome agree);

2. stable unit treatment value assumption (i.e., a subject's outcome is not influenced by other subjects' treatment allocation); and

3. no unmeasured confounders.

The first assumption requires that it is possible for all treatment options to be assigned to all individuals in the population. The second assumption requires that there are no interpersonal dynamics between the population's individuals that could influence the outcomes. The third assumption always holds under randomization and may be approximately true in observational studies when all confounders have been measured. Additional assumptions might also be necessary depending on the context and the methodology.

We will now introduce some notation that will be reused in the following subsections. Assume there are K stages. Let A_j denote the random variable (RV) describing the treatment/action taken at stage j. Let O_j denote the RV describing the observation/state at stage j. Let $H_j \triangleq (O_1, A_1, O_2, \ldots, A_{j-1}, O_j)$ denote the RV describing the history at stage j. Let Y_j denote the RV describing the outcome/reward at stage j. For all the previous RVs, let a_j, o_j, h_j and y_j denote their respective observed values. A DTR is often represented by a policy $d \triangleq (d_1, d_2, \ldots, d_K)$ that gives the action a_j that needs to be taken at each stage j according to the decision rules d_j that depends on the history H_j, i.e. $a_j \triangleq d_j(H_j)$. The optimal policy is denoted by d_{opt}. Finally, for each RVs or observed values, let $\bar{x}_j \triangleq (x_1, x_2, \ldots, x_j)$, $\underline{x}_j \triangleq (x_j, x_{j+1}, \ldots, x_K)$ and $\hat{x}$ denote the estimated value of x.

TABLE 2.2: Notation reference for Section 2.4.

Notation	Description
A_j	Action/treatment at stage j
O_j	State/observation of a patient at stage j
$H_j = (\bar{O}_j, \bar{A}_{j-1})$	History at stage j
$Y_j = Y_j(H_j, A_j, O_{j+1}) = Y_j(\bar{O}_j, \bar{A}_j, O_{j+1})$	Reward/outcome at stage j
$d_j = d_j(H_j)$	Decision rule at stage j that depends on history H_j
$d = (d_1, d_2, \ldots, d_K)$	Policy
d_{opt}	Optimal policy

Indirect Methods

Indirect methods were the first methods to be proposed for DTRs. These are called indirect because they use a model in order to evaluate the value of regimes. These can be split in three categories: (1) quality learning, (2) advantage learning and (3) G-computation. The first two are semi-parametric while the last one is fully-parametric. We review all of them and give a short comparison of the first two.

Quality Learning Quality learning (Q-learning) was brought from the reinforcement learning literature. It was originally proposed by Watkins (1989) as a method to solve multi-stage decision problems from sample data trajectories. In contrast with dynamic programming (Bellman, 1957), it doesn't require the complete knowledge of the system dynamics and it is able to deal with the curse of dimensionality with the use of function approximations. Q-learning can be viewed as an ADP approach.

As described in Chakraborty and Moodie (2013), Q-learning consists in learning a Q-function at each stage, j, that is parameterized by an action, a_j, and an history, h_j. The Q-function computes the total expected reward-to-go starting from the history h_j at stage j, using the action a_j and then following the policy d until the end of the horizon:

$$Q_j^d(h_j, a_j) = E[Y_j(H_j, A_j, O_{j+1}) + V_{j+1}^d(H_{j+1})|H_j = h_j, A_j = a_j]$$

where $V_{j+1}^d(H_{j+1}) = E_d[\sum_{k=j+1}^{K} Y_k(H_k, A_k, O_{k+1})|H_{j+1}]$. If this policy d is optimal, we then called this Q-function the optimal Q-function, Q_j^{opt}. Once the optimal Q-functions have been learned by backward induction for all the stages, finding the optimal policy consists only in solving at each stage and for all histories the action that maximize the corresponding Q-function.

It is rarely the case that we can represent this Q-function exactly, especially when it depends on the full history as described here. We must therefore approximate this function by a model. Almost any model can be used to approximate this function. There are many types of models that are proposed in the reinforcement literature such as trees (Ernst, Geurts, and Wehenkel, 2005) and kernels (Ormoneit and Sen, 2002). Refer to Vincent (2014) for other examples in the RL literature. However, the most often used model for the Q-functions in the DTR literature is the linear parametric model. Only few papers in the DTR literature have applied different models. Moodie, Dean, and Sun (2013) used the generalized additive model (GAM) framework for the Q-functions and compared this approach to Q-learning with linear models and other variants on simulated data. The GAM approach showed low bias and low variance compared to these other models. Pineau et al. (2007) used kernel regression as proposed by Ormoneit and Sen (2002) on data from the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial that is described below. The concept of kernel regression is to regress on the patients instead of their variables. This is why it is called an instance-based approach. You regress by putting more weights on the patients that are similar to the one you are trying to explain. Similarity is measured as the distance between the states of a pair of patients with a same treatment history. If the treatment histories are different, then the similarity measure is equal to

zero. Unfortunately, this paper was not able to identify a clear drug winner because there weren't enough data or enough difference between the different drug effects.

Advantage Learning The Q-learning approach models the conditional mean outcomes. This section describes methods that model the contrast of conditional means, i.e., advantage learning (A-learning) methods. Most of the following methods were originally developed in the causal inference literature in contrast to the Q-learning approach that were developed in the AI literature.

Murphy (2003) proposed to model the regret function at each stage by a parametric model. The regret function captures the increase in the total expected reward-to-go that we forego by taking action a_j instead of the optimal action at stage j from history h_j:

$$\mu_j(h_j, a_j) = \max_{a_j} Q_j^{opt}(h_j, a_j) - Q_j^{opt}(h_j, a_j) = V_j^{opt}(h_j) - Q_j^{opt}(h_j, a_j).$$

The proposed parametric model for $\mu_j(h_j, a_j)$ use (1) a link function that is parameterized by the difference between the chosen action and the optimal action and, (2) a scaling parameter applied to this link function. By changing the shape of the link function, it is possible to influence differently the regret function when the action chosen deviates from the optimal decision. Murphy (2003) also proposed an iterative procedure to estimate the parameters of these regret functions. This procedure is based on an equation that the regret functions should respect.

Robins (2004) extended the earlier work done by Robins and colleagues (Vansteelandt and Joffe, 2014) on structural nested mean models (SNMMs) and G-estimation to the DTR setting. Like Murphy (2003), the proposed approach models the contrast of conditional means. However, this contrast is defined differently. It is defined by the optimal blip-to-reference function:

$$\gamma_j(h_j, a_j) = E[Y(\bar{a}_j, \underline{d}_{j+1}^{opt}) - Y(\bar{a}_{j-1}, d_j^{ref}, \underline{d}_{j+1}^{opt})|H_j = h_j]$$

where optimal refers to the policy, $\underline{d}_{j+1}^{opt}$, followed after stage j and blip refers to the single-stage change in treatment at stage j for the reference regime d_j^{ref} instead of a_j. The most often used reference regime is a "zero regime" such as placebo or standard care in which case we call the function the optimal blip-to-zero function. We can easily see that the approach proposed by Robins (2004) captures the model proposed by Murphy (2003) by taking the negative of the optimal blip-to-reference function where the reference regime is

the optimal regime. Robins (2004) proposed G-estimation in order to find the parameters of the optimal blip functions.

Q-Learning vs. A-Learning Due to their differences, Q-learning and A-learning show different properties. Blatt, Murphy, and Zhu (2004) showed that A-learning is less subject than Q-learning to exhibit bias due to the function approximation. However, A-learning was shown to produce more variability than Q-learning. On a similar note, Schulte et al. (2014) found that A-learning offer robustness to model misspecification compared to Q-learning but it may not be as good when the model is correctly specified. Schulte et al. (2014) also mentions that many diagnostic tools are available for Q-learning and that A-learning increases in complexity rapidly when increasing the number of treatments at each stage.

G-Computation G-computation (Dawid and Didelez, 2010; Robins, 1986) is a fully parametric approach that models and then simulates data forward in time. It thus solves the dynamic program going forward in time. It is able to compute the value of a policy, d, by fitting a model $\phi_j(h_j, a_j; \theta_j)$ for the inner conditional expectation of the following equation (Chakraborty and Moodie, 2013):

$$V^d = E\Bigg\{ \sum_{\{(h_j,a_j):1\leq j\leq K\}} \mathbb{1}[d_1(h_1) = a_1, \ldots, d_K(h_K) = a_K] \times E\left[\sum_{j=1}^{K} Y_j | H_j = h_j, A_j = a_j\right] \Bigg\}.$$

The main advantage of G-computation is that it doesn't require knowledge or estimation of the exploration policy π which is unknown in observational data. It requires, however, to keep track of many trajectories in complex settings with many stages and dimensions, and it is also prone to model mispecification (Chakraborty and Moodie, 2013). G-computation easily extends to a Bayesian framework instead of the frequentist framework.

Direct Methods

Direct methods, also known as value maximization methods or policy search methods, evaluate directly the value function of dynamic treatment regimes V^d. Thus, in order to find the optimal DTR, we only need to find the policy, d, that maximizes the estimator, $\hat{V}^d$. These approaches suffer less bias than the indirect methods but may lead to more

variance (Zhao and Laber, 2014). These approaches also lack a prognostic model but are suitable when we want to find an optimal DTR among a restricted set. There are currently three main approaches to directly estimate the value of an arbitrary regime: (1) inverse probability of treatment weighting, (2) classification-based methods and (3) dynamic marginal structural mean models. We will now discuss these.

Inverse Probability of Treatment Weighting Inverse probability of treatment weighting (IPTW) is used when computing a statistics for a population different than the one we observed. In our case, we are trying to compute V^d under policy d for data that has been generated according to an exploration policy π. This can be done with:

$$V^d = E_d Y = \int Y dP_d = \int w_{d,\pi} Y dP_\pi$$

where

$$w_{d,\pi} = \prod_{j=1}^{K} \frac{\mathbb{1}[A_j = d_j(H_j)]}{\pi_j(A_j|H_j)},$$

$Y = \sum_{j=1}^{K} Y_j(H_j, A_j, O_{j+1})$ and $\pi_j(A_j|H_j)$ denotes the probability of selecting A_j under history H_j for an exploration policy π_j (Chakraborty and Moodie, 2013). You can refer to Hernán et al. (2006) for a comparison between IPTW and G-estimation.

In order to accommodate the case where the exploration policy, π, is unknown (i.e., observational data). Zhang et al. (2013a) extended the augmented IPTW method proposed by Zhang et al. (2012a) to the multi-stage setting. The approach focuses on a restricted class of regimes and offers comparable performance against Q- and A-learning under correctly specified models. However, it is computationally burdensome due to the maximization of a discontinuous objective function and there are no theoretical results available for this approach.

Classification-Based Methods Zhao et al. (2015) extended the approach proposed by Zhao et al. (2012), a particular case of the framework proposed by Zhang et al. (2012b), to the multi-stage setting. It consist in transforming the search of an optimal DTR into a sequence of weighted classification problems or into a single classification problem. In order to work, these nonparametric approaches require data generated from a SMART design (see Section 2.4.1) with a known generative policy that respects certain conditions. However, Zhao et al. (2015) note that extension of the framework to observational data

should be possible. Zhao et al. (2015) showed that their proposed approaches have better or equal empirical performance than Q-learning and A-learning.

Dynamic Marginal Structural Mean Models Orellana, Rotnitzky, and Robins (2010) extended the marginal structural mean (MSM) approach to the dynamic multi-stage setting. This approach consists in augmenting the dataset with a replicate of an individual's trajectory for each regime of interest with which it (partially) agrees up to some stage. These replicates are then censored at the time when they stop agreeing with this particular regime. Finally, these censored trajectories are weighted with IPTW in order to compute the corrected value V^d for each regime d of interest. This approach can be easily applied to a small set of regimes like identifying a threshold for switching treatment (Shortreed and Moodie, 2012). Recent developments on this approach have been done from a Bayesian perspective for non-dynamic treatment (Saarela et al., 2015).

Other Interesting Aspects

Most of the previously presented approaches were developed for a discrete action space with a single continuous outcome where the outcome usually represents an health state. Depending on the problem at hand, however, it might be interesting to look at variants of this context. For instance, Rich, Moodie, and Stephens (2016) applied SNMMs with a continuous action space while Moodie, Dean, and Sun (2013) worked on the aspect of discrete-valued outcomes.

Laber, Lizotte, and Ferguson (2014) focused instead on the issue of multiple reward functions (e.g., symptoms and side effects). Past works have usually compounded multiple outcomes into a single score. While the work by Lizotte, Bowling, and Murphy (2012) was able to estimate an optimal regime for all linear trade-offs simultaneously, it still assumed that the composite score was a linear combination of all the outcomes. The set-valued DTRs approach developed by Laber, Lizotte, and Ferguson (2014) differs. It constructs a set of DTRs that are dominating all the other regimes by at least a predefined margin over the different outcomes. A physician can then select a regime within this set.

Zhao et al. (2011) worked on a different definition of the outcome. The outcome in this case is the survival time. This outcome can be described as a censored time-to-event outcome. They used support vector regression with a Gaussian kernel in Q-learning in order to solve this problem.

Finally, Shortreed et al. (2011) applies multiple imputation (i.e., fully conditional specification and Bayesian mixed effects methods) to fill in the missing values of their dataset

in the context of Q-learning. Missing data is an issue that must usually be addressed in both observational and controlled studies. If unaddressed, it can lead to biases.

Confidence Sets

One of the last methodological subject that we cover is the construction of confidence sets. These can be either constructed for (1) the parameters or (2) the value of a regime. These two types of confidence sets are discussed in the context of Q- and A-learning. You can consult Orellana, Rotnitzky, and Robins (2010) for information regarding confidence sets for parameters in dynamic MSM models.

Confidence Intervals for the Parameters of a Regime Confidence intervals (CIs) for the parameters of a regime is important because it can help with variable selection (e.g., which part of the history does not need to be collected) and can help to determine whether there is sufficient support to recommend one treatment over another one. A major complication with constructing these CIs consists in the phenomenon of non-regularity. Non regularity results from the non-smooth maximization operation at each stage and usually happens when the optimal decision rule is not unique (Chakraborty and Moodie, 2013). Non-regularity causes an asymptotic bias to the CIs computed with standard methods. It must thus be addressed. A number of approaches have been developed to devise CIs for the parameters in the context of non-regularity.

Robins (2004) used the idea of projection CIs and improved it for DTRs. This approach construct a joint CI for all of the parameters and then projects this CI to obtain the CIs of interest. It is however conservative and computationally expensive. Hence, it has not yet been implemented (Chakraborty and Murphy, 2014).

Chakraborty, Murphy, and Strecher (2010) proposed hard- and soft-thresholds estimators for Q-learning. These consist in setting to zero (for the hard-threshold) or shrinking to zero (for the soft-threshold) the contribution of the treatment effect for a particular history if we can't reject the null hypothesis of the treatment effect with this history. The resulting functions are still non-smooth but they can be assumed to be less non-regular (Chakraborty and Moodie, 2013). We can thus use the usual bootstrap approach to devise CIs for the parameters. Chakraborty, Murphy, and Strecher (2010) proposed a data-driven approach to select the tuning parameters in the soft-threshold estimator but none was proposed to select the tuning parameters for the hard-threshold estimator. A variant of this approach for G-estimation (i.e., Zeroing Instead of Plugging In) has been proposed by Moodie and Richardson (2009).

Song et al. (2015) proposed a similar shrinking framework called penalized Q-learning. The big difference with this new approach is that the shrinking happens during the fitting of the parameters with a penalized least squares optimization. This new approach allows the computation of CIs explicitly (i.e., without relying to bootstraps) and is hence less computationally expensive. The drawback of this approach is that it can only deals with discrete observations; a concern that has been addressed in the improved version proposed by Goldberg, Song, and Kosorok (2013). Finally, this approach can be easily adapted for G-estimation.

Different bootstrap variants have also been proposed in order to compute CIs in the face of non-regularity. Chakraborty, Murphy, and Strecher (2010) implemented the double bootstrap CI for Q-learning. It was empirically found to offer valid CIs in the face of non-regularity. It can also be applied for G-computation. It is however computationally very intensive. Laber and Murphy (2011) proposed an adaptive bootstrap procedure. It is theoretically based and provide valid (but potentially conservative) CIs. It is however again computationally expensive. It requires to solve a difficult nonconvex optimization problem. Finally, Chakraborty, Laber, and Zhao (2013) proposed the m-out-of-n bootstrap scheme for Q-learning with a data driven approach to select the resample size m. It also provides valid (but potentially conservative) CIs. However, it is conceptually and computationally simple.

Confidence Intervals for the Value of a Regime Little work has been done on the construction of CIs for the value of an estimated regime, even though this challenge has been addressed by multiple studies for RCTs specifically designed for pre-specified regimes (Chakraborty and Murphy, 2014). The only work that almost accomplish this goal is the work by Laber and Murphy (2011). They proposed a CI in the context of classification, which is close to the direct estimation of DTR as we saw earlier. More work is however needed in order to apply this method to the DTR setting (Chakraborty and Murphy, 2014).

Sequential Multiple Assignment Randomized Trial Design

These previously discussed methodologies can be applied to data originating from either longitudinal observational studies or sequentially randomized trials (Chakraborty and Moodie, 2013). While we decided to focus mainly on observational data in this paper, we will briefly introduce the sequential multiple assignment randomized trial (SMART) design that is frequently used to develop optimal DTRs. Murphy (2005) proposed the general framework of the SMART design for the development of adaptive treatment strategies. A

SMART is a multi-stage trial where randomization occurs at each stage usually conditional on some past observations (i.e., past outcomes and/or actions). This type of trial design conforms better to the way clinical practice for chronic disorders occurs while retaining the advantages of randomization (e.g., the elimination of confounders) (Chakraborty and Moodie, 2013). The methodologies developed around this design allows, for example, the designer of a study to compute the optimal sample size in order to answer its primary research question with confidence.

One SMART that is of particular interest to us is the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) (Fava et al., 2003; Rush et al., 2004). It is a multi-site, multi-level randomized controlled trial designed to assess the comparative effectiveness of different treatment regime for patients with MDD (Chakraborty and Moodie, 2013). There were 4,041 patients enrolled in this study. There were several clinic visits during each treatment level of the study. If a patient was considered unsuccessful with his assigned treatment (i.e., a score of more than 5 on the Quick Inventory of Depressive Symptomatology (QIDS) scale), he was rerandomized to the next level according to his preferences (i.e., switch or augment) if these apply. If a patient was successful with his treatment, he entered the follow-up phase. An overview of the different levels of the STAR*D trial is available in Figure 2.3 of Chakraborty and Moodie (2013).

2.4.2 Applications

We will now review some applications of these methodologies that fit with our inclusion/exclusion criteria. Rosthøj et al. (2006) focused on the problem of dynamically dosing warfarin to control the risk of blood clothing and excessive bleeding. It is one of the first studies to apply the regret approach proposed by Murphy (2003). They looked at the first 14 clinic visits of 303 patients and started their analysis after the fourth visit. The state is defined as a deviation measure from the standard INR range. The action is the change in the warfarin dosage; a discrete variable determined by the 0.5, 1, 3 and 5 mg tablet sizes. They did not take into account the timing of visits. The outcome is the estimated percentage time in range between the visits. Their semi-parametric model for the regret only depends on the current state and selected action, i.e. it does not depend on the previous history, and it forces the parameters to be the same for all stages. With the use of a simple bootstrap estimate, they justified that all the parameters of the model were necessary. They were able to get a 10% improvement over the current dosing policy.

Cain et al. (2010) explored DTRs of the form "initiate treatment within m months after the recorded CD4 cell count first drops below x cells/mm^3" in the context of HIV-infected

patients. They used a dynamic MSM method to find the optimal DTR. They allowed x to take values between 200 and 500 in increments of 10 and, they fixed m at zero and then at three. There were 4,237 HIV-infected individuals in their database. The outcome of interest is clinical acquired immune deficiency syndrome (AIDS) or death. In the end, they found the same optimal DTR with $m = 0$ and $m = 3$.

Cotton and Heagerty (2011) searched for the optimal hematocrit range that does not require a change of epoetin dosage in order to maximize survival time for patients with end-stage renal disease. Their approach uses a dynamic MSM model and is based on Medicare claims for hemodialysis of 7,495 subjects. The DTRs that they looked at are of the form (Chakraborty and Moodie, 2013):

$$A_j \in \begin{cases} A_{j-1} \times (0, 0.75) & \text{if } O_j \geq \psi - 3 \\ A_{j-1} \times (0.75, 1.25) & \text{if } O_j \in (\psi - 3, \psi + 3) \\ A_{j-1} \times (1.25, \infty) & \text{if } O_j \leq \psi + 3 \end{cases}$$

where A_j and A_{j-1} are the epoetin dose respectively at stage j and $j - 1$, O_j is most recent hematocrit measurement, and ψ is the middle value of the target range. They did pairwise comparisons between the regime defined by $\psi = 33$ and the ten regimes defined by $\psi = 31, \ldots, 40$. They found that the range $(34, 40)$ significantly improves survival over the range $(30, 36)$ for this population.

Li et al. (2014) compared a fixed set of pre-specified DTRs for antidepressants on a database maintained at the Department of Veterans Affairs' Serious Mental Illness Treatment Research and Evaluation Center. There were 100,517 records of veterans taking antidepressant medication. Their outcome of interest is the adherence time. They used IPTW to adjust for censoring.

Laan and Petersen (2007) applied a dynamic MSM model to data from the Study of the Consequences of the Protease Inhibitor Era in order to determine when to switch antiretroviral therapy. They analyzed 100 subjects within this cohort study. Their results shows that "immediately following loss of viral suppression, individuals with high CD4 T-cell counts can wait to switch, while individuals with low CD4 T cell counts should switch immediately". The dependance on CD4 T cell count is less important at later time points.

2.5 Discussion

In this section, we reflect over the papers that we saw in the last two sections. We start by going over the papers that directly addressed the treatment of major depressive disorder (MDD). We then discuss some of the differences in the methodologies of the last two sections. Finally, we discuss the natural synergy between these two fields.

2.5.1 Applications to Major Depressive Disorder

Among the previously discussed papers, there are seven papers (see Table 2.3) who focused on MDD. All of them were described in Section 2.4. No studies applied to depression and fitting our criteria were found in the OR/MS literature. Five of these papers used the data from the STAR*D trial described in Section 2.4.1 and analyzed some different aspects of the data. One paper (Zhao et al., 2012) used the data from the Nefazodone-CBASP trial, a trial comparing the use of Nefazodone, cognitive behavioral-analysis system of psychotherapy and the combination of both. Finally, a last paper (Li et al., 2014) used observational data from the Department of Veterans Affairs' Serious Mental Illness Treatment Research and Evaluation Center in order to see the effects of different antidepressants on adherence time.

While all these applications used different methodologies, the issues they addressed were limited to the framework of the trials except for the study by Li et al. (2014) that used observational data. Thus, these studies only explored some limited decision set that might not be relevant for every day's practice. However, these studies might be insightful, for example, when trying to decide on which variables to include within the history and how to define the outcome. We can't necessarily define them the same way that it was done in the studies, due to data or goals, but these studies will give us a rough idea of what should be done.

2.5.2 Methodologies

We now discuss some of the differences between the methodologies proposed in Section 2.3 and 2.4.

Discrete State vs. Continuous State

In Section 2.3, most papers used a MDP or POMDP model in order to optimize the treatment. To do so, they discretized their continuous state space into a number of intervals.

TABLE 2.3: References focusing on MDD within the inclusion/exclusion criteria.

Reference	Methodology	Issue	Data source
Pineau et al. (2007)	Q-learning with kernel regression	Sequencing of the optimal treatments according to the prefered categories (switch or augment)	STAR*D trial
Song et al. (2015)	Penalized Q-learning	Sequencing of the optimal treatments within a reduced set	STAR*D trial
Zhao et al. (2012)	Weighted support vector machine	Selection of the optimal treatment Nefazodone, CBASP or Nefazodone + CBASP	Nefazodone-CBASP clinical trial
Chakraborty, Laber, and Zhao (2013)	Q-learning with a linear model and adaptive m-out-of-n bootstrapping	Sequencing of the SSRI/non-SSRI treatments	STAR*D trial
Zhang et al. (2013a)	Doubly robust augmented IPTW	Sequencing of the switch/augment decisions	STAR*D trial
Li et al. (2014)	IPTW logrank test	Effects of antidepressant selection on adherence time	Department of Veterans Affairs' Serious Mental Illness Treatment Research and Evaluation Center
Schulte et al. (2014)	Q- and A-learning with linear models	Sequencing of the switch/augment decisions	STAR*D trial

They generally selected these intervals to be the same as in the medical guidelines. This lead to a small number of intervals with two advantages. First, this small number of intervals is more computationally tractable because it limits the curse of dimensionality. Second, this small number of intervals also leads to a TPM that suffers less from sampling error because there are more observations within each intervals.

However, the lumping of these states introduces a lumping error and may lead to a lost of the Markov property in the reduced model (if this Markov property even existed at first) (Regnier and Shechter, 2013). Regnier and Shechter (2013) showed that the lumping error is generally bigger than the sampling error. They also found that the number of intervals should depend on the amount of available data and a modeler should consider a large state space even if there are few observations in each intervals. Otherwise, this *bad* predictive model will lead to *bad* treatment recommendations.

On the other hand, in Section 2.4, most of the papers dealt with a continuous state. They used models that accept continuous values so they didn't need to discretize the state like in a Markov chain. These models were carefully postulated for the problem at hand and tested for appropriateness when possible. Even though the selection of these models might seems like a big assumption that can lead to a bias, a similar assumption is made when selecting a Markov chain as the underlying predictive model (see next section).

In the end, we believe that a modeler should select the most appropriate model for the data at hand. In that sense, we believe that continuous data should be kept as continuous data when possible even though it might requires an increase in the computational burden.

Markovian State vs. History-Dependent State

Another interesting difference between Section 2.3 and 2.4 is the fact that models within the former section generally assume the Markov property while the models within the latter section generally use a state that is history dependent. The Markov property simplifies the model by reducing the size of the state space. It can be relevant for different issues like, for example, the initiation of treatment. However, it might not work for other contexts. For instance, for the sequencing of treatments, we might decide on the new treatment according to the previously given treatments, their effects and the order by which they were given.

Hence, the Markov property might not be appropriate for all contexts. It is therefore important to carefully evaluate its appropriateness when modeling and not only assume it because of its computational tractability.

2.5.3 Synergy Between the Fields

When looking at the methodologies used in Section 2.3 and 2.4, it is easy to see the potential for synergy between these fields. There are many aspects that needs to be looked at when designing a new model to resolve an issue. On one hand, it is necessary that this model is appropriate for the issue at hand and is computationally tractable. This is probably the biggest strengths of the OR/MS literature. On the other hand, it is also necessary to validate the results of this model before using them in every day's practice; particularly when this model is used on observational data. This is probably the biggest strength of the DTR literature. Hence, it is easy to see the potential for synergy between these fields.

A nice example of this synergy is the paper of Nikolaev et al. (2013). In their paper, they used optimization in order to open a new light to the causal inference literature. In particular, they used discrete optimization in order to balance the covariates distribution of the treatment and control groups in observational data. By doing so, they are able to minimize the bias of observed confounders. This proposed method compared well to the matching methods and showed to be superior over the IPTW approaches.

We believe that this synergy between the fields is a great avenue for future research and for the applications to the treatment of MDD.

2.6 Conclusion

This chapter did a review of some relevant papers to the treatment of major depressive disorder in the operations research and management science, artificial intelligence and statistics literature. A wide array of issues, models and methodologies were shown from papers respecting our inclusion/exclusion criteria. We also showed and discussed some of the significant differences between the OR/MS literature, and the AI and statistics literature. Finally, we discussed the strong potential for synergy between these fields.

We believe that the complexity of the treatment of MDD requires methodologies from all these fields in order to obtain results that will be useful for the psychiatrists. We also believe that methodologies at the intersection of OR/MS, AI and statistics are a great avenue for future research.

Chapter 3

Causal Inference from Observational Data: A Case Study on Treatment-Resistant Depression

3.1 Introduction

With the randomization of the treatment assignment among the patients, randomized controlled trials (RCTs) ensure that the treated and control groups are similar in the distribution of their pre-treatment (observed as well as unobserved) characteristics. For example, a RCT on the efficacy of a new drug to treat major depressive disorder (MDD) splits participants randomly between a treated group (i.e., the group receiving the new drug) and a control group (i.e., the group receiving the current standard treatment). If large enough, these two groups are homogeneous with respect to gender (observed) and genes (unobserved). This similarity in the distribution of the pre-treatment characteristics is necessary in order to make an unbiased comparison of these two groups. Unfortunately, it is not always practical (i.e., time and funding) or deemed ethical to run such RCTs. In these cases, using the available observational data is the only viable option but it should be used carefully. Given that the distribution of the pre-treatment characteristics might differ between treatment groups in observational data, it is necessary to adjust for this potential bias using methods from causal inference, the field addressing inference when making decisions. Refer to Appendix A.1 for an illustrative example of why this imbalance is problematic in the case of medical decision making.

Within the healthcare operations research and management science (OR/MS) literature, however, it appears that not all prescriptive studies explicitly discuss causal inference issues when using observational data; an important issue to address in order to obtain the buy-in of the healthcare professionals that are used to the gold standard of RCTs (Wagner

and Jopling, 2017). In particular, it appears that no *prescriptive optimization modeling studies* using observational data discuss these issues (e.g., Ibrahim et al. (2016), Mason et al. (2014), Kurt et al. (2011), He, Zhao, and Powell (2010), Lee, Chertow, and Zenios (2008), Shechter et al. (2008), and Alagoz et al. (2007a)) while some *prescriptive empirical studies* using observational data discuss these issues (e.g., Staats, Kc, and Gino (2018), Ramdas et al. (2018), Staats et al. (2017), Chan, Farias, and Escobar (2017), Kim et al. (2015), Song, Tucker, and Murrell (2015), KC and Terwiesch (2011), and KC and Terwiesch (2009)). In fact, even within the much wider scope of the operations management literature, it appears that less than 37% of the prescriptive empirical studies using observational data directly address the causal inference issues (Ho et al., 2017).

An objective of this work is to bring to light the issue of using observational data within prescriptive optimization models without considering causal inference. To our knowledge, this issue is not addressed in the OR/MS literature and hence not also in the healthcare OR/MS literature. In particular, Nikolaev et al. (2013) and Sauppe, Jacobson, and Sewell (2014), the first and only prescriptive optimization modeling papers discussing causal inference in the OR/MS literature, did not address this issue. They have however shown how optimization methods from OR/MS can improve the causal inference methods. We are also motivated by Van De Klundert (2016) that recently encouraged the OR/MS literature to evaluate empirically their proposed policies in order to improve the level of evidence provided to the healthcare sector and, hence, the impact of these proposed policies. While causal inference is not a substitute for empirical evaluation, we do believe that its use will reduce the gap between a policy's predicted value and its empirical evaluation and hence improve the proportion of useful policies proposed in the OR/MS literature.

The contributions of this work applies both to the methods and the domain. On the methodological side, we revisit the kernel matching with probability weights approach of Kallus (2017) (referred to as kernel mean matching for causal inference in this chapter). In particular, we rederive this approach from the kernel mean matching approach used for the covariate shift problem in machine learning; in doing so, we prove for the first time a link between kernel mean for causal inference and stabilized inverse probability of treatment weighting. We also show for the first time how kernel mean matching for causal inference can be used to compute different treatment effects over multiple treatment groups. Finally, we propose a new tuning approach that applies to the approaches for causal inference and the covariate shift problem. On the domain side, we discuss and highlight for the first time the particular challenges related to treatment optimization of a

mental illness (i.e., treatment-resistant depression) in the OR/MS literature.

Several links can be made between the methods developed for causal inference for the Markov decision process (MDP) setting (i.e., dynamic treatment regime (Chakraborty and Moodie, 2013)) and methods developed for off-policy MDPs; for example, refer to Paduraru (2013) for links regarding policy evaluation. However, an important difference between these methods is the explicitly stated assumptions of the causal inference methods. "Causal conclusions are only as valid as the causal assumptions upon which they rest" (Pearl, 2009a) and thus these assumptions should be made explicit so that decision makers can understand the basis of conclusions. Yet, it is unclear how the framework of causal inference translates to other optimization modeling methods from the OR/MS literature such as mathematical programming. For a mathematical programming model to be valid, it is required that both the qualitative information (i.e., the structure of the problem such as the objective function and the constraints) and the quantitative information (i.e, the parameters of this model) be representative of the problem addressed. While a good discussion is generally given with respect to the structure of the problem, there also needs to be a discussion around the issues of using observational data to determine the value of these parameters. This discussion of both the chosen structure of the problem and the fitting of the parameters is somewhat equivalent to the causal inference assumptions.

The chapter is organized as follows. Section 3.2 describes the fundamentals of causal inference, followed by the related work in the causal inference literature in Section 3.3. Next, Section 3.4 presents the kernel mean matching approach for causal inference, Section 3.5 presents a new tuning approach for kernel mean matching, Section 3.6 presents a comparative analysis of the kernel mean matching method with the state-of-the-art approaches in causal inference, and Section 3.7 illustrates the use of the method with data from patients suffering from treatment-resistant depression. We conclude in Section 3.8.

3.2 The Fundamentals

In this work, we adopt the Neyman-Rubin potential-outcome framework (Splawa-Neyman, Dabrowska, and Speed, 1923; Rubin, 1974). While this framework is subsumed by the structural causal model (Pearl, 2009b), it is the most prevalent framework in the health sciences and the one for which many well-known methods have been developed to compute treatment effects. In this section, we will explain the framework for the single stage multiple discrete treatments but this framework also holds for more general settings (e.g., continuous treatments or multi-stage treatments).

3.2.1 Notation

Let the tuple of random variables (RVs) (X, Z, Y) be the generative model of a population of interest where $X \in \mathcal{X}$ denotes a k-dimensional vector of pre-treatment RVs, $Z \in \mathcal{Z} \triangleq \{1, 2, \ldots, T\}$ is a RV indicating the treatment used amongst the T possible treatments and $Y \in \mathcal{Y}$ is a RV denoting the value of the post-treatment variable (i.e., the outcome). Let $\{(x_i, z_i, y_i)\}_{i=1}^{n}$ denote the set of the n iid observed realizations of this tuple of RVs, i.e., a sample of n individuals from the population of interest. Let the RVs $Y^{(1)}, Y^{(2)}, \ldots, Y^{(T)}$ represent the potential outcomes, i.e., $Y^{(t)}$ represents the outcome if treatment t had been used. It is important to note that although the RVs $Y^{(t)}$ are defined for all $t \in \mathcal{Z}$, we only observe the value of one after the treatment. If a potential outcome is observed, it is called a factual; otherwise, it is called a counterfactual.

A measure of interest is the distribution of the pre-treatment variables conditional on the treatment variable $z \in \mathcal{Z}$, i.e., $\Pr(X \mid Z = z)$.[1] We refer to this distribution, unless otherwise noted, as the distribution of pre-treatment characteristics for treatment group z. It is a "joint" distribution if the full vector of X is taken into account, otherwise it is a "marginal" distribution if only a subvector of X is taken into account.

In this work, a covariate denotes a pre-treatment variable that is not affected by treatment while an outcome denotes a post-treatment variable that may be affected by treatment.

3.2.2 Treatment Effects

We are interested in the effect of using one treatment versus another, i.e., a causal effect. For an individual, a causal effect is defined as the difference between two potential outcomes: $D(u, v) = Y^{(u)} - Y^{(v)}$ for $u \in \mathcal{Z}$ and $v \in \mathcal{Z} \backslash \{u\}$. If there are many treatment options (i.e., $T > 2$), then many such causal effects exist. Unfortunately, we only observe one potential outcome per individual and thus can't compute these causal effects: an issue referred to as the fundamental problem of causal inference (Holland, 1986). We can however compute the expected causal effects. Three popular expected causal effects are:

1. Average treatment effect (ATE) (Imbens and Wooldridge, 2009): The average treatment effect of treatment u relative to treatment v is defined as

$$\begin{aligned} ATE_{u,v} &\triangleq \mathbb{E}[D(u, v)] \\ &= \mathbb{E}[Y^{(u)}] - \mathbb{E}[Y^{(v)}]. \end{aligned}$$

[1] For the sake of clarity, we abuse notation for probability density functions.

There are $T(T-1)/2$ ATEs.

2. Conditional average treatment effect (CATE)[2] (Abrevaya, Hsu, and Lieli, 2015): The conditional average treatment effect of treatment u relative to treatment v conditional on the covariates e is defined as

$$\begin{aligned} CATE_{u,v,e} &\triangleq \mathbb{E}[D(u,v) \mid E = e] \\ &= \mathbb{E}[Y^{(u)} \mid E = e] - \mathbb{E}[Y^{(v)} \mid E = e] \end{aligned}$$

where E is a feature summary of X that takes on values of e_1 through e_l, i.e., l possible discrete values of a subset of covariates X. There are $T(T-1)/2$ CATEs per value e.

3. Average treatment effect among the treated (ATT) (McCaffrey et al., 2013): The average treatment effect among the treated of treatment u relative to treatment v is defined as[3]

$$\begin{aligned} ATT_{u,v} &\triangleq \mathbb{E}[D(u,v) \mid Z = u] \\ &= \mathbb{E}[Y^{(u)} \mid Z = u] - \mathbb{E}[Y^{(v)} \mid Z = u]. \end{aligned}$$

There are $T(T-1)$ ATTs.

ATE is used to compute the effect of a treatment on the population. CATE is used to compute the effect of a treatment on a subpopulation characterized by some covariate values. Finally, ATT is used to compute the effect of a treatment on a subpopulation that received some treatment. ATT makes sense in a setting such as a social program in which only a subpopulation might come forward to benefit from this program.

In the context of a randomized experiment, these expected causal effects can be computed directly because all treatment groups are similar with respect to their covariates due to the randomization of treatment. In the context of observational data, we cannot take for granted this similarity and hence need to balance the treatment groups prior to the computation of these expected causal effects. This balancing procedure requires assumptions that are described next.

[2]The literature has a variety of definitions for CATE. We use the definition of Abrevaya, Hsu, and Lieli (2015).

[3]While it is possible to compute the more general definition $\mathbb{E}[D(u,v) \mid Z = w]$, this definition has limited utility in practice and this is why we won't consider it.

3.2.3 Assumptions

Two fundamental assumptions are required to express the treatment effects in terms of data. If these assumptions do not hold, causal inference is still possible in theory but more complicated assumptions are required.

Assumption 1 (Consistency Assumption (Robins, 1986))**.** *An individual's potential outcome under a hypothetical treatment that happened to materialize is precisely the outcome experienced by that individual. Formally,*

$$Z = z \Rightarrow Y^{(z)} = Y.$$

Assumption 2 (Strong Ignorability (Rosenbaum and Rubin, 1983))**.** *Conditional on the covariates, the treatment assignment strategy used is independent of the potential outcomes. In addition, conditional on the covariates, all treatments are possible. Formally, the following holds for all treatment $z \in \mathcal{Z}$ and covariate vector $x \in \mathcal{X}$:*

$$\textit{(a) } Z \perp\!\!\!\perp Y^{(1)}, Y^{(2)}, \ldots, Y^{(T)} \mid X \textit{ and (b) } \Pr(Z = z \mid X = x) > 0$$

with $\perp\!\!\!\perp$ denoting independence.

These assumptions are required in order to balance the treatment groups with respect to the covariates (i.e., obtain $\Pr(X \mid Z = u) \approx \Pr(X \mid Z = v)\ \forall u, v \in \mathcal{Z}$). Assumption 1 is required to define the observed potential outcomes, while Assumption 2 is required to infer the unobserved potential outcomes. In particular, Assumption 2a means that all the confounders are included in X and therefore the covariates to include within X must be carefully selected. With a causal diagram of the problem at hand, it suffices to select a set of covariates X that satisfies the back-door criterion (Pearl, 2009b). Refer to Appendix A.2 for a discussion.

Unfortunately, the available subject-matter knowledge is often inadequate to construct a causal graph. In these cases, one must rely on the available literature and on expert opinions to select the covariates; it is unfortunately not possible to statistically test which covariates to include (Pearl, 2009b). There is an ongoing debate on whether it is worse to include in X or exclude from it too many covariates, and if it is even possible to answer such a question (Myers et al., 2011a; Pearl, 2011; Myers et al., 2011b). This leads to many heuristic strategies on how to select these covariates (e.g., Schneeweiss et al. (2009) and Brookhart et al. (2010)).

3.3 Related Work

There exist many different types of methods for determining causal effects from observational data such as instrumental variables and regression discontinuity design; refer to (Hernán and Robins, 2018) for more information. Among these, the *matching* and *inverse probability of treatment weighting (IPTW)* methods are two traditional types of methods related to this work. *Matching* methods use the most similar individuals in a treatment group u to infer the counterfactual $Y^{(u)}$ of an individual in a treatment group v. Different variants of matching are constructed around the definition of similarity and the numbers of individuals in a group that can be matched to a individual in another group. *IPTW* methods use a function of the propensity score (i.e., the probability of treatment assignment conditional on the observed covariates) in order to weight the individuals in a treatment group u. The expectation of the missing counterfactuals $Y^{(u)}$ for the individuals in treatment group v is then inferred using the weighted group u. Many variants of IPTW methods are constructed around the the estimation of the propensity score and the computation of the weights from it. Both matching and IPTW usually requires several iterations of parameters tuning before obtaining treatment groups that are sufficiently balanced. To these traditional types of methods, we now add the *direct balancing* methods.

Direct balancing methods are newer methods that try to balance directly the covariates between two treatment groups while eliminating the need for several iterations and limiting the number of required user inputs. To the best of our knowledge, five such approaches exist in the causal inference literature. Four of them are now briefly described. The last one is described in more details in Section 3.4.

Nikolaev et al. (2013) introduced the balance optimization subset selection (BOSS) approach that consists in minimizing the distance between the discretized empirical covariate density of the treatment groups by selecting only the relevant individuals in one of the treatment groups. They solve this combinatorial search problem for the ATT with a simulated annealing algorithm and restrict themselves to balancing marginal distributions to limit the size of the problem. This method proves to be comparable to the existing matching methods, when exact matching is not possible, by only balancing with respect to the marginal distributions. For a discussion of the relationship between the BOSS approach and matching, refer to Sauppe, Jacobson, and Sewell (2014). The two papers mentioned above constitute the only OR/MS work that we know of on causal inference and their focus is purely methodological.

Ratkovic (2015a) introduced support vector machine matching (SVMMatch) that uses support vector machines in order to determine the largest subset of individuals in a

treatment group v that is balanced with a treatment group u. This subset corresponds to the individuals in the treatment group v that are difficult to classify as whether they belong to treatment group u or v based on the centered covariates. By doing so, they can compute the ATT of treatment u relative to treatment v.

Hainmueller (2012) introduced entropy balancing (EBal). This approach optimizes weights over the individuals in order to satisfy empirical moment constraints while remaining as close as possible (according to the Kullback-Leibler divergence (KL divergence)) to a set of uniform weights. With an iterative algorithm that is globally convergent, they solve this mathematical program to compute the ATT. In some cases, this mathematical program might be unfeasible if the sample moment constraints are too restrictive.

Zubizarreta (2015) introduced stable balancing weights (SBW). This approach optimizes weights over the individuals in order to satisfy empirical moment constraints up to a certain precision while minimizing the variance of the weights. They solve this convex program using an optimization solver. They focus on the issue of incomplete outcome data but discuss how their approach can easily be used to compute treatment effects.

Direct balancing methods were shown superior to matching or IPTW for correcting the treatment imbalance error in observational data in the previous papers. The approaches of Nikolaev et al. (2013) and Ratkovic (2015a) are *selecting* individuals to include into the balanced data set. The approaches of Hainmueller (2012) and Zubizarreta (2015) are instead *weighting* the different individuals. While the previous approaches other than Zubizarreta (2015) focused on the ATT, they can all be adapted to any of the previously described treatment effects. None of these approaches, however, balance all moments of the joint distributions of the covariates; an issue addressed by the approach in the next section.

3.4 Kernel Mean Matching for Causal Inference

In this section, we revisit the kernel matching with probability weights approach of Kallus (2017). In particular, we rederive this approach from the kernel mean matching (KMM) approach, an approach used for the covariate shift problem in machine learning, we show for the first time how to use this approach to compute different treatment effects over multiple treatment groups, and we prove for the first time a link between this approach and stabilized IPTW.

The covariate shift problem, a special case of domain adaptation (Daume and Marcu, 2006; Jiang, 2008), happens when there is a shift of the independent variables' distribution

between a training and a test data set (i.e., $\Pr_{train}(X) \neq \Pr_{test}(X)$), while the conditional distribution of the dependent variables given the independent variables remains the same (i.e., $\Pr_{train}(Y \mid X) = \Pr_{test}(Y \mid X)$). This problem is similar to the primary challenge in the potential-outcome framework where the covariates distribution of each treatment group are not equal in general (i.e., $\Pr(X \mid Z = u) \neq \Pr(X \mid Z = v)$). However, in the potential-outcome framework, the conditional distributions $\Pr(Y \mid X, Z = u)$ and $\Pr(Y \mid X, Z = v)$ are not assumed to be equal. In fact, assuming them to be equal would be equivalent to assuming a null causal effect. The similarity between the covariate shift problem and the primary challenge in the potential outcome framework motivated us to adapt KMM, originally presented in Huang et al. (2007) for the covariate shift problem.[4] While the link between covariate shift and causal inference has been previously noted in Schölkopf et al. (2012) and Johansson, Shalit, and Sontag (2016), in this section we formalize this link in the context of KMM.

Our approach is along the lines of Hainmueller (2012) and Zubizarreta (2015) but differs from them because we minimize the difference in the distributions' moments using a reproducing kernel Hilbert space (RKHS) (Aronszajn, 1950). This allows us to simultaneously balance all moments of the full joint distribution. Our approach also always converges to the best possible solution; it cannot suffer from inconsistent balance constraints like the approach of Hainmueller (2012).

We now give a sketch of the population and empirical versions of KMM for causal inference described below. Given a causal effect of interest and two treatment groups u and v, KMM is executed on each treatment group with the respective treatment group (denoted prime) and some other group that we are trying to replicate (denoted double prime) as inputs. The output of these two KMM problems gives weights β_u and β_v that are then used with the respective treatment outcomes to compute the causal effect of interest.

3.4.1 Population Version

The KMM involves minimizing the mean discrepancy between two distributions $\Pr'(X = x)$ and $\Pr''(X = x)$ in a RKHS by optimizing a weight function $\beta(x)$, subject to normalization and non-negativity constraints. In the following definition, the feature function $\Phi : \mathcal{X} \to \mathcal{F}$ denotes a map into a feature space $\mathcal{F}$ that can be used to define a universal kernel $k(x_i, x_j) = \Phi(x_i)^\top \Phi(x_j)$. In the sense of Steinwart (2002), a universal kernel can induce any continuous function to arbitrary precision. In this work, we focus on the Gaussian radial basis function (RBF) kernel $k(x_i, x_j) = \exp\big(-\|x_i - x_j\|^2 / 2\sigma^2\big)$, where σ is

[4] We discovered the work of Kallus (2017) after our adaptation of KMM for causal inference.

the bandwidth parameter, that is known to be universal (Steinwart, 2002). This kernel allows us to simultaneously balance all moments of the full joint distributions, instead of only a finite number of moments, because a Gaussian RBF kernel is an inner product of two feature functions that are proportional to an infinite polynomial expansion.

Definition 1 (Population Kernel Mean Matching (Huang et al., 2007))**.**

$$\begin{aligned} \underset{\beta}{\textit{minimize}} \quad & \left\| \mathbb{E}_{x\sim \Pr'(X=x)} \left[\beta(x)\Phi(x)\right] - \mathbb{E}_{x\sim \Pr''(X=x)} \left[\Phi(x)\right] \right\| \\ \textit{subject to} \quad & \mathbb{E}_{x\sim \Pr'(X=x)}[\beta(x)] = 1, \\ & \beta(x) \geq 0. \end{aligned}$$

It is known that under certain conditions KMM converges to a unique β as explained in the following Lemma from Huang, Smola, and Schölkopf (2006).

Lemma 1. *Assume that* $\Pr''(X = x)$ *is absolutely continuous with respect to* $\Pr'(X = x)$ *(i.e.,* $\Pr'(X = x) = 0$ *implies* $\Pr''(X = x) = 0$*) and that the kernel* k *is universal. Then the solution of the population KMM (Definition 1) is* $\beta^*(x) = \Pr''(X = x)/\Pr'(X = x)$*.*

Proof. See Lemma 2 of Huang, Smola, and Schölkopf (2006). □

In addition to Lemma 1, in order to compute the treatment effects of Section 3.2.2 using KMM, it is necessary to define how these treatment effects can be computed using balancing weights. The following Lemma is an adaptation of previous proofs for a different approach (i.e., stabilized IPTW (Robins, 1998; Robins, Hernán, and Brumback, 2000)). The proof is given in Appendix A.3 for the sake of completeness.

Lemma 2 (Balancing Weights)**.** *Under Assumptions 1 and 2, the expected treatment effects can be computed using Table 3.1*

TABLE 3.1: Definition of the weighting functions for population kernel mean matching.

Causal effect	$\beta_u(x)$	$\beta_v(x)$
$ATE_{u,v}$	$\frac{\Pr(X=x)}{\Pr(X=x\|Z=u)}$	$\frac{\Pr(X=x)}{\Pr(X=x\|Z=v)}$
$CATE_{u,v,e}$	$\frac{\Pr(X=x\|E=e)}{\Pr(X=x\|Z=u)}$	$\frac{\Pr(X=x\|E=e)}{\Pr(X=x\|Z=v)}$
$ATT_{u,v}$	1	$\frac{\Pr(X=x\|Z=u)}{\Pr(X=x\|Z=v)}$

and the following causal effect equation

$$\begin{aligned} TE^{pop}_{u,v} = &\int \mathbb{E}[Y \mid Z = u, X = x]\beta_u(x) \Pr(X = x \mid Z = u)dx \\ &- \int \mathbb{E}[Y \mid Z = v, X = x]\beta_v(x) \Pr(X = x \mid Z = v)dx. \end{aligned} \tag{3.1}$$

We are now ready to put together the above building blocks to state the main theorem of this chapter, i.e., how to define $\Pr'(X = x)$ and $\Pr''(X = x)$ as inputs to KMM to enable its use for causal inference.

Theorem 1 (Population KMM for Causal Inference)**.** *Assume that Assumptions 1 and 2 hold, and that k is universal. Then, for a causal effect of interest in Table 3.1 (i.e., a row), each cell corresponds to an associated KMM (Definition 1) that takes as inputs the denominator for $\Pr'(X = x)$ and the numerator for $\Pr''(X = x)$. Using the obtained $\beta_u(x)$ and $\beta_v(x)$, the estimation of the causal effect of interest follows with Equation 3.1.*

Proof. It is easy to see that Assumption 2b implies that $\Pr''(X = x)$ is absolutely continuous with respect to $\Pr'(X = x)$ because it imposes $\Pr'(X = x) > 0$ for all treatment effects through Bayes' theorem. The rest of the proof follows from Lemmas 1 and 2. □

Additional results proving a link between KMM and stabilized IPTW (Robins, 1998; Robins, Hernán, and Brumback, 2000) are given in Appendix A.4.

3.4.2 Empirical Version

We now turn to the empirical version of KMM. Let $\mathcal{T}'$, $\mathcal{T}''$, be samples of individuals for whom the distribution of the covariates, X, is given by $\Pr'(X = x)$ and $\Pr''(X = x)$, respectively. If we want to balance the covariates of the individuals in $\mathcal{T}'$ to the covariates of the individuals in $\mathcal{T}''$ using KMM, we have to solve the following quadratic problem that now gives a *vector* β^*.

Definition 2 (Empirical Kernel Mean Matching (Huang et al., 2007))**.**

$$\begin{aligned} \underset{\beta}{\textit{minimize}} \quad & \beta^\top K\beta - 2\kappa^\top\beta \\ \textit{subject to} \quad & \left|\frac{1}{|\mathcal{T}'|}\mathbf{1}^\top\beta - 1\right| \leq \epsilon, \\ & \beta_i \in [0, B],\ \forall i \in \mathcal{T}'. \end{aligned}$$

with $|\cdot|$ denoting the absolute value for the left-hand side of the first constraint and the cardinality of a set everywhere else, $K_{ij} = k(x_i, x_j)$ for $i, j \in \mathcal{T}'$ and $\kappa_i = \frac{|\mathcal{T}'|}{|\mathcal{T}''|} \sum_{j \in \mathcal{T}''} k(x_i, x_j)$ for $i \in \mathcal{T}'$. The first constraint brings $\frac{\beta_i}{|\mathcal{T}'|} \Pr'(X = x_i)$ close to a probability distribution and the second constraint limits the scope of the distribution change.

Knowing that the dot product of two feature functions results in a kernel function (in our case, the Gaussian RBF kernel $k(x_i, x_j) = \exp\left(-\|x_i - x_j\|^2 / 2\sigma^2\right)$), the objective function in Definition 2 is obtained by squaring the empirical version of the objective function in Defintion 1 (i.e., the norm of the discrepency between the two empirical means):

$$\left\| \frac{1}{|\mathcal{T}'|} \sum_{i \in \mathcal{T}'} \beta_i \Phi(x_i) - \frac{1}{|\mathcal{T}''|} \sum_{i \in \mathcal{T}''} \Phi(x_i) \right\|^2 = \beta^\top K \beta - 2\kappa^\top \beta + \text{constant}.$$

The constraints are obtained by introducing a normalization slack ϵ and an upper bound B to reduce variability in the resulting β_i. This convex quadratic program can be solved efficiently using interior point methods (Boyd and Vandenberghe, 2004).

In order to compute a causal effect of interest using empirical KMM, it suffices to follow the procedure given in Corollary 1 that directly follows from Theorem 1.

Corollary 1 (Empirical KMM for Causal Inference)**.** *Assume that Assumptions 1 and 2 hold, and that k is universal. Then, for a causal effect of interest in Table 3.2, the associated KMM models (Definition 2) are defined using the corresponding samples of individuals $\mathcal{T}'$ and $\mathcal{T}''$ as inputs.*

TABLE 3.2: Definition of the samples of individuals used within empirical kernel mean matching to compute the weighting vectors. $^\dagger e_i$ denotes the observed feature summary of x_i while e denotes one of its l discrete values.

Causal effect	β_u		β_v	
	$\mathcal{T}'$	$\mathcal{T}''$	$\mathcal{T}'$	$\mathcal{T}''$
$ATE_{u,v}$	$\{i : z_i = u\}$	$\{i = 1, \ldots, n\}$	$\{i : z_i = v\}$	$\{i = 1, \ldots, n\}$
$CATE_{u,v,e}$	$\{i : z_i = u\}$	$\{i : e_i = e\}^\dagger$	$\{i : z_i = v\}$	$\{i : e_i = e\}^\dagger$
$ATT_{u,v}$			$\{i : z_i = v\}$	$\{i : z_i = u\}$

Then, the causal effect of interest is computed using the following equation

$$TE^{emp}_{u,v} = \frac{1}{|\mathcal{T}'_u|} \sum_{i \in \mathcal{T}'_u} \beta_{ui} y_i - \frac{1}{|\mathcal{T}'_v|} \sum_{i \in \mathcal{T}'_v} \beta_{vi} y_i \tag{3.2}$$

where $\mathcal{T}'_u$ and $\mathcal{T}'_v$ correspond to the samples $\mathcal{T}'$ used to compute respectively β_u and β_v.

3.5 Tuning of Kernel Mean Matching

As per definition, solving KMM requires the specification of an upper bound B, a normalization slack ϵ and a kernel function k. No work has been done on the identification of the optimal value for B. This upper bound B is however related to weight trimming in the IPTW literature (Lee, Lessler, and Stuart (2011)); the upper bound B is applied directly on the weights while weight trimming constrains the propensity score (e.g., $\Pr(Z = u \mid X = x)$) that is used to compute the weights. In this regard, Lee, Lessler, and Stuart (2011) found that the optimal level of trimming is difficult to identify and "analysts should focus on the procedures leading to the generation of weights (i.e., proper specification of the propensity score model) rather than relying on ad-hoc methods such as weight trimming." In our setting, this translates to finding the proper kernel function k instead of optimizing the value B.

For the normalization slack ϵ, Huang et al. (2007) showed that it should be $\mathcal{O}(B/\sqrt{|\mathcal{T}'|})$ when given the true weight function $\beta(x)$ that lies within $[0, B]$. For their application, they proposed to use $\epsilon = (\sqrt{|\mathcal{T}'|} - 1)/\sqrt{|\mathcal{T}'|}$ and $B = 1000$; values that were reused in most KMM applications. In practice, we believe that there is no way to set ϵ correctly for a given data set and, if it is set incorrectly, it might increase the bias.

For the kernel function k, Yu and Szepesvari (2012) showed analytically that its selection highly affects the convergence rate of the KMM. In this regard, Miao, Farahat, and Kamel (2013) proposed an approach that automatically selects the kernel function or tunes the kernel function parameters in KMM using an independent objective function, the normalized mean squared error (NMSE). They tuned the bandwidth σ of a Gaussian RBF kernel $k(x_i, x_j) = \exp\big(-\|x_i - x_j\|^2/2\sigma^2\big)$ with a grid search over the range $[0.1 : 0.1 : 3, 4 : 1 : 10] * \sigma_{med}$, where σ_{med} is the median of the individual's pairwise distances.

In our own experiments (see Section 3.6), the NMSE tuning approach appears highly variable—probably because it is based on the prediction of the weights for the individuals in $\mathcal{T}''$. Hence, in this study, we propose a new approach to tune the bandwidth σ of a Gaussian RBF kernel (or the parameters of any other kernel function) by minimizing the entropy of the normalized KMM solution, i.e., the scaled solution that integrates/sums to one. In particular, we select the bandwidth with the lowest entropy for a given B when doing a grid search over a pre-defined discretized set of parameter values as defined in Definition 3.

Definition 3 (Entropy Tuning). *For a given B, select the parameter values of the kernel k for which the normalized KMM solution, i.e., $\tilde{\beta}_i = \beta_i / \sum_{i \in \mathcal{T}'} \beta_i$, has the lowest entropy $H(\tilde{\beta})$:*

$$H(\tilde{\beta}) = -\sum_{i \in \mathcal{T}'} \tilde{\beta}_i \log \tilde{\beta}_i.$$

It is important to note that the normalization slack ϵ in Definition 2 can significantly worsen the solution because of the normalization requirement of the entropy function. Therefore, the proposed approach sets ϵ to zero, i.e., the first constraint of Definition 2 is $\mathbf{1}^\top \beta = |\mathcal{T}'|$.

Refer to Section 3.6 for an empirical comparison of the different tuning approaches for KMM with the direct balancing approaches of Section 3.3. We now provide some intuition regarding this entropy tuning approach.

Remark 1. *Let $\Pr''$ denote the distribution that we are trying to replicate and let $\Pr'$ be our initial distribution. In addition, let $\widehat{\Pr''} = \tilde{\beta}(x) \Pr'(x)$ denote our approximation with $\tilde{\beta}(x)$ corresponding to the normalized solution of the population version of KMM, i.e., $\tilde{\beta}(x) = \beta(x) / \int \beta(x) dx$. Then, assume that we want to minimize the KL divergence (Kullback, 1968) between the distribution $\Pr''$ and our approximation $\widehat{\Pr''}$:* [5]

$$\begin{aligned}
\underset{\tilde{\beta}}{\textit{minimize}} \quad D_{KL}\left(\Pr'' \parallel \widehat{\Pr''}\right) &= \underset{\tilde{\beta}}{\textit{minimize}} \quad \mathbb{E}_{x \sim \Pr''}\left[\log \frac{\Pr''(x)}{\widehat{\Pr''}(x)}\right] \\
&= \underset{\tilde{\beta}}{\textit{minimize}} \quad \mathbb{E}_{x \sim \Pr''}\left[\log \frac{\Pr''(x)}{\tilde{\beta}(x) \Pr'(x)}\right] \\
&= \underset{\tilde{\beta}}{\textit{minimize}} \quad \mathbb{E}_{x \sim \Pr''} \log \frac{\Pr''(x)}{\Pr'(x)} - \mathbb{E}_{x \sim \Pr''} \log \tilde{\beta}(x).
\end{aligned}$$

In this KL divergence expression, only the second term of the last equation is relevant given that it is the only term that our tuning affects. Unfortunately, we do not have access to the distribution

[5] Minimizing $D_{KL}\left(\Pr'' \parallel \widehat{\Pr''}\right)$ instead of $D_{KL}\left(\widehat{\Pr''} \parallel \Pr''\right)$, the usual approach, rewards a $\widehat{\Pr''}$ that has high probability where $\Pr''$ has high probability. In other words, no probability mass of $\Pr''$ is left out even if it requires $\widehat{\Pr''}$ to put some probability mass where there is none in $\Pr''$.

$\Pr''$ *to compute this second term. We can however approximate it as:*

$$\begin{aligned}
\underset{\tilde{\beta}}{\textit{minimize}} \quad -\mathbb{E}_{x\sim\Pr''}\log\tilde{\beta}(x) &= \underset{\tilde{\beta}}{\textit{minimize}} \quad -\int \log\Big(\tilde{\beta}(x)\Big)\Pr''(x)dx \\
&\approx \underset{\tilde{\beta}}{\textit{minimize}} \quad -\int \log\Big(\tilde{\beta}(x)\Big)\widehat{\Pr''}(x)dx \\
&= \underset{\tilde{\beta}}{\textit{minimize}} \quad -\int \tilde{\beta}(x)\log\Big(\tilde{\beta}(x)\Big)\Pr'(x)dx
\end{aligned}$$

where the last term is the entropy of $\tilde{\beta}(x)$*.*

3.6 Comparative Analysis

3.6.1 Simulation Model

The comparative analysis is done with a simulation model adapted from Hill (2011) which consists of experimental data from the Infant Health and Development Program (IHDP) where an imbalance is created by removing a nonrandom portion of the treated group: all children with nonwhite mothers. This imbalanced data set consists of 608 control and 139 treated individuals with 25 covariates. These 25 covariates consist of 17 distinct covariates (i.e., six continuous, nine binary and two categorical covariates) where each of the categorical covariates is replaced by $n-1$ dummy covariates. Using this data, we fit a propensity scoring function that consists of a L2-regularized logistic regression with the treatment indicator variable as the dependent variable and a polynomial expansion of degree three of the 25 covariates as the independent variables.

To form a data set, we sample covariates X; in particular, we sample with replacement 700 individuals from the data set of Hill (2011) to generate $\mathbf{X} \in \mathbb{R}^{700\times 25}$. Next, we generate $\mathbf{Z} \in \{0,1\}^{700}$ by sampling from the above logistic regression model using as covariates each row of $\mathbf{X}$. Then, as in Hill (2011), we sample the potential outcomes from $y_i^{(0)} \sim N(\exp\big(\psi^\top(1, x_i^{std}+W)\big), 1)$ and $y_i^{(1)} \sim N(\psi^\top(1, x_i^{std}) - \omega, 1)$ where ψ is a 26-dimensional vector with each element randomly sampled from the set $\{0, 0.1, 0.2, 0.3, 0.4\}$ with probabilities $(0.6, 0.1, 0.1, 0.1, 0.1)$, x_i^{std} is a row of the matrix $\mathbf{X}^{std}$ (i.e., the matrix $\mathbf{X}$ with the continuous covariates standardized to a unit Gaussian), W is an offset vector of the same dimension as x_i^{std} with every value equal to 0.5 and ω is set such that the expected ATT is 4, i.e.,

$$\omega \triangleq \frac{\sum_{i=1}^{700} z_i\left[\psi^\top(1, x_i^{std}) - \exp\big(\psi^\top(1, x_i^{std}+W)\big)\right]}{\sum_{i=1}^{700} z_i} - 4.$$

Both ψ and ω are fixed for the 700 individuals but vary across the data sets. Finally, we set $y_i \triangleq y_i^{(z_i)}$ for $i = 1, \ldots, 700$ to form $\mathbf{Y} \in \mathbb{R}^{700}$.

This simulation model respects Assumptions 1 and 2. Thus, it is possible to compute the causal effect from its generated data. We refer to the data sets sampled from this simulation model as the *data sets with no unmeasured confounding*.

Since Assumption 2a might not always hold in practice, we also present comparative results where this assumption does not hold. These data sets, referred to as the *data sets with hidden bias*, are obtained by removing four random covariates out of the 17 distinct covariates from $\mathbf{X}$ after computing $\mathbf{Z}$ and $\mathbf{Y}$.[6]

3.6.2 Approaches

The different approaches compared in this analysis are:

- Difference in Means (Diff. in Means), i.e., the unbalanced difference between the two treatment groups.
- KMM with a Gaussian RBF kernel using different tuning approaches for the kernel bandwidth σ (N for NMSE or E for entropy) and different values for B (B1 for 1000 or B0 for no upper bound) and ϵ (e1 for $(\sqrt{|\mathcal{T}'|} - 1)/\sqrt{|\mathcal{T}'|}$ or e0 for no slack). As discussed in Section 3.5, $B = 1000$ and $\epsilon = (\sqrt{|\mathcal{T}'|} - 1)/\sqrt{|\mathcal{T}'|}$ have been the usual values in the past KMM applications. All KMM approaches use the same range of kernel bandwidths as in Miao, Farahat, and Kamel (2013). The data for these approaches is also preprocessed by using one-hot encoding for categorical variables and by normalizing variables to the $[0, 1]$ range.
- BOSS (Nikolaev et al., 2013) using the default parameters and encoding each categorical covariate as one variable with integer values.
- SVMMatch (Ratkovic, 2015b) using the default parameters.
- EBal (Hainmueller, 2014) using the default parameters.
- SBW (Zubizarreta and Allouah, 2016) using a value of 1×10^{-4} for the parameter `bal_tols` since there is no default value for this parameter and the value of 1×10^{-4}

[6] If any of these covariates are categorical, then all binary indicators of these covariates are removed. In other words, removing one covariate can result in removing multiple columns of $\mathbf{X}$. Removing a covariate can also have no effect since its associated ψ value can be zero; if the ψ values of all four removed covariates are zero, then the associated data set is still with no unmeasured confounding.

corresponds to the smallest tolerance used in Zubizarreta (2015). We use the default values for the other parameters.

Note that, except Diff. in Means and KMM, the above approaches correspond to the direct balancing approaches described in Section 3.3 that have been previously shown to be superior to matching or IPTW.

KMM is implemented in Python 3 using CVXOPT (Andersen, Dahl, and Vandenberghe, 2015) and the BOSS approach with the code in Nikolaev et al. (2013). SVMMatch, EBal and SBW use the previously cited R packages (R Core Team, 2017). We use the default parameters for BOSS, SVMMatch, EBal and SBW because there is no available way to tune the parameters without the ground-truth; some balance measures have been proposed but they are all incomplete (e.g., they look only at the marginal distributions). Finally, note that we use KMM only with the Gaussian RBF kernel (i.e., the kernel function that is the most used with the KMM approach); using other kernel functions might alter the results of the comparative analysis but these are not explored in this study.

3.6.3 Results

We now compare the performance of the previously described approaches with respect to the bias, RMSE, range and time. The bias is the difference between the estimated ATT and the expected ATT of 4, averaged over the different simulations. The root mean squared error (RMSE) is the square root of the average value of $(\widehat{ATT} - 4)^2$ where $\widehat{ATT}$ is the estimated ATT. The time is the average wall clock time on a single core, i.e., no parallelization is used. These comparisons are done on data sets with no unmeasured confounding and on data sets with hidden bias as described previously.

First, we compare all approaches on 1000 data sets with no unmeasured confounding, generated as described above. The results of Table 3.3 show that our proposed approach (KMM-E-B0-e0) obtains the lowest RMSE.[7] Also, the mean computation time of our proposed approach is good; it can be easily reduced by parallelizing the grid search over the 37 bandwidth values. Note that, as expected, KMM-E-B1-e1 and KMM-E-B0-e1 give inferior solutions due to the normalization requirement of the entropy function (see Section 3.5); these approaches are shown for completeness but should not be used. In addition, note that SVMMatch is not able to balance all 1000 data sets.

[7]The other variants of KMM-E-* are shown in the table for completeness even if we believe that KMM-E-B0-e0 is the best approach.

TABLE 3.3: Bias, RMSE, range and mean computation time of the 1000 estimated ATTs for each approach on data sets with no unmeasured confounding. Proposed approach is in bold font. †SVMMatch only succeeded on 967 data sets.

Approach	Bias	RMSE	Range		Time
			Minimum	Maximum	s
Diff. in Means	0.01	1.16	−2.73	11.46	
KMM-N-B1-e1	0.00	0.29	1.68	6.30	5.2
KMM-N-B1-e0	0.00	0.28	1.83	6.03	3.3
KMM-N-B0-e1	0.00	0.31	1.72	6.30	4.6
KMM-N-B0-e0	−0.00	0.28	1.67	6.28	2.9
KMM-E-B1-e1	4.49	5.54	4.48	32.58	5.5
KMM-E-B1-e0	0.00	0.22	2.95	5.58	3.4
KMM-E-B0-e1	4.49	5.55	4.48	32.58	4.8
KMM-E-B0-e0	0.00	0.22	2.94	5.62	2.9
BOSS	0.00	0.37	2.06	6.86	50.5
SVMMatch†	0.00	0.41	1.15	6.90	2.2
EBal	0.00	0.33	1.79	6.62	0.4
SBW	0.01	0.33	1.90	6.83	0.6

Second, we compare all approaches on 1000 data sets with hidden bias, generated as described above.[8] The results of Table 3.4 show that our proposed approach (KMM-E-B0-e0) performs reasonably well in the presence of hidden bias in comparison with the other approaches, since it obtains again the lowest RMSE. Note however that all approaches have a higher RMSE in the presence of hidden bias.

Third, note that additional results, for both the data sets with no unmeasured confounding and the data sets with hidden bias, are given in Table A.2 and A.3 of Appendix A.5. These results were generated by increasing the standard deviations of $y_i(0)$ and $y_i(1)$ to 4, i.e., by decreasing the signal-to-noise ratio from 4 to 1. In Table A.2, the proposed approach does not obtain the lowest RMSE anymore, but it has however the tightest range after KMM-E-B1-e0. For Table A.3, it appears that the proposed approach suffers somewhat more than the other approaches from the effect of hidden bias and a low signal-to-noise ratio.

Finally, note that additional results on the *estimation of the ATE*, that are unreported for the sake of clarity, provide a different conclusion with respect to the ranking of the

[8]These data sets differ from the data sets with no unmeasured confounding with respect to the rows of $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{Y}$, and not only with respect to the missing columns of $\mathbf{X}$.

TABLE 3.4: Bias, RMSE, range and mean computation time of the 1000 estimated ATTs for each approach on data sets with hidden bias. Proposed approach is in bold font. †SVMMatch only succeeded on 960 data sets.

Approach	Bias	RMSE	Range		Time
			Minimum	Maximum	s
Diff. in Means	0.04	1.26	−3.74	14.35	
KMM-N-B1-e1	0.01	0.55	−2.94	7.54	5.6
KMM-N-B1-e0	−0.01	0.52	−2.64	7.49	3.9
KMM-N-B0-e1	0.00	0.51	0.01	7.51	5.4
KMM-N-B0-e0	−0.01	0.50	−1.10	7.08	3.5
KMM-E-B1-e1	3.56	4.71	2.91	29.96	5.8
KMM-E-B1-e0	−0.00	0.48	1.02	11.56	3.6
KMM-E-B0-e1	3.62	4.73	2.56	29.96	5.0
KMM-E-B0-e0	−0.01	0.47	1.30	11.57	3.1
BOSS	0.00	0.60	−0.96	12.99	42.4
SVMMatch†	−0.01	0.62	−3.60	7.86	2.0
EBal	0.01	0.57	−0.10	10.40	0.4
SBW	0.01	0.52	1.35	8.79	0.6

approaches. In these results, KMM-E-B0-e0 is just behind SBW and EBal in terms of bias and RMSE.

3.7 Treatment-Resistant Depression Case Study

MDD, is amongst the top ten causes of the global burden of disease and is predicted to become the leading cause by 2030 (World Health Organization, 2008). Up to 15% of the population affected by MDD remains significantly depressed despite the aggressive use of multiple pharmacological and psychotherapeutical approaches. These patients are generally referred to as suffering from treatment-resistant depression (TRD). Although there is no consensus regarding the definition of TRD, a patient suffering from MDD is usually considered treatment-resistant (or refractory) when at least two trials with antidepressants from different pharmacologic classes (adequate in dose, duration, and compliance) fail to produce a significant clinical improvement (Berlim and Turecki, 2007).

TRD patients are quite hard to treat by definition and necessitate a referral to a specialized mental health clinic where pharmacotherapy, psychotherapy and neurostimulation therapy are all possible treatment options. The psychiatrist equipped with these many

options has to determine the next best option for his/her patient. Given that TRD patients are already following a treatment, the "next best option" refers to the treatment selected at the initial visit to the specialized clinic. In particular, the five options related to pharmacotherapy are (1) the optimization of current treatment in dosage or duration, (2) the augmentation with a non-antidepressant drug, (3) the combination with another antidepressant, (4) the switch to a different antidepressant or (5) watchful waiting.[9] There are different types of psychotherapy but a psychiatrist usually only decides on whether to initiate it or not and psychotherapy usually only begins a couple of months after the initial visit. Hence, its effect is less immediate. There are also different types of neurostimulation therapy but it is mostly used as a last resort. While future changes to the patient's treatment will generally be made, the next best option is of high importance because it can initiate a response to a patient that might be at risk of suicide and that has been suffering from MDD for a significant time (i.e., duration of previous treatments and waiting time for initial visit at the clinic).

Selecting the next best option is challenging because it requires taking a decision with important consequences for the patient (e.g., side effects). The medical literature regarding this next best option is unclear. While some of the literature covers the treatment of TRD, few of the studies compare multiple treatments, and it is hard to reconcile the prevailing studies because of their different inclusion and exclusion criteria and the multiple definitions of TRD (Berlim, Fleck, and Turecki, 2008). In addition, the guidelines (Kennedy et al., 2016) are primarily designed to treat MDD, i.e., they are mostly concerned in identifying a good initial treatment. Thus, they are of limited use to treat patients suffering from TRD who have been following non-effective treatments for some time.

The goal of this case study is to provide medical decision making guidance to physicians regarding these five pharmacotherapy options (i.e., optimization, augmentation, combination, switch and watchful waiting) at the patient's initial visit to the specialized outpatient clinic; we focus on these five options since they are the ones used in medical guidelines such as Kennedy et al. (2016). In particular, we quantify the effects and the uncertainties regarding these different treatment options using a data set collected at the depressive and suicide disorders program (DSDP) of the Douglas Mental Health University Institute in Montreal.

The OR/MS literature is starting to become a mature field with respect to medical decision making for physical diseases (e.g., acquired immune deficiency syndrome (AIDS), diabetes, cancer, infertility, thrombosis). In particular, with respect to pharmacological

[9]This categorization of the options comes from the medical literature on MDD and TRD; for example, Kennedy et al. (2016) use this categorization.

treatments, there is some work on treatment initiation (Shechter et al., 2008), switching (Jiang and Powell, 2015), sequencing (Kahruman et al., 2010) and dosage (He, Zhao, and Powell, 2010; He, Zhao, and Powell, 2012; Ibrahim et al., 2016).

However, mental illnesses are generally quite different than the physical diseases and complications previously addressed by the OR/MS literature. For example, in contrast with the previously listed diseases and complications, the pathophysiology of MDD is currently unknown (Hasler, 2010) as is also the case for other mental illnesses. *Hence, models for these mental illnesses might differ substantially from models for physical diseases and complications.* With this case study, we expose for the first time, to our knowledge, the challenges related to treatment optimization of mental illnesses in the OR/MS literature.

3.7.1 Problem Setting

Our data set consists of all the unarchived medical records for adult patients suffering from TRD who started receiving treatment between August 2006 and August 2015 at the DSDP. A patient file is archived when it has not been used for more than a year so that additional storage space is available in the front office for new patient files. Of these 463 patient files, we analyzed the 87 with no missing values in the covariates, treatment and outcome described below. In addition, when several different values were available for a patient, we kept the last value since this value was often found to be a correction of the previous values.

We selected the first `QIDS-SR-16` total score (Rush et al., 2003) measured in the 30 to 90 days period after the initial visit as the outcome of interest. The `QIDS-SR-16` consists of the 16-item quick inventory of depressive symptomatology self-reported score, a score often used in the medical literature to evaluate the severity of MDD. This score lies between 0 and 27 with a higher score denoting a worse outcome. A delay of at least 30 days is given in order to let the treatment show its full potential.

Note that the `QIDS-SR-16` total score corresponds to only one very specific outcome; the broader objective of treatment consists in the general well-being of patients. Thus, when making treatment modifications, the physicians need to take into considerations additional elements such as the preferences of the patients with respect to the side-effects of the treatments (e.g., weight gain), the restrictions associated with the treatments (e.g., low tyramine diet) and an acceptable frequency of commuting between his home and the clinic for various appointments (e.g., drug dosage using blood tests, psychotherapy). It is however not possible to capture this broader perspective in this study without additional

features describing these patients' preferences; hence, this work focuses on the very specific QIDS-SR-16 total score as the outcome.

The treatments consist of the five options discussed above. The treatment strategy category is determined with Algorithm 3.1 from the current drugs taken by the patient at the time of the initial visit and the new drugs prescribed at that time. Because there could be multiple prescriptions given to a patient within 30 days following the initial visit, Algorithm 3.2 is used to determine the overall strategy that was used on the patient, i.e., the strategy with the most potential effect on the outcome.

Algorithm 3.1: Get strategy category

```
Input  : CurrentDrugs, NewDrugs
Output : StrategyCategory
begin
    Strip CurrentDrugs & NewDrugs of all drugs that are non-related to TRD.
    Change all drugs' names in CurrentDrugs & NewDrugs to their chemical names.
    Categorize all drugs in CurrentDrugs & NewDrugs as either antidepressant or
     add-on.                                                  /* Table A.4 */
if NewDrugs is empty or NewDrugs' drugs and dosages are the same as the ones in
 CurrentDrugs then
    StrategyCategory ← WatchfulWaiting
else if all NewDrugs' antidepressants exist in CurrentDrugs then
    if all NewDrugs' add-ons exist in CurrentDrugs then
        StrategyCategory ← Optimization
    else
        StrategyCategory ← Augmentation
else if CurrentDrugs' antidepressants exist in NewDrugs then
    StrategyCategory ← Combination
else
    StrategyCategory ← Switch
```

We consider several different treatment effects in this study. In particular, for these five treatment options, we compute the corresponding ATEs and ATTs. In addition, we compute two sets of CATEs. The first one is conditional on whether the treatment groups suffer from severe MDD at the initial visit. Here, severe MDD is defined as a 17-item Hamilton depression rating scale (HAM-D-17) total score greater than or equal to 24 (Zimmerman et al., 2013). Multiple versions of the original Hamilton depression rating scale (Hamilton, 1960) exist. Ours is consistent with the one used in Zimmerman et al. (2013); its score lies between 0 and 52 with a higher score denoting again a worse outcome. There are 65 individuals with HAM-D-17 < 24 and 22 individuals with HAM-D-17 ≥ 24.

Refer to Appendix A.6 for a histogram of the HAM-D-17 total score per treatment group. The second set of CATEs is conditional on the gender. There are 37 males (i.e., IsMale == True) and 50 females (i.e., IsMale == False).

3.7.2 Discussion of Assumptions

We believe that Assumption 1 holds for our data set because the only interactions between the individuals occur under group therapies and these are unlikely to begin (if prescribed) until a couple of months after the initial appointment, rendering the significance of the effect negligible within less than 90 days. Hence, when a treatment is prescribed, we only observe the effect of this treatment on an individual.

We also believe that Assumption 2 holds because, except for extreme cases, each treatment can be prescribed to any patient (Assumption 2b) and our data set contains an extensive number of covariates which increases the probability of observing all confounders (Assumption 2a). In addition, our problem consists of only one stage which limits the range of possible imbalances, i.e., here, the imbalance is only a consequence of the covariates. However, we do not have access to a causal diagram due to the unknown pathophysiology of MDD which complicates the selection of the covariates as described in Section 3.2.3. We thus decided, like other authors working on the effects of pharmacotherapy from medical record data have done (Schneeweiss et al., 2009; Brookhart et al., 2010), to include a covariate within our covariates X as long as it resembles a potential confounder, i.e., a covariate that predicts both treatment Z and outcome Y (Brookhart et al., 2010). In this work, like in Schneeweiss et al. (2009) and Brookhart et al. (2010), we are more inclusive than exclusive with respect to the covariates.

3.7.3 Covariates

With these previous considerations in mind and with the active involvement of the DSDP chief, we selected the following covariates. Age indicates the age at the initial visit. IsMale indicates whether the patient is male or otherwise female. Education is an ordinal variable indicating the highest education level obtained by the patient among four levels: (1) less than secondary school graduation, (2) secondary school diploma or equivalent, (3) some postsecondary education or (4) postsecondary certificate, diploma or degree. Abused indicates whether the patient has been a victim of abuse in the past. FamPsyHx indicates whether the patient has first-degree relatives (i.e., parents, siblings or offsprings) with psychiatric disorders. HAM-A consists in the Hamilton anxiety rating

scale (Hamilton, 1959) score that lies between 0 and 56, a measure of the anxiety's severity. HAM-D-17 consists of the 17-item Hamilton depression rating scale (Zimmerman et al., 2013) score that lies between 0 and 52, a measure of the severity of MDD. SCID1Dx is a binary vector indicating the comordibities evaluated through the Structured Clinical Interview for DSM Axis I Disorders (SCID-I) (First et al., 2002) where DSM denote the Diagnostic and Statistical Manual of Mental Disorders. We consider three diagnoses of these comorbidities, namely SCID1Dx4, SCID1Dx26 and SCID1Dx32. PastEpi indicates whether the patient had past MDD episodes. SCID2 indicates whether the patient suffers from comorbidities evaluated with the Structured Clinical Interview for DSM Axis II Personality Disorders (SCID-II) (First et al., 1997). PastSuiAtt indicates whether the patient has committed past suicidal attempts. Finally, SSI consists in the scale of suicide ideation (Beck, Kovacs, and Weissman, 1979) score that lies between 0 and 38. An higher score for HAM-A, HAM-D-17 or SSI denotes a worse outcome.

A description per treatment group of the unbalanced previous covariates and outcome is available in Appendix A.6.

3.7.4 Results

To compute the ATEs, CATEs and ATTs, we applied KMM-E-B0-e0 (see Section 3.6.2 for the definition) to the 87 patients. The results are given in Table 3.5. In each cell, the expected treatment effect and its 95% confidence interval are given. The latter is estimated with the percentile method (Efron, 1981) from 5000 bootstrap replications. Following the notation of Section 3.2.2, Table 3.5A gives the average treatment effects $ATE_{u,v}$, Table 3.5B and C give the average treatment effects $CATE_{u,v,e}$ conditional on the depression severity and Table 3.5D gives the average treatment effects among the treated $ATT_{u,v}$; for the sake of space, the results for the average treatment effects $CATE_{u,v,e}$ conditional on gender are given in Table A.6.[10] If $ATE_{u,v}$ is negative, then treatment u is more effective than treatment v to reduce the QIDS-SR-16 total score. If $CATE_{u,v,e}$ is negative, then treatment u is more effective than treatment v to reduce the QIDS-SR-16 total score for patients with characteristics e. If $ATT_{u,v}$ is negative, then treatment u is more effective than treatment v to reduce the QIDS-SR-16 total score for patients that received treatment u.

The estimated treatment effects in the 1st row of Table 3.5A indicate that on average across patients an optimization of treatment is best. Also focusing on patients with lower HAM-D-17 scores, the strategy of treatment optimization is best (see 1st row of Table 3.5B).

[10] Within Table 3.5A–C and Table A.6, the results in the lower triangular are the opposite of the results in the upper triangular and are given for ease of exposition.

However this same finding does not hold for patients with greater pre-treatment severity; here augmentation generally appears to perform best (see 2nd row of Table 3.5C). This is consistent with the Canadian Network for Mood and Anxiety Treatments (CANMAT) recommendation (Kennedy et al., 2016) to use adjunctive medication (i.e., augmentation and combination) when depression is more severe. When focusing on gender (see Table A.6), it appears that treatment optimization is best for females and males. When we consider the ATT, it appears that among patients whose treatment was optimized, this treatment indeed appears to be the most effective option (indicated by all negative numbers in the 1st row of Table 3.5D). In addition, optimization does appear to be more effective than the selected treatment within all other treatment groups (indicated by all positive numbers in the 1st column of Table 3.5D). In fact, it is the most effective treatment for patients under any treatment other than augmentation.

Unfortunately, none of the above treatment effects are statistically significant with respect to the 95% confidence intervals. This is most likely due to the relatively small number of patients with complete data out of the 463 patients that were included in this case study; note that 376 patients were excluded from this case study because of missing data. Yet, another possible reason for the absence of statistically significant results might be the inappropriateness of the five strategies; remember that these strategies are broadly defined, i.e., each strategy might contain several different drug modifications. Thus, interestingly, the absence of statistically significant results might call into question the use of these five strategies within the medical literature.

Finally, the focus of this work is on establishing the importance of balancing observational data prior to the assessment of intervention alternatives. To this end, we estimated the ATEs on the unbalanced observational data as well (see Table 3.6). In this case, it appears that augmenting the treatment is the most preferable strategy; a different result than the result obtained from the balanced data. This may partially explain the medical communities reluctance to follow advice derived on unbalanced observational data, that has been the common practice among the OR/MS community.

Algorithm 3.2: Get overall strategy

```
Input  : EvalDate, CurrentDrugs, Prescriptions
Output : OverallStrategy
begin
    Sort Prescriptions by ascending date.
    Remove items from Prescriptions that precedes EvalDate or aren't within 30
     days after EvalDate
    Store the number of days between each prescription's date of Prescriptions and
     EvalDate in Prescriptions.diff
    Store length of Prescriptions in l
    Set OverallStrategy.length to 0
if Prescriptions(1).diff != 0 then
    Strategy.category ← WatchfulWaiting
    Strategy.length ← Prescriptions(1).diff
    Append Strategy to Strategies
for i = 1 to l − 1 do
    NewDrugs ← Prescriptions(i)
    Strategy.category ← GetStrategy(CurrentDrugs, NewDrugs)
     /* Algorithm 1 */
    Strategy.length ← Prescriptions(i+1).diff-Prescriptions(i).diff
    Append Strategy to Strategies
NewDrugs ← Prescriptions(l)
Strategy.category ← GetStrategy(CurrentDrugs, NewDrugs)   /* Algorithm 1
 */
Strategy.length ← 30 - Prescriptions(l).diff
Append Strategy to Strategies
Group Strategy objects in Strategies by category and compute the cumulative
 lengths for each group
Set OverallStrategy to the category with the longest cumulative length; if ties exist,
 set OverallStrategy to the category, out of the ties, occurring first in Strategies
```

TABLE 3.5: Treatment effects with 95% confidence intervals for the TRD case study. The row is u and the column is v. 1st row/column is Optimization, 2nd row/column is Augmentation, 3rd row/column is Combination, 4th row/column is Switch and 5th row/column is Watchful Waiting.

(A) $ATE_{u,v}$

		−0.70 (−3.57,	3.62)	−2.99 (−5.77,	1.43)	−3.06 (−6.49,	2.32)	−1.14 (−4.33,	4.18)
0.70 (−3.62,	3.57)			−2.29 (−5.14,	0.57)	−2.36 (−5.67,	1.67)	−0.44 (−3.85,	3.19)
2.99 (−1.43,	5.77)	2.29 (−0.57,	5.14)			−0.07 (−3.60,	4.00)	1.85 (−1.45,	5.52)
3.06 (−2.32,	6.49)	2.36 (−1.67,	5.67)	0.07 (−4.00,	3.60)			1.92 (−2.64,	6.24)
1.14 (−4.18,	4.33)	0.44 (−3.19,	3.85)	−1.85 (−5.52,	1.45)	−1.92 (−6.24,	2.64)		

(B) $CATE_{u,v,e}$; e is `HAM-D-17` < 24

		−2.09 (−3.63,	3.65)	−3.47 (−5.95,	1.50)	−3.07 (−6.49,	2.44)	−0.42 (−4.35,	4.19)
2.09 (−3.65,	3.63)			−1.38 (−5.19,	0.69)	−0.99 (−5.68,	1.73)	1.67 (−3.86,	3.27)
3.47 (−1.50,	5.95)	1.38 (−0.69,	5.19)			0.40 (−3.75,	4.15)	3.05 (−1.56,	5.56)
3.07 (−2.44,	6.49)	0.99 (−1.73,	5.68)	−0.40 (−4.15,	3.75)			2.66 (−2.68,	6.29)
0.42 (−4.19,	4.35)	−1.67 (−3.27,	3.86)	−3.05 (−5.56,	1.56)	−2.66 (−6.29,	2.68)		

(C) $CATE_{u,v,e}$; e is `HAM-D-17` ≥ 24

		1.76 (−4.08,	3.95)	−2.98 (−6.42,	1.81)	−2.68 (−6.97,	2.72)	−1.59 (−5.19,	4.40)
−1.76 (−3.95,	4.08)			−4.74 (−5.60,	1.17)	−4.44 (−5.97,	2.03)	−3.35 (−4.22,	3.42)
2.98 (−1.81,	6.42)	4.74 (−1.17,	5.60)			0.30 (−4.14,	4.54)	1.39 (−2.01,	5.90)
2.68 (−2.72,	6.97)	4.44 (−2.03,	5.97)	−0.30 (−4.54,	4.14)			1.09 (−3.25,	6.49)
1.59 (−4.40,	5.19)	3.35 (−3.42,	4.22)	−1.39 (−5.90,	2.01)	−1.09 (−6.49,	3.25)		

(D) $ATT_{u,v}$

		−0.26 (−2.84,	3.19)	−2.42 (−5.44,	0.81)	−3.14 (−6.39,	2.50)	−1.13 (−4.54,	3.31)
0.01 (−4.60,	3.94)			−0.65 (−5.18,	3.59)	−1.52 (−5.88,	2.72)	2.95 (−3.49,	4.95)
3.18 (−4.28,	6.96)	−0.15 (−3.89,	4.55)			−3.25 (−7.04,	2.30)	−0.52 (−3.94,	5.30)
5.31 (−1.82,	9.94)	2.59 (−0.79,	6.60)	−0.53 (−4.50,	4.09)			1.82 (−3.25,	6.62)
2.62 (−2.50,	5.60)	1.27 (−2.08,	4.27)	−1.89 (−4.94,	1.57)	1.16 (−4.63,	4.13)		

TABLE 3.6: Unbalanced ATE with 95% confidence intervals for the TRD case study. For the unbalanced $ATE_{u,v}$, the row is u and the column is v. 1st row/column is Optimization, 2nd row/column is Augmentation, 3rd row/column is Combination, 4th row/column is Switch and 5th row/column is Watchful Waiting.

		0.25 (−3.00,	3.60)	−0.82 (−4.60,	3.14)	−2.56 (−6.74,	1.55)	−0.66 (−3.75,	2.59)
−0.25 (−3.60,	3.00)			−1.07 (−5.42,	3.24)	−2.81 (−7.21,	1.71)	−0.91 (−4.61,	2.88)
0.82 (−3.14,	4.60)	1.07 (−3.24,	5.42)			−1.74 (−6.69,	3.19)	0.16 (−4.08,	4.43)
2.56 (−1.55,	6.74)	2.81 (−1.71,	7.21)	1.74 (−3.19,	6.69)			1.90 (−2.53,	6.28)
0.66 (−2.59,	3.75)	0.91 (−2.88,	4.61)	−0.16 (−4.43,	4.08)	−1.90 (−6.28,	2.53)		

3.8 Conclusion

In this work, we described the fundamentals of the Neyman-Rubin potential-outcome framework, rederived kernel matching with probability weights (Kallus, 2017) from kernel mean matching and provided a new tuning approach for kernel mean matching. Next, we explicitly stated and justified our assumptions, and used the kernel mean matching approach to compute the treatment effects from observational data on treatment-resistant depression, a challenging setting given the unknown pathophysiology of depression. While some of these assumptions in our work might seem strong at first, it is important to note that these would have been even stronger in some of the prevailing healthcare OR/MS papers due to the more elaborate methods used (e.g., our single-stage policy evaluation vs. their multi-stage policy optimization) if only these assumptions had been acknowledged. We hope that this work will increase awareness within the healthcare OR/MS community of the implicit assumptions that are made when optimizing decisions over observational data.

There are several limitations to our treatment-resistant depression case study. First, we used observational data that was not purposely collected for research and thus contained several missing values. This resulted in the analysis of a subset of patients that may be different from the larger set of patients in ways that lead to bias in the estimation of the treatment effects. Second, we focused exclusively on pharmacotherapy and did not consider other types of therapy (e.g., psychotherapy, nutrition). Therapies other than pharmacotherapy are known to have an effect on remission of MDD (Lam et al., 2016a). Finally, we considered strategies that are composed of multiple drug treatments. Because the types of medications and dosing of these medications are evolving, the strategies considered here will not include newly available medications nor will include new approaches to dosing of presently available medications.

Future research areas regarding kernel mean matching include (1) the generalization of kernel mean matching to continuous and multi-stage treatments and (2) the further analysis of the use of entropy for tuning kernel mean matching.

Chapter 4

When to Set the Next Appointment of Patients Suffering from Treatment-Resistant Depression

4.1 Introduction

Setting the frequency of appointments for patients suffering from treatment-resistant depression (TRD), a severe form of major depressive disorder (MDD), is an important decision due to the trade-off that high-frequency and low-frequency appointments entails. On one hand, high-frequency appointments can lead to a waste of highly-specialized resources with limited availability such as specialized outpatient clinic, which might decrease the access to these resources. On the other hand, low-frequency appointments can lead to degradation of patients' health and to further adverse effects; for example, a patient who suffers from side effects can decide to stop the ongoing treatment if the concerns of this patient are not captured in time, and can then suffer from low quality of life and be at risk of suicide. However, even if this decision is important, it remains unclear what should be the optimal frequency of appointments since different factors can affect this decision such as (1) the patient's characteristics (e.g., health state, treatment, availability), (2) the physician's characteristics (e.g., experience, availability) and (3) the clinic's characteristics (e.g., staff availability, opening hours, waiting list's length, budget).

While the Canadian Network for Mood and Anxiety Treatments (CANMAT) guidelines (Lam et al., 2016a) do not indicate what should be the time between appointments, the Kaiser Permanente guidelines (National Guideline Directors, 2016) recommend, for patients starting an antidepressant treatment for MDD, one follow-up contact within the first month and then one other follow-up contact four to eight weeks afterwards. After remission, they recommend one follow-up contact five or six months afterwards. Finally

for asymptomatic patients with MDD who are continuing a pharmacological treatment, they recommend at least one contact per year. In contrast, Trangle et al. (2016) recommend weekly contacts first to ensure engagement and then monthly contacts for mild MDD. For moderate MDD, they recommend weekly contacts to ensure engagement and then one contact every two to four weeks. Finally, for severe MDD, they recommend weekly contacts until the severity decreases. Thus, it appears that there exists a disparity in the recommended times between appointments in the literature (when any recommendations are given) and, thus, it appears that no single best practice exists. Note that the previous recommendations (National Guideline Directors, 2016; Trangle et al., 2016) are for contacts that could be done over the phone by a care manager and not necessarily in-person with a psychiatrist.

The goal of this study is to identify which factors affect the time to the next appointment and the relative importance of these factors in this decision. To address this goal, we first elicit a set of potential features from physicians (i.e., the main decision makers in this decision) with semi-structured interviews and then use machine learning (ML) methods on data to estimate the importance of each of these features. In particular, we use methods from the field of imitation learning (IL) that are used to *reproduce* the behavior of an expert from demonstrations; we justify the use of such methods to *characterize* the behavior by assuming that a model used to reproduce a behavior is a potential model to explain the behavior. The data set used in this work, collected prior to the semi-structured interviews, consists of 463 adult patients suffering from treatment-resistant depression (TRD) with an initial visit at the depressive and suicide disorders program (DSDP) of the Douglas Mental Health University Institute in Montreal between August 2006 and August 2015. Each of these patients is followed by one of the four psychiatrists working at this specialized outpatient clinic.

This two-stage framework (i.e., semi-structured interviews and IL methods) is selected for three reasons. First, this framework allows the modeling of the potentially complex timing decision by requiring less interventions and time from the physicians in comparison to fully eliciting this knowledge from them. Second, we believe that this framework provides results which are subject to less biases than a full expert elicitation method, since the second stage does not require the interventions of physicians and is fully data-driven. For an overview of expert elicitation methods and associated biases, refer to Meyer and Booker (2001). Third, since data generally require pre-processing before it can be used, it is not practically possible to only have a fully data-driven stage without any expert knowledge; this would be too much time consuming for large data sets and would lead to

the omission of relevant features which are engineered from raw features with the help of expert knowledge.

This chapter is organized as follows. Section 4.2 recalls the basics of Markov decision processes (MDPs) and semi-Markov decision processes (SMDPs), followed by an overview of imitation learning (IL) in Section 4.3. Next, Section 4.4 introduces our proposed models to identify the potential relevant factors, and Section 4.5 presents the results with respect to our application, i.e., the timing of appointments for patients suffering from TRD. We conclude in Section 4.6.

4.2 Preliminaries

4.2.1 Markov Decision Process

A Markov decision process (MDP) (Puterman, 2005) is a tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, T, R, \gamma, \alpha \rangle$ where $\mathcal{S} = \{s_1, \ldots, s_n\}$ is a finite set of states, $\mathcal{A} = \{a_1, \ldots, a_k\}$ is a finite set of actions, $T : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the stochastic transition function where $T(s, a, s') = \Pr(s_{\delta+1} = s' \mid s_\delta = s, a_\delta = a)$ is the probability that action a in state s at decision epoch δ will lead to state s' at decision epoch $\delta + 1$, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the bounded reward function where $R(s, a)$ is the reward obtained after taking action a in state s, $\gamma \in [0, 1)$ is the discount factor and α is the initial state distribution. Using matrix notations, the transition function is denoted as an $|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|$ transition probability matrix (TPM) $\mathbf{T}$ and the reward function as an $|\mathcal{S}||\mathcal{A}|$-dimensional vector $\mathbf{R}$.

A deterministic and stationary policy π is defined as a mapping π: $\mathcal{S} \to \mathcal{A}$. The value of a policy π is the expected discounted sum of rewards and is defined as $V^\pi = \mathbb{E}\left[\sum_{\delta=0}^{\infty} \gamma^\delta R(s_\delta, a_\delta) \mid \alpha, \pi\right]$ where s_0 is distributed according to α and the action a_δ is $\pi(s_\delta)$. The value function of a policy π for state s is computed using the Bellman equation (Puterman, 2005) $V^\pi(s) = R(s, \pi(s)) + \gamma \sum_{s' \in \mathcal{S}} T(s, \pi(s), s') V^\pi(s')$ so that $V^\pi = \sum_{s \in \mathcal{S}} \alpha(s) V^\pi(s)$. Similarly, the Q-value function $Q^\pi(s, a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ of a state-action pair when following the policy π afterwards is computed as $Q^\pi(s, a) = R(s, a) + \gamma \sum_{s' \in \mathcal{S}} T(s, a, s') V^\pi(s')$. Under matrix notation, these equations are

$$\mathbf{V}^\pi = \mathbf{R}^\pi + \gamma \mathbf{T}^\pi \mathbf{V}^\pi$$
$$\mathbf{Q}_a^\pi = \mathbf{R}^a + \gamma \mathbf{T}^a \mathbf{V}^\pi$$

where $\mathbf{V}^\pi$ is an $|\mathcal{S}|$-dimensional vector with the sth element being $V^\pi(s)$, $\mathbf{R}^\pi$ is an $|\mathcal{S}|$-dimensional vector with the sth element being $R(s, \pi(s))$, $\mathbf{T}^\pi$ is an $|\mathcal{S}| \times |\mathcal{S}|$ matrix with

the (s, s') element being $T(s, \pi(s), s')$, $\mathbf{Q}_a^\pi$ is an $|\mathcal{S}|$-dimensional vector with the sth element being $Q^\pi(s, a)$, $\mathbf{R}^a$ is an $|\mathcal{S}|$-dimensional vector with the sth element being $R(s, a)$, and $\mathbf{T}^a$ is an $|\mathcal{S}| \times |\mathcal{S}|$ matrix with the (s, s') element being $T(s, a, s')$.

An optimal policy π^* is defined as the policy with the largest expected discounted sum of rewards, i.e., $V^{\pi^*}(s) \geq V^\pi(s),\ \forall s \in \mathcal{S}$ for all possible deterministic and stationary policies π. The value and Q-value functions of this optimal policy $\pi^*(s)$ are computed as

$$V^*(s) = \max_{a \in \mathcal{A}} \left[R(s, a) + \gamma \sum_{s' \in \mathcal{S}} T(s, a, s') V^*(s') \right]$$
$$Q^*(s, a) = R(s, a) + \gamma \sum_{s' \in \mathcal{S}} T(s, a, s') V^*(s')$$

with $V^*(s)$ and $Q^*(s, a)$ denoting respectively $V^{\pi^*}(s)$ and $Q^{\pi^*}(s, a)$.

4.2.2 Semi-Markov Decision Process

Semi-Markov decision processeses (SMDPs) (Puterman, 2005) generalize MDPs by allowing action choices at random times in contrast to predefined equidistant time points. In the most general form of this framework, the system state is allowed to change several times between decision epochs and is continuously modeled throughout time by the natural process.

This framework goes as follows. As a consequence of choosing action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$, the next decision epoch occurs within σ time units after the current decision epoch, and the system state becomes $s' \in \mathcal{S}$ with probability $U(s, a, \sigma, s')$.

For the sake of this work, we only consider, from now on, a special SMDP variant. In this variant, after taking an action in a given state, the system remains in this state for a random amount of time before transitioning to a new state at the next decision epoch. In this case, the joint probability $U(s, a, \sigma, s')$ can be conveniently expressed as $U(s, a, \sigma, s') = T(s, a, s')F(s, a, \sigma)$ where $T(s, a, s')$ denotes the probability that the state at the next decision epoch is s' if action a is taken in state s at the current decision epoch, and $F(s, a, \sigma)$ denotes the probability that the next decision epoch occurs within σ time units after the current decision epoch when action a is taken in state s at the current decision epoch. In this variant, the delay σ is independent of the next state s' to which the system transitions.

Using these previously defined quantities, it is possible to compute the expected reward $R(s, a)$ which is composed of a lump sum reward $r(s, a)$ and a continuous reward obtained

at rate $c(s,a)$. The lump sum reward $r(s,a)$ is obtained when taking action a in state s at the current decision epoch while the continuous reward is obtained at a rate $c(s,a)$ when action a is taken in state s at the previous decision epoch.[1] This expected reward $R(s,a)$ is computed as

$$R(s,a) = r(s,a) + c(s,a) \int_0^\infty \int_0^u e^{-\alpha\sigma} \, \mathrm{d}\sigma \, F(s,a,\mathrm{d}u)$$

where $\alpha > 0$ is the continuous-time discounting rate and $F(s,a,\mathrm{d}u)$ denotes the time-differential of F.

Similarly to MDPs, the value function of a deterministic and stationary policy π under this SMDP variant is

$$V^\pi(s) = R(s,\pi(s)) + \sum_{s' \in \mathcal{S}} P(s,\pi(s),s') V^\pi(s')$$

where

$$P(s,\pi(s),s') = T(s,\pi(s),s') \int_0^\infty e^{-\alpha\sigma} F(s,\pi(s),\mathrm{d}\sigma).$$

Furthermore, the Q-value function is

$$Q^\pi(s,a) = R(s,a) + \sum_{s' \in \mathcal{S}} P(s,a,s') V^\pi(s').$$

Under matrix notation, these equations are

$$\mathbf{V}^\pi = \mathbf{R}^\pi + \mathbf{P}^\pi \mathbf{V}^\pi$$
$$\mathbf{Q}_a^\pi = \mathbf{R}^\pi + \mathbf{P}^a \mathbf{V}^\pi$$

where $\mathbf{P}^\pi$ is an $|\mathcal{S}| \times |\mathcal{S}|$ matrix with the (s,s') element being $P(s,\pi(s),s')$, and $\mathbf{P}^a$ is an $|\mathcal{S}| \times |\mathcal{S}|$ matrix with the (s,s') element being $P(s,a,s')$.

The value and Q-value functions of the optimal policy π^* are computed as

$$V^*(s) = \max_{a \in \mathcal{A}} \left[R(s,a) + \sum_{s' \in \mathcal{S}} P(s,a,s') V^*(s') \right]$$
$$Q^*(s,a) = R(s,a) + \sum_{s' \in \mathcal{S}} P(s,a,s') V^*(s').$$

[1] Remember, that in the variant described, the natural process remains in the state of the previous decision epoch until the next decision epoch. Thus, we simplified the usual reward rate $c(s,a,s')$, that also depends on the the state of the natural process s', to $c(s,a)$.

4.3 Imitation Learning

As previously described, imitation learning (IL) is used to reproduce an expert's behavior. It generally consists of five key ingredients: (1) an access to pre-collected demonstrations or to an interactive demonstrator (i.e., the expert to imitate), (2) a model of the environment (i.e., a simulator)[2], (3) a policy class to search[3], (4) a loss function to evaluate the candidate policy against the agent policy and (5) a learning algorithm (i.e., a method to optimize the policy or reward). Formally, the general IL problem can be formulated as

$$\arg\min_{\theta} \mathbb{E}_{(s,a)\sim D_{\pi_\theta}} L(\pi^*(s), a)$$

where D_{π_θ} is the state-action distribution under the candidate policy π_θ and L is a loss function (Yue and Le, 2018).

Depending on the goal and requirements as shown in Table 4.1, the IL approaches can be split into one of three categories: (1) behavioral cloning (BC), (2) interactive direct policy learning (IDPL) and (3) inverse reinforcement learning (IRL). We now briefly describe each of them.

TABLE 4.1: Overview of the goals and requirements for the behavioral cloning (BC), interactive direct policy learning (IDPL) and inverse reinforcement learning (IRL) approaches (adapted from Yue and Le (2018)). ✓denotes a goal/requirement while [✓] denotes an optional requirement.

	Goal		**Requirement**		
	Direct policy learning	**Reward learning**	**Model of environment**	**Pre-collected demonstrations**	**Interactive demonstrator**
BC	✓			✓	
IDPL	✓		✓	[✓]	✓
IRL		✓	✓	✓	

4.3.1 Behavioral Cloning

BC corresponds to a reduction of imitation learning to supervised learning by considering that the state-action distribution is provided exogenously and that the observations under

[2]This item is not required for behavioral cloning as described below.

[3]Or, somewhat equivalently, a reward class to search in the case of some inverse reinforcement learning approaches, as described below.

this distribution are iid. There are two possible interpretations for this reduction. First, this reduction can be interpreted as the minimization of the 1-step deviation error along the expert trajectories. Second, it can be interpreted as, by assuming perfect imitation so far, it learns to continue imitating perfectly. Formally, BC is formulated as the following supervised learning problem

$$\arg\min_{\theta} \mathbb{E}_{(s,a^*)\sim D_{\pi^*}} L(a^*, \pi_\theta(s))$$

where D_{π^*} is the state-action distribution under the policy of the demonstrator π^* (Yue and Le, 2018).

A notable result regarding BC is that its worst-case error grows quadratically in the trajectory length H. More formally, this result is defined as

$$V^{\pi^\theta} \geq V^{\pi^*} - \epsilon H^2 R^{\max}$$

where ϵ is an upper bound on the expected 0-1 loss under the state-action distribution of the demonstrator's policy, i.e.,

$$\mathbb{E}_{(s,a^*)\sim D_{\pi^*}} [\mathbb{1}[a^* \neq \pi_\theta(s)]] \leq \epsilon,$$

and $R^{\max}$ is an upper-bound on the absolute value of the expected reward $R(s, a)$. Refer to Theorem 2.1 in Ross and Bagnell (2010) or Lemma 3 in Syed and Schapire (2010) for the proof of this result.

While such theoretical result appears to discredit the use of behavioral cloning for imitation learning, it is important to note that this framework is much simpler than the following frameworks; it consists only in using any off-the-shelf classifier or regressor. Thus, it is still useful in practice. However, it is important to note that the policy found with BC might lead to catastrophic errors, since an expert rarely makes mistakes and thus the pre-collected demonstrations contain few examples of how to recover from them (Pomerleau, 1989). In addition, BC does not generalize well to new environment since it only replicates the expert's policy in the current environment and thus doesn't learn the intrinsic motivation of the expert. Finally, BC requires lots of data if the agent or the environment is stochastic to cover all possible state-action pairs (Ho and Ermon, 2016).

4.3.2 Interactive Direct Policy Learning

IDPL corresponds to a reduction of imitation learning to a sequence of supervised learning

problems; note that BC is a special case of this framework with only one iteration. An iteration of this framework works as the following. First, a policy is used to generate some state trajectories. Second, the optimal action for each of these states is queried from the expert to generate state-action trajectories. Finally, a new candidate policy is obtained by minimizing a loss function over the state-action trajectories. Formally, IDPL is formulated as minimizing a sequence of loss functions

$$L_i(\pi) = \mathbb{E}_{(s,a^*)\sim D^i} L(a^*, \pi(s))$$

where D^i is the state-action distribution of iteration i (more on this below) (Yue and Le, 2018). The loss function used at the first iteration is generally the loss over the pre-collected demonstrations, i.e.,

$$L_1(\pi) = \mathbb{E}_{(s,a^*)\sim D_{\pi^*}} L(a^*, \pi(s)).$$

The optimization of the policy with respect to these losses can be done by two options, which ensures convergence to an optimal policy by slowly modifying the candidate policy. These options are (1) policy aggregation (e.g., SMILe (Ross and Bagnell, 2010)) and (2) data aggregation (e.g., DAgger (Ross, Gordon, and Bagnell, 2011)).

In policy aggregation, the state trajectories used to construct the state-action distribution D^i are generated using a policy that is a mixture of expert queries and policies of the preceding iterations. For example, this mixture in SMILe (Ross and Bagnell, 2010) puts exponentially decaying weights on the policies of the preceding iterations and the expert queries (considered the initial policy). This approach returns a stationary stochastic policy that never queries the expert.

In data aggregation, the state trajectories are generated using a policy that is a mixture of expert queries and the previous policy. Then, these state trajectories as well as all the previous trajectories are used with the corresponding expert's actions to construct the state-action distribution D^i. This approach returns a stationary deterministic policy.

As can be seen, this framework requires an interactive demonstrator which is a major limitation for our application. However, when this interactive demonstrator is available, this framework enables better theoretical guarantees and practical performances than BC (Ross, Gordon, and Bagnell, 2011; Ross and Bagnell, 2010), since the policy learns to recover from deviations of the expert's policy by exploring relevant states in an efficient manner. Finally, note that, as for BC, IDPL does not generalize well to new environment since it only learns to replicate an expert's policy in the current environment and not the expert's intrinsic motivation.

4.3.3 Inverse Reinforcement Learning

IRL consists in learning a reward function R such that the optimal policy under this reward function is the policy executed by the expert (Yue and Le, 2018), i.e., find R such that

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{(s,a)\sim D_\pi} R(s,a).$$

An iteration of this framework goes as the following. First, the reward function is updated. Second, the reinforcement learning problem is solved with this updated reward function to obtain a candidate policy. Finally, this candidate policy is compared with the expert policy to determine how to update the reward function.

The IRL approaches can be model-given or model-free. In the model-given case, the dynamics T is known and the reward function R is generally assumed to be linear. In addition, since the dynamics needs to be stored, this type of approach is generally limited to MDPs with small discrete state and action spaces. In the model-free case, the dynamics T is unknown, but there exists a simulator. Thus, since there is no need to store the dynamics, this model-free case allows large and continuous spaces where the reward function is often modeled using a derivable function (e.g., deep neural net).

The major advantage of IRL is that it accounts for transfer and generalizability by first learning the reward function (i.e., the intrinsic motivation of the expert (Ng and Russell, 2000)), and computing the policy only *a posteriori*. A major disadvantage of IRL is that this framework requires solving, fully or partially, a reinforcement learning problem within each iteration; thus, it requires a simulator or the dynamics, and it is more computationally intensive.

In this work, we focus on the model-given IRL problem since it appears easier and more sensible to derive the dynamics T than to construct a simulator with limited data. In addition, since we are interested in interpretable results (which we discuss in the next section), the simpler reward and policy functions of model-given IRL are more appropriate. Hence, we now refer only to model-given IRL whenever we discuss IRL. A formal definition of model-given IRL and an overview of the literature are now given.

Formally, model-given IRL consists in recovering a reward function R from a set of trajectories $\mathcal{D} = \{\zeta_1, \ldots, \zeta_M\}$ given a MDP without the reward function $\mathcal{M} \setminus R \triangleq \langle \mathcal{S}, \mathcal{A}, T, \gamma, \alpha \rangle$ (Ng and Russell, 2000). In this framework, these trajectories are assumed to be generated by executing an (unknown) optimal policy π^* with respect to the (unknown) reward function R where ζ_m is a state-action pairs sequence of length H_m, i.e., $\zeta_m = \{(s_1^m, a_1^m), \ldots, (s_{H_m}^m, a_{H_m}^m)\}$.

Ng and Russell (2000) proved that the IRL problem is ill-posed, i.e., there exists multiple reward functions R for which the observed policy π is optimal. The region defining these reward functions is given in the following condition.

Condition 1 (Reward Optimality Condition (Choi and Kim, 2011; Ng and Russell, 2000))**.** *Given $\mathcal{M} \setminus R$, policy π is optimal if and only if reward function $\mathbf{R}$ satisfies*

$$\left[\mathbf{I} - (\mathbf{I}^A - \gamma\mathbf{T})(\mathbf{I} - \gamma\mathbf{T}^\pi)^{-1}\mathbf{E}^\pi\right]\mathbf{R} \leq \mathbf{0},$$

where $\mathbf{I}$ denotes an identity matrix, $\mathbf{I}^A$ is an $|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|$ matrix constructed by stacking the $|\mathcal{S}| \times |\mathcal{S}|$ identity matrix $|\mathcal{A}|$ times, and $\mathbf{E}^\pi$ is an $|\mathcal{S}| \times |\mathcal{S}||\mathcal{A}|$ matrix with the $(s, (s', a'))$ element being 1 if $s = s'$ and $\pi(s') = a'$.

Proof. See Corollary 1 of Choi and Kim (2011) for proof. □

Since the IRL problem is ill-posed, multiple approaches were proposed in the literature to identify a unique reward function R among the region defined in Condition 1; these different approaches could be seen as some form of "regularization" over the region. For example, the approaches of Ng and Russell (2000), Abbeel and Ng (2004), and Ratliff, Bagnell, and Zinkevich (2006) search for a reward function where the value under the expert's policy (i.e., the policy creating the trajectories) is larger than or equal to the value under any other policy. In particular, the approach of Ng and Russell (2000) optimizes for a reward function that maximizes the expected gap between the values of the expert's policy and the other policies; the approach of Abbeel and Ng (2004) optimizes for a reward function that maximizes the gap between the values of the expert's policy and the second-best policy; and the maximum margin planning (MMP) approach of Ratliff, Bagnell, and Zinkevich (2006) optimizes for a reward function where the value under the expert's policy is larger than or equal to the value under any other policy by a predefined margin that depends on the state-action pairs. As another example, the method of Ziebart et al. (2008) uses the principle of maximum entropy to resolve the ambiguity over the distribution of trajectories, i.e., they assume that two trajectories are as probable if they have the same value and that trajectories with higher values are exponentially more preferred; using this likelihood, they identify with maximum likelihood the distribution over the trajectories that is parameterized with a linear reward function. Taking a different perspective, the multiplicative weights for apprenticeship learning (MWAL) method (Syed and Schapire, 2007) approaches IRL from a game-theoretic perspective, i.e., this method identifies a policy for which its value approaches the one of the expert or even surpass it for the worst-case reward function. Finally, in contrast to the previous non-Bayesian

approaches, Ramachandran and Amir (2007) introduced a Bayesian IRL model that outputs a distribution over the reward functions. This approach was later modified by Choi and Kim (2011) with the maximum a posteriori inference for Bayesian inverse reinforcement learning (MAP-BIRL) that recovers a single reward function with the maximum a posteriori estimation. In addition to being easier to implement than the approach of Ramachandran and Amir (2007), MAP-BIRL was shown to generalize many of the previous non-Bayesian IRL approaches (Choi and Kim, 2011).

In this work, we focus on the MAP-BIRL approach (Choi and Kim, 2011) and a variant of this approach proposed by Kim and Pineau (2016) since they subsume previous approaches and are easy to implement. Hence, we now describe in more details these two approaches.

The MAP-BIRL algorithm goes as follows. Assuming that (1) the agent is attempting to maximize the total accumulated rewards (i.e., the agent is not acting completely at random), (2) the policy used by the agent is stationary, (3) the initial belief about the expected rewards are i.i.d.[4], and (4) the likelihood of the set of trajectories is an independent exponential distribution, the prior and likelihood are given as

$$\Pr(\mathbf{R}) = \prod_{s\in\mathcal{S}, a\in\mathcal{A}} \Pr(\mathbf{R}(s,a))$$

$$\Pr(\mathcal{D} \mid \mathbf{R}) = \prod_{m=1}^{M}\prod_{h=1}^{H_m} \Pr(a_h^m \mid s_h^m, \mathbf{R}) = \prod_{m=1}^{M}\prod_{h=1}^{H_m} \frac{\exp(\beta Q^*(s_h^m, a_h^m; \mathbf{R}))}{\sum_{a\in\mathcal{A}} \exp(\beta Q^*(s_h^m, a; \mathbf{R}))}$$

where β is a parameter representing our confidence in the agent choosing the optimal action.[5]

The goal of MAP-BIRL is to identify the reward function that maximizes the log-posterior distribution $\mathbf{R}_{MAP} = \arg\max_{\mathbf{R}} [\log \Pr(\mathbf{R} \mid \mathcal{D})]$ where $\Pr(\mathbf{R} \mid \mathcal{D}) \propto \Pr(\mathcal{D} \mid \mathbf{R}) \Pr(\mathbf{R})$. This optimization (Choi and Kim, 2011) is done using a gradient method with the following update rule

$$\mathbf{R}_{i+1} \leftarrow \mathbf{R}_i + \Delta_i \nabla_{\mathbf{R}_i} \log \Pr(\mathbf{R}_i \mid \mathcal{D}) \tag{4.1}$$

where Δ_i is the learning rate at iteration i and $\nabla_{\mathbf{R}_i} \log \Pr(\mathbf{R}_i \mid \mathcal{D})$ is the gradient of the log-posterior with respect to $\mathbf{R}_i$.

Since the gradient with respect to $\mathbf{R}$ of the log-posterior is proportional, up to a constant that does not depend on $\mathbf{R}$, to the sum of the gradients of the log-prior and the

[4]Note that, in infinite-horizon MDPs, rewards are assumed to be identically distributed for tractability, and independent since the process is Markov.

[5]The parameter β is strictly greater than zero whenever we are confident that the agent is not acting completely at random.

log-likelihood, this log-posterior is equivalently computed as

$$\nabla_{\mathbf{R}_i} \log \Pr(\mathbf{R}_i \mid \mathcal{D}) \approx \nabla_{\mathbf{R}_i} \log \Pr(\mathcal{D} \mid \mathbf{R}_i) + \nabla_{\mathbf{R}_i} \log \Pr(\mathbf{R}_i) \tag{4.2}$$

by which the same reward $\mathbf{R}_{MAP}$ is obtained. Assuming a differentiable log-prior with known gradient $\nabla_{\mathbf{R}} \log \Pr(\mathbf{R})$, the only piece missing is the gradient of the log-likelihood $\nabla_{\mathbf{R}} \log \Pr(\mathcal{D} \mid \mathbf{R})$. The log-likelihood is given as

$$\log \Pr(\mathcal{D} \mid \mathbf{R}) = \sum_{m=1}^{M} \sum_{h=1}^{H_m} \left[\beta Q^*(s_h^m, a_h^m; \mathbf{R}) - \log \sum_{a \in \mathcal{A}} \exp(\beta Q^*(s_h^m, a; \mathbf{R})) \right], \tag{4.3}$$

while its gradient is

$$\nabla_{\mathbf{R}} \log \Pr(\mathcal{D} \mid \mathbf{R}) = \beta \sum_{m=1}^{M} \sum_{h=1}^{H_m} \left[\nabla_{\mathbf{R}} Q^*(s_h^m, a_h^m; \mathbf{R}) - \sum_{a \in \mathcal{A}} \psi(a, s_h^m; \mathbf{R}) \nabla_{\mathbf{R}} Q^*(s_h^m, a; \mathbf{R}) \right] \tag{4.4}$$

with $\psi(a, s; \mathbf{R}) = \frac{\exp(\beta Q^*(s,a;\mathbf{R}))}{\sum_{a' \in \mathcal{A}} \exp(\beta Q^*(s,a';\mathbf{R}))}$.

Given $\nabla_{\mathbf{R}} \mathbf{Q}^*(\mathbf{R})$ where $\mathbf{Q}^*(\mathbf{R})$ is an $|\mathcal{S}||\mathcal{A}|$-dimensional vector with the (s, a) element being $Q^*(s, a; \mathbf{R})$, it is possible to compute the previous gradient. Choi and Kim (2011) proved the following two properties for $\mathbf{Q}^*(\mathbf{R})$:

Lemma 3. *Each element of $\mathbf{Q}^*(\mathbf{R})$ is convex.*

Proof. See Theorem 2 of Choi and Kim (2011) for proof. □

Lemma 4. *Each element of $\mathbf{Q}^*(\mathbf{R})$ is differentiable almost everywhere. In particular, for $\mathbf{R} \in C(\pi)$ where $C(\pi)$ is the reward optimality region with respect to π, $\mathbf{Q}^*(\mathbf{R})$ is differentiable with $\nabla_{\mathbf{R}} \mathbf{Q}^*(\mathbf{R}) = \nabla_{\mathbf{R}} \mathbf{Q}^{\pi}(\mathbf{R}) = (\mathbf{I} - \gamma \mathbf{T} \mathbf{E}^{\pi})^{-1}$ strictly inside reward optimality regions and $\nabla_{\mathbf{R}} \mathbf{Q}^{\pi}(\mathbf{R})$ is a subgradient of $\mathbf{Q}^*(\mathbf{R})$ on the boundaries.*

Proof. See Theorem 3 of Choi and Kim (2011) for proof. □

Given that multiple rewards give the same optimal policy, Choi and Kim (2011) proposed Algorithm 4.1 to reduce the number of MDPs to solve and the number of gradients of the Q-value function to compute.[6] This algorithm uses Condition 1 (i.e., $\mathbf{H}^{\pi}\mathbf{R} \leq \mathbf{0}$) in order to reuse past results (i.e., $\pi, \nabla_{\mathbf{R}} \mathbf{Q}$) when possible.

If R is a linear parametric function (i.e., $\mathbf{R} = \mathbf{\Phi}\boldsymbol{\omega}$ with $\mathbf{\Phi} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times d}$) and that we want to optimize its weights $\boldsymbol{\omega}$, the previous results hold by replacing $\mathbf{R}$ by $\mathbf{\Phi}\boldsymbol{\omega}$. For example,

[6] It is easy to see in Lemma 4 that the computation of the gradient requires π to formulate $\mathbf{E}^{\pi}$ but does not require $\mathbf{R}$.

Algorithm 4.1: MAP-BIRL algorithm (Choi and Kim, 2011).

Input : $\mathcal{M} \setminus R$, trajectories $\mathcal{D}$, step-size sequence $\{\Delta_i\}$, number of iterations ν
Output: $\mathbf{R}$

Initialize $\mathbf{R}$
$\pi \leftarrow \text{solveMDP}(\mathbf{R})$ /* Compute the optimal policy */
$\mathbf{H}^\pi \leftarrow \text{computeRewardOptRgn}(\pi)$ /* Condition 1 */
$\nabla_{\mathbf{R}}\mathbf{Q} \leftarrow \text{computeQGradient}(\pi)$ /* Lemma 4 */
$\Pi \leftarrow \{\langle \pi, \mathbf{H}^\pi, \nabla_{\mathbf{R}}\mathbf{Q} \rangle\}$ /* Store these results */
for $i = 1$ **to** ν **do**
  $\nabla_{\mathbf{R}} \log \Pr(\mathbf{R} \mid \mathcal{D}) \leftarrow \text{computeLPGrad}(\mathbf{R}, \pi, \nabla_{\mathbf{R}}\mathbf{Q}, \mathcal{D})$ /* Equation 4.2 */
  $\mathbf{R}_{new} \leftarrow \mathbf{R} + \Delta_i \nabla_{\mathbf{R}} \log \Pr(\mathbf{R} \mid \mathcal{D})$ /* Equation 4.1 */
  if $\mathbf{H}^\pi \mathbf{R}_{new} > 0$ **then**
    /* If $\mathbf{R}_{new}$ and $\mathbf{R}$ don't lead to the same optimal policy π and Q-value function gradient $\nabla_{\mathbf{R}}\mathbf{Q}$, then search in the past results. */
    $\langle \pi, \mathbf{H}^\pi, \nabla_{\mathbf{R}}\mathbf{Q} \rangle \leftarrow \text{findRewardOptRgn}(\mathbf{R}_{new}, \Pi)$
    if $\text{isEmpty}(\langle \pi, \mathbf{H}^\pi, \nabla_{\mathbf{R}}\mathbf{Q} \rangle)$ **then**
      /* If no past reward optimality region $\mathbf{H}^\pi$ gives $\mathbf{H}^\pi \mathbf{R}_{new} \leq 0$, then compute the relevant quantities. */
      $\pi \leftarrow \text{solveMDP}(\mathbf{R}_{new})$ /* Compute the optimal policy */
      $\mathbf{H}^\pi \leftarrow \text{computeRewardOptRgn}(\pi)$ /* Condition 1 */
      $\nabla_{\mathbf{R}}\mathbf{Q} \leftarrow \text{computeQGradient}(\pi)$ /* Lemma 4 */
      $\Pi \leftarrow \Pi \cup \{\langle \pi, \mathbf{H}^\pi, \nabla_{\mathbf{R}}\mathbf{Q} \rangle\}$ /* Store these results */
  $\mathbf{R} \leftarrow \mathbf{R}_{new}$

the derivation of the Q-value function changes to $\nabla_{\boldsymbol{\omega}}\mathbf{Q}^*(\boldsymbol{\omega}) = (\mathbf{I} - \gamma \mathbf{T}\mathbf{E}^\pi)^{-1}\boldsymbol{\Phi}$. In this setting, even if the $\boldsymbol{\Phi}$ matrix is problematic for large state and action spaces, learning $\boldsymbol{\omega}$ instead of $\mathbf{R}$ does however allow to interpolate $R(s, a)$ for unseen states and/or actions.

Finally, to conclude on IRL, there exist approaches that have been proposed to regularize the weight vector $\boldsymbol{\omega}$, without having to specify a prior $\Pr(\boldsymbol{\omega})$, when R is a linear parametric function. This is the approach taken by Kim and Pineau (2016) that we focus on in this work. Their approach maximizes the L1-regularized log-likelihood

$$\boldsymbol{\omega}^* = \underset{\boldsymbol{\omega} \in \mathbb{R}^d}{\arg\max} \left[\log \Pr(\mathcal{D} \mid \boldsymbol{\omega}) - \lambda \|\boldsymbol{\omega}\|_1 \right]$$

where $\lambda > 0$ is the regularization parameter[7]. This allows them to obtain a sparse weight vector. In their approach, they reuse the same definition of the log-likelihood for a linear

[7]Note that if $\boldsymbol{\Phi}$ contains a bias term (which is generally the case) then the corresponding element of $\boldsymbol{\omega}$ is not regularized; this particularity is assumed everywhere within this study.

reward function as in Choi and Kim (2011).

4.4 Proposed Models

We now discuss the proposed models which are adaptations of existing models to our problem. The proposed models were selected for their capacity to (1) identify the relevant factors used in making the timing decisions, and (2) estimate the relative importance of these factors. While there are many different interpretable models which could have been selected for these tasks, we focus in this study on sparse linear models since we believe that this class is an adequate proxy to interpretability with respect to the goal of this study. Note that the literature on the adequacy of interpretable machine learning models relative to the tasks is still in its infancy (see, e.g., Doshi-Velez and Kim (2017)).

The proposed models consist of IRL and BC models; we omit IDPL approaches since we don't have access to the interactive demonstrator required by these approaches. The IRL model, referred to as semi maximum likelihood inverse reinforcement learning (SMLIRL), recovers a linear reward function from a set of trajectories that contains static, dynamic and time dependency information based on the model of Kim and Pineau (2016) and SMDP. In particular, SMLIRL maximizes the L1-regularized log-likelihood as in Kim and Pineau (2016) for a discrete version of SMDP. On the one hand, we select the approach of Kim and Pineau (2016) since we want to recover a small subset of relevant features. On the other hand, we use SMDP with a discrete state and action spaces due to the challenges of working with continuous spaces.

The BC models consist of off-the-shelf supervised models, i.e., a multiclass logistic regression model and least absolute shrinkage and selection operator (LASSO). These models offer another perspective on the behavior of the physicians and can be seen as myopic variants of the IRL model.

In this section, we first describe the components of the SMLIRL model, the formulation of the associated Bellman equations in matrix notation and the SMLIRL algorithm. Then, we analyze the myopic version of SMLIRL and show that it is equivalent to BC approaches that we plan to use. Next, we present our procedures for the discretization of the state and action spaces, and our model selection procedure. Finally, we discuss how the proposed models differ when they are used to characterize instead of reproducing a behavior.

4.4.1 Components and Notation of SMLIRL Model

We now describe the different components of the semi maximum likelihood inverse reinforcement learning (SMLIRL) model and the formulation of the associated Bellman equations in matrix notation.

State

The state of the system is defined as $s \in \mathcal{S}$ and can be decomposed as $s = (b, g, t) \in \mathcal{B} \times \mathcal{G} \times \mathcal{T}$ where b is the static information on the patient, physician and clinic, g is the dynamic information on the patient, physician and clinic, and t is the time since the first appointment. We assume that the sets $\mathcal{B}$, $\mathcal{G}$ and $\mathcal{T}$ are all discrete sets. In addition, note that the time dependency t is added to the state in order to capture whether a physician spaces out the appointments after some time.

Action

The action set, assumed discrete, consists of only one action, i.e., the delay before the next appointment. Thus, we denote the action as $a \in \mathcal{A} = \{a_1, a_2, \ldots, a_k\}$.

Observations

The set of observations is defined as $\mathcal{D} = \{\zeta_1, \ldots, \zeta_M\}$ with trajectories $\zeta_m = \{b^m, (g_1^m, t_1^m, a_1^m), \ldots, (g_{H_m}^m, t_{H_m}^m, a_{H_m}^m), (g_{H_m+1}^m, t_{H_m+1}^m)\}$ where $b^m \in \mathcal{B}$ is the static information for the trajectory m, $g_h^m \in \mathcal{G}$ is the dynamic information for the trajectory m at decision epoch h, $t_h^m \in \mathcal{T}$ is the time since the initial appointment for the trajectory m at decision epoch h, and $a_h^m = t_{h+1}^m - t_h^m$ is the observed action (i.e., delay before next appointment) for the trajectory m at decision epoch h. Note that the last decision epoch of the sequence (i.e., $H_m + 1$) does not contain an action as we do not observe this value.

Transition Functions

We now define the functions $F(s, a, \sigma)$, $T(s, a, s')$ and $P(s, a, s')$ that are needed to define this SMDP. The function $F(s, a, \sigma)$ is defined as

$$F(s, a, \sigma) = \begin{cases} 0 & \text{if } \sigma < a \\ 1 & \text{if } \sigma \geq a, \end{cases}$$

and its time differential is $F(s, a, \mathrm{d}\sigma) = \delta(\sigma - a)\,\mathrm{d}\sigma$ where $\delta(\cdot)$ is the Dirac delta function.

The function $T(s, a, s')$ is defined as

$$T((b, g, t), a, (b', g', t')) = \begin{cases} W_b(g, a, g') & \text{if } t' = t + a \text{ and } b' = b \\ 0 & \text{otherwise} \end{cases}$$

where $s = (b, g, t)$ and $W_b(g, a, g') = \Pr(g_{\delta+1} = g' \mid g_\delta = g, a_\delta = a, b_\delta = b)$. It is important to note here that, while the dynamic information g transition stochastically to g', the static information b stays the same and the time t since the initial appointment transition deterministically to $t' = t + a$. Also, note that we assume that W_b does not depend on t nor t'.

Finally, the function $P(s, a, s')$ is defined as

$$\begin{aligned} P(s, a, s') &= T(s, a, s') \int_0^\infty e^{-\alpha\sigma} F(s, a, \mathrm{d}\sigma) \\ &= T(s, a, s') \int_0^\infty e^{-\alpha\sigma} \delta(\sigma - a)\,\mathrm{d}\sigma \\ &= e^{-\alpha a} T(s, a, s') \end{aligned}$$

since $\int_0^\infty f(x)\delta(x - a)\,\mathrm{d}x = f(a)$ when $a \in [0, \infty)$. When expanding the state s, it reduces to

$$P((b, g, t), a, (b', g', t')) = \begin{cases} e^{-\alpha a} W_b(g, a, g') & \text{if } t' = t + a \text{ and } b' = b \\ 0 & \text{otherwise.} \end{cases}$$

Note that in this model, we assume that $\alpha > 0$ is given. If it is not, then we have to search for the best value of α from a predictive performance point of view like for the other parameters.

Reward Function

The reward function is defined as $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ where $R(s, a)$ is the reward obtained after choosing action a in state $s = (b, g, t)$. In this model, it is assumed that the reward function $R(s, a)$ is composed only of the lump sum reward $r(s, a)$, i.e., the continuous reward $c(s, a)$ is zero. In addition, it is assumed that this reward function is a linear function of a set of features, i.e., $R(s, a) = \sum_{i=1}^{d} \omega_i \phi_i(s, a)$ with $\{\omega_i\}_{i=1}^d$ denoting the parameters and $\{\phi_i\}_{i=1}^d$ denoting the feature transformation functions $\phi_i : \mathcal{S} \times \mathcal{A} \to [0, 1]$. Inferring this reward function (i.e., the parameters $\{\omega_i\}_{i=1}^d$) is the goal of this model in order to understand the important features (i.e., the features with $\omega_i \neq 0$) and their importance (i.e., the magnitude

of $|\omega_i|$ relative to $\min_{j=1,\dots,d} |\omega_j|$). This reward function is written in matrix notation as $\mathbf{R} = \mathbf{\Phi}\boldsymbol{\omega}$ where $\mathbf{\Phi}$ is an $|\mathcal{S}||\mathcal{A}| \times d$ matrix with element $((s,a),i)$ equal to $\phi_i(s,a)$, and $\boldsymbol{\omega}$ is an d-dimensional vector.

In this work, we assume that the feature transformation functions are given by

$$\boldsymbol{\phi}(s,a) = \begin{bmatrix} \mathbb{1}[a = a_1]\boldsymbol{\phi}(s) \\ \vdots \\ \mathbb{1}[a = a_k]\boldsymbol{\phi}(s) \end{bmatrix} \tag{4.5}$$

where $\boldsymbol{\phi}(s,a)^\top$ is a row of $\mathbf{\Phi}$ and $\boldsymbol{\phi}(s)$ is composed of the features characterizing the state s and a bias term.[8]

Matrix Notation

For the previously defined model components, we denote the Bellman equation and the Q-value function in the following matrix notation:

$$\mathbf{V}^\pi = \mathbf{\Phi}^\pi\boldsymbol{\omega} + \mathbf{P}^\pi\mathbf{V}^\pi$$
$$\mathbf{Q}_a^\pi = \mathbf{\Phi}^a\boldsymbol{\omega} + \mathbf{P}^a\mathbf{V}^\pi$$

where $s = (b,g,t)$, $\mathbf{V}^\pi$ is an $|\mathcal{S}|$-dimensional vector with element s equal to $V^\pi(s)$, $\mathbf{\Phi}^\pi$ is an $|\mathcal{S}| \times d$ matrix with row s equal to row $(s,\pi(s))$ in $\mathbf{\Phi}$, $\mathbf{P}^\pi$ is an $|\mathcal{S}| \times |\mathcal{S}|$ matrix with element (s,s') equal to $P(s,\pi(s),s')$, $\mathbf{Q}_a^\pi$ is an $|\mathcal{S}|$-dimensional vector with element s equal to $Q^\pi(s,a)$, $\mathbf{\Phi}^a$ is an $|\mathcal{S}| \times d$ matrix with row s equal to row (s,a) in $\mathbf{\Phi}$, and $\mathbf{P}^a$ is an $|\mathcal{S}| \times |\mathcal{S}|$ matrix with element (s,s') equal to $P(s,a,s')$.

4.4.2 SMLIRL Algorithm

We now present the proposed semi maximum likelihood inverse reinforcement learning (SMLIRL) model that extends the approach of Kim and Pineau (2016) with the SMDP framework to take into account decision epochs of variable lengths. We describe it with the equivalent subroutines to Algorithm 4.1.

[8]Note that we use ϕ and $\boldsymbol{\phi}$ to denote several different components. However, the meaning should be clear from the arguments and indexes of this component, and whether this component is in bold.

Computing the Optimal Policy: solveMDP($\boldsymbol{\omega}$)

Using value iteration, the proposed model is solved as

$$V^*(s) = \max_{a \in \mathcal{A}} Q^*(s, a)$$

where $s = (b, g, t)$ and

$$Q^*(s, a) = \sum_{i=1}^{d} \omega_i \phi_i(s, a) + \sum_{s' \in \mathcal{S}} P(s, a, s') V^*(s').$$

Then, the optimal policy is obtained with

$$\pi^*(s) = \arg\max_{a \in \mathcal{A}} Q^*(s, a).$$

This subroutine replaces the subroutine solveMDP($\mathbf{R}$) in Algorithm 4.1.

Computing the Reward Optimality Region: computeRewardOptRgn(π)

For the proposed model, Condition 1 is rewritten as the following:

Condition 2 (Reward Optimality Condition for Proposed Model)**.** *Given $\langle \mathcal{S}, \mathcal{A}, P, \boldsymbol{\Phi} \rangle$, policy π is optimal if and only if reward parameters $\boldsymbol{\omega}$ satisfy*

$$\left[\boldsymbol{\Phi} - (\mathbf{I}^A - \mathbf{P})(\mathbf{I} - \mathbf{P}^\pi)^{-1}\boldsymbol{\Phi}^\pi\right] \boldsymbol{\omega} \leq \mathbf{0}$$

where $\mathbf{I}^A$ is an $|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|$ matrix constructed by stacking the $|\mathcal{S}| \times |\mathcal{S}|$ identity matrix $|\mathcal{A}|$ times, $\mathbf{P}$ is an $|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|$ matrix with element $((s, a), s')$ equal to $P(s, a, s')$, and $\mathbf{I}$ is an identity matrix.

Proof.

Policy π is optimal

$$\begin{aligned}
&\Leftrightarrow \mathbf{Q}_a^\pi \leq \mathbf{V}^\pi, && \forall a \in \mathcal{A} \\
&\Leftrightarrow \boldsymbol{\Phi}^a\boldsymbol{\omega} + \mathbf{P}^a\mathbf{V}^\pi \leq \boldsymbol{\Phi}^\pi\boldsymbol{\omega} + \mathbf{P}^\pi\mathbf{V}^\pi, && \forall a \in \mathcal{A} \\
&\Leftrightarrow \boldsymbol{\Phi}^a\boldsymbol{\omega} + \mathbf{P}^a(\mathbf{I} - \mathbf{P}^\pi)^{-1}\boldsymbol{\Phi}^\pi\boldsymbol{\omega} \leq \boldsymbol{\Phi}^\pi\boldsymbol{\omega} + \mathbf{P}^\pi(\mathbf{I} - \mathbf{P}^\pi)^{-1}\boldsymbol{\Phi}^\pi\boldsymbol{\omega}, && \forall a \in \mathcal{A} \\
&\Leftrightarrow \boldsymbol{\Phi}^a\boldsymbol{\omega} - (\mathbf{I} - \mathbf{P}^a)(\mathbf{I} - \mathbf{P}^\pi)^{-1}\boldsymbol{\Phi}^\pi\boldsymbol{\omega} \leq \boldsymbol{\Phi}^\pi\boldsymbol{\omega} - (\mathbf{I} - \mathbf{P}^\pi)(\mathbf{I} - \mathbf{P}^\pi)^{-1}\boldsymbol{\Phi}^\pi\boldsymbol{\omega}, && \forall a \in \mathcal{A} \\
&\Leftrightarrow \left[\boldsymbol{\Phi}^a - (\mathbf{I} - \mathbf{P}^a)(\mathbf{I} - \mathbf{P}^\pi)^{-1}\boldsymbol{\Phi}^\pi\right] \boldsymbol{\omega} \leq \mathbf{0}, && \forall a \in \mathcal{A}
\end{aligned}$$

The third equivalence holds by $\mathbf{V}^\pi = (\mathbf{I} - \mathbf{P}^\pi)^{-1}\mathbf{\Phi}^\pi\boldsymbol{\omega}$. The fifth equivalence holds because the right-hand side is $\mathbf{0}$. Stacking up the last equivalence for all $a \in \mathcal{A}$ gives the condition. □

This subroutine replaces the subroutine computeRewardOptRgn(π) in Algorithm 4.1.

Computing the Q-Value Function Gradient: computeQGradient(π)

Similarly to Lemma 4, the Q-value function gradient with respect to $\boldsymbol{\omega}$ is computed as

$$\nabla\mathbf{Q}^\pi = (\mathbf{I} - \mathbf{P}\mathbf{E}^\pi)^{-1}\mathbf{\Phi}$$

where $\nabla\mathbf{Q}^\pi$ is a $|\mathcal{S}||\mathcal{A}| \times d$ matrix with element $((s,a),i)$ equal to $\frac{\partial Q^\pi(s,a)}{\partial\omega_i}$, since $\mathbf{Q}^\pi = \mathbf{\Phi}\boldsymbol{\omega} + \mathbf{P}\mathbf{E}^\pi\mathbf{Q}^\pi$ where $\mathbf{E}^\pi$ is an $|\mathcal{S}| \times |\mathcal{S}||\mathcal{A}|$ matrix with element $(s,(s',a'))$ equal to 1 if $s = s'$ and $\pi(s') = a'$, and $\mathbf{Q}^\pi$ is an $|\mathcal{S}||\mathcal{A}|$-dimensional vector with element (s,a) equal to $Q^\pi(s,a)$. This subroutine replaces the subroutine computeQGradient(π) in Algorithm 4.1.

Computing the Log-Likelihood and its Derivatives: computeLLDeriv($\boldsymbol{\omega}, \pi, \mathbf{Q}, \nabla_{\boldsymbol{\omega}}\mathbf{Q}, \mathcal{D}$)

Let s_h^m denote the state (b^m, g_h^m, t_h^m). Then, similarly to Equation 4.3, the log-likelihood is

$$\begin{aligned} L(\boldsymbol{\omega}) &\triangleq \log \Pr(\mathcal{D} \mid \boldsymbol{\omega}) \\ &= \sum_{m=1}^{M}\sum_{h=1}^{H_m}\left[\beta Q^*(s_h^m, a_h^m; \boldsymbol{\omega}) - \log\sum_{a\in\mathcal{A}}\exp(\beta Q^*(s_h^m, a; \boldsymbol{\omega}))\right]. \end{aligned}$$

Similarly to Equation 4.4, the gradient of this log-likelihood with respect to $\boldsymbol{\omega}$ is

$$\nabla L(\boldsymbol{\omega}) = \beta\sum_{m=1}^{M}\sum_{h=1}^{H_m}\Big[\nabla Q^*(s_h^m, a_h^m; \boldsymbol{\omega}) - \sum_{a\in\mathcal{A}}\psi(s_h^m, a; \boldsymbol{\omega})\nabla Q^*(s_h^m, a; \boldsymbol{\omega})\Big],$$

where $\nabla L(\boldsymbol{\omega})$ is a d-dimensional vector with element i equal to $\frac{\partial L(\boldsymbol{\omega})}{\partial\omega_i}$ and $\psi(s_h^m, a; \boldsymbol{\omega}) = \frac{\exp\left(\beta Q^*(s_h^m, a; \boldsymbol{\omega})\right)}{\sum_{a'\in\mathcal{A}}\exp\left(\beta Q^*(s_h^m, a'; \boldsymbol{\omega})\right)}$.

Finally, the hessian of this log-likelihood with respect to $\boldsymbol{\omega}$ is

$$\nabla^2 L(\boldsymbol{\omega}) = \beta^2 \sum_{m=1}^{M} \sum_{h=1}^{H_m} \left[\left(\sum_{a \in \mathcal{A}} \psi(s_h^m, a; \boldsymbol{\omega}) \nabla Q^*(s_h^m, a; \boldsymbol{\omega}) \right) \left(\sum_{a' \in \mathcal{A}} \psi(s_h^m, a'; \boldsymbol{\omega}) \nabla Q^*(s_h^m, a'; \boldsymbol{\omega}) \right)^\top \right.$$
$$\left. - \sum_{a \in \mathcal{A}} \psi(s_h^m, a; \boldsymbol{\omega}) \left(\nabla Q^*(s_h^m, a; \boldsymbol{\omega}) \right) \left(\nabla Q^*(s_h^m, a; \boldsymbol{\omega}) \right)^\top \right]$$

where $\nabla^2 L(\boldsymbol{\omega})$ is a $d \times d$ matrix with element (i, j) equal to $\frac{\partial^2 L(\boldsymbol{\omega})}{\partial \omega_i \partial \omega_j}$.

This subroutine replaces the subroutine computeLPGrad$(\mathbf{R}, \pi, \nabla_{\mathbf{R}} \mathbf{Q}, \mathcal{D})$ in Algorithm 4.1.

Minimizing the L1-Regularized Negative Log-Likelihood

In this proposed model, the goal is to minimize the L1-regularized negative log-likelihood (as in Kim and Pineau (2016))

$$\boldsymbol{\omega}^* = \underset{\boldsymbol{\omega} \in \mathbb{R}^d}{\arg\min} \left[-L(\boldsymbol{\omega}) + \lambda \|\boldsymbol{\omega}\|_1 \right] \tag{4.6}$$

where $\|\cdot\|_1$ is the L1-norm and λ is the regularization parameter.

However, we cannot use the standard gradient descent approach (e.g., Equation 4.1) to optimize this objective since the L1-norm is not differentiable at zero. Fortunately, there exists approaches to minimize general loss function with L1-regularization (e.g., Schmidt (2010)). This optimization replaces the update rule in Algorithm 4.1.

4.4.3 Myopic Model Analysis

We now analyze the myopic variant of the SMLIRL model and show that this variant is equivalent to BC. A myopic model considers only the immediate rewards, i.e., it consists of the model obtained in the limit $\alpha \to \infty$; hence, it is an interesting approach since it does not need to model the discounted dynamics P. In this model, the Q-value function reduces to $Q(s, a) = \sum_{i=1}^{d} \omega_i \phi_i(s, a)$. Hence, the gradient of the Q-value function is given by $\nabla \mathbf{Q} = \boldsymbol{\Phi}$ and it does not depend on the policy. So, it is easy to see that this model is much easier to optimize since we only need to minimize the L1-regularized negative log-likelihood without having to solve for the optimal policy as a subroutine at each iteration. In addition, this model is equivalent to a classification model over the state-action pairs as shown in Proposition 1; thus, this myopic model is equivalent to a BC approach.

Proposition 1. *Let $\boldsymbol{\phi}(s^{(i)}, a_j)$ be defined as in Equation 4.5, and let the features transformations $\boldsymbol{\phi}(s^{(i)})$ and $\boldsymbol{\phi}(x^{(i)})$ contain a bias term and be equivalent. Then, the myopic model loss function (without regularization)*

$$J(\boldsymbol{\omega}) = -\sum_{i=1}^{N}\sum_{j=1}^{k} \mathbb{1}[a^{(i)} = a_j] \log \frac{\exp\big(\beta \boldsymbol{\omega}^\top \boldsymbol{\phi}(s^{(i)}, a_j)\big)}{\sum_{l=1}^{k} \exp(\beta \boldsymbol{\omega}^\top \boldsymbol{\phi}(s^{(i)}, a_l))} \tag{4.7}$$

for the trajectory $\{(s^{(1)}, a^{(1)}), \ldots, (s^{(N)}, a^{(N)})\}$ and action set $a^{(i)} \in \{a_1, \ldots, a_k\}$ is equivalent to the loss function of the multiclass logistic regression (Bishop, 2006)

$$J(\boldsymbol{\Theta}) = -\sum_{i=1}^{N}\sum_{j=1}^{k} \mathbb{1}[y^{(i)} = y_j] \log \frac{\exp\big(\boldsymbol{\theta}_j^\top \boldsymbol{\phi}(x^{(i)})\big)}{\sum_{l=1}^{k} \exp\big(\boldsymbol{\theta}_l^\top \boldsymbol{\phi}(x^{(i)})\big)} \tag{4.8}$$

for the training set $\{(x^{(1)}, y^{(1)}), \ldots, (x^{(N)}, y^{(N)})\}$ and the class labels $y^{(i)} \in \{y_1, \ldots, y_k\}$.

Proof. Since $\boldsymbol{\phi}(s^{(i)})$ contains a bias term, β is redundant and can be omitted. Then it is easy to see that Equation 4.7 with Equation 4.5 is equivalent to Equation 4.8 since the corresponding kd-dimensional vector $\boldsymbol{\omega}$ is the vectorization of the $k \times d$ matrix $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k]^\top$ with all rows set to zero except row j. □

We now provide two remarks regarding this myopic model.

Remark 2. *Because of this equivalence between Equations 4.8 and 4.7, we have that the myopic case cannot be solved in closed-form in general.*

Remark 3. *Under the features in Equation 4.5, the myopic case is overparameterized. Thus, even though it is convex, it contains multiple minimizers. Fortunately this overparametrization is easily addressed with regularization.*

Finally, since the myopic case does not require to solve for the optimal policy as a subroutine and is equivalent to a BC model, it can also be addressed without discretizing the action values. Hence, we propose to use LASSO (i.e., L1-regularized regression) (Hastie, Tibshirani, and Friedman, 2009) as well as the previously described multiclass logistic regression (MLR) model to recover relevant features and their relative importance.

4.4.4 Discretization of the Observations

We now describe several discretization procedures that are used for the SMLIRL model and can be used for the BC models. Let the set of raw extracted observations be denoted by $\bar{\mathcal{D}} =$

$\{\bar{\zeta}_1, \ldots, \bar{\zeta}_M\}$ with trajectories $\bar{\zeta}_m = \{\bar{b}^m, (\bar{g}_1^m, \bar{t}_1^m, \bar{a}_1^m), \ldots, (\bar{g}_{H_m}^m, \bar{t}_{H_m}^m, \bar{a}_{H_m}^m), (\bar{g}_{H_m+1}^m, \bar{t}_{H_m+1}^m)\}$. Since the raw values $\bar{b}^m, \bar{g}_h^m, \bar{t}_h^m, \bar{a}_h^m$ may constitute large or uncountable sets, it may be required (e.g., to model the dynamics) or preferable (e.g., for computational performance) to discretize these values (or discretize them further) for the proposed SMLIRL and BC models.

In particular, since the action a and the time since the initial appointment t are real numbers, these values need to be discretized for the SMLIRL model. However, noting that t is linked to a through $t_{\delta+1} = t_\delta + a_\delta$, the discrete set $\mathcal{T}$ arises directly from the discrete set $\mathcal{A}$. Thus, it is only necessary to define the set $\mathcal{A}$ to obtain the set $\mathcal{T}$. It is important to note that the size of the set $\mathcal{A}$ shouldn't be too big to limit the size of the set $\mathcal{T}$; see Proposition 2 for an upper bound on the size of the set $\mathcal{T}$ given the size of the set $\mathcal{A}$. For example, according to Proposition 2, for $k = 5$ and $H = 10$, the upper bound is $\frac{15!}{10!\,5!} = 3003$.

Proposition 2. *For a particular discrete action set $\mathcal{A}$, the size of the associated discrete time set $\mathcal{T}$ is upper bounded by $\binom{k+H}{H} = \frac{(k+H)!}{H!\,k!}$ where $k = |\mathcal{A}|$ and $H = \max_{m\in\{1,\ldots,M\}} H_m$.*

Proof. There are $\binom{k+H}{H}$ ways to choose from 0 to H actions from a set of k actions if repetitions are allowed. Since different supersets of 0 to H actions can sum to the same value, this result is an upper bound. □

There exist a variety of discretization approaches that can be classified as global vs. local, supervised vs. unsupervised, and static vs. dynamic (Dougherty, Kohavi, and Sahami, 1995). Global discretization methods apply the same discretization procedure on the full instance space while local discretization methods apply different discretization procedures to the different regions of the instance space. Supervised discretization methods use additional information, such as the instance labels in supervised learning, to discretize the instance space while unsupervised discretization methods do not use such information. Finally, static discretization methods are applied at the outset of the learning algorithm while dynamic discretization methods are used within the learning algorithm.

In this work, we focus on global, unsupervised and static discretization approaches. In particular, we test how equal width, equal frequency and k-means discretization approaches compare for the proposed models. Note that in past IRL models (e.g., Abbeel and Ng (2004), Choi and Kim (2011), and Kim and Pineau (2016)), discretization was generally done according to the equal width discretization method or heuristically without any considerations to other potential discretization approaches. Yet, different discretization approaches can lead to different intervals and discretized values. Thus, the selection of the discretization approach can affect the predictive accuracy of the SMLIRL and BC models;

an example of the equal width, equal frequency and k-means discretization approaches applied to the times between consecutive appointments (i.e., the action of our application) for three intervals is given in Figure 4.1.

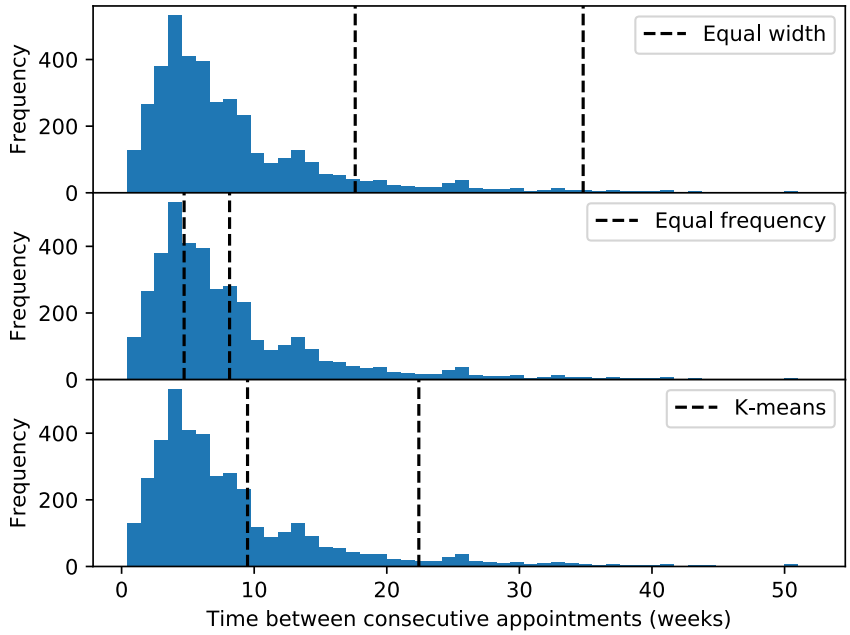

FIGURE 4.1: Equal width, equal frequency and k-means discretization of the times between consecutive appointments in three intervals.

We now describe the equal width, equal frequency and k-means discretization approaches for the discretization of the raw action values $\bar{a}_h^m$. The same methods can also be used to discretize the raw state values $\bar{b}^m$ and $\bar{g}_h^m$. Note that the raw state values $\bar{b}^m$ and $\bar{g}_h^m$ can be discretized within each dimension or across their dimensions.

Equal Width Discretization

Equal width discretization (Dougherty, Kohavi, and Sahami, 1995) splits the interval of the raw action values $\bar{a}_h^m$ into k intervals of equal width where the interval boundaries are given by $\bar{a}_{min} + i\xi$ for $i = 0, \ldots, k$ with $\xi = (\bar{a}_{max} - \bar{a}_{min})/k$, $\bar{a}_{max} = \max_{m,h} \bar{a}_h^m$, and $\bar{a}_{min} = \min_{m,h} \bar{a}_h^m$. Yet, since these intervals are determined using only the trajectories in the training set (see Section 4.4.5), some action values in the validation or testing set might be smaller than the left-boundary of the first interval or larger than the right-boundary of the last interval. Thus, each raw value $\bar{a}_h^m$ is replaced by the *closest* interval center value, where the interval center value is defined as the mean of the interval boundary values, to obtain the discretized values a_h^m.

Equal Frequency Discretization

Equal frequency discretization (Dougherty, Kohavi, and Sahami, 1995) divides the sorted raw action values $\bar{a}_h^m$ into k intervals so that each interval contains approximately the same number of instance, i.e., each interval contains approximately N/k instances where $N = \sum_{m=1}^{M} H_m$. Similarly to the equal width discretization method, each raw value $\bar{a}_h^m$ is then replaced by the closest interval center value to obtain the discretized values a_h^m.

K-Means Discretization

It is also possible to use a clustering algorithm, such as k-means clustering (Lloyd, 1982), to discretize data. In this case, since k-means clustering identify k clusters that minimize the average Euclidean distance from the cluster's observations to the centroid (i.e., the mean of the cluster's observations), it is coherent to replace each raw value $\bar{a}_h^m$ by the label or centroid of the corresponding cluster to obtain the discretized values a_h^m. Note that in this work, we use the algorithm k-means++ (Arthur and Vassilvitskii, 2007) which improves the initialization step of the k-means clustering algorithm.

4.4.5 Model Selection

Because of the different parameters and subroutines used (e.g., size of action set k, continuous-time discounting rate α, regularization parameter λ, discretization procedure) within the proposed approaches, it is required to have a way to validate which choices are the best. For this task, we propose to use cross-validation.

Since we only have access to a batch of data and cannot run the models online, we evaluate the models' results by comparing the estimated optimal policy with the observed policy over one decision epoch at a time; we cannot evaluate the full trajectory defined by the estimated optimal policy since we cannot query the expert for this trajectory.

In addition, since our state definition may not capture all the dependencies between the different decision epochs, we construct the training, validation and testing sets by splitting randomly the M trajectories ζ_m from the data $\mathcal{D}$. Thus, no complete nor partial trajectories contained in the validation and testing sets have been used during the training.

The validation procedure for a model that requires discretized state and action sets goes as follows:

1. With the knowledge elicited from the semi-structured interviews, define the variables that should be part of the state b and g, and define the features matrix Φ.

2. Extract the relevant data to obtain the raw trajectories $\bar{\mathcal{D}}$, and the raw sets $\bar{\mathcal{B}}, \bar{\mathcal{G}}, \bar{\mathcal{T}}$ and $\bar{\mathcal{A}}$.

3. Leave out a number of trajectories from $\bar{\mathcal{D}}$ for the test set such that $\bar{\mathcal{D}} = \bar{\mathcal{D}}_{cv} \cup \bar{\mathcal{D}}_{test}$ with $\bar{\mathcal{D}}_{cv} \cap \bar{\mathcal{D}}_{test} = \emptyset$.

4. For each fold of the cross-validation procedure, split the trajectories $\bar{\mathcal{D}}_{cv}$ randomly across a training set $\bar{\mathcal{D}}_{train}$ and a validation set $\bar{\mathcal{D}}_{val}$.

   (a) For each potential models (e.g., selected model parameters, discretization procedure), do the following:

      i. Discretize the trajectories in the training set $\bar{\mathcal{D}}_{train}$ to obtain $\mathcal{D}_{train}$.

      ii. Train the model on the training set $\mathcal{D}_{train}$.

      iii. Discretize the observed states in the validation set $\bar{\mathcal{D}}_{val}$ using the same state set as for the training set $\mathcal{D}_{train}$.

      iv. Evaluate the model expected error using $\sum_i (\hat{\pi}(s_i) - \bar{a}_i)^2$ where $\hat{\pi}$ is the estimated optimal policy, s_i is the discretized state from the validation set and $\bar{a}_i$ is a raw observed action from the validation set.

5. Identify the best model according to the expected validation error.

6. Train this best model on the full discretized cross-validation data $\mathcal{D}_{cv}$.

7. Discretize the observed states in the test set $\bar{\mathcal{D}}_{test}$ using the same state set as for the previous step.

8. Evaluate the best model expected error $\sum_i (\hat{\pi}(s_i) - \bar{a}_i)^2$ on the test set and report the results.

For a model that can take raw values as input, the validation procedure simply skips over the discretization steps and always use the raw values.

4.4.6 From Imitation to Understanding

While the proposed SMLIRL and BC (i.e., multiclass logistic regression and LASSO) approaches all predict an action given a state, their underlying algorithm works quite differently. Thus, when trying to explain a behavior (i.e., the goal of this study), each model provides a different perspective. We now discuss these differences and how they provide different perspectives.

First, SMLIRL fits a reward function and thus returns a function that isolates the intrinsic motivation/goal of the expert, i.e., it returns a function in which the discounting and environmental effects are removed. With this model, we obtain information regarding how one action is preferable to another one in a particular state. In contrast, the BC models fit a policy function and thus provide a description of how the actions are taken under these effects. With these models, we just obtain information regarding which action is taken under a particular state.

Second, in SMLIRL, the linearity constraint is on the reward function and not on the policy function; in fact, the policy function in SMLIRL is a table mapping a discrete state to a discrete action. In contrast, the linearity constraint within multiclass logistic regression and LASSO is on the policy function; in particular, in the case of multiclass logistic regression there is one linear model per discretized action.

Third, the approaches use discretization to a different degree: SMLIRL discretizes the states and actions, multiclass logistic regression discretizes only the actions and LASSO does not discretize.

In the end, these differences lead to results that have different meanings. Within SMLIRL, the nonzero weights identify the relevant discretized state-action pairs within the reward function, i.e., the combination of states and actions that are tried to be avoided or sought. Within multiclass logistic regression, the nonzero weights identify the relevant state variables that trigger a particular discretized action. Within LASSO, the nonzero weights identify the relevant state variables that make the action increase or decrease.

Finally, these differences in the models should also result in different models' capacities and predictive performances. It is however hard to compare these models' capacities analytically since these models' classes are so distinct. In addition, it is not possible to compare these models' predictive performances analytically since they are highly data dependent. For example, while it appears much more natural to use LASSO with the raw targets than multiclass logistic regression with the discretized targets for a prediction task, some applications have shown that classification with a discretized target is sometimes superior to regression with a raw target (e.g., Oord, Kalchbrenner, and Kavukcuoglu (2016), Bogucki (2016), and Kaggle Team (2015)).

4.5 Application

As discussed previously, setting the frequency of appointments for patients suffering from treatment-resistant depression (TRD) is an important decision; yet, the guidelines

are unclear with respect to this decision and it appears to the best of our knowledge that this decision has not been studied in the literature. Hence, later in this section, we use our proposed two-stage framework to better understand how this decision is made. In particular, we apply our framework to a data set collected at the depressive and suicide disorders program (DSDP) of the Douglas Mental Health University Institute in Montreal, which is an outpatient clinic treating patients suffering from TRD. However, prior to the framework, we provide some context to the application by describing the process of an appointment and the collected data set.

At the DSDP, an appointment with a psychiatrist generally consists in the four following steps:

1. The patient answers four computerized tests that evaluates his health state (i.e., quality of life, depression, side effects and suicide ideation).

2. The patient meets and discusses with the psychiatrist.

3. The psychiatrist makes treatment modifications if necessary.

4. The psychiatrist decides on the time of the next appointment and this time is generally respected.

Note however that the process at the initial appointment with the psychiatrist differs since a patient passes a set of baseline tests prior to this visit instead of the computerized tests. In addition, on some rare occasions, the process might differ due to a patient that does not pass the computerized tests (e.g., when a patient arrives late at an appointment) or that is seen earlier than planned (e.g., expedited appointment due to too much side effects). In this work, we ignore the initial appointment and the appointments where no computerized tests have been logged, and we treat the unscheduled/off schedule appointments as if they have been initially scheduled on this date.

For this application, the data set consists of 463 adult patients suffering from treatment-resistant depression (TRD) with an initial visit at the depressive and suicide disorders program (DSDP) between August 2006 and August 2015. This data set is composed of clinical and research data. The clinical variables include the treating psychiatrist (i.e., one of four psychiatrist), patient gender, age, date of initial visit, origin of referral (i.e., ED, internal, external), medical file closing date and reason, whether there are comorbidities on the first axis (i.e., major mental disorders) and second axis (i.e., personality disorders), the medications (i.e., drugs and dosages) taken at the initial visit, the prescribed medications (i.e., drugs and dosages) at all the following appointments and the computerized tests

scores (i.e., tests for quality of life, depression, side effects and suicide ideation) done prior to the appointments. On the other hand, the research data is composed of questionnaires on the socio-demographics and suicidal behavior history, of standardized tests for the evaluation of anxiety, depression and suicide ideation, and of diagnostic tests for the first and second axis.

In the rest of this section, we discuss the results of the semi-structured interviews with the psychiatrists (i.e., the first stage of the framework). Then, we describe the ensuing data set that is used in the second stage of the framework. Finally, we provide the parameters used for the proposed models of the second stage, and discuss the main and additional results.

4.5.1 Semi-Structured Interviews

The potential features used to determine the time between consecutive appointments are elicited from the four psychiatrists at the DSDP with semi-structured interviews. With these interviews, we also collect information on the experience of the physicians, the decisions they find challenging, and the typical and maximal times between appointments. Note that the results in this subsection were validated with the chief of the DSDP. Refer to Appendix B.1 for the interview guide.

Experience

For confidentiality reasons, it was decided not to associate, in this text, the experience with a specific physician (i.e., Dr. A, B, C or D). Also note that these experiences are given with respect to 2015 as the reference date. One physician worked three years in general psychiatry and then 13 years on MDD at the DSDP. A second physician worked 11 years on mood disorders after his specialty degree and then five years on MDD at the DSDP. Then, a third physician worked 13 years in general psychiatry and then five years on MDD at the DSDP. Finally, the last physician worked for five years on MDD at the DSDP after his specialty degree.

Challenging Decisions

Dr. A identified the decisions related to patient's rights and autonomy as the most challenging; for example, the decision of whether to call the police or not if a patient seems determined to commit a suicide. In contrast, this physician said that the timing decision between appointments was "automatic".

Dr. B mentioned that the diagnostic of depression, the exclusion of a bipolar disorder, the identification of an imminent suicidal risk and the identification of the somatic aspects (e.g., age, diseases, treatments) that can influence the treatment are the most challenging decisions. This physician also thinks that the timing decision between appointments is not difficult. It might only be if there are not enough availabilities.

Dr. C identified the diagnostic (e.g., understanding of the problem, reason that the patient is here) and the treatment (e.g., pharmacological, psychological) as the difficult decisions. This physician did not identify the decision of whether to intervene with respect to a risk of suicide as a difficult decision. This physician did not think either that the frequency of appointments is a difficult decision. He generally consults the patient to understand his preferences with respect to this decision.

Finally, Dr. D identified the evaluation of the health state, the intervention with respect to a risk of suicide and the choice of the treatment as the difficult decisions. This physician did not find that the timing decision between the appointments was a difficult decision. This physician said that, while there are no precise scientific markers for this decision, it is common sense.

Note that, even though the four physicians did not identify the timing decision as a challenging decision, it still remains an important one. In fact, there is currently a waiting list to access this specialized outpatient clinic, and this timing decision most probably has an effect on the size of this waiting list.

Potential Features

Dr. A identified the following important variables for the timing decision: (1) major treatment modifications (e.g., adding an antidepressant or certain add-on agents such as lithium) and (2) health state of the patient (e.g., suicidal risk, symptoms' severity, side effects).

Dr. B identified the following important variables for the timing decision: (1) suicidal risk, (2) doubts on whether the patient is suffering from a bipolar disorder, (3) tolerance to treatment and (4) fragility of patient with respect to the somatic aspects.

Dr. C identified the following important variables for the timing decision: (1) severity of symptoms, (2) treatment modifications (e.g., adding a drug, changing dosage), (3) stability of the health state and (4) patient preferences. This physician also highlighted that a patient might misinterpreted a frequency decision and this is why it is important to discuss it with the patient.

Finally, Dr. D identified the following important variables for the timing decision: (1) risk of suicide, (2) tolerance to treatment (i.e., side effects), (3) presence of anxiety and panic attacks, (4) presence of sleeping disorders, (5) modification of treatment (e.g., 4-6 weeks for antidepressants, 1 week after stable dosage for lithium). This physician also highlighted as an important variable the differences in the availabilities of the psychiatrists, and the number and severity of patients they are following.

Typical and Maximal Times Between Appointments

The usual timing decisions for Dr. A are: (1) two to six weeks for an unstable patient, (2) three months for an almost stable patient, (3) four months for a more stable patient, (4) six months for a truly stable patient on which we need to make some minor treatment modifications (e.g., remove unnecessary drugs as a second or third antidepressant) and, finally, (5) one year for a truly stable patient that doesn't need treatment modifications but still need to have follow-up at the clinics since, for example, this patient does not have a family physician. Finally, this physician mentioned that there should not be more than 18 months between two consecutive appointments if these are scheduled.

The usual timing decisions for Dr. B are: (1) one week for a patient with high suicidal risk, (2) two to four weeks for a patient with a lower suicidal risk and with some treatment modifications, (3) three months for a stabilized patient, and finally (4) six months for a patient that is going very well. This physician does not schedule appointments beyond six months; if a patient is this healthy, then he is told to call on need.

The usual timing decisions for Dr. C are: (1) one week when a new drug that has been never tried is introduced, the patient is severely depressed or the patient is at a high risk of suicide without needing an intervention, (2) two to three weeks for a dosage modification, (3) two months to remove medication with the possibility of a sooner appointment on call if necessary and (4) two to three months if the patient is stable and does not have a family physician. This physician does not schedule appointments beyond three months.

The usual timing decisions for Dr. D are: (1) five to seven days if the patient poses a risk towards his security, (2) one to two weeks if the patient is highly unstable or has a bad tolerance to treatment, (3) four to six weeks for most of the patients (e.g. unstable, good social support, good tolerance to treatment, modification of antidepressants), (4) two months for more stable patients and (5) up to three to four months for patients who will have their medical leave in the next three to six months or patients who are chronically depressed. This physician does not schedule appointments beyond four months. If a patient is not seen within six months, his file is closed.

4.5.2 Data Set

With the help of these interviews, a list of features characterizing the state was selected. These are described in Table 4.2; the first 5 rows are for the static information b, the next 16 rows are for the dynamic information g and the last row is the time since the initial appointment t.

A few remarks are in order regarding these features. First, we believe that these features are reasonable to describe the state since the physicians claimed to use them in the interviews. Yet, for the IRL models, these features might be insufficient to fully characterize the transitions. Thus, the IRL models could show more variability than the BC models due to this issue. Second, some of these features are intentionally designed more specific than others; for example, some features focus on specific drugs (e.g., lithium carbonate) and drug classes (e.g., tricyclic antidepressants (TCAs), monoamine oxidase inhibitors (MAOIs), anticonvulsants) since these drugs and drug classes require a tighter follow-up. On the other hand, we also include features which are more general and might include other features as a subset (e.g., `ADAdded_Any` includes `ADAdded_TCA`) since they might account for other characteristics. The goal of these specific and general features is to try to capture all potential relevant characteristics in a relatively small number of features due to the limited available data (discussed next). Third, note that some of these features (e.g., `Age`) may need to be discretized (or further discretized) depending on the IL approach. Fourth, note that 4 binary values indicating the physicians are included within the features. Thus, each model is trained over the full data; in other words, we do not train a model per physician.[9] There are two motivations behind this approach. Since the size of the data set is small, this approach allows us to use more data per model. In addition, this approach allows us to compare the *physician* factor against the other factors. Finally, note that the action a is computed in weeks since the feature `FollowingTime` is also in weeks. We believe that this granularity provides enough precision from a practical point of view.

These selected features were then extracted of the DSDP data set as follows. For each patient, a raw trajectory $\bar{\zeta}_m$ is built by setting $\bar{b}^m$ to the static information (e.g., gender, treating physician) and creating a new decision epoch with $(\bar{g}^m_h, \bar{t}^m_h, \bar{a}^m_h)$ for each time a computerized test is passed where $\bar{g}^m_h$ is the dynamic information (e.g., treatment modifications, computerized tests), $\bar{t}^m_h$ is the time since the initial appointment and $\bar{a}^m_h = \bar{t}^m_{h+1} - \bar{t}^m_h$ is the observed timing decision (i.e., timing of next appointment). The last decision epoch consists in $(\bar{g}^m_{H_m+1}, \bar{t}^m_{H_m+1})$ since we do not observe the action $\bar{a}^m_{H_m+1}$. If the

[9] The *expert* being replicated can now be considered as a meta physician that can switch between different physician's policies or reward functions with the help of a bias.

TABLE 4.2: List of state features.[1]See Table B.1.

Feature	Type	Description
MD_A, MD_B, MD_C, MD_D	4 binary values	Indicators of the patient's treating psychiatrist
IsMale	Binary value	Patient gender
Age	Real value	Patient age (in years) at the initial appointment
FirstAxis	Binary value	Indicator of comorbidities on the first axis at the initial appointment
SecondAxis	Binary value	Indicator of comorbidities on the second axis at the initial appointment
ADDosageIncrease_TCA, ADDosageIncrease_Any	2 binary values	Indicators of increased dosage for antidepressants from the tricyclic antidepressant (TCA) class, or for any antidepressants since last appointment[1]
ADAdded_TCA, ADAdded_MAOI, ADAdded_Any	3 binary values	Indicators of the addition of antidepressant drugs from the TCA or monoamine oxidase inhibitor (MAOI) classes, or of any antidepressants since last appointment[1]
AODosageIncrease_Li, AODosageIncrease_AED, AODosageIncrease_Any	3 binary values	Indicators of increased dosage for add-on drugs that are either lithium carbonate or from the anticonvulsant class, or for any add-on drugs since last appointment[1]

Continued on next page...

TABLE 4.2 – continued from previous page.

Feature	Type	Description
AOAdded_Li, AOAdded_AED, AOAdded_Any	3 binary values	Indicators of the addition of add-on drugs that are either lithium carbonate or from the anticonvulsant class, or of any add-on drugs since last appointment[1]
FIBSERScore	Integer value between 0 and 18	Current Frequency, Intensity, and Burden of Side Effects Rating (FIBSER) score
FIBSERTrend_Inc, FIBSERTrend_Dec	2 binary values	Indicator of strict increase or decrease in FIBSERScore with respect to the previous score
OverallFIBSERTrend_Inc, OverallFIBSERTrend_Dec	2 binary values	Indicator of strict increase or decrease in FIBSERScore with respect to the initial score
QIDSScore	Integer value between 0 and 27	Current 16-item Quick Inventory of Depressive Symptomatology (QIDS) self-reported score
QIDSTrend_Inc, QIDSTrend_Dec	2 binary values	Indicator of strict increase or decrease in QIDSScore with respect to the previous score
OverallQIDSTrend_Inc, OverallQIDSTrend_Dec	2 binary values	Indicator of strict increase or decrease in QIDSScore with respect to the initial score

Continued on next page...

TABLE 4.2 – continued from previous page.

Feature	Type	Description
SSIScore	Integer value between 0 and 38	Current Scale for Suicide Ideation (SSI) score
SSITrend_Inc, SSITrend_Dec	2 binary values	Indicator of strict increase or decrease in SSIScore with respect to the previous score
OverallSSITrend_Inc, OverallSSITrend_Dec	2 binary values	Indicator of strict increase or decrease in SSIScore with respect to the initial score
QLDSScore	Integer value between 0 and 34	Current Quality of Life in Depression Scale (QLDS) score
QLDSTrend_Inc, QLDSTrend_Dec	2 binary values	Indicator of strict increase or decrease in QLDSScore with respect to the previous score
OverallQLDSTrend_Inc, OverallQLDSTrend_Dec	2 binary values	Indicator of strict increase or decrease in QLDSScore with respect to the initial score
FollowingTime	Real value	Time (in weeks) since the initial appointment

delay between two consecutive decision epochs for a trajectory m is longer than 52 weeks (i.e., $\bar{t}^m_{h+1} - \bar{t}^m_h \geq 52$), then we end trajectory m with $(\bar{g}^m_h, \bar{t}^m_h)$, and start a new trajectory $m+1$ with $\bar{b}^{m+1} = \bar{b}^m$, $\bar{g}^{m+1}_1 = \bar{g}^m_{h+1}$ and $\bar{t}^{m+1}_1 = \bar{t}^m_{h+1}$.

Within the extracted data set, there are 316 out of the 463 patients with at least one usable state-action pair, where a usable state is defined as a state with no missing values in the features and a usable action is defined as an action with a value less or equal to 52 weeks. These 316 patients yield a total of 3949 usable state-action pairs and 3949 usable state-action-state tuples, where the latter is the data available to estimate the transition function. Descriptive statistics of the states features and actions are given in Appendix B.3 for the observations that are part of the usable state-action pairs.

4.5.3 IL Models

The different IL models used on the extracted data set are:

- Least absolute shrinkage and selection operator (LASSO) from the *scikit-learn* library (Pedregosa et al., 2011) where the regularization weight is optimized with a grid search over 30 values spaced evenly on a log scale between 10^{-2} and 10^2.

- Multiclass logistic regression with L1-regularization (L1-MLR) from the *scikit-learn* library (Pedregosa et al., 2011). The parameters are optimized with a grid search over 30 regularization weights spaced evenly on a log scale between 10^{-2} and 10^2, and over the 3 previously discussed method of discretization (i.e., equal width, equal frequency and k-means) with either 2 or 3 discrete action values.

- Semi maximum likelihood inverse reinforcement learning (SMLIRL) implemented with an adaptation of some of the code of Choi and Kim (2013) and the *L1General2* library (Schmidt, 2010). The parameters are again optimized with a grid search over the regularization weights, the discounting values and the discretization parameters. The grid for the regularization weights consists in 10 values spaced evenly on a log scale between 10^{-2} and 10^2 while the grid for the discounting values consists in 10 values of $e^{-\alpha}$ spaced evenly on a linear scale between 0.1 and 0.95. The grid for the discretization of the action consists in the 3 discretization approaches with 3 discrete values. Then, the discretization of the state is done using one of two approaches. The first approach consists in discretizing (1) the static information values into 2 values using k-means, (2) the dynamic information value into 3 values using k-means and (3) the `FollowingTime` values into 5 values using the equal frequency method.

The second method consists in discretizing the state values (i.e., static and dynamic information, and the `FollowingTime` values) into 20 values using k-means. Finally, the confidence parameter β is set to a value of 1.

Note that we scale the features prior to running the models since we want the weights to be on a similar scale and be comparable. In particular, we scale all features to the $[0, 1]$ interval which allows the binary features to keep their meaning and allows all features to be comparable in their effect.[10] As with the other preprocessing approaches (e.g., discretization), the same procedure used for training is used on the data to predict.

Also note that we select the previous two discretization procedures for the state of the SMLIRL model in order to do a trade-off between having enough state values (i.e., a meaningful state) and having a state space size that is manageable (i.e., a state space that is small enough to obtain many observations per state-action-state value in order to fit the TPM). We do a uniform initialization of the TPM before fitting it to the data in order to limit the impact of having too few observations per state-action-state tuple (i.e., we add a bias to the TPM in order to limit its variance).[11]

To conclude this section, note that 90% of the patients (rounded down) are used for cross-validation while the others are used for the testing. Cross-validation is done with 5 folds and the mean of the root mean squared error (RMSE) across the validation sets is used to select the best parameters for each model.

4.5.4 Main Results

In order to characterize the behavior, we first determine the best parameters for each model using cross-validation. Then we retrain these models on the data set used for cross-validation and compare these best models on a test set, in order to obtain a sense of the performance of each model. Finally, we retrain these best models on the full data set (which includes the test set) to obtain the weights associated with each feature; note that these weights are provided without confidence intervals since these are not trivial to obtain for these particular regularized models.

According to cross-validation, the best regularization weight for the LASSO model is 0.01, the lowest value of the range. For the L1-MLR model, the best parameters are 2.21 for the regularization weight and a discretization of the action in 3 values using the equal

[10] This scaling is done with the *MinMaxScaler* procedure of the *scikit-learn* library.

[11] This uniform initialization consists in setting a fake count of 1 to each state-action-state value before counting the real state-action-state observations. The TPM is then obtained by scaling this 2-dimensional matrix of counts such that each of its rows sums to 1.

width approach. For the SMLIRL model, a discretization of the action values with the equal width approach. This model is indifferent to the other parameters.

The root mean squared error (RMSE) and mean absolute error (MAE) on the test set for the best LASSO, L1-MLR and SMLIRL models are given in Table 4.3. These results show that LASSO is better than L1-MLR and SMLIRL in this case to replicate the behavior (i.e., lower RMSE and MAE) and, thus, LASSO should provide better explanations.

TABLE 4.3: Root mean squared error (RMSE) and mean absolute error (MAE) of the LASSO, L1-MLR and SMLIRL models on the test set.

Model	**RMSE**	**MAE**
LASSO	6.25	4.32
L1-MLR	6.85	5.08
SMLIRL	6.85	5.08

The weights obtained by training the best LASSO model on the full data set are given in Table 4.4. Most features are used by LASSO with the exception of 7 features (i.e., `MD_D`, `FirstAxis`, `ADDosageIncrease_TCA`, `ADAdded_TCA`, `ADAdded_MAOI`, `AODosageIncrease_Li`, `AODosageIncrease_AED`); in fact, the LASSO model might have used all features if we allowed the regularization weight to go below 0.01. The two most important features are `FollowingTime` and `MD_A` and their weights are somewhat above the others. It feels logical that `FollowingTime` is an important feature since over time, a physician gets to know his patient and might be more willing to spread out the appointments; in addition, this patient might feel better after some time, motivating again appointments which are more distant. Rather interestingly, however, it appears that `MD_A` is the second most important feature. This finding implies that Dr. A sets appointments which are much more spaced out than the other physicians; the other physicians' weights are -1.336, 0.684 and zero respectively for `MD_B`, `MD_C` and `MD_D`.

The weights obtained by training the best L1-MLR model on the full data set are given in Table 4.5. In the case of a small action value (i.e., action value close to 9.024), there are many factors that are taken into account with the most important ones being `FollowingTime`, `FIBSERScore`, `SSIScore` and `MD_A` (i.e., a large `FollowingTime` value and a true `MD_A` value decrease the chance of a small action value while large `FIBSERScore` and `SSIScore` values increase the chance of a small action value). Then, in the case of a moderate action value (i.e., action value close to 26.214), there are a lot less factors taken into account with the most important ones being `OverallQLDSTrend_Dec` and `IsMale` (i.e., a true `OverallQLDSTrend_Dec` value decreases the chance of a moderate action

value while a true IsMale value increases the chance of a moderate action value). Finally, in the case of a large action value (i.e., action value close to 43.405), the most important factors are ADAdded_Any and MD_B (i.e., true ADAdded_Any and MD_B values decrease the chance of a large action value). A few remarks are in order. First, note that there are a lot less moderate and large action values than small action values (see Figure 4.1). Thus, since these weights were trained with less data, these results should be taken with caution. Second, rather interestingly, it appears that the features used to predict a small, moderate or large action value are not the same (e.g., Age plays a role mostly in the moderate action value); this effect might be due to the preceding remark. Finally, note that the top factors found in LASSO are similar to the ones found for the small action value in L1-MLR. This is probably due in part to the large number of small action values.

The weights obtained by training the best SMLIRL model (i.e., a discounting factor $e^{-\alpha} = 0.1$, a regularization factor $\lambda = 0.01$, a discretization of the state using the first previously discussed approach and a discretization of the action using the equal width approach) on the full data set are given in Table 4.6; note that several other SMLIRL models are also considered best. In the case of a small action value (i.e., action value close to 9.024), the most important factors taken into account are MD_B and FollowingTime (i.e., a large FollowingTime value decreases the reward associated with a small action value while a true MD_B value increases the reward associated with a small action value). Then, in the case of a moderate action value (i.e., action value close to 26.214), the most important factor is FollowingTime (i.e., a large FollowingTime value increases the reward associated with a moderate action value). Finally, in the case of a large action value (i.e., action value close to 43.405), the most important factors are MD_B and QLDSScore (i.e., a true MD_B value decreases the reward associated with a large action value while a large QLDSScore value increases the reward associated with a large action value). A few remarks are in order. First, note again that the results for moderate and large action values should be taken with caution. Second, these weights are only for the reward function; the optimal policy also takes into account the environment (i.e., the transition function). Yet, the optimal policy consists in the smallest action value for all state values, i.e., it is static and does not depend on the state. Third, there appears to be some inconsistencies in these weights. For example, it is unclear why a large QLDSScore value provides a high reward with any action value. This might be due to the limited number of observations for large action values, to the association of the QLDSScore to other features across the discrete state values or just that these weights are so small that they should be ignored. Finally, note that the top factors found in LASSO (e.g., MD_A) are not necessarily present in SMLIRL. This might be due to

the discretization of both the state and action values, and to the estimation of the TPM with limited data in SMLIRL.

In summary, it appears that both `FollowingTime` and `MD_A` have a large effect on the time to the next appointment. In addition, it appears that features capturing the side effects, suicide ideation and depression severity (i.e., `FIBSERScore`, `SSIScore` and `QIDSScore`) are important. Finally, it appears that the cost of discretization is quite high, and, probably because in part from this, LASSO appears better than L1-MLR and SMLIRL.

TABLE 4.4: Trained weights for the LASSO model sorted by decreasing absolute values. Null values are not shown. The intercept is 11.643.

Feature	Weights
FollowingTime	4.896
MD_A	3.708
FIBSERScore	-2.512
QIDSScore	-1.762
SSIScore	-1.470
QLDSScore	-1.429
AOAdded_AED	1.339
MD_B	-1.290
OverallQLDSTrend_Inc	-1.287
AOAdded_Li	-1.253
AOAdded_Any	-1.205
AODosageIncrease_Any	-1.189
Age	1.074
QLDSTrend_Inc	-1.021
ADAdded_Any	-0.962
OverallQLDSTrend_Dec	-0.912
QLDSTrend_Dec	-0.841
OverallFIBSERTrend_Inc	-0.819
MD_C	0.749
OverallSSITrend_Inc	-0.529
SecondAxis	0.473
FIBSERTrend_Dec	-0.448
IsMale	0.395
QIDSTrend_Inc	-0.372
OverallSSITrend_Dec	-0.368
ADDosageIncrease_Any	-0.341
OverallQIDSTrend_Dec	0.319
OverallQIDSTrend_Inc	0.251
FIBSERTrend_Inc	0.223
SSITrend_Inc	0.197
QIDSTrend_Dec	-0.153
SSITrend_Dec	0.132
OverallFIBSERTrend_Dec	0.038

TABLE 4.5: Trained weights for the L1-MLR model. Each column corresponds to a discrete action value. The action values are 9.024, 26.214 and 43.405. The intercepts are 1.518, 0.106 and -1.624. The 8 features with only null weights are not shown.

Feature	Action 1 weights	Action 2 weights	Action 3 weights
MD_A	-0.866		0.345
MD_B	0.513		-0.513
MD_C	-0.005		
MD_D	0.045		-0.246
IsMale	-0.018	0.313	
Age		-0.037	
SecondAxis	-0.154		0.121
ADDosageIncrease_Any	0.090	-0.144	
ADAdded_Any	0.195		-0.607
AODosageIncrease_Any	0.372		-0.108
AOAdded_AED	-0.176		
AOAdded_Any	0.303	-0.076	
FIBSERScore	1.338	-0.022	
FIBSERTrend_Inc	-0.219		
FIBSERTrend_Dec	0.099		-0.184
OverallFIBSERTrend_Inc	0.056		
OverallFIBSERTrend_Dec	-0.118		0.063
QIDSScore	0.484		
QIDSTrend_Inc	0.054		
QIDSTrend_Dec		-0.022	
OverallQIDSTrend_Inc	-0.008		
SSIScore	0.885		
SSITrend_Inc	-0.137		
SSITrend_Dec	-0.060	0.090	
OverallSSITrend_Inc	0.143		-0.114
OverallSSITrend_Dec	0.075		-0.097
QLDSScore	0.283		
QLDSTrend_Inc	0.454		
QLDSTrend_Dec	0.366	-0.075	
OverallQLDSTrend_Inc	0.319		
OverallQLDSTrend_Dec	0.038	-0.355	
FollowingTime	-1.491		

TABLE 4.6: Trained weights for the SMLIRL model. Each column corresponds to a discrete action value. The action values are 9.024, 26.214 and 43.405. The intercepts are 1.666, 0.790 and -0.536. The 9 features with all null weights are not shown.

Feature	Action 1 weights	Action 2 weights	Action 3 weights
MD_A	0.002	-0.062	0.004
MD_B	0.979	0.020	-0.810
MD_C		-0.001	
MD_D		-0.146	-0.023
IsMale	0.012	-0.115	0.014
Age	0.007	0.007	-0.075
FirstAxis	0.468	-0.012	-0.003
SecondAxis	0.008	-0.013	-0.085
ADAdded_Any	0.002		
AOAdded_Any			-0.003
FIBSERScore	0.047	-0.023	
FIBSERTrend_Inc	0.155		
FIBSERTrend_Dec	-0.100	-0.007	
OverallFIBSERTrend_Inc	0.026		-0.002
OverallFIBSERTrend_Dec	0.001	-0.001	
QIDSScore	0.209	-0.051	-0.334
QIDSTrend_Inc	0.272		-0.464
QIDSTrend_Dec	0.013	-0.004	-0.294
OverallQIDSTrend_Inc	0.292		
OverallQIDSTrend_Dec	0.018	-0.007	-0.006
SSIScore	0.001		
SSITrend_Inc	0.002	-0.026	-0.053
SSITrend_Dec	0.003	0.001	
OverallSSITrend_Inc	0.103		-0.010
OverallSSITrend_Dec	-0.071	0.058	-0.004
QLDSScore	0.399	-0.024	0.634
QLDSTrend_Inc	0.391	-0.005	-0.185
QLDSTrend_Dec	0.689		
OverallQLDSTrend_Inc	0.172	-0.024	-0.279
OverallQLDSTrend_Dec	-0.020	-0.138	0.029
FollowingTime	-0.820	0.427	0.458

4.5.5 Additional Results

In this section, additional results are provided. These complement the main results of the application.

How good can SMLIRL become when using only the "best" features?

This first additional section tests if it is possible to improve the fit of the SMLIRL method when using only the "best" features identified by the LASSO method in the state definition; if the SMLIRL method can still not improve over the LASSO method in this case, then it is clearly not the best approach for this context. Two discretization schemes are again tried for this experiment. In the first discretization scheme, the top 3 features of LASSO are discretized individually as follows:

- `FollowingTime` is discretized in 4 bins using equal frequency,
- `MD_A` is kept as is (i.e., binary), and
- `FIBSERScore` is discretized in 2 bins using equal frequency.

In the second discretization scheme, these 3 features are discretized together into 20 values using k-means. For both schemes, the action is discretized in 3 values using equal frequency, equal width and k-means. All the other parameters are the same as in the main results section.

It appears that this reduced set of features leads to a SMLIRL model that is again indifferent to all parameters except the discretization of the action; the best models use a discretization of the action in 3 values using the equal width approach. In Table 4.7, it is shown that the SMLIRL model with the best 3 features from LASSO performs worst on the test set than the SMLIRL model with all features. In summary, this procedure does not appear to improve the predictive performance of the SMLIRL model, and thus the SMLIRL model still does not appear as the appropriate model for this context.

TABLE 4.7: Root mean squared error (RMSE) and mean absolute error (MAE) of the SMLIRL model on the test set with 3 state features.

Model	RMSE	MAE
SMLIRL (3 state features)	8.0736	5.7507

How does the feature weights differ across physicians?

In the main results section, it was found that if a patient is followed by MD_A then its time between appointments is larger. Thus, it is interesting to check how these weights differ across physicians. Yet, to make sure that these four models are still good from a predictive point of view since they are trained only on a subset of the data, the cross-validation and validation on a test set procedures were redone for these models with the same parameters as before. The RMSE and MAE of these four models are given in Table 4.8. It appears that the models for MD_A and MD_D are worst than the full model while the models for MD_B and MD_C are better or equivalent. Thus, the results of MD_A and MD_D should be interpreted with caution.

TABLE 4.8: Root mean squared error (RMSE) and mean absolute error (MAE) of the LASSO model for each physician on the test sets.

Model	RMSE	MAE
MD_A	9.49	7.64
MD_B	4.94	3.32
MD_C	6.41	5.01
MD_D	7.72	4.99

In Table 4.9, the trained weights of the four models are given. For MD_A, it appears that the most important features are FollowingTime, QIDSScore, and then to a smaller degree FIBSERScore, SSIScore, AOAdded_Any, QLDSScore, QLDSTrend_Inc, AODosageIncrease_Any and Age. For MD_B, it appears that the most important features are FIBSERScore and QLDSScore. For MD_C and MD_D, it appears that no features have a weight larger than 2 or smaller than -2; thus, the measured features do not affect significantly (from a practical point of view) the time between appointments.

In summary, with limited confidence due to data limitations, it appears that these weights do differ across the physicians. Yet, additional data and features should be collected to establish the correct weights for each physician with appropriate confidence.

Which physician's patients are better off?

At the DSDP, the assignment of new patients to physicians is done at random every week; unless a patient has some history with a particular physician. Hence, it is possible to compare the outcomes of each physician panel without having to deal with confounders.

As shown in Table 4.10, the panel of patients of each physician appears similar in their baseline characteristics (i.e., `IsMale`, `Age`, `FirstAxis`, `SecondAxis`). For example, the mean age is close to 44 years old for all panels; a one-way analysis of variance (ANOVA) on this covariate leads to a p-value of 0.0458.

Yet, these panels might differ in their outcomes due to the combined effect of the treatments' selection and the selection of the timing between appointments by the different physicians. Results of one-way ANOVAs indicate a p-value of < 0.001, 0.108, 0.002 and < 0.001 respectively for the `FIBSERScore`, `QIDSScore`, `SSIScore` and `QLDSScore` covariates. These results indicate that for each of these scores, except `QIDSScore`, at least one panel score mean is statistically significantly different than the other panel means. Since these differences are however insignificant from a practical point of view, it appears that no physician is better than the others with respect to these patients' outcomes.

TABLE 4.9: Trained weights for the LASSO model for each physician. Null values are not shown. The intercepts are respectively 19.587, 9.236, 9.035 and 8.326 for the model of MD_A, MD_B, MD_C and MD_D.

Feature	MD_A	MD_B	MD_C	MD_D
IsMale	1.113	0.203		
Age	2.216	0.081		
FirstAxis	-1.107		1.597	
SecondAxis	0.512			
ADDosageIncrease_Any	-0.236	-0.550		
ADAdded_Any	-1.308	-0.716		
AODosageIncrease_Any	-2.365	-0.793		
AOAdded_Any	-2.585	-1.065	-0.281	
FIBSERScore	-2.968	-2.533	-1.026	
FIBSERTrend_Inc	0.318			
FIBSERTrend_Dec	-1.108	-0.273		
OverallFIBSERTrend_Inc	-1.764	-0.440		
OverallFIBSERTrend_Dec	0.486			
QIDSScore	-6.302	-0.130		
QIDSTrend_Inc	-0.379	-0.049		
QIDSTrend_Dec	-0.624			
OverallQIDSTrend_Inc	0.288			
OverallQIDSTrend_Dec				0.130
SSIScore	-2.690			
SSITrend_Inc	0.381			
OverallSSITrend_Inc	-1.501			
OverallSSITrend_Dec	-1.331			
QLDSScore	-2.517	-2.492		
QLDSTrend_Inc	-2.452			-0.351
QLDSTrend_Dec	-1.233	-0.129		
OverallQLDSTrend_Inc		-0.732	-1.340	
OverallQLDSTrend_Dec	-0.751			
FollowingTime	9.522	1.985		

TABLE 4.10: Descriptive statistics of the state features and action per physician. The following means and standard deviations (in parentheses) are computed using respectively 876, 1689, 517 and 867 observations for `MD_A`, `MD_B`, `MD_C` and `MD_D`.

Feature	**MD_A**	**MD_B**	**MD_C**	**MD_D**
IsMale	0.51 (0.25)	0.27 (0.20)	0.38 (0.24)	0.37 (0.23)
Age	44.44 (10.10)	43.62 (10.74)	44.92 (10.23)	44.17 (9.68)
FirstAxis	0.49 (0.25)	0.68 (0.22)	0.44 (0.25)	0.76 (0.18)
SecondAxis	0.73 (0.20)	0.57 (0.25)	0.47 (0.25)	0.90 (0.09)
ADDosageIncrease_TCA	0.01 (0.01)	0.02 (0.02)	0.05 (0.05)	0.03 (0.03)
ADDosageIncrease_Any	0.13 (0.11)	0.18 (0.15)	0.15 (0.13)	0.25 (0.19)
ADAdded_TCA	0.02 (0.02)	0.01 (0.01)	0.06 (0.06)	0.02 (0.02)
ADAdded_MAOI	0.01 (0.01)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
ADAdded_Any	0.13 (0.11)	0.10 (0.09)	0.19 (0.15)	0.15 (0.13)
AODosageIncrease_Li	0.01 (0.01)	0.03 (0.03)	0.01 (0.01)	0.00 (0.00)
AODosageIncrease_AED	0.00 (0.00)	0.03 (0.03)	0.00 (0.00)	0.00 (0.00)
AODosageIncrease_Any	0.17 (0.14)	0.20 (0.16)	0.09 (0.08)	0.18 (0.15)
AOAdded_Li	0.02 (0.02)	0.01 (0.01)	0.02 (0.02)	0.01 (0.01)
AOAdded_AED	0.01 (0.01)	0.03 (0.03)	0.00 (0.00)	0.01 (0.01)
AOAdded_Any	0.22 (0.17)	0.22 (0.17)	0.19 (0.15)	0.19 (0.15)
FIBSERScore	6.58 (5.17)	6.13 (5.49)	5.40 (4.73)	6.04 (5.19)
FIBSERTrend_Inc	0.35 (0.23)	0.30 (0.21)	0.33 (0.22)	0.31 (0.21)
FIBSERTrend_Dec	0.38 (0.24)	0.31 (0.21)	0.38 (0.24)	0.37 (0.23)
OverallFIBSERTrend_Inc	0.39 (0.24)	0.35 (0.23)	0.29 (0.21)	0.31 (0.21)
OverallFIBSERTrend_Dec	0.50 (0.25)	0.44 (0.25)	0.57 (0.25)	0.50 (0.25)
QIDSScore	13.62 (5.91)	13.19 (6.10)	12.95 (6.57)	13.58 (6.52)
QIDSTrend_Inc	0.40 (0.24)	0.39 (0.24)	0.41 (0.24)	0.40 (0.24)
QIDSTrend_Dec	0.49 (0.25)	0.46 (0.25)	0.47 (0.25)	0.47 (0.25)
OverallQIDSTrend_Inc	0.22 (0.17)	0.25 (0.19)	0.23 (0.18)	0.23 (0.18)
OverallQIDSTrend_Dec	0.72 (0.20)	0.67 (0.22)	0.70 (0.21)	0.67 (0.22)
SSIScore	5.52 (7.45)	5.03 (8.18)	4.00 (5.67)	4.73 (6.37)
SSITrend_Inc	0.31 (0.21)	0.26 (0.19)	0.28 (0.20)	0.30 (0.21)
SSITrend_Dec	0.33 (0.22)	0.29 (0.21)	0.27 (0.20)	0.34 (0.22)
OverallSSITrend_Inc	0.30 (0.21)	0.25 (0.19)	0.25 (0.19)	0.26 (0.19)
OverallSSITrend_Dec	0.47 (0.25)	0.42 (0.24)	0.37 (0.23)	0.49 (0.25)
QLDSScore	19.32 (10.08)	17.05 (10.69)	18.30 (10.00)	20.38 (10.81)
QLDSTrend_Inc	0.41 (0.24)	0.38 (0.24)	0.42 (0.24)	0.39 (0.24)
QLDSTrend_Dec	0.48 (0.25)	0.46 (0.25)	0.43 (0.25)	0.43 (0.25)
OverallQLDSTrend_Inc	0.35 (0.23)	0.32 (0.22)	0.28 (0.20)	0.35 (0.23)
OverallQLDSTrend_Dec	0.60 (0.24)	0.60 (0.24)	0.62 (0.24)	0.55 (0.25)
FollowingTime	100.82 (85.20)	67.71 (51.69)	64.82 (50.89)	93.08 (80.83)
Action	11.94 (9.91)	6.65 (5.53)	9.01 (7.47)	8.28 (6.40)

4.6 Conclusion

In this work, we provided an overview of the field of imitation learning (IL) and proposed a new inverse reinforcement learning (IRL) approach (i.e., semi maximum likelihood inverse reinforcement learning (SMLIRL)). Next, we discuss the difference between these IL approaches when trying to explain a behavior instead of imitating it. Finally, we applied a two-stage framework (i.e., semi-structured interviews and IL) in order to understand the factors that affect the time to the next appointment. In particular, we applied our framework to a data set collected from a specialized outpatient clinic treating patients suffering from treatment-resistant depression (TRD). By doing so, we found out that LASSO is the most appropriate method for this context, even if the SMLIRL method is better from a theoretical point of view.

There are several limitations to our TRD case study. First, we used observational data that was not purposely collected for research and thus contained several missing values. This resulted in the analysis of a subset of patients that may be different from the larger set of patients in ways that lead to a mischaracterization of the timing decisions. Second, we omitted several features that contained too many missing values or that were not part of our data set. These features might have been more predictive that the features used in this study. For example, non-pharmacological treatments (e.g., psychotherapy) are not part of this data set but should be predictive of the time between appointments. Third, we used linear models which might not be the most appropriate model's class. Other classes (e.g., decision trees) might have lead to better predictions and a different characterization of the timing decisions. Finally, we did not validate that the identified features were the true cause of the actions.

Future research areas include (1) the further characterization of these timing decisions and (2) an analysis to determine which timing policies is best.

Chapter 5

Using Recommender Systems to Improve the Treatment of Treatment-Resistant Depression

5.1 Introduction

Major depressive disorder (MDD), is amongst the top ten causes of the global burden of disease and is predicted to become the leading cause by 2030 (World Health Organization, 2008). Up to 15% of the population affected by MDD remains significantly depressed despite the aggressive use of multiple pharmacological and psychotherapeutical approaches. These patients are generally referred to as suffering from treatment-resistant depression (TRD). Although there is no consensus regarding the definition of TRD, a patient suffering from MDD is usually considered treatment-resistant (or refractory) when at least two trials with antidepressants from different pharmacologic classes (adequate in dose, duration, and compliance) fail to produce a significant clinical improvement (Berlim and Turecki, 2007).

The treatment of patients suffering from treatment-resistant depression (TRD) is a hard task requiring a referral to a specialized clinic. Due to the limited evidence on predictors of differential response to alternative treatments (Simon and Perlis, 2010) and to the vague medical guidelines (e.g., Lam et al. (2016a)), the psychiatrists, at this specialized outpatient clinic, usually rely on their past experiences on similar patients to decide the next treatment to prescribe. However, this large amount of information cannot necessarily be fully processed by their human brains (Miller, 1994). Thus, the successful treatment (if any) is found through a trial-and-error process probably much longer than needed; a long and difficult time for the patients.

This work addresses this issue by building a recommender system (RS) for the pharmacological treatments of patients. This RS is trained on an observational data set collected at the depressive and suicide disorders program (DSDP) of the Douglas Mental Health University Institute in Montreal, a specialized outpatient clinic treating TRD. This work consists somewhat in an extension of Chapter 3 where the drug treatments over the full treatment path are now considered instead of only the five treatment modification strategies at the initial appointment. In addition, in this chapter, we assume that the sequence of treatments that have been administered to the patient does not affect the efficacy of the current treatment, a medically reasonable assumption according to our medical collaborator.

After correcting for potential confounders in the observational data with an approach from causal inference, we use this system to try to address the following questions:

- Is it possible to accurately predict the outcome of a particular treatment on a particular patient?
- Do some treatments consistently provide good response for all patients?
- Do some treatments consistently provide good response within a particular patient subgroup?

Note that the first question consists in the main objective of this study and that the following questions are somewhat dependent on the results to this first question. Also note that further replications of this study, ideally in randomized controlled trials (RCTs), are needed to fully answer these questions.

This chapter is organized as follows. Section 5.2 introduces some preliminaries and Section 5.3 describes the related work in the RSs literature. Then, Section 5.4 presents the results of our treatment-resistant depression case study. We conclude in Section 5.5.

5.2 Preliminaries

5.2.1 Notation

Let $a_{ij} \in \{0, 1\}$ denote whether an outcome score $r_{ij} \in \mathbb{R}$ is observed for patient $i \in \mathcal{I} \triangleq \{1, \ldots, m\}$ and treatment $j \in \mathcal{J} \triangleq \{1, \ldots, n\}$. In addition to these variables, let $y_i \in \mathbb{R}^o$ and $z_j \in \mathbb{R}^p$ denote vectors of features characterizing respectively a patient $i \in \mathcal{I}$ and a treatment $j \in \mathcal{J}$. The full data set is denoted by $\mathcal{D} \triangleq \{(y_i, z_j, a_{ij}, r_{ij})\}$ and the observed entries as $\mathcal{O} \triangleq \{(i, j) \in \mathcal{I} \times \mathcal{J} \mid a_{ij} = 1\}$. Note that in this work, we will refer interchangeably to user/patient, item/treatment and rating/score with a prevalence of

the former terminology; the former coming from the RS literature and the latter from our application.

5.2.2 Baseline Model

In this work, we assume that all the methods are applied to the normalized ratings $r_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j$ where $\hat{\mu}$ is the estimated overall mean, $\hat{\alpha}_i$ is the estimated user effect and $\hat{\beta}_j$ is the estimated item effect. Thus, to obtain the predicted ratings, it is necessary to compute $\hat{r}_{ij} + \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j$ where $\hat{r}_{ij}$ is the predicted rating. However, to simplify notation, we omit this normalization step in the following discussion. In addition, note that this baseline model can sometime be directly included into another model, for example matrix factorization (Equation 5.1). We omit however these variants in the following discussion for ease of exposition.

5.3 Related Work

In this work, we focus on the prediction of the ratings to understand whether it is possible to accurately predict them from the available data. In particular, we focus on models that can predict for each patient the outcomes of all treatments (observed as well as unobserved), so that we can evaluate the effect of standard as well as non-standard treatments. In addition, another goal of this work is to understand what data is the most relevant for these predictions. Hence, we omit, from this work and literature review, contextual multi-armed bandit models that focus on the operationalization of the best decisions and models that directly rank the ratings. Finally, we keep for future research multi-criteria models (e.g., models that predict multiple types of ratings such as depression severity, quality of life, suicide ideation and side effects).

We now discuss the related literature with a focus on hybrid recommender systems (RSs) (Burke, 2002) because these appear to be the most appropriate for our application. In particular, the most appropriate system appears to be an hybrid of collaborative filtering models (i.e., models using the similarity in ratings r_{ij} to predict new ratings) and features based models (i.e., models using the vectors of features y_i and z_j to predict new ratings), since an hybrid system using these two models appears to be able to recommend treatments to patients with observed scores (i.e., patients with an history at the clinic) as well as to patients with no observed scores (i.e., new patients). Note that our system needs to address new patients but doesn't often encounter new treatments. Also note that the cold-start

aspect (i.e., predicting ratings to patients with no observed ratings) is quite important to us in order to avoid multiple treatment attempts on a patient. See Table 5.1 for an overview of the background information that the collaborative filtering and features based models use, the input they require and the process they use to predict scores. In the rest of this section, we describe in more details the related models in the literature and other noteworthy aspects.

TABLE 5.1: Recommendation models (adapted from Burke (2002)).

Model	**Background**	**Input**	**Process**
Collaborative Filtering	Ratings from $\mathcal{I}$ of items in $\mathcal{J}$.	Ratings from i of items in $\mathcal{J}$.	Identify users in $\mathcal{I}$ similar to i, and extrapolate from their ratings of j.
Features Based	Auxiliary information on $\mathcal{I}$ and $\mathcal{J}$, and the ratings from $\mathcal{I}$ of items in $\mathcal{J}$.	Auxiliary information on i and j.	Learn parameters for auxiliary information on $\mathcal{I}$ and $\mathcal{J}$, and extrapolate from these parameters to predict rating of j by user i.

5.3.1 Collaborative Filtering Model

Collaborative filtering models can be seen as the generalization of the regression/classification models since we do not have anymore a clear distinction between the dependent and independent variables. There exist two types of methods within collaborative filtering models. One one hand, there are memory-based methods, also known as neighborhood-based methods, which predict the rating of an item j for an user i from the ratings of the neighborhood; this neighborhood can either be defined among the users/rows (i.e., user-based collaborative filtering) or the items/columns (i.e., item-based collaborative filtering). These methods are often used for their simplicity and interpretability. However, they do not address well sparsity in the ratings.

On the other hand, there are model-based methods which creates a summarized model of the data first and then use this model to make predictions. These methods are generally found to better address sparsity in the ratings than the memory-based methods, at the expense of interpretability. While most regression/classification methods can be

implemented as a collaborative filtering model, it appears that the latent factor models (also known as matrix factorization models within the collaborative filtering literature) are the most accurate and the most popular. We focus on them for the rest of this section.

One of these popular matrix factorization models consists in the following formulation where the predicted score is given by

$$\hat{r}_{ij} = \sum_{s=1}^{k} u_{is} v_{js} \tag{5.1}$$

with u_i as the user latent factors and v_j as the item latent factors. It is important to note that k is generally selected such that $k \ll m$ and $k \ll n$, i.e., it is a low-rank matrix factorization.

The coefficients $U \in \mathbb{R}^{m \times k}$ and $V \in \mathbb{R}^{n \times k}$ are fitted by minimizing the following regularized least squares loss function

$$L_{MF} \triangleq \sum_{(i,j) \in \mathcal{O}} \left(r_{ij} - \sum_{s=1}^{k} u_{is} v_{js} \right)^2 + \lambda_1 \|U\| + \lambda_2 \|V\|$$

where $\|A\| \triangleq \|\mathrm{vec}(A)\|$ can be a L1 or L2 norm, and $\mathrm{vec}(A)$ represent the vectorization of a matrix A. There exist a variety of approaches to solve the previous optimization problem such as gradient descent and iterated least squares.

To conclude this section, note that previous work by Koren (2010) has combined memory- and model-based methods with great success, improving the accuracy with respect to both approaches.

5.3.2 Features Based Model

While there exist an extensive literature on models that only use the features of the items z_j, there is limited literature on models that use both the features of the users y_i and the items z_j. To our knowledge, one notable work combining both types of features is the work of Ansari, Essegaier, and Kohli (2000).

Ansari, Essegaier, and Kohli (2000) proposed to do a linear regression of the user features, item features and user-item feature interactions. The predicted score, in a similar model, is given by

$$\hat{r}_{ij} = \alpha_1 + \alpha_2^\top y_i + \alpha_3^\top z_j + \alpha_4^\top \mathrm{vec}(y_i \otimes z_j) \tag{5.2}$$

where $\alpha_1 \in \mathbb{R}, \alpha_2 \in \mathbb{R}^o, \alpha_3 \in \mathbb{R}^p$ and $\alpha_4 \in \mathbb{R}^{op}$ are coefficient vectors, and $\otimes$ represent the Kronecker product of two vectors (i.e., all possible cross-product combinations). The

coefficient vectors $\alpha_1, \alpha_2, \alpha_3$ and α_4 can be fitted again by minimizing a regularized least squares loss function with gradient descent.

5.3.3 Hybrid Model

The different approaches described earlier (i.e., Equations 5.1 and 5.2) each have their strengths and weaknesses. For example, matrix factorization (Equation 5.1) tends to make good predictions but it is not able to predict ratings for new users; an issue that is resolved with a features based model (Equation 5.2). In addition, it is highly probable that the matrix factorization model and the features based model do not capture the same patterns in the data; a difference due to the models and/or the data used. Hence, combining the strengths of these models and their variants, and the different types of data within an hybrid model seem to be a promising avenue. This claim is supported by the Netflix Prize contest, the most popular RS competition, where the winning approach consisted of a blending of 800 models (Feuerverger, He, and Khatri, 2012).

Thus, we now describe different types of hybrid models. In particular, we describe five types of hybrid models out of the seven types of Burke (2002); we omit meta-level and mixed hybrids since they cannot be easily used with the earlier models (i.e., Equations 5.1 and 5.2). These five hybrids are described in Table 5.2. They are also discussed below with examples of how they can be used with our proposed models and their variants. The reader can refer to Jahrer, Töscher, and Legenstein (2010) for other hybrid models that were used in the Netflix Prize contest.

TABLE 5.2: Hybridization methods (adapted from Burke (2002) and Aggarwal (2016)).

Hybridization methods	**Description**
Weighted	The system weights scores of several models to produce one score.
Switching	The system switches between models depending on the situation.
Cascade	The system uses a series of models where each model improves the predictions of the previous models.
Feature Augmentation	The system uses a series of models where the previous models augment the features of the next models.
Feature Combination	The system uses several inputs in an unified representation.

Weighted Hybrids

Weighted hybrids use a weighting of several scores to produce one score. For example, if we have q predictive models, then it is possible to weight the predictions to get

$$\hat{r}_{ij} = \sum_{h=1}^{q} \beta_h \hat{r}_{ij}^h \tag{5.3}$$

where $\hat{r}_{ij}^h$ is the prediction for user-item pair (i, j) of model h and $\beta_1, \ldots, \beta_q$ are parameters to fit. These parameters can be fitted in order to minimize a loss function on a holdout set. For example, these parameters could be set to minimize the following mean squared error (MSE) or mean absolute error (MAE) on a holdout set $\mathcal{E}$:

$$MSE(\beta) \triangleq \frac{\sum_{(i,j)\in\mathcal{E}} (r_{ij} - \hat{r}_{ij})^2}{|\mathcal{E}|},$$
$$MAE(\beta) \triangleq \frac{\sum_{(i,j)\in\mathcal{E}} |r_{ij} - \hat{r}_{ij}|}{|\mathcal{E}|}.$$

If the MSE is used, then this problem becomes a linear least-squares regression. If the MAE is used, then this problem can be solved using gradient descent.

Unfortunately, this type of hybrid model (with Equations 5.1 and 5.2) cannot address cold-starts since matrix factorization by itself cannot address cold-starts. Variants of this hybrid model could be formulated to address this issue. For example, the weights could depend on the number of observed ratings of a user such that a weight of zero is given to matrix factorization if there are not enough observed ratings. This is somewhat similar to the switching hybrid models discussed next.

Weighted hybrids also consist of approaches such as bagging and randomness injection that can improve the predictions of a model (Bar et al., 2013).

Switching Hybrids

Switching hybrids switch between models depending on the situation and are thus good to address cold-start issues. For example, a switching hybrid can use one model when few ratings are observed for a user and then switch to another model once there is sufficient data:

$$\hat{r}_{ij} = \begin{cases} \hat{r}_{ij}^1, \text{ if } \sum_{j=1}^{m} a_{ij} \leq \tau, \\ \hat{r}_{ij}^2, \text{ otherwise}, \end{cases}$$

where $\hat{r}_{ij}^1$ and $\hat{r}_{ij}^2$ are predictions coming from two different models, and τ is a threshold parameter that can be fitted to minimize the MSE or MAE on a holdout set.

It is also possible to construct a switching hybrid model that makes a smoother transition between the models according to the number of observed ratings.

Cascade Hybrids

Cascade hybrids consist in a series of models where each model improves the predictions of the previous models. For example, it is possible to adapt boosting to RSs (Bar et al., 2013). The adapation of boosting is as follows. First, the weights b_{ij}^1 are initialized to $1/|\mathcal{O}|$ for all tuples $(i,j) \in \mathcal{O}$. Then, for each model $h = 1, \ldots, q$:

1. the model h learns from the weighted observations,
2. the absolute error is computed for its predictions as $AE_{ij}^h = |r_{ij} - \hat{r}_{ij}^h|$,
3. the error rate of model h is computed as $\beta_h = \sum_{(i,j)\in\mathcal{O}:AE_{ij}^h>\Delta} b_{ij}^h$,
4. the next weights are computed as

$$b_{ij}^{h+1} = \frac{b_{ij}^h}{\bar{Z}_h} \times \begin{cases} \beta_h \text{ if } AE_{ij}^h \leq \Delta, \\ 1 \text{ otherwise}, \end{cases}$$

where the normalization factor $\bar{Z}_h$ assures that the weights sum to 1.

Finally, the prediction from the boosting model is given as

$$\hat{r}_{ij} = \left(\sum_{h=1}^{q} \log\left(\frac{1}{\beta_h}\right) \hat{r}_{ij}^h \right) / \left(\sum_{h=1}^{q} \log\left(\frac{1}{\beta_h}\right) \right).$$

Feature Augmentation Hybrids

Feature augmentation hybrids are similar to stacking ensemble in classification. They use previous models to augment the features of the next models. For example, it is possible to use a features-based model to fill the missing values in the rating matrix. Then, the matrix factorization model learns from this dense matrix that consists of observed ratings and pseudo-ratings (i.e., predicted ratings). This approach allows the use of matrix factorization for new user with no ratings, i.e., it addresses the cold-start problem. This example is similar to Melville, Mooney, and Nagarajan (2002) that used a content-based model prior to a neighborhood-based model.

Since the the pseudo-ratings are only predictions, the second model should put less weight on them in order to improve accuracy. Assuming that the observed ratings are given weights of one, the weight b_{ij} of a pseudo-rating should approach one when $\sum_{i=1}^{m} a_{ij} \to m$ and $\sum_{j=1}^{n} a_{ij} \to n$ since the features-based model will give better predictions.

Feature Combination Hybrids

Feature combination hybrids use input data from various sources within a unified representation. For example, it is possible to use the features y_i and z_j within matrix factorization using either alignment-based collaborative filtering or regression-constrained factorization. In addition, it is possible to use the features y_i and z_j along latent features in factorization machines. Finally, it is possible to incorporate meta-features.

First, alignment-based collaborative filtering uses similarity conditions. For binary vectors y_i and z_j, these are defined as $\mathcal{S}_c(i) \triangleq \{i' \mid i' \neq i \text{ and } y_i^\top y_{i'} \geq c\}$ and $\mathcal{T}_d(j) \triangleq \{j' \mid j' \neq j \text{ and } z_j^\top z_{j'} \geq d\}$. Then, using these similarity conditions, alignment-based CF (an extension of Nguyen and Zhu (2013)) minimizes the following loss function

$$L_{AB} \triangleq L_{MF} - \lambda_3 \sum_{i=1}^{m} \sum_{i' \in \mathcal{S}_c(i)} \frac{u_i^\top u_{i'}}{|\mathcal{S}_c(i)|} - \lambda_4 \sum_{j=1}^{n} \sum_{j' \in \mathcal{T}_d(j)} \frac{v_j^\top v_{j'}}{|\mathcal{T}_d(j)|}.$$

It is also possible to formulate a smoothed-generalization of alignment-based collaborative filtering (as in Nguyen and Zhu (2013)) that minimizes the following loss function

$$L_{gAB} \triangleq L_{MF} - \lambda_3 \sum_{i=1}^{m} \sum_{i' \in \mathcal{S}_c(i)} \phi(i,i') u_i^\top u_{i'} - \lambda_4 \sum_{j=1}^{n} \sum_{j' \in \mathcal{T}_d(j)} \psi(j,j') v_j^\top v_{j'}$$

with $\phi(i,i') \propto \sigma(\gamma y_i^\top y_{i'} - c)$, $\psi(j,j') \propto \sigma(\delta z_j^\top z_{j'} - d)$ and $\sigma(x) = 1/(1+\exp(-x))$. The weights $\phi(i,i')$ and $\psi(j,j')$ are generally normalized to sum to one, i.e., $\sum_{i' \in \mathcal{S}_c(i)} \phi(i,i') = 1, \forall i = 1, \ldots, m$ and $\sum_{j' \in \mathcal{T}_d(j)} \psi(j,j') = 1, \ \forall j = 1, \ldots, n$.

Second, regression-constrained factorization constrains the latent features within matrix factorization using observed features. For example, regression-constrained factorization using user and item features (an extension of Nguyen and Zhu (2013)) minimizes the following loss function

$$L_{RC} \triangleq \sum_{(i,j) \in \mathcal{O}} \left(r_{ij} - y_i^\top U V^\top z_j\right)^2 + \lambda_1 \|U\| + \lambda_2 \|V\|$$

where U is a $o \times k$ matrix and V is a $p \times k$ matrix.

Third, factorization machines (FMs) (Rendle, 2010) associate each rating r with a $d \triangleq (m+n+o+p)$-dimensional vector x where m is the number of users, n is the number of items, o is the number of user features and p is the number of item features; this vector x is used to indicate the user and item corresponding to rating r, and the associated feature values. Then, using this vector x, the predicted unnormalized[1] rating $\hat{r}(x)$ of a second-order factorization machine is

$$\hat{r}(x) = \mu + \sum_{s=1}^{d} \alpha_s x_s + \sum_{s=1}^{d} \sum_{s'=s+1}^{d} u_s^\top u_{s'} x_s x_{s'} \tag{5.4}$$

where μ is the global bias, α_s is the bias associated with the element x_s, and u_s is a k-dimensional latent vector associated with the element x_s. Note that although the number of interaction terms in FMs might appear large, most of them evaluate to zero when x is sparse; in the case of an unknown user and/or item, even more entries of x are 0. Also note that FMs are optimized with methods similar to the ones used for matrix factorization models.

Finally, it is possible to use meta-features (e.g., number of items rated by a user, number of users that rated this item) extracted from the ratings or other features in order to improve the approaches. For example, it is possible to use these meta-features to refine the weights in Equation 5.3. Since introducing weights β_{ij}^h for each user-item pair (i,j) might lead to overfitting, these weights are constrained to be a function of meta-features, i.e., $\beta_{ij}^h = \sum_{s=1}^{l} v_{hs} x_{ij}^s$ where v_{hs} is the importance of each meta-feature s toward the weight β_{ij}^h and x_{ij}^s is the value of the meta-feature s for the user-item pair (i,j). Thus, the refined Equation 5.3 is given by

$$\hat{r}_{ij} = \sum_{h=1}^{q} \sum_{s=1}^{l} v_{hs} x_{ij}^s \hat{r}_{ij}^h$$

where v_{hs} can again be learned to minimize, for example, the MSE or MAE on a holdout set.

5.3.4 Evaluation of the System

The evaluation of the system is a critical step, especially if we need to do model selection or adjust hyperparameters. While this evaluation can generally be done online or offline, in this work, we focus on offline evaluation on a batch of data. We now describe different

[1] Refer to Section 5.2.2 for a reminder of what is meant by an unnormalized/normalized rating.

evaluation goals, the design of the evaluation process and different evaluation metrics which we categorize as accuracy, ranking and decision support metrics.

Evaluation Goals

There exist several evaluation goals. While the most popular and the most objective goal is the accuracy of the predictions, other evaluation goals such as novelty, confidence, trust, coverage or serendipity are covered in the literature (Aggarwal, 2016). In addition to accuracy, two important evaluation goals with respect to this work are confidence and trust.

On one hand, confidence measures the "system's faith in the evaluations" (Aggarwal, 2016). For example, if the predictions are given with confidence intervals, then it is possible to evaluate the coverage and the width of these confidence intervals and compare them between different systems. These confidence intervals can be obtained by bootstrapping the system (Efron, 1981). Otherwise, they can also be obtained if we use a probabilistic graphical model instead of a discriminative model. For example, there exist several probabilistic models for matrix factorization: Salakhutdinov and Mnih (2007), Xin and Steck (2011), and Hernandez-Lobato, Houlsby, and Ghahramani (2014).

On the other hand, trust measures the "user's faith in the evaluations" (Aggarwal, 2016). This goal is much harder to evaluate objectively but it can be improved if, for example, logical explanations of the predictions are given. While interpreting the features based model is relatively straightforward, it is not as easy to do with matrix factorization. Fortunately several approaches have been proposed that either explain the matrix factorization solutions or that replace standard matrix factorization with an explainable matrix factorization model: Hyvönen, Miettinen, and Terzi (2008), Brun, Aleksandrova, and Boyer (2014), Sanchez et al. (2015), Carmona and Riedel (2015), and Heckel et al. (2017).

Evaluation Design

When evaluating a system offline, it is important not to evaluate it on the same data as it has been trained. Generally, different systems (e.g., different models and different hyperparameters such as regularization parameters) are trained on a training data set, then the best system is determined using a validation data set and finally this best system is evaluated on a testing data set. These training, validation and testing data sets are disjoint data sets. They are generally splitted along a 2:1:1 ratio for training, validation and testing with respect to the observed ratings. It is important that the testing data set is used

only one time at the end to evaluate the performance of the final system on unseen data, otherwise it won't give an unbiased estimate. On the contrary, the data used for training and validation can be reused multiple times. For example, with k-fold cross-validation, the data without the testing set is splitted into k folds. Then, $k-1$ folds are used to train the systems while the remaining fold is used for validation. After repeating this process for the k different validation sets, the results are averaged. The best system is selected with respect to the average performance and its variability.

Accuracy Metrics

The accuracy metrics are generally based on the following equation:

$$\bar{e} = \left[\frac{\sum_{(i,j)\in\mathcal{E}} w_{ij} |e_{ij}|^{\gamma}}{\sum_{(i,j)\in\mathcal{E}} w_{ij}} \right]^{1/\gamma}$$

where $e_{ij} = r_{ij} - \hat{r}_{ij}$ is the prediction error (also known as the residual), γ is a constant, w_{ij} is a weight and $\mathcal{E}$ is a data set on which we evaluate (e.g., test set).

Two popular values for γ are 1 and 2 which gives the following metrics when $w_{ij} = 1 \; \forall (i,j) \in \mathcal{E}$:

$$MAE = \frac{1}{|\mathcal{E}|} \sum_{(i,j)\in\mathcal{E}} |e_{ij}| \tag{5.5}$$

$$RMSE = \sqrt{\frac{1}{|\mathcal{E}|} \sum_{(i,j)\in\mathcal{E}} |e_{ij}|^2} \tag{5.6}$$

where *MAE* denote the mean absolute error and *RMSE* denote the root mean squared error. Note that here we discuss RMSE instead of MSE since the units of RMSE are the same as MAE. RMSE is good to penalize large deviations while MAE is robust to outliers.

It is also possible to normalize these metrics such that their values are within the range $[0, 1]$:

$$NMAE = \frac{MAE}{r_{max} - r_{min}}$$

$$NRMSE = \frac{RMSE}{r_{max} - r_{min}}$$

where *NMAE* denote the normalized MAE and *NRMSE* denote the normalized RMSE.

It is also possible to average these metrics on each item (or user) to correct for the impact of the long tail, i.e., to give the same weight in the metric to items (users) with few ratings than to items (users) with many ratings. Let $\mathcal{E}_{\mathcal{J}} \triangleq \{j \in \mathcal{J} \mid \exists i, (i,j) \in \mathcal{E}\}$ and $\mathcal{E}_{\mathcal{I}}(j) \triangleq \{i \in \mathcal{I} \mid (i,j) \in \mathcal{E}\}$, then these average metrics with respect to the items are:

$$\textit{Average MAE} = \sum_{j \in \mathcal{E}_{\mathcal{J}}} \frac{1}{|\mathcal{E}_{\mathcal{I}}(j)|} \sum_{i \in \mathcal{E}_{\mathcal{I}}(j)} |e_{ij}|$$

$$\textit{Average RMSE} = \sum_{j \in \mathcal{E}_{\mathcal{J}}} \sqrt{\frac{1}{|\mathcal{E}_{\mathcal{I}}(j)|} \sum_{i \in \mathcal{E}_{\mathcal{I}}(j)} |e_{ij}|^2}.$$

Note that it is also possible to put more weight on some items (users) depending on the importance or utility of them.

Finally, if needed, it is possible to weight asymmetrically the error, i.e., to penalize more when the predicted rating is smaller or larger than the true rating. For example, similarly to the loss function used within expectile matrix factorization (Zhu et al., 2017), it is possible to define the following asymmetric metrics:

$$\textit{Asymmetric MAE} = \frac{1}{|\mathcal{E}|} \sum_{(i,j) \in \mathcal{E}} |e_{ij}||\omega - I_{e_{ij}<0}|$$

$$\textit{Asymmetric RMSE} = \sqrt{\frac{1}{|\mathcal{E}|} \sum_{(i,j) \in \mathcal{E}} |e_{ij}|^2 |\omega - I_{e_{ij}<0}|}$$

where $\omega \in (0,1)$ is a chosen constant and $I_{e<0}$ is the indicator function that returns 1 if $e < 0$ and 0 otherwise. In these metrics, the positive errors e_{ij} are weighted by ω and the negative errors by $1 - \omega$. Note that if $\omega = 0.5$ then these metrics are symmetric. Also note that, if such metrics are desirable, then it would also be beneficial to make the loss functions associated with Equations 5.1 and 5.2 asymmetric.

These asymmetric metrics are useful with skewed data set (here, with respect to the rating value) since a symmetric metric, such as RMSE, is impacted by the outliers in the long tail. In addition, using these asymmetric metrics is useful when we want to improve our predictions in the lower or upper range. For example, in this work, we want to recommend treatments which lead to a low severity score (i.e., effective treatments) and it happens that there are few effective and many ineffective treatments. In this case, we do not necessarily care about good predictions of the high scores. So, we could set $\omega = 0.1$ in order to get better predictions for the low scores since we penalize more the negative residuals (Zhu et al., 2017). This setting would also make our recommendations more

conservative since we do not want to predict a lower score than the true score. We can also obtain somewhat similar results by using the ranking and decision support metrics described next.

Ranking Metrics

One interesting and intuitive ranking metric is the fraction of concordant pairs (FCP) (Koren and Sill, 2013) which looks at the number of concordant and discordant pairs between the predicted and true ratings. Let $\mathcal{E}_\mathcal{I} \triangleq \{i \in \mathcal{I} \mid \exists j, (i,j) \in \mathcal{E}\}$ and $\mathcal{E}_\mathcal{J}(i) \triangleq \{j \in \mathcal{J} \mid (i,j) \in \mathcal{E}\}$, then the numbers of concordant, n_c^i, and discordant, n_d^i, pairs for user i are:

$$n_c^i = |\{(j,j') \mid j,j' \in \mathcal{E}_\mathcal{J}(i),\ \hat{r}_{ij} > \hat{r}_{ij'} \text{ and } r_{ij} > r_{ij'}\}|$$
$$n_d^i = |\{(j,j') \mid j,j' \in \mathcal{E}_\mathcal{J}(i),\ \hat{r}_{ij} \geq \hat{r}_{ij'} \text{ and } r_{ij} < r_{ij'}\}|$$

Summing over all users, we obtain $n_c = \sum_{i \in \mathcal{E}_\mathcal{I}} n_c^i$ and $n_d = \sum_{i \in \mathcal{E}_\mathcal{I}} n_d^i$ and can finally compute the FCP as:

$$FCP = \frac{n_c}{n_c + n_d}. \tag{5.7}$$

Note that the FCP always lies in the range $[0, 1]$.

However, this previous metric places the same emphasis on all pairs whatever their relevance. It also only looks at the pairs within a list and not at the list itself. One popular ranking that also looks at the relevance of the items in a listwise fashion is the normalized discounted cumulative gain (NDCG) (Järvelin and Kekäläinen, 2002). There exists several variants of this metric in the literature. The variant we use in this work is the following:

$$NDCG = \frac{1}{|\mathcal{E}_\mathcal{I}|} \sum_{i \in \mathcal{E}_\mathcal{I}} \frac{DCG_i}{IDCG_i} \tag{5.8}$$

where the discounted cumulative gain (DCG) for user i is defined as

$$DCG_i = \sum_{j \in \mathcal{E}_\mathcal{J}(i)} \frac{2^{rel_{ij}} - 1}{\log_2(\hat{\pi}_{ij} + 1)}.$$

In DCG, rel_{ij} consists in the true relevance of an item j to user i (e.g., often the true rating r_{ij} if it is non-negative and if a higher rating value denotes a higher relevance), and $\hat{\pi}_{ij}$ is the estimated rank of the item j in $\mathcal{E}_\mathcal{J}(i)$ (e.g., according to the predicted rating $\hat{r}_{ij}$).[2] The ideal discounted cumulative gain (IDCG) consists in replacing $\hat{\pi}_{ij}$ by the ideal ranking π_{ij}

[2] In the case of ties, we follow the approach of McSherry and Najork (2008).

that is obtained using the relevances rel_{ij}. Note that the NDCG does not penalize directly the ranking of bad items (i.e., items with a relevance of 0); it focuses on ranking correctly the highly relevant items in the top ranking positions. Also note that the NDCG always lies in the range $[0, 1]$.

Decision Support Metrics

Let τ be some threshold that differentiates the good items from the bad ones, and let $t_{ij} = 1$ (respectively $t_{ij} = 0$) indicate a good (bad) item according to

$$t_{ij} = \begin{cases} 1 \text{ if } r_{ij} \geq \tau, \\ 0 \text{ otherwise .} \end{cases}$$

Also let $\hat{t}_{ij}$ be defined similarly using $\hat{r}_{ij}$ instead of r_{ij}. Then, it is possible to define the precision and recall as:

$$\begin{aligned} \textit{precision} &= \frac{tp}{tp + fp} \\ \textit{recall} &= \frac{tp}{tp + fn} \end{aligned}$$

where the number of true positives, false positives and false negatives are computed as

$$\begin{aligned} tp &= \sum_{(i,j)\in\mathcal{E}} I_{t_{ij}=1} I_{\hat{t}_{ij}=1}, \\ fp &= \sum_{(i,j)\in\mathcal{E}} I_{t_{ij}=0} I_{\hat{t}_{ij}=1}, \\ fn &= \sum_{(i,j)\in\mathcal{E}} I_{t_{ij}=1} I_{\hat{t}_{ij}=0}. \end{aligned}$$

These two metrics focus on different aspects; focusing on precision removes bad items from our recommendations while focusing on recall includes good items into our recommendations. Note that these are not the typical definitions of precision and recall in the RS literature; for the typical definitions, refer to Section 7.5.4 of Aggarwal (2016).

Since it is generally not preferable to focus on only one of these two metrics, precision and recall are often combined using the following F1 measure (Chinchor, 1992)

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}. \tag{5.9}$$

This measure puts equal emphasis on both the precision and the recall. Note that this measure doesn't care about the ranks, it just care about identifying the good items from the bad ones. Also note that we set $F_1 = 0$ when $tp + fp = 0$ and/or $tp + fn = 0$. The F1 measure lies in the range $[0, 1]$.

5.3.5 Confounding

An issue with offline training and evaluation of a recommender system on observational data consists in the biases introduced by confounders. For example, users might be exposed only to items which they tend to like, or patients might only be provided treatments with a high probability of success. Thus, in these cases, most items with observed ratings should have better ratings than the (unobserved) ratings of the other items and, hence, RSs trained naively on this observational data provide poor predictions for these unobserved ratings, the ones that we are interested in. Yet, the RS literature addressing this issue is relatively new (e.g., Marlin and Zemel (2009), Hernandez-Lobato, Houlsby, and Ghahramani (2014), Liang, Charlin, and Blei (2016), Schnabel et al. (2016), and Wang et al. (2018)).

In this work, we follow the *deconfounded recommender* approach of Wang et al. (2018). For all user-item pairs (i.e., $\forall (i, j) \in \mathcal{I} \times \mathcal{J}$), let respectively $r_{ij}(0)$ and $r_{ij}(1)$ denote the potential outcomes of the Neyman-Rubin potential-outcome framework (Splawa-Neyman, Dabrowska, and Speed, 1923; Rubin, 1974); here, $r_{ij}(0)$ denotes the rating that user i would provide to item j if this user *isn't exposed* to item j while $r_{ij}(1)$ denotes the rating that user i would provide to item j if this user *is exposed* to item j. Note that these potential outcomes are not all necessarily observed. In fact, this RS setting is in contrast to most causal inference applications where exactly one of the two potential outcomes is observed. Here, we only observe a few of the potential outcomes $r_{ij}(1)$ (i.e., if $a_{ij} = 1$, then we observe $r_{ij}(1) = r_{ij}$) and none of the potential outcomes $r_{ij}(0)$.

Unfortunately, not observing the potential outcomes $r_{ij}(0)$ removes, among other things, the ability to identify whether a patient would be better off without any treatments. Yet, it is still possible to identify the best treatment for a patient. The identification of this best treatment consists in finding the best individualized treatment effect (ITE) for a particular patient i over all treatment j where the ITE is defined as $ITE_{ij} = r_{ij}(1) - r_{ij}(0)$. Now, since $r_{ij}(0)$ is equivalent to the outcome under no treatment, it is reasonable to assume the same outcome whatever the omitted treatment for a particular patient i, i.e., $r_{ij}(0) = r_{ij'}(0),\ \forall j, j' \in \mathcal{J}$. Hence, the identification of the best treatment for a particular patient i can be done by only looking at the potential outcomes $r_{ij}(1)$, and, fortunately, the

unobserved potential outcomes $r_{ij}(1)$ can be recovered with the following theorem since several patients received several treatments.

Theorem 2 (Theorem 1 of Wang et al. (2018))**.** *Assume stable unit treatment value assumption (SUTVA), single ignorability, consistency of substitute confounders and overlap of a subset of causes. Then, the deconfounded recommender forms unbiased causal inference*

$$\mathbb{E}\left[r_{ij}(a)\right] = \mathbb{E}\left[\mathbb{E}\left[r_{ij}(a_{ij}) \mid a_{ij} = a, \eta_i^\top \theta_j\right]\right], \quad \forall (i,j) \in \mathcal{I} \times \mathcal{J}$$

when $a_{ij} \sim Poisson(\eta_i^\top \theta_j)$ *for some independent random vectors* η_i*'s and* θ_j*'s.*

In other words, under the previous assumptions, the deconfounded recommender estimates the unobserved potential outcomes $r_{ij}(1)$ with the use of the propensities $\hat{a}_{ij}$, i.e., the probabilities of exposure; these propensities can be obtained with any factor model such as the previous Poisson factorization (PF) model.[3] We now describe the different steps of our deconfounded recommender implementation:

1. We do cross-validation on the observed exposure for all users in the training set, $\mathbf{a}$, to identify the best hyper-parameters of a propensity factor model.

2. After retraining the factor model on the full exposure training set $\mathbf{a}$, the propensities $\hat{a}_{ij}$ are estimated for all entries in the training set.

3. Finally, the outcome model is fitted by minimizing the following weighted loss:

$$L = \sum_{(i,j) \in \mathcal{O}} \frac{1}{\hat{a}_{ij}} l(r_{ij}, \hat{r}_{ij})$$

   where $\hat{r}_{ij}$ is the output the outcome model. Note that this weighted procedure differs from the procedure used in Wang et al. (2018) for the outcome model; it is however acknowledged as an alternative approach in Wang and Blei (2018). Note also that this procedure is similar to inverse probability of treatment weighting (IPTW) in the causal inference literature; it mostly differs from the traditional causal inference approaches by using a factor model (i.e., latent features) to compute the weights.

[3] Refer to Wang et al. (2018) and Wang and Blei (2018) for the proof of Theorem 2 and details about the assumptions.

5.4 Case Study

We now present our TRD case study, in the following order. First, we present the setting and characterize our data set. Then, we describe the methods. Finally, we discuss the associated results.

5.4.1 Setting

In this section, we introduce the setting, i.e., we define the treatment and outcome, and we describe the patient and treatment features.

Definition of Treatment

Our application is different than the standard RS applications with respect to many aspects. In particular, one important difference consists in the definition of the treatment. While it is clear that the patient is the user, it is unclear what the treatment (i.e., the item) is. Should a treatment take into account the antidepressants and the add-on drugs (i.e., drugs that are used to complement the antidepressants)? Should it also take into account the dosage of these drugs and the total time they were taken? While most of these aspects are probably important, it is not practically possible to all take them into account, since we would end up with too many treatments and too few observations for each of them. Thus, with the help of our medical collaborator, we selected two treatment definitions which make sense from the clinical point of view. Note that both of these treatment definitions will be pursued in the remainder of the chapter.

Our treatment definitions consist of two parts: (1) the antidepressants and (2) the add-on drugs. Within our treatment definitions, the antidepressants can be identified as coming from the monoamine oxidase inhibitors (MAOIs) (`AD_MAOI`), serotonin-norepinephrine reuptake inhibitors (SNRIs) (`AD_SNRI`), selective serotonin reuptake inhibitors (SSRIs) (`AD_SSRI`) or tricyclic antidepressants (TCAs) (`AD_TCA`) class, or can be identified as bupropion (`AD_Bupropion`) or mirtazapine (`AD_Mirtazapine`). Note that one or many of these six binary variables can have a true value at the same time for a particular treatment. Note also that a treatment with two SNRIs is considered the same, with these treatment definitions, as a treatment with only one SNRI. Then, our treatment definitions include some of the add-on drugs used. In particular, we plan on trying two alternatives to indicate these add-on drugs. On the one hand, one alternative consists in identifying many of these add-on drugs by categories. In this case, we have the antipsychotic (`AO_Antipsychotic`),

anxiolytic (AO_Anxiolytic), hypnotic (AO_Hypnotic) and stimulant (AO_Stimulant) categories, and the aripiprazole (AO_Aripiprazole), liothyronine (AO_Liothyronine) and lithium (AO_Lithium) drugs that are identified individually. On the other hand, another alternative consists in identifying only the most prescribed add-on drugs without any categories, i.e., aripiprazole (AO_Aripiprazole), atomoxetine (AO_Atomoxetine), liothyronine (AO_Liothyronine), lithium (AO_Lithium), lurasidone (AO_Lurasidone), methylphenidate (AO_Methylphenidate), modafinil (AO_Modafinil), pramipexole (AO_Pramipexole), quetiapine (AO_Quetiapine) and trazodone (AO_Trazodone). It is important to note that we categorize trazodone in this study as an add-on drug, even tough it is an antidepressant, since it is mostly used as an hypnotic at our collaborating clinic and not as an antidepressant. Also note that many binary variables indicating add-on categories and drugs can again be present at the same time. See Appendix C.1 for the list of drugs within each antidepressant class and add-on category.

In summary, our definitions of the treatment can be seen as a vector of binary values where each of these binary values indicate the presence or absence of a component within the treatment. Then, two treatments are considered to be the same if they have the same binary representations. A list of the binary variables considered within the two treatment definition alternatives is given in Table 5.3.

Definition of Outcome

In this work, we define the outcome for a patient-treatment pair as the minimum Quick Inventory of Depressive Symptomatology (QIDS) score (Rush et al., 2003) for this patient in the period that begins 28 days after the treatment and ends at the next *different* treatment. The QIDS score used in this study is the 16-item self-reported score, a score often used in the medical literature to evaluate the severity of MDD and TRD. This score lies between 0 and 27 with a higher score denoting a worse outcome; remission is often defined as a QIDS score less or equal to 5 (Rush et al., 2006). Also, the severity of depression is often determined with respect to different thresholds of this score (Rush et al., 2006): normal (0–5), mild (6–10), moderate (11–15), severe (16–20) and very severe (21–27). Note that the period of observation begins 28 days after the treatment to let the treatment shows its full potential. Also note that we use the minimum score in the period instead of the mean or median since we want to predict the best potential outcome under a treatment. By using the mean or median score, we would obtain a prediction that is influenced by the period of time that a patient undergoes a particular treatment; it is often the case that a patient relapses after responding to this same treatment. Finally, note that the QIDS score corresponds to

TABLE 5.3: List of binary variables considered in the two treatment definition alternatives.

Variable	Alternative 1	Alternative 2
AD_MAOI	✓	✓
AD_SNRI	✓	✓
AD_SSRI	✓	✓
AD_TCA	✓	✓
AD_Bupropion	✓	✓
AD_Mirtazapine	✓	✓
AO_Antipsychotic	✓	
AO_Anxiolytic	✓	
AO_Hypnotic	✓	
AO_Stimulant	✓	
AO_Aripiprazole	✓	✓
AO_Atomoxetine		✓
AO_Liothyronine	✓	✓
AO_Lithium	✓	✓
AO_Lurasidone		✓
AO_Methylphenidate		✓
AO_Modafinil		✓
AO_Pramipexole		✓
AO_Quetiapine		✓
AO_Trazodone		✓

only one very specific outcome; the broader objective of treatment consists in the general well-being of patients. Thus, when making treatment selection, the physicians need to take into considerations additional elements such as the preferences of the patients with respect to the side-effects of the treatments (e.g., weight gain), the restrictions associated with the treatments (e.g., low tyramine diet) and an acceptable frequency of commuting between his home and the clinic for various appointments (e.g., drug dosage using blood tests). It is however not possible to capture this broader perspective in this study without additional features describing these patients' preferences; hence, this work focuses on the very specific QIDS score as the outcome.

Patient Features

The patient features consist in 4 features (i.e., 1 real number variable and 10 binary variables associated with the 3 other categorical features) collected at the initial appointment of the patient to the clinic and that correspond, to our knowledge, to good predictors of

the response. In particular, the features are the age (Pt_Age) and gender (Pt_Gender_F and Pt_Gender_M) of the patient, and the indication of comorbidities on the first axis (Pt_FirstAxis_N, Pt_FirstAxis_TBI, Pt_FirstAxis_Y and Pt_FirstAxis_NA) and second axis (Pt_SecondAxis_N, Pt_SecondAxis_TBI, Pt_SecondAxis_Y and Pt_SecondAxis_NA). The first and second axis are evaluated respectively through the Structured Clinical Interview for DSM Axis I Disorders (SCID-I) (First et al., 2002) and the Structured Clinical Interview for DSM Axis II Personality Disorders (SCID-II) (First et al., 1997), where DSM denote the Diagnostic and Statistical Manual of Mental Disorders. Note that they are no missing values for age and gender while they are some for the first and second axes (indicated by _NA). Also note that _TBI indicates a possible diagnosis of a comorbidity on the first or second axis that needs to be further investigated. Finally, note that this one-hot encoding of all features (even the binary gender feature) is necessary to improve the predictive performance of our proposed model.[4]

Treatment Features

The treatment features consist in the binary variables of the selected alternative in Table 5.3 and in five additional binary variables that define the desired effects of the add-on categories and drugs. The desired effects for an add-on category or drug can be to increase the effect of antidepressants (AODE_ADBooster), and can be antipsychotic (AODE_Antipsychotic), anxiolytic (AODE_Anxiolytic), hypnotic (AODE_Hypnotic) and stimulant (AODE_Stimulant). Note that while there might be several add-on drugs prescribed, the features for the desired effects are only binary, i.e., the features are not counting the number of add-on drugs given for a particular effect but are rather indicating whether these effects are desired at all or not within the treatment. Also note that to again improve the predictive performance of our proposed model all these binary variables are further one-hot encoded (in _False and _True variables).[5] Finally, note that, in total, there are 36 binary treatment features for alternative 1 and 42 binary treatment features for alternative 2 when counting the one-hot encodings. See Appendix C.1 for the desired effects associated with the add-on categories and drugs.

[4]This one-hot encoding leads to the same number of active binary variables whether you are male or female for example.

[5]For example, the binary variable AD_MAOI $= 0$ is one-hot encoded with AD_MAOI_False $= 1$ and AD_MAOI_True $= 0$.

5.4.2 Data Set

Our data set consists of all the unarchived medical records for adult patients suffering from TRD who started receiving treatment between August 2006 and August 2015 at the DSDP. A patient file is archived when it has not been used for more than a year so that additional storage space is available in the front office for new patient files. Of these 463 patient files, we kept the 364 (respectively 365) patient files with at least one outcome subsequent to a treatment according to our first (second) treatment definition.

The characterization of the 364 patients relevant to our first treatment definition is provided in Table 5.4. Since the characterization of the patients relevant to our second treatment definition is similar, it is provided in Appendix C.2. As detailed in the tables, an average patient at this clinic consists in a female patient of around 40 years old with first- and second-axis comorbidities.

TABLE 5.4: Sample means and standard deviations of the patient features for the first treatment definition. The one-hot encoding of binary variables has been removed to improve readability. †These means and standard deviations are computed using proportions.

Variable	Mean	Std
Pt_Age	43.558	10.615
Pt_Gender_M†	0.368	0.482
Pt_FirstAxis_N†	0.365	0.482
Pt_FirstAxis_TBI†	0.047	0.211
Pt_FirstAxis_Y†	0.453	0.498
Pt_FirstAxis_nan†	0.135	0.341
Pt_SecondAxis_N†	0.297	0.457
Pt_SecondAxis_TBI†	0.239	0.426
Pt_SecondAxis_Y†	0.327	0.469
Pt_SecondAxis_nan†	0.137	0.344

In Table 5.5, the characterization of the treatment variables is provided by taking into account the frequency of the different treatments in the data set. For both treatment definitions, it appears that the most prescribed antidepressants are from the SSRI and SNRI classes. Then, under the first treatment definition, it appears that antipsychotics and hypnotics are often prescribed as add-on drugs. It also appears that the most common desired effects of the add-on drugs are to manage anxiety (anxiolytic) and psychosis (antipsychotic). Finally, under the second treatment definition, it appears that quetiapine, an antipsychotic, is the most prescribed add-on drug, and that the antipsychotic and stimulant effects of the add-on drugs are the most desired. Note that the difference

between these results are due to the treatment definitions; there are 393 different observed treatments and 1401 patient-treatment pair's observed outcomes under the first definition of treatment while there are 287 different observed treatments and 1271 patient-treatment pair's observed outcomes under the second definition of treatment within the data set.

TABLE 5.5: Sample means and standard deviations of the treatment features in the two treatment definition alternatives. The one-hot encoding of binary variables has been removed to improve readability. All means and standard deviations are computed using proportions, and take into account the frequency of treatments in the data set.

(A) Alternative 1

Variable	Mean	Std
AD_MAOI	0.012	0.109
AD_SNRI	0.348	0.476
AD_SSRI	0.435	0.496
AD_TCA	0.138	0.345
AD_Bupropion	0.269	0.443
AD_Mirtazapine	0.110	0.313
AO_Antipsychotic	0.398	0.490
AO_Anxiolytic	0.198	0.398
AO_Hypnotic	0.258	0.438
AO_Stimulant	0.183	0.386
AO_Aripiprazole	0.146	0.353
AO_Liothyronine	0.024	0.152
AO_Lithium	0.105	0.306
AODE_ADBooster	0.126	0.332
AODE_Antipsychotic	0.490	0.500
AODE_Anxiolytic	0.516	0.500
AODE_Hypnotic	0.258	0.438
AODE_Stimulant	0.304	0.460

(B) Alternative 2

Variable	Mean	Std
AD_MAOI	0.011	0.104
AD_SNRI	0.345	0.475
AD_SSRI	0.430	0.495
AD_TCA	0.141	0.348
AD_Bupropion	0.265	0.441
AD_Mirtazapine	0.110	0.313
AO_Aripiprazole	0.149	0.357
AO_Atomoxetine	0.009	0.097
AO_Liothyronine	0.026	0.159
AO_Lithium	0.108	0.310
AO_Lurasidone	0.009	0.093
AO_Methylphenidate	0.149	0.357
AO_Modafinil	0.025	0.157
AO_Pramipexole	0.020	0.139
AO_Quetiapine	0.286	0.452
AO_Trazodone	0.105	0.307
AODE_ADBooster	0.131	0.338
AODE_Antipsychotic	0.394	0.489
AODE_Anxiolytic	0.292	0.455
AODE_Hypnotic	0.105	0.307
AODE_Stimulant	0.314	0.464

With respect to the outcomes, the density of the rating matrix is 0.979% (respectively 1.213%) with a mean outcome score of 11.8 (11.7) and a standard deviation of 6.5 (6.6) for alternative 1 (alternative 2). Thus, only few patient-treatment pairs are observed with respect to all possible combinations of the observed patients and treatments, and on average the best observed outcomes are around the middle of the QIDS range.[6] For both

[6]Remember that the outcomes in the ratings matrix are the minimum for each patient-treatment pair.

alternatives, about 20% of all observed patient-treatment outcomes are less or equal to 5, i.e., corresponds to remission. In addition, about 51% of the users and 34% of the items have at least one observed outcome that corresponds to remission. This demonstrates that most patient-treatment matches do not lead to remission under the current practice. In addition, only half of the patients experience remission at some point and only one third of the treatments lead at least one patient to remission.

When looking at the most frequent treatments (see Figure 5.1 and C.1), there is a wide range in their associated outcomes; thus the most prescribed treatments do not necessarily lead to remission. This is the case even though all results in this section are biased towards better patient-treatment matches, when assuming reasonably that the treatment selection done by physicians is better than uniformly random treatment selection.

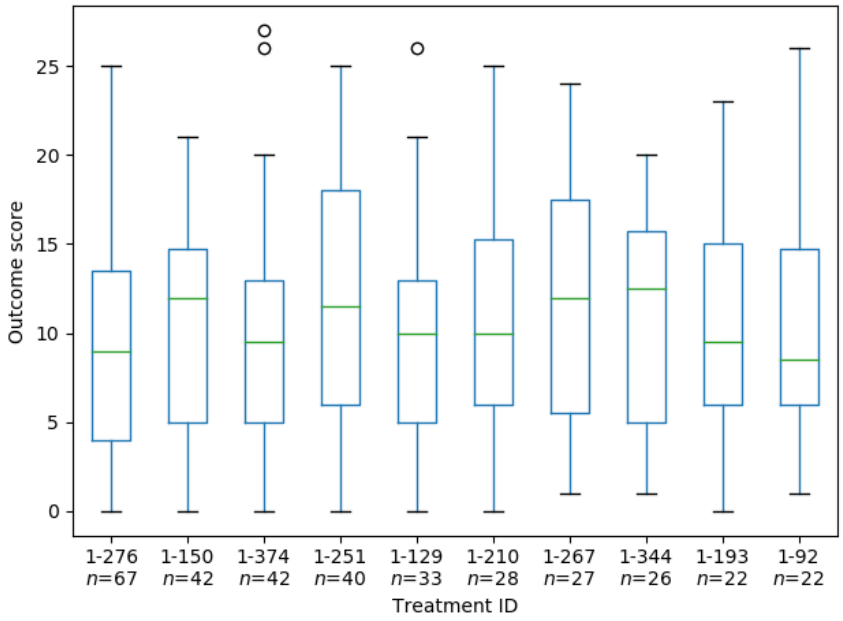

FIGURE 5.1: Boxplot of outcome score for the 10 most frequent treatments under the first definition of treatment.

Even when looking at the treatments with lowest 95^{th} percentile for outcome score and at least three observations (see Figure 5.2 and C.2), treatments do not consistently lead to remission. Hence, no treatment appears to consistently lead to remission for all patients, even under the bias towards better patient-treatment matches; this partially addresses our third research question. Histograms showing the frequency of treatment for different number of observed outcomes are provided in Appendix C.2; these histograms show that most treatments are prescribed only a few times.

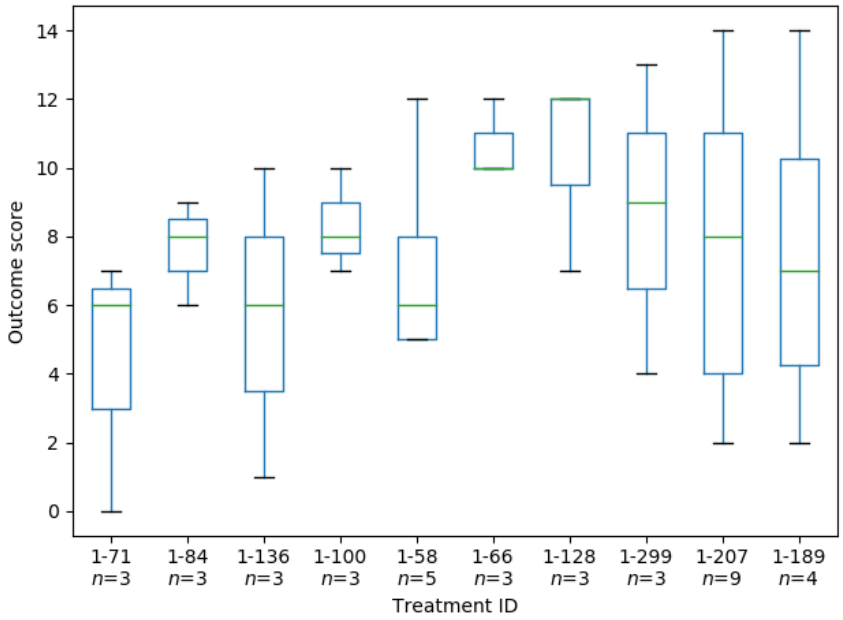

FIGURE 5.2: Boxplot of outcome score for the 10 treatments with lowest 95th percentile for outcome score under the first definition of treatment.

Finally, it appears that among the patients which experienced remission (i.e., have at least one outcome score less or equal to 5) about 5 patients required up to seven times more ineffective treatments than the number of effective treatments received while others received an effective treatment on their first trial (see Figure 5.3 and C.5); here effective (ineffective) denotes a treatment that leads to remission (doesn't lead to remission). This could be explained by some patients being easier to treat (or just having good luck with their treatment selection). Figure 5.4 and C.6 also reemphasize this point; it is easy to see that some patients are still not able to get into remission even after many treatment trials. Histograms showing the frequency of patient for different number of observed outcomes are provided in Appendix C.2; these histograms show that most users are prescribed only a few treatments.

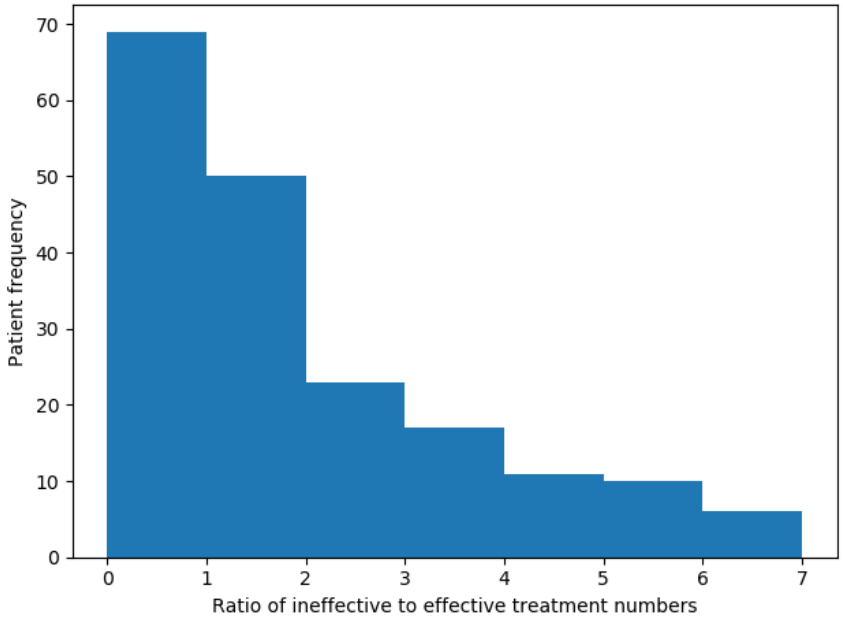

FIGURE 5.3: Histogram of the ratio of ineffective to effective treatment numbers for patients with at least one remission under the first definition of treatment.

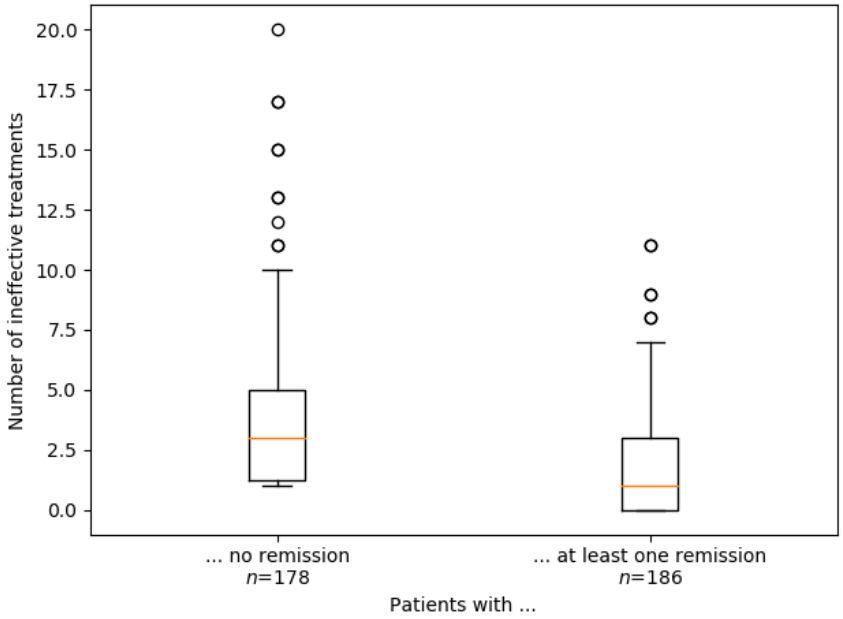

FIGURE 5.4: Boxplot of the number of ineffective treatments for patients with no remission and at least one remission under the first definition of treatment.

5.4.3 Methods

We now describe the procedure to obtain the main results of this chapter, i.e., whether it appears possible to accurately predict the outcome of a particular treatment on a particular patient.

First, we randomly split the data in a training and test set with 75% of the ratings in the training set and 25% of the ratings in the test set. Hence, there might be some users and items from the test set that are unobserved in the training set.

Second, since we believe that the assumptions of Theorem 2 hold in our setting, we compute the IPTW weights as described in section 5.3.5 on the full training set. To compute the propensities, we use a FM model (Equation 5.4) with a sigmoid on the output

$$\hat{a}_{ij} = \sigma\left(\mu + \sum_{s=1}^{d} \alpha_s x_s + \sum_{s=1}^{d} \sum_{s'=s+1}^{d} u_s^\top u_{s'} x_s x_{s'}\right)$$

where $\sigma(x) = 1/(1+\exp(-x))$. The weights obtained by taking the inverse of the propensities are then used to fit the IPTW variants of the outcome models in the cross-validation phase and the retraining phase on the full training set; both phases are described below. The implementation details of the propensity model are provided in Appendix C.3.

Third, we do cross-validation on the training set in order to identify the best hyperparameters for each of the following outcome models.

- Constant: Model that returns the estimated global mean, $\hat{\mu}$, of the training set as the prediction.
- Baseline: Model that uses the global mean with the user and item biases, i.e., $\hat{r}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j$, of the training set as the prediction (see section 2.1 of Koren, 2010).
- FM-base: Second-order factorization machine (Equation 5.4) with one-hot vectors identifying the corresponding user and item in the feature vector x; this is equivalent to biased matrix factorization. Note that the size of the feature vector x is dependent on the training set.
- FM-features: FM-base model with the addition of patient and treatment features (see Section 5.4.1) to the feature vector x.
- FM-outcomes: FM-base model with the addition of a vector providing the other treatment outcomes divided by the number of other treatment outcomes to the feature

vector x (i.e., for user i and item j, the additional features consist of $(\bar{r}_{i1}, \bar{r}_{i2}, \ldots, \bar{r}_{in})$ where $\bar{r}_{ij'} = r_{ij'} / \left(\sum_{j=1}^{n} a_{ij} - 1\right)$ if $j' \neq j$ and $(i, j') \in \mathcal{O}$, otherwise $\bar{r}_{ij'} = 0$).[7]

- FM-full: FM-base model with the additional features of both FM-features and FM-outcomes to the feature vector x.

We mostly use the factorization machine models since they generalize several models of the literature by only modifying the feature vector x and since they can easily take into account external data. Note that each of these models are trained using the uncorrected training set (i.e., no prefix) and the IPTW weighted training set (i.e., IPTW prefix); in the case of the IPTW-Constant model it is the weighted global mean that is returned while in the case of the other models it is the weighted loss that is optimized. The implementation details of these models as well details regarding the random search over their hyper-parameters are provided in Appendix C.3.

Finally, we compare all of these models on different test sets using the following metrics.

- Root mean squared error (RMSE): Equation 5.6.
- Mean absolute error (MAE): Equation 5.5.
- Fraction of concordant pairs (FCP): Equation 5.7.
- Normalized discounted cumulative gain (NDCG): Equation 5.8. Note, that in our setting, a higher rating denotes a worst outcome. Thus, for the relevance function, we use $rel_{ij} = 27 - r_{ij}$. We also order the items in increasing order of $\hat{r}_{ij}$ to obtain $\hat{\pi}_{ij}$.
- F1 score: Equation 5.9. The good items are the ones with a rating score less than or equal to 5 while the other items are considered bad.

In total, three different test sets of the same size are used: 1 random test set and 2 intervened test sets. The random test set consists in the test set described at the beginning of this section. The first intervened test set is constructed by sampling with replacement each (i, j) entry from the random test set according to the probability $p(i, j) \propto 1/(1 + freq_i)$ where $freq_i$ is the frequency of item i in the training set. The second intervened test set is constructed by sampling with replacement each (i, j) entry from the random test set according to the probability $p(i, j) \propto 1/(1 + r_{ij})$. These additional intervened test sets are used to analyze the models' performance on infrequent items and items with a low rating; remember that a low rating is preferable.

[7] Note that all algorithms use the ratings offsetted to the 1–28 scale. Thus, setting $\bar{r}_{ij'} = 0$ is equivalent to no rating instead of a rating of zero.

5.4.4 Main Results

Following the steps of the previous section, we split the data into a training set of 1050 (respectively 953) ratings and a random test set of 351 (318) ratings for the first (second) definition of treatment. We also compute the IPTW weights for the training sets of both definitions of treatment. Then, after doing cross-validation, we obtain the following results on the different test sets for each of the best models. Table 5.6, 5.7 and 5.8 provide the results for the first definition of treatment on the random test set and the two intervened test sets. The results for the second definition of treatment are provided in Table C.4, C.5 and C.6.

With respect to the results on the random test set (see Table 5.6 and C.4), it appears that different metrics benefit different models, and these models that benefit are not necessarily the same in both definitions of treatment. Second, it appears that Constant and IPTW-Constant do relatively well on the NDCG for both definitions of treatment even if their predictions are always the same. Note also that the range of values for NDCG is not that big even for differences of 1.5 on the RMSE. Thus, it appears that NDCG, as defined in this case study, is not that useful as a metric; we omit it from later discussions. Third, it appears that most of the models under both treatment definitions obtain a value of zero for F1. This is probably due to a threshold with a value that is too extreme in order to define the good/bad classes in F1. Thus, it appears that F1, as defined in this case study, is not that useful as a metric; we also omit it from later discussions. Finally, for the RMSE, MAE and FCP metrics, it appears that FM-full obtains some of the best results under the first definition of treatment while IPTW-FM-full obtains some of the best results under the second definition of treatment. Yet, the absolute results obtained under the second definition of treatment are somewhat worst for the RMSE and MAE but somewhat better for the FCP than under the first treatment definition.

With respect to the results on the first intervened test set (see Table 5.7 and C.5), it appears that IPTW-Baseline performs best; FM-full (under the first treatment definition) and IPTW-FM-full (under the second treatment definition) are however not too far with respect to RMSE and MAE while providing a better FCP metric. Second, note that IPTW-FM-full still appears not to perform well under the first treatment definition as is also the case on the random test set. Finally, it appears that the IPTW variant of most of the models performs better on the RMSE and MAE metrics; this is logical given the definition of this first intervened test set. Yet, the IPTW variants appear to perform worst on the FCP.

With respect to the results on the second intervened test set (see Table 5.8 and C.6), it appears that no model performs best overall. However, the IPTW variants appear

TABLE 5.6: RMSE, MAE, FCP, NDCG and F1 metrics of the Constant, Baseline, FM-base, FM-features, FM-outcomes and FM-full models, and their IPTW variants on the random test set under the first definition of treatment.

Model	RMSE	MAE	FCP	NDCG	F1
Constant	6.47	5.32	0.000	0.911	0.000
IPTW-Constant	6.47	5.32	0.000	0.911	0.000
Baseline	5.17	4.17	0.422	0.905	0.000
IPTW-Baseline	5.15	4.11	0.405	0.906	0.054
FM-base	5.94	4.80	0.427	0.905	0.000
IPTW-FM-base	5.92	4.81	0.379	0.902	0.000
FM-features	5.43	4.44	0.470	0.905	0.027
IPTW-FM-features	5.42	4.44	0.474	0.910	0.080
FM-outcomes	7.05	5.52	0.418	0.915	0.457
IPTW-FM-outcomes	6.38	4.83	0.405	0.906	**0.460**
FM-full	**4.96**	**3.96**	0.453	**0.914**	0.365
IPTW-FM-full	5.30	4.21	**0.483**	0.905	0.000

somewhat better than the non-IPTW variants. This is again logical since ratings with small values are infrequent in the data set. Yet, note that IPTW-FM-full still performs badly under the first treatment definition. Finally, note that the absolute results are worst in the second intervened test set than in other test sets. This might be due to the ratings with small values being harder to predict.

Overall, we would like to highlight that FM-full (under the first treatment definition) and IPTW-FM-full (under the second treatment definition) appear as reasonable models. We hypothesize that IPTW-FM-full (under the first treatment definition) is not performing well due to incorrect IPTW weights; note that the model fitting the weights deals with about a 100 times more data and it is thus computationally expensive to execute several randomized search iterations. While these models do not improve the RMSE and MAE metrics so much with respect to the Constant and IPTW-Constant models, they do provide some ordering of the treatments; even if the FCP values are not that great, these orderings can still make sense when the ratings values are low and the differences in ratings are significant. In addition, while the RMSE and MAE values appears somewhat high, they correspond to about the range of one QIDS severity category. Remember that these categories are often defined as: normal (0–5), mild (6–10), moderate (11–15), severe (16–20) and very severe (21–27).

To conclude this section, we would like to try to answer the following question: Is it possible to accurately predict the outcome of a particular treatment on a particular patient?

TABLE 5.7: RMSE, MAE, FCP, NDCG and F1 metrics of the Constant, Baseline, FM-base, FM-features, FM-outcomes and FM-full models, and their IPTW variants on the first intervened test set under the first definition of treatment. This first intervened test set is sampled proportionally to the inverse frequency of items in the training set.

Model	**RMSE**	**MAE**	**FCP**	**NDCG**	**F1**
Constant	6.63	5.39	0.000	0.928	0.000
IPTW-Constant	6.63	5.39	0.000	0.928	0.000
Baseline	5.08	4.06	0.243	0.938	0.000
IPTW-Baseline	**4.99**	**3.88**	0.226	**0.940**	0.000
FM-base	6.12	4.95	0.241	0.929	0.000
IPTW-FM-base	6.11	4.95	0.201	0.920	0.000
FM-features	5.29	4.35	**0.589**	0.933	0.029
IPTW-FM-features	5.25	4.31	0.519	**0.940**	0.058
FM-outcomes	6.57	5.31	0.180	0.919	**0.273**
IPTW-FM-outcomes	6.01	4.78	0.223	0.924	0.234
FM-full	5.17	4.18	0.444	0.928	0.141
IPTW-FM-full	5.52	4.36	0.373	0.910	0.000

While it appears somewhat possible to improve over simply predicting the mean outcome, this task is not easy. In particular, with our limited data set, we improve by about 1 to 1.5 points the RMSE and MAE. Thus, additional data and research are necessary in order to obtain a model that can be used as a decision aid. Still, we proceed cautiously with additional results in the following section using the IPTW-FM-full model.

TABLE 5.8: RMSE, MAE, FCP, NDCG and F1 metrics of the Constant, Baseline, FM-base, FM-features, FM-outcomes and FM-full models, and their IPTW variants on the second intervened test set under the first definition of treatment. This second intervened test set is sampled proportionally to the inverse rating.

Model	**RMSE**	**MAE**	**FCP**	**NDCG**	**F1**
Constant	8.36	7.46	0.000	0.958	0.000
IPTW-Constant	8.39	7.49	0.000	0.958	0.000
Baseline	7.08	6.12	0.441	0.953	0.000
IPTW-Baseline	6.98	5.99	0.315	0.955	0.068
FM-base	7.50	6.53	0.409	0.958	0.000
IPTW-FM-base	7.45	6.52	0.291	0.957	0.000
FM-features	7.13	6.16	0.339	0.950	0.010
IPTW-FM-features	6.91	5.98	0.346	0.952	0.113
FM-outcomes	6.20	4.56	0.425	**0.961**	**0.687**
IPTW-FM-outcomes	6.09	**4.39**	0.291	0.947	0.636
FM-full	**5.78**	4.87	0.457	0.953	0.500
IPTW-FM-full	6.98	5.98	**0.504**	0.953	0.000

5.4.5 Secondary Results

Using the IPTW-FM-full model from the previous section (still fitted on the full training set), we now try to answer the following additional questions. Again note that these results should be taken cautiously given the limited quality of the results in the previous section. Also note that further replications of this study, ideally in RCTs, are needed to fully answer these questions.

Do some treatments consistently provide good response for all patients?

To answer this question, the missing entries of the rating matrix are predicted with the IPTW-FM-full-model. Then, after filling the missing entries with the predictions and keeping the true ratings for the other entries, we obtain the results in Figure 5.5 and 5.6 where the top 10 treatments with respect to the lowest 95th percentile are shown. In these figures, it appears that no treatment consistently provides good response for all patients, i.e., no treatments provide an outcome less or equal to five to all patients. In fact, the best median of all the treatments (not only these 10 treatments) is around 10 for both treatment definitions. This variability in the response to treatment shows that a non-personalized approach for the treatment of TRD doesn't work.

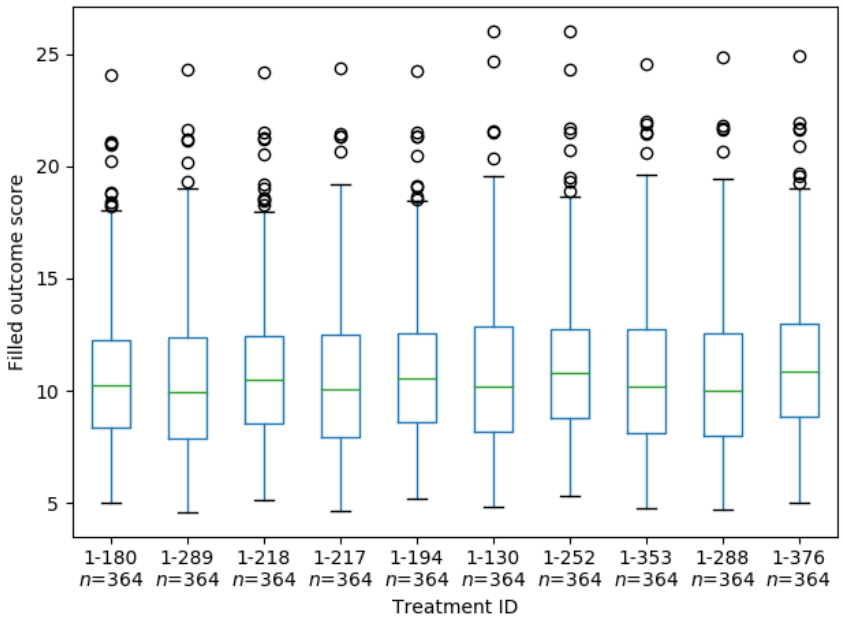

FIGURE 5.5: Boxplot of filled outcome score for the 10 treatments with lowest 95th percentile for filled outcome score under the first definition of treatment.

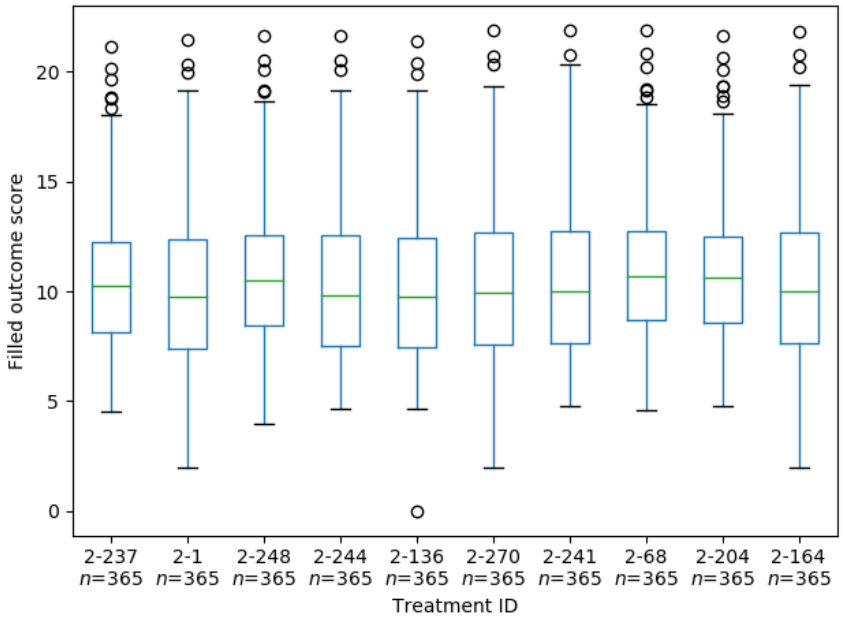

FIGURE 5.6: Boxplot of filled outcome score for the 10 treatments with lowest 95th percentile for filled outcome score under the second definition of treatment.

These results are consistent with the comments provided in Section 5.4.2 for the positively biased data, i.e., data that is biased toward better patient-treatment matches. When looking at the average of the true ratings from the original data set, we should expect a better average than if we observed all ratings; otherwise, the physicians are acting worst than uniformly random treatment selection.

Finally, note that these results support the claim that the physicians are better than random since the boxplots in the current section are worst than the ones in Section 5.4.2, i.e., the additional ratings we observe are bad.

Do some treatments consistently provide good response within a particular patient subgroup?

To answer this last question, we count, for each treatment, the number of ratings less or equal to five (i.e., the number of normal ratings) from the previous filled rating matrix. We then keep the treatments (i.e., the clusters) with at least 10 of these normal ratings. For the first treatment definition, seven such treatments are identified. In Table 5.9, the composition of these treatments is provided; the frequency of these treatments in the original data (`Freq_Data`), the number of normal ratings (`Cluster_Size`) and

the number of predicted entries among all these normal ratings (`Filled_Nb`) are also provided for each treatment.[8] Among the seven treatments, three of them are more frequent in the original data and all their normal ratings consist of true ratings. For the other four treatments, all the normal ratings are predicted. In addition, the three previous treatments consist of a single drug category (i.e., `AD_SNRI`, `AD_SSRI` or `AO_Antipsychotic`). Thus, the identified patients in remission (i.e., the normal ratings) for these three treatments may actually correspond to patients not suffering from TRD since they appear easy to treat. Finally, for the other four treatments, these appear to consist of many categories and drugs; from one to two antidepressant indicators and from three to four add-on indicators. Still, these can be considered realistic treatments since they are prescribed one time each in the original data. Additional psychiatric expertise is required to further understand why (or why not) these treatments appear promising.

When looking at the patients in remission in these clusters, it appears that respectively 44 patients belong to only one treatment cluster, 4 patients belong to 2 clusters, 16 patients belong to 4 clusters and 2 patients belong to 5 clusters. The sample means and standard deviations of these patients are provided in Table 5.11. A first cluster than stands out with respect to these numbers is the 1-374 since it appears to consist of young patients (`Pt_Age`); this cluster also contains more male patients than the other patients (`Pt_Gender_M`). With respect to the first axis, it appears that most patients in the cluster 1-276 do not have these disorders (`Pt_FirstAxis_N`), while most patients in clusters 1-150 and 1-374 have them (`Pt_FirstAxis_Y`). For the second axis, the numbers are much more uniformly distributed. Thus, overall, it is at the moment quite difficult to understand the composition of these clusters with the limited amount of information we have describing these patients; a retrospective analysis of these patients' medical records is needed.

A similar analysis is also done for the second definition of treatment. However, since there exist 16 treatments with at least 10 normal ratings, we increase that threshold to 20 normal ratings and obtain 8 corresponding treatments (see Table 5.10); note that overall the clusters are larger than under the first treatment definition. Contrary to the results under the first treatment definition, only one treatment is frequent in the original data and its identified normal ratings all consist of true ratings. Note that the identified patients in this cluster could also correspond to patients not suffering from TRD since this treatment appears simplistic, i.e., only `AD_SSRI`. The other treatments again consist of several drugs and are observed only one time in the original data; further expertise is needed to understand how these treatments work. With respect to the the responsive patients

[8] Remember that the predicted ratings of these treatments are not all normal ratings as shown in the boxplots of the previous question.

TABLE 5.9: Description of the identified clusters under the first treatment definition. The cluster IDs correspond to the treatment IDs. ✓corresponds to a variable with true value.

Variable	1-150	1-217	1-276	1-288	1-289	1-353	1-374
AD_MAOI							
AD_SNRI	✓						
AD_SSRI		✓	✓				
AD_TCA				✓	✓		
AD_Bupropion							
AD_Mirtazapine		✓		✓	✓	✓	
AO_Antipsychotic				✓			✓
AO_Anxiolytic		✓		✓	✓	✓	
AO_Hypnotic		✓		✓	✓	✓	
AO_Stimulant		✓		✓	✓	✓	
AO_Aripiprazole		✓					
AO_Liothyronine							
AO_Lithium		✓		✓	✓	✓	
AODE_ADBooster		✓		✓	✓	✓	
AODE_Antipsychotic		✓		✓			✓
AODE_Anxiolytic		✓		✓	✓	✓	✓
AODE_Hypnotic		✓		✓	✓	✓	
AODE_Stimulant		✓		✓	✓	✓	
Freq_Data	42	1	67	1	1	1	42
Cluster_Size	12	22	22	18	23	17	12
Filled_Nb	0	22	0	18	23	17	0

in these clusters (see Table 5.12), it is again hard to identify a pattern. It does however appear that the cluster 2-241 consists of younger patients while the cluster 2-200 consists of patients with fewer first axis disorders than the other clusters. Note also that 26 patients belong to one or two clusters while the other 37 patients are members of three to seven clusters. In particular, 19 patients belong to seven clusters. Hence, there appears to be more overlap in these clusters than in the ones under the first definition of treatment.

To conclude this section, it does appear that some treatments work well on some patients' subgroups. These subgroups are however somewhat small; a maximum of 23 patients out of 364 patients (6.3%) under the first treatment definition and a maximum of 38 patients out of 365 patients (10.4%) under the second treatment definition. Additionally, the number of treatments is also small; there are respectively 7 and 16 treatments with at least 10 normal ratings under the first and second definition of treatment. It is thus important to correctly understand these subgroups to obtain interesting outcomes. Unfortunately, with

TABLE 5.10: Description of the identified clusters under the second treatment definition. The cluster IDs correspond to the treatment IDs. ✓corresponds to a variable with true value.

Variable	**2-1**	**2-136**	**2-164**	**2-200**	**2-238**	**2-241**	**2-244**	**2-270**
AD_MAOI	✓							
AD_SNRI								
AD_SSRI		✓	✓	✓				
AD_TCA								
AD_Bupropion		✓			✓	✓	✓	
AD_Mirtazapine								
AO_Aripiprazole			✓		✓	✓		
AO_Atomoxetine							✓	✓
AO_Liothyronine								
AO_Lithium	✓	✓	✓		✓			
AO_Lurasidone								
AO_Methylphenidate		✓			✓	✓	✓	
AO_Modafinil						✓	✓	
AO_Pramipexole								
AO_Quetiapine			✓		✓			
AO_Trazodone								
AODE_ADBooster	✓	✓	✓		✓			
AODE_Antipsychotic			✓		✓	✓		
AODE_Anxiolytic			✓		✓			
AODE_Hypnotic								
AODE_Stimulant		✓	✓		✓	✓	✓	✓
Freq_Data	1	1	1	94	1	1	1	1
Cluster_Size	39	37	25	31	27	21	36	32
Filled_Nb	38	36	24	0	27	20	36	31

the currently available data, it is not possible to characterize these patients' subgroups to a sufficient level to construct a decision aid.

TABLE 5.11: Sample means and standard deviations (in parentheses) of the patients in remission in the identified clusters under the first treatment definition. The cluster IDs correspond to the treatment IDs. †These means and standard deviations are computed using proportions.

Variable	1-150	1-217	1-276	1-288	1-289	1-353	1-374
Pt_Age	43.33 (9.89)	42.95 (12.66)	42.86 (12.62)	40.22 (11.56)	43.26 (12.46)	39.47 (11.46)	35.17 (9.59)
Pt_Gender_M†	0.42 (0.49)	0.36 (0.48)	0.36 (0.48)	0.39 (0.49)	0.35 (0.48)	0.41 (0.49)	0.50 (0.50)
Pt_FirstAxis_N†	0.17 (0.37)	0.36 (0.48)	0.59 (0.49)	0.33 (0.47)	0.35 (0.48)	0.29 (0.46)	0.08 (0.28)
Pt_FirstAxis_TBI†	0.08 (0.28)	0.09 (0.29)	0.05 (0.21)	0.11 (0.31)	0.09 (0.28)	0.12 (0.32)	0.00 (0.00)
Pt_FirstAxis_Y†	0.58 (0.49)	0.32 (0.47)	0.18 (0.39)	0.28 (0.45)	0.35 (0.48)	0.29 (0.46)	0.67 (0.47)
Pt_FirstAxis_nan†	0.17 (0.37)	0.23 (0.42)	0.18 (0.39)	0.28 (0.45)	0.22 (0.41)	0.29 (0.46)	0.25 (0.43)
Pt_SecondAxis_N†	0.33 (0.47)	0.18 (0.39)	0.41 (0.49)	0.17 (0.37)	0.17 (0.38)	0.12 (0.32)	0.33 (0.47)
Pt_SecondAxis_TBI†	0.17 (0.37)	0.41 (0.49)	0.23 (0.42)	0.39 (0.49)	0.39 (0.49)	0.41 (0.49)	0.17 (0.37)
Pt_SecondAxis_Y†	0.33 (0.47)	0.18 (0.39)	0.18 (0.39)	0.17 (0.37)	0.22 (0.41)	0.18 (0.38)	0.25 (0.43)
Pt_SecondAxis_nan†	0.17 (0.37)	0.23 (0.42)	0.18 (0.39)	0.28 (0.45)	0.22 (0.41)	0.29 (0.46)	0.25 (0.43)

TABLE 5.12: Sample means and standard deviations (in parentheses) of the patients in remission in the identified clusters under the second treatment definition. The cluster IDs correspond to the treatment IDs. †These means and standard deviations are computed using proportions.

Variable	2-1	2-136	2-164	2-200	2-238	2-241	2-244	2-270
Pt_Age	42.62 (12.63)	42.35 (12.83)	37.12 (12.13)	42.23 (12.97)	38.81 (12.58)	34.19 (10.13)	42.44 (13.00)	40.53 (12.62)
Pt_Gender_M†	0.41 (0.49)	0.41 (0.49)	0.40 (0.49)	0.35 (0.48)	0.41 (0.49)	0.33 (0.47)	0.42 (0.49)	0.41 (0.49)
Pt_FirstAxis_N†	0.44 (0.50)	0.43 (0.50)	0.36 (0.48)	0.52 (0.50)	0.44 (0.50)	0.38 (0.49)	0.44 (0.50)	0.44 (0.50)
Pt_FirstAxis_TBI†	0.05 (0.22)	0.05 (0.23)	0.04 (0.20)	0.03 (0.18)	0.04 (0.19)	0.05 (0.21)	0.06 (0.23)	0.03 (0.17)
Pt_FirstAxis_Y†	0.33 (0.47)	0.35 (0.48)	0.40 (0.49)	0.29 (0.45)	0.33 (0.47)	0.38 (0.49)	0.36 (0.48)	0.38 (0.48)
Pt_FirstAxis_nan†	0.18 (0.38)	0.16 (0.37)	0.20 (0.40)	0.16 (0.37)	0.19 (0.39)	0.19 (0.39)	0.14 (0.35)	0.16 (0.36)
Pt_SecondAxis_N†	0.33 (0.47)	0.32 (0.47)	0.28 (0.45)	0.42 (0.49)	0.30 (0.46)	0.24 (0.43)	0.33 (0.47)	0.31 (0.46)
Pt_SecondAxis_TBI†	0.18 (0.38)	0.19 (0.39)	0.20 (0.40)	0.26 (0.44)	0.19 (0.39)	0.19 (0.39)	0.19 (0.40)	0.16 (0.36)
Pt_SecondAxis_Y†	0.31 (0.46)	0.32 (0.47)	0.32 (0.47)	0.16 (0.37)	0.33 (0.47)	0.38 (0.49)	0.33 (0.47)	0.38 (0.48)
Pt_SecondAxis_nan†	0.18 (0.38)	0.16 (0.37)	0.20 (0.40)	0.16 (0.37)	0.19 (0.39)	0.19 (0.39)	0.14 (0.35)	0.16 (0.36)

5.5 Conclusion

In this work, we try using recommender systems (RSs) to improve the selection of pharmacological treatments for patients suffering from treatment-resistant depression. After describing the setting of our case study and the acquired observational data, several RS models are tested. Since the best RMSE found on the test set is around 5, additional research and training data are necessary before using any of these models as a medical decision aid. Yet, this degree of accuracy still allows some insights. In particular, it is found that no treatments appear to provide good response for all patients. In addition, it is found that some subgroups of patients respond to the same treatments.

There are several limitations to our case study. First, we use observational data that is not purposely collected for research and thus contains several missing values. Thus, we have to infer the drugs the patients are taking since we only have access to the prescriptions made at this hospital. Second, we make some (informed) choices regarding the preprocessing of data (e.g., treatment definition, time-window for outcome). Doing other choices could lead to different results as is observed in this case study with our two treatment definitions. Finally, we focus only on the pharmacological treatments while other types of treatments such as psychotherapy are well known to help towards remission; psychotherapy is often prescribed to patients of the DSDP.

Future research areas include (1) testing other RS models and (2) analyzing which patient's and treatment's features are good predictors and mediators of outcome.

Chapter 6

Concluding Remarks and Future Research

In this thesis, we use and develop several analytics methods to improve medical decision making with respect to the management of patients suffering from treatment-resistant depression (TRD). We focus on TRD, a severe form of the major depressive disorder (MDD), since this is a growing global health concern that is both complex and not well understood; in addition, it appears that mental illnesses (in particular, TRD) are not addressed by the operations research and management science (OR/MS) literature.

The analytics methods in this thesis expand over the traditional OR/MS methods by borrowing methods from other other fields (e.g., artificial intelligence, statistics, causal inference). The goal of these non-traditional methods is to expand the problems' types that can be addressed. In particular, this research focus on medical decision making (MDM) using observational data. While there exists OR/MS literature on MDM that uses observational data, most of this literature does not explicitly address the limitations of using observational data. In this thesis, we take into account explicitly these limitations whenever we are using the observational data collected at the depressive and suicide disorders program (DSDP) of the Douglas Mental Health University Institute in Montreal.

6.1 Summary of Research Findings

In Chapter 2, a survey of MDM methods relevant for treating depression is done across the OR/MS literature, and the artificial intelligence and statistics literature (in particular, the dynamic treatment regimes (DTRs) literature). This survey highlights that no OR/MS studies are applied to MDD while several studies exist in the DTR literature. In addition, this survey highlights two major differences between these fields. First, while most OR/MS literature uses Markov decision process (MDP) or partially observable Markov

decision process (POMDP) models with discrete states, the DTR literature mostly deals with continuous states. Second, the OR/MS literature assumes Markovian states while the DTR literature uses history-dependent states. These two differences exist probably because of the focus of these two fields. On the one hand, the OR/MS literature typically focuses on devising models with good computational tractability to be able to address complex decisions. By doing so, they however somewhat neglect the validity of these models. On the other hand, the DTR literature is typically limited to models that address simple decisions to increase the probability that these models are valid even when using observational data. By not limiting itself to traditional OR/MS methods, this research tries to exploit the synergy between these fields.

In Chapter 3, an improved approach for causal inference is proposed; this approach is able to balance all distributions' moments by balancing the treatment groups in a reproducing kernel Hilbert space (RKHS). When compared to other similar causal inference methods on different simulation models, this approach is shown to obtain similar results and in many cases the best results. This approach is also used in a TRD case study where the goal is to determine which of five treatment modification strategies is best at the initial visit. Unfortunately, while some findings are consistent with the medical literature and guidelines, the obtained treatment effects are not statistically significant with respect to the 95% confidence intervals. This is most likely due to the small data set obtained after the exclusion of missing data. Another possible explanation could be the inappropriateness of the five strategies; while these are often used in the medical literature, they could be too broadly defined. Still, this case study consists in a helpful tutorial to causal inference for the OR/MS community. In particular, it demonstrates, with a case study, that omitting to address the causal inference issues can lead to opposite results.

In Chapter 4, different imitation learning (IL) methods are used to identify the relevant variables among the patient's, physician's and clinic's characteristics for the timing decision between appointments. This timing decision is important since it implies a trade-off between the consequences of low- and high-frequency appointments. Yet, it appears that there is a disparity in the medical recommendations regarding this decision. A two-stage framework is used in this study to identify the relevant variables. First, with the use of semi-structured interviews, potential features used to determine the time between consecutive appointments are elicited from the four psychiatrists at the DSDP. Unsurprisingly, it appears that similar features are used by each of these psychiatrists; yet, the importance of these features to each psychiatrist cannot be captured by these interviews, the reason for the existence of the framework's second stage. Still, these interviews also capture the

variable experience and variable typical times between appointments for each of these psychiatrists. Then, the second stage of the framework consists in the different IL methods. After a brief review of the existing IL methods, three methods are selected for the case study, with one method being a proposed extension. The results of the case study first show that the cost of discretization is quite high for this setting; the proposed approach that uses discretization doesn't perform well even when trying to favor this approach. Thus, it appears that the best approach to identify the relevant variables is a simple least absolute shrinkage and selection operator (LASSO) model. In the case study, the top two identified features are the time since the initial appointment and the indication of one particular treating psychiatrist; both increasing the time between appointments. Additional results also tend to show that the identified features vary across the physicians. Yet, additional data and features are needed to further establish this claim. Finally, it appears that the outcomes of each physician panel are statistically significantly different with respect to several outcome scales. This leads to believe that the compounding of the treatment and timing decisions is better done by some of the physicians. Yet, these outcomes do not appear to significantly differ from a practical point of view.

In Chapter 5, different recommender system (RS) models are used to try to provide personalized recommendations of treatments to patients suffering from TRD. In order to do so, we assume that the sequence of treatments that have been administered to the patient does not affect the efficacy of the current treatment. In addition, we define two different treatment definitions which make sense from a medical point of view. On these two treatment definitions, we then fit models that use different features available in the data set such as features describing the patient, the treatment and the outcomes resulting from other treatments. Note also that some models use weights over the training observations (i.e., inverse probability of treatment weighting (IPTW)) to correct for potential confounders in the observational data. According to different metrics, it appears that the models using the most features from the data provide the best results. Thus, the limited number of features describing the patient and the treatment does contain some relevant information. Yet these models are not performing well enough to be used as decision aids. In this work, it is also found that no treatment consistently leads to remission for all patients. This result is found by imputing the unobserved patient-treatment outcomes with the best RS model found. Finally, some patients' subgroups are found to respond to the same treatments. These particular treatments could then be assumed as being better than the other treatments which appear to only work for one or two patients. Further characterization of these patients' subgroups is however necessary to be able to identify

whether a particular patient belongs to one of these subgroups.

These above studies show different ways in which methods, that are not traditionally used in OR/MS, can be used to address MDM problems; there also exists a large number of other methods that could be used to address other types of relevant problems. In addition, these above studies highlight the importance and difficulties of addressing causal inference issues when using observational data for prescriptive tasks.

To conclude, note that the quality of data is of uttermost importance; as the saying goes "garbage in, garbage out". Unfortunately, this quality is not always present, especially in observational data that is not purposely collected for research, as is the case in this research. Thus, several of the above findings could be improved with additional and better quality data. Better quality data would also have helped a lot during this journey. Clinics, hospitals and other institutions that would like to accomplish research using observational data should define and document strict processes for the collection of data; by thinking about the questions they would like to answer in the future and the data that could be necessary to address these tasks. They should also ideally store this data with the use of a common data format. This would help tremendously the researchers and would also ensure to use this data to its full potential.

6.2 Future Research

While this thesis addresses several questions around the management of TRD using different approaches, there is still a lot of research to be done. First and foremost, replication of the above results in similar and different contexts is key to obtain what can be called scientific evidence. These replications can also answer additional questions such as: Do the treatment effects found in Chapter 3, the features found in Chapter 4 and the recommended treatments found in Chapter 5 only apply to the DSDP or they also apply to other healthcare organizations? These replications can be done using observational data but ideally some of these replications should also be done in randomized controlled trials (RCTs). An interesting opportunity for replication (and even new research) consists in the use of patient-level data from past RCTs; this data is now available to researchers through different platforms. Another interesting opportunity lies in the combination of this randomized patient-level data with observational data.

Second, future research could also be done on additional questions related to TRD. In particular, in most of this research (e.g., Chapter 3 and 5), the focus is on pharmacological treatments. Yet, the medical guidelines and literature acknowledge the importance of

psychological treatments as well as other alternative treatments (Lam et al., 2016a). Thus, it would be interesting to see how these other treatments alter the current results; for example, could the use of psychotherapy reduce the need to use certain drugs? In addition, these additional treatments also raise additional questions on how to better select them. For example, when prescribing psychotherapy, one needs to select the type of therapy (e.g., cognitive-behavioural therapy, interpersonal therapy), the mode of delivery (e.g., online therapy, group therapy, individual therapy) and the frequency (e.g., weekly, monthly) (Lam et al., 2016a); it is also necessary to decide at first whether psychotherapy is beneficial at all.

Third, additional methods could be also be explored in future research. For example, exploring probabilistic graphical models for causal inference in Chapter 3 and 5 could be interesting in order to directly obtain confidence intervals around the estimates. In addition, the use of a model-free inverse reinforcement learning (IRL) model in Chapter 4 could be interesting in order to avoid discretizing the states and actions; yet, a procedure would then be needed to interpret the results since the goal of this chapter is to characterize and not to imitate the timing decisions. Then, as shown in Chapter 5, there exist a vast literature on RSs. There are thus many other approaches that could be tested in order to recommend treatments (e.g., probabilistic graphical models, deep learning models); exploring such other models could lead to the identification of a model that outperforms the ones in Chapter 5. Finally, an interesting avenue to explore consists in the use of generative adversarial nets (GANs) (Goodfellow et al., 2014) for data augmentation. This data augmentation procedure could enable the use of complex models in settings with limited data, as is the case in this research. Note that there already exist some successful GANs applications to medical data (Choi et al., 2017; Esteban, Hyland, and Rätsch, 2017). Yet, it remains to be seen to which point the medical literature would accept the models fitted with these augmented data sets.

Finally, future research could be done on how to best implement decision aids based on the above and future related results; for example, these decision aids could suggest the best treatment modification strategy, the best delay before the next appointment and/or the treatments with potentially the best outcomes. This implementation phase isn't as easy as it could seem, even when a correct model is available for the task at hand. It requires research on the design of the decision aid's interface, a topic addressed in the field of human-computer interaction (HCI) (Shneiderman et al., 2016). In particular, this design should lead to a good user experience in order for the physician to keep using this decision aid in the future and to not make mistakes. In addition, this design could take into account

the patient perspective since it is the patient that is affected by the decisions, and engaging the patient in the decision process could improve its adherence to the resulting decisions.

Appendix A

Appendix to Chapter 3

A.1 Causal Inference Illustrative Example

We have a population of patients in which we want to compute the effect of a treatment. In this population, suppose that patients with high social support do better and that their response is independent of whether they are treated ($Z = 1$) or not ($Z = 0$), i.e., the treatment has no effect. Furthermore suppose that patients with high social support ($X = 1$) have a 60% response rate and that patients with low social support ($X = 0$) have a 40% response rate. Thus, $E[Y \mid Z = z, X = 1] = 0.6$ and $E[Y \mid Z = z, X = 0] = 0.4$ for $z = 0, 1$. Finally, suppose that 50% of the patients have high social support ($X = 1$).

If we randomize the treatment to the patients, we obtain a treated and a control group that contains an equal proportion of patients with low social support ($X = 0$) and high social support ($X = 1$). We thus obtain the following response rates for the controls

$$\begin{aligned}\mathbb{E}[Y \mid Z = 0] &= \sum_{x=0}^{1} \mathbb{E}[Y \mid Z = 0, X = x] \Pr(X = x \mid Z = 0) \\ &= (.4)(.5) + (.6)(.5) \\ &= 0.5\end{aligned}$$

and for the treated

$$\begin{aligned}\mathbb{E}[Y \mid Z = 1] &= \sum_{x=0}^{1} \mathbb{E}[Y \mid Z = 1, X = x] \Pr(X = x \mid Z = 1) \\ &= (.4)(.5) + (.6)(.5) \\ &= 0.5.\end{aligned}$$

In a randomized experiment, the data provides an unbiased estimator of the true treatment effect (i.e., a null effect).

Lets now suppose that it is not possible to run a randomized experiment and that instead we are given (observational) data by a clinician. Suppose that this clinician (incorrectly) suspects that patients with strong social support ($X = 1$) benefit more from the treatment. Thus this clinician treats 90% of his patients with high social support ($\Pr(Z = 1 \mid X = 1) = 0.9$) and only 50% of his patients with low social support ($\Pr(Z = 1 \mid X = 0) = 0.5$).

TABLE A.1: Propensity for treatment given social support covariate.

Treatment	**Low social support** $X = 0$	**High social support** $X = 1$
$Z = 0$	$\Pr(Z = 0 \mid X = 0) = 0.5$	$\Pr(Z = 0 \mid X = 1) = 0.1$
$Z = 1$	$\Pr(Z = 1 \mid X = 0) = 0.5$	$\Pr(Z = 1 \mid X = 1) = 0.9$

If we blindly use this data to estimate the response rates, we obtain the following response rates for the controls

$$\begin{aligned}\mathbb{E}[Y \mid Z = 0] &= \frac{\sum_{x=0}^{1} \mathbb{E}[Y \mid Z = 0, X = x] \Pr(Z = 0 \mid X = x) \Pr(X = x)}{\sum_{x=0}^{1} \Pr(Z = 0 \mid X = x) \Pr(X = x)} \\ &= \frac{(.4)(.5)(.5) + (.6)(.1)(.5)}{(.5)(.5) + (.1)(.5)} \\ &= 0.43\end{aligned}$$

and for the treated

$$\begin{aligned}\mathbb{E}[Y \mid Z = 1] &= \frac{\sum_{x=0}^{1} \mathbb{E}[Y \mid Z = 1, X = x] \Pr(Z = 1 \mid X = x) \Pr(X = x)}{\sum_{x=0}^{1} \Pr(Z = 1 \mid X = x) \Pr(X = x)} \\ &= \frac{(.4)(.5)(.5) + (.6)(.9)(.5)}{(.5)(.5) + (.9)(.5)} \\ &= 0.53.\end{aligned}$$

We easily see that we do not obtain the true treatment effect. We obtain a treatment effect of 0.1 while the true effect is null.

If we instead balance the treatment groups with respect to X, we obtain in the balanced groups that the treatment is independent of the social support, i.e., $\Pr(Z = 0|X = x) = \Pr(Z = 0)$ and $\Pr(Z = 1|X = x) = \Pr(Z = 1)$. We thus obtain the following corrected

response rates for the controls

$$\begin{aligned}\mathbb{E}[Y \mid Z=0] &= \frac{\sum_{x=0}^{1}\mathbb{E}[Y \mid Z=0, X=x]\Pr(Z=0 \mid X=x)\Pr(X=x)}{\sum_{x=0}^{1}\Pr(Z=0 \mid X=x)\Pr(X=x)} \\ &= \frac{\Pr(Z=0)[(.4)(.5)+(.6)(.5)]}{\Pr(Z=0)[(.5)+(.5)]} \\ &= 0.5\end{aligned}$$

and for the treated

$$\begin{aligned}\mathbb{E}[Y \mid Z=1] &= \frac{\sum_{x=0}^{1}\mathbb{E}[Y \mid Z=1, X=x]\Pr(Z=1 \mid X=x)\Pr(X=x)}{\sum_{x=0}^{1}\Pr(Z=1 \mid X=x)\Pr(X=x)} \\ &= \frac{\Pr(Z=1)[(.4)(.5)+(.6)(.5)]}{\Pr(Z=1)[(.5)+(.5)]} \\ &= 0.5.\end{aligned}$$

This corresponds to the true treatment effect. It is now easy to see the benefit of using balancing approaches when computing a treatment effect from observational data.

Using the potential outcome notation of Section 3.2, here we have that $Y^{(0)}, Y^{(1)}$ are independent of Z conditional on X. It is easy to see here because the response rate depends only on X. For example, $\mathbb{E}[Y^{(0)} \mid Z=0, X=0] = \mathbb{E}[Y^{(0)} \mid X=0] = 0.4$. However, $Y^{(0)}, Y^{(1)}$ are not marginally independent of Z. For example, $\mathbb{E}[Y^{(0)} \mid Z=0] = \mathbb{E}[Y|Z=0] = 0.43$ while $\mathbb{E}[Y^{(0)}] = 0.5*0.4 + 0.5*0.6 = 0.5$. After balancing using X, $Y^{(0)}, Y^{(1)}$ are marginally independent of Z. In the balanced data, we have $\mathbb{E}[Y^{(0)} \mid Z=0] = \mathbb{E}[Y \mid Z=0] = 0.5$ which is the same as $\mathbb{E}[Y^{(0)}] = 0.5$.

A.2 Covariates Selection

A causal graph is a useful representation of causality assumptions. In a causal graph, nodes represent variables while directed edges represent direct causal effects between variables. Thus, the causal assumptions are represented by missing edges. For example, in Figure A.1, A has a direct effect on Z while A only has an indirect effect on Y through Z and B.

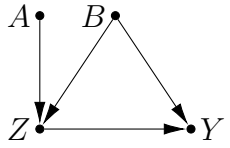

FIGURE A.1: Simple causal graph. Z denote treatment, Y denote outcome, and A and B denote covariates.

As discussed in Section 3.2.3, in order for Assumption 2 to hold and for the computed treatment effects to be unbiased, it suffices to select a set of covariates X that satisfies the following back-door criterion:

Definition 4 (Back-Door Criterion (Pearl, 2009b))**.** *A set of variables X satisfies the back-door criterion relative to an ordered pair of variables (Z, Y) in a causal diagram if:*

(a) no node in X is a descendant of Z; and

(b) X blocks every path between Z and Y that contains an arrow into Z.

With respect to Figure A.1, X should then be constituted of either $\{A, B\}$ or $\{B\}$. If it consists of only $\{A\}$, the back-door criterion will not be respected and the estimated treatment effect might be biased.

In practice, it is important to note however that selecting $\{A, B\}$ instead of $\{B\}$ might amplify the residual bias due to an additional unobserved covariate (Pearl, 2010).

A.3 Proof of Lemma 2

Proof.

1. $$\begin{aligned} ATE_{u,v} &\triangleq \mathbb{E}[Y^{(u)}] - \mathbb{E}[Y^{(v)}] \\ &= \mathbb{E}[\mathbb{E}[Y^{(u)} \mid X]] - \mathbb{E}[\mathbb{E}[Y^{(v)} \mid X]] \\ &= \mathbb{E}[\mathbb{E}[Y^{(u)} \mid Z = u, X]] - \mathbb{E}[\mathbb{E}[Y^{(v)} \mid Z = v, X]] \\ &= \mathbb{E}[\mathbb{E}[Y \mid Z = u, X]] - \mathbb{E}[\mathbb{E}[Y \mid Z = v, X]] \\ &= \int \mathbb{E}[Y \mid Z = u, X = x] \Pr(X = x) dx \\ &\quad - \int \mathbb{E}[Y \mid Z = v, X = x] \Pr(X = x) dx \\ &= \int \mathbb{E}[Y \mid Z = u, X = x] \Pr(X = x) \frac{\Pr(X = x \mid Z = u)}{\Pr(X = x \mid Z = u)} dx \\ &\quad - \int \mathbb{E}[Y \mid Z = v, X = x] \Pr(X = x) \frac{\Pr(X = x \mid Z = v)}{\Pr(X = x \mid Z = v)} dx \\ &= \int \mathbb{E}[Y \mid Z = u, X = x] \beta_u(x) \Pr(X = x \mid Z = u) dx \\ &\quad - \int \mathbb{E}[Y \mid Z = v, X = x] \beta_v(x) \Pr(X = x \mid Z = v) dx \end{aligned}$$

2. $$\begin{aligned} CATE_{u,v,e} &\triangleq \mathbb{E}[Y^{(u)} \mid E = e] - \mathbb{E}[Y^{(v)} \mid E = e] \\ &= \mathbb{E}[\mathbb{E}[Y^{(u)} \mid X] \mid E = e] - \mathbb{E}[\mathbb{E}[Y^{(v)} \mid X] \mid E = e] \\ &= \mathbb{E}[\mathbb{E}[Y^{(u)} \mid Z = u, X] \mid E = e] - \mathbb{E}[\mathbb{E}[Y^{(v)} \mid Z = v, X] \mid E = e] \\ &= \mathbb{E}[\mathbb{E}[Y \mid Z = u, X] \mid E = e] - \mathbb{E}[\mathbb{E}[Y \mid Z = v, X] \mid E = e] \\ &= \int \mathbb{E}[Y \mid Z = u, X = x] \Pr(X = x \mid E = e) dx \\ &\quad - \int \mathbb{E}[Y \mid Z = v, X = x] \Pr(X = x \mid E = e) dx \\ &= \int \mathbb{E}[Y \mid Z = u, X = x] \Pr(X = x \mid E = e) \frac{\Pr(X = x \mid Z = u)}{\Pr(X = x \mid Z = u)} dx \\ &\quad - \int \mathbb{E}[Y \mid Z = v, X = x] \Pr(X = x \mid E = e) \frac{\Pr(X = x \mid Z = v)}{\Pr(X = x \mid Z = v)} dx \\ &= \int \mathbb{E}[Y \mid Z = u, X = x] \beta_u(x) \Pr(X = x \mid Z = u) dx \\ &\quad - \int \mathbb{E}[Y \mid Z = v, X = x] \beta_v(x) \Pr(X = x \mid Z = v) dx \end{aligned}$$

3. $$\begin{aligned} ATT_{u,v} &\triangleq \mathbb{E}[Y^{(u)} \mid Z = u] - \mathbb{E}[Y^{(v)} \mid Z = u] \\ &= \mathbb{E}[\mathbb{E}[Y^{(u)} \mid Z = u, X] \mid Z = u] - \mathbb{E}[\mathbb{E}[Y^{(v)} \mid Z = u, X] \mid Z = u] \end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}[\mathbb{E}[Y^{(u)} \mid Z = u, X] \mid Z = u] - \mathbb{E}[\mathbb{E}[Y^{(v)} \mid Z = v, X] \mid Z = u] \\
&= \mathbb{E}[\mathbb{E}[Y \mid Z = u, X] \mid Z = u] - \mathbb{E}[\mathbb{E}[Y \mid Z = v, X] \mid Z = u] \\
&= \int \mathbb{E}[Y \mid Z = u, X = x] \Pr(X = x \mid Z = u) dx \\
&\quad - \int \mathbb{E}[Y \mid Z = v, X = x] \Pr(X = x \mid Z = u) dx \\
&= \int \mathbb{E}[Y \mid Z = u, X = x] \Pr(X = x \mid Z = u) dx \\
&\quad - \int \mathbb{E}[Y \mid Z = v, X = x] \Pr(X = x \mid Z = u) \frac{\Pr(X = x \mid Z = v)}{\Pr(X = x \mid Z = v)} dx \\
&= \int \mathbb{E}[Y \mid Z = u, X = x] \beta_u(x) \Pr(X = x \mid Z = u) dx \\
&\quad - \int \mathbb{E}[Y \mid Z = v, X = x] \beta_v(x) \Pr(X = x \mid Z = v) dx
\end{aligned}
$$

□

A.4 Link with Stabilized Weights

Corollary 2. *When kernel mean matching (KMM) is used to compute treatment effects as in Theorem 1, it is equivalent to stabilized IPTW (Robins, 1998; Robins, Hernán, and Brumback, 2000).*

Proof. We will only prove it for average treatment effect (ATE). It is easy to prove for conditional average treatment effect (CATE) and average treatment effect among the treated (ATT) using a similar process. For $ATE_{u,v}$, the first set of weights that KMM finds gives

$$\begin{aligned}\beta_u(x) &= \frac{\Pr(X = x)}{\Pr(X = x \mid Z = u)} \\ &= \frac{\Pr(X = x)\Pr(Z = u)}{\Pr(X = x, Z = u)} \\ &= \frac{\Pr(Z = u)}{\Pr(Z = u \mid X = x)}\end{aligned}$$

while the second gives

$$\beta_v(x) = \frac{\Pr(Z = v)}{\Pr(Z = v \mid X = x)}.$$

These weights correspond to stabilized IPTW for $ATE_{u,v}$. The numerator corresponding to the stabilization weight and the denominator corresponding to the propensity score. □

Remark 4. *While Corollary 2 states that KMM and stabilized IPTW are equivalent in theory, we do believe that KMM is superior to stabilized IPTW empirically because it directly computes the weights instead of fitting a propensity function and then using a transformation of it to compute the weights.*

A.5 Additional Results to the Comparative Analysis

The results in this section are obtained by increasing the standard deviations of $y_i(0)$ and $y_i(1)$ to 4, i.e., decreasing the signal-to-noise ratio from 4 to 1.

TABLE A.2: Bias, RMSE, range and mean computation time of the 1000 estimated ATTs for each approach on data sets with no unmeasured confounding and a signal-to-noise ratio of 1. Proposed approach is in bold font. †SVMMatch only succeeded on 967 data sets.

Approach	**Bias**	**RMSE**	**Range**		**Time**
			Minimum	**Maximum**	s
Diff. in Means	-0.01	1.22	-2.54	11.50	
KMM-N-B1-e1	-0.02	0.51	1.75	5.86	4.8
KMM-N-B1-e0	-0.02	0.51	1.89	6.24	3.0
KMM-N-B0-e1	-0.01	0.52	1.79	6.28	4.3
KMM-N-B0-e0	-0.03	0.51	1.72	6.39	2.7
KMM-E-B1-e1	4.47	5.54	3.92	32.24	5.3
KMM-E-B1-e0	-0.01	0.54	2.45	5.91	3.2
KMM-E-B0-e1	4.47	5.54	3.92	32.24	4.6
KMM-E-B0-e0	-0.01	0.54	2.46	5.96	2.7
BOSS	-0.02	0.59	1.33	6.98	49.2
SVMMatch†	-0.02	0.57	0.79	6.98	2.1
EBal	-0.02	0.51	1.79	6.75	0.4
SBW	-0.01	0.51	1.61	6.93	0.6

TABLE A.3: Bias, RMSE, range and mean computation time of the 1000 estimated ATTs for each approach on data sets with hidden bias and a signal-to-noise ratio of 1. Proposed approach is in bold font. †SVMMatch only succeeded on 960 data sets.

Approach	**Bias**	**RMSE**	**Range**		**Time**
			Minimum	**Maximum**	*s*
Diff. in Means	0.03	1.32	−3.27	14.15	
KMM-N-B1-e1	0.01	0.71	−3.21	7.86	4.9
KMM-N-B1-e0	−0.01	0.68	−2.91	7.27	3.2
KMM-N-B0-e1	−0.00	0.67	−0.15	7.87	5.0
KMM-N-B0-e0	−0.01	0.66	−1.35	6.86	3.1
KMM-E-B1-e1	3.55	4.71	2.81	30.10	5.5
KMM-E-B1-e0	−0.01	0.70	0.81	11.23	3.3
KMM-E-B0-e1	3.60	4.73	2.83	30.10	4.7
KMM-E-B0-e0	−0.01	0.70	0.95	11.24	2.7
BOSS	−0.01	0.77	−0.71	12.81	41.8
SVMMatch†	−0.01	0.74	−3.48	7.53	1.9
EBal	0.00	0.69	0.34	10.15	0.4
SBW	0.00	0.65	1.05	8.51	0.5

A.6 Characterization of the DSDP Data Set and Additional Results

TABLE A.4: List of the antidepressants and add-on drugs used at the clinic.

Antidepressant	**Add-on**	
Amitriptyline	Alprazolam	Methotrimeprazine
Bupropion	Amphetamine aspartate	Methylphenidate
Citalopram	Aripiprazole	Modafinil
Clomipramine	Atomoxetine	Naltrexone
Desipramine	Bromazepam	Nitrazepam
Desvenlafaxine	Buspirone	Olanzapine
Doxepin	Carbamazepine	Oxazepam
Duloxetine	Chlorpromazine	Paliperidone
Escitalopram	Clonazepam	Paliperidone inj
Fluoxetine	Dextroamphetamine	Periciazine
Fluvoxamine	Diazepam	Pimozide
Imipramine	Diphenhydramine	Pramipexole
Mirtazapine	Docusate	Pregabalin
Moclobemide	Flupentixol	Propranolol
Nortriptyline	Flurazepam	Quetiapine
Paroxetine	Gabapentin	Riluzole
Sertraline	Haloperidol	Risperidone
Tranylcypromine	Hydroxyzine	Temazepam
Trazodone	Lamotrigine	Topiramate
Venlafaxine	Liothyronine	Valproic acid
	Lisdexamfetamine	Ziprasidone
	Lithium carbonate	Zolpidem
	Lorazepam	Zopiclone
	Loxapine	Zuclopenthixol
	Lurasidone	

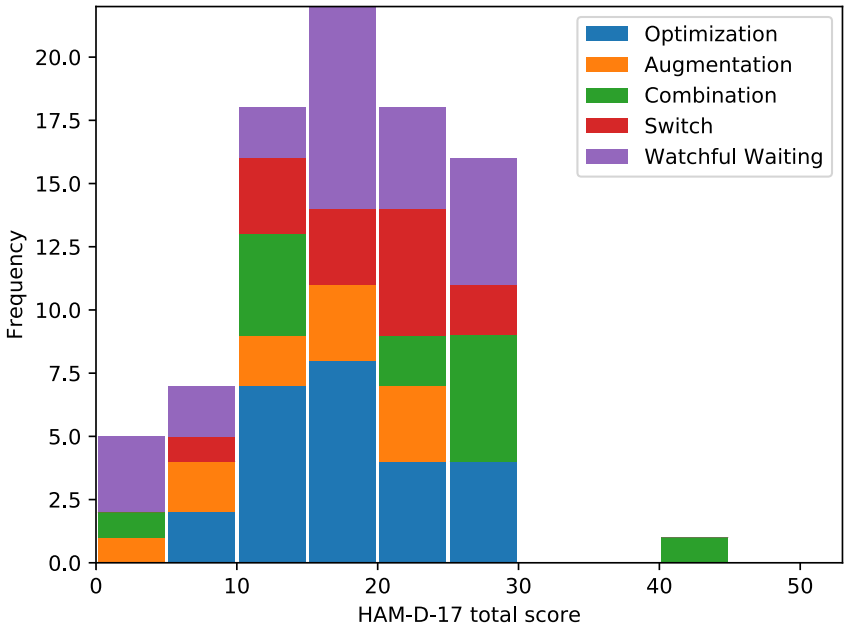

FIGURE A.2: Histogram of the HAM-D-17 total score per treatment group.

TABLE A.5: Sample means and standard deviations of the unbalanced covariates and outcome for each treatment category. †These means and standard deviations are computed using proportions.

Variable	**Optimization** ($n = 25$)	**Augmentation** ($n = 11$)	**Combination** ($n = 13$)	**Switch** ($n = 14$)	**Watchful waiting** ($n = 24$)
Age	44.1 ± 7.9	46.1 ± 10.2	46.6 ± 11.7	44.3 ± 10.9	44.0 ± 11.6
IsMale†	0.48 ± 0.50	0.55 ± 0.50	0.38 ± 0.49	0.14 ± 0.35	0.50 ± 0.50
Education1†	0.12 ± 0.32	0.27 ± 0.45	0.15 ± 0.36	0.00 ± 0.00	0.04 ± 0.20
Education2†	0.08 ± 0.27	0.09 ± 0.29	0.31 ± 0.46	0.29 ± 0.45	0.12 ± 0.33
Education3†	0.08 ± 0.27	0.09 ± 0.29	0.15 ± 0.36	0.00 ± 0.00	0.04 ± 0.20
Education4†	0.72 ± 0.45	0.55 ± 0.50	0.38 ± 0.49	0.71 ± 0.45	0.79 ± 0.41
Abused†	0.56 ± 0.50	0.27 ± 0.45	0.54 ± 0.50	0.36 ± 0.48	0.29 ± 0.45
FamPsyHx†	0.72 ± 0.45	0.55 ± 0.50	0.77 ± 0.42	0.50 ± 0.50	0.46 ± 0.50
HAM-A	21.0 ± 6.9	20.7 ± 8.8	24.0 ± 12.1	23.1 ± 9.1	18.3 ± 9.9
HAM-D-17	17.0 ± 6.1	14.8 ± 7.0	21.1 ± 9.5	18.6 ± 5.8	16.8 ± 7.9
SCID1Dx4†	0.92 ± 0.27	0.73 ± 0.45	0.69 ± 0.46	0.86 ± 0.35	0.71 ± 0.45
SCID1Dx26†	0.12 ± 0.32	0.18 ± 0.39	0.08 ± 0.27	0.21 ± 0.41	0.17 ± 0.37
SCID1Dx32†	0.24 ± 0.43	0.27 ± 0.45	0.15 ± 0.36	0.07 ± 0.26	0.12 ± 0.33
PastEpi†	1.00 ± 0.00	0.91 ± 0.29	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
SCID2Dx†	0.36 ± 0.48	0.27 ± 0.45	0.15 ± 0.36	0.43 ± 0.49	0.42 ± 0.49
PastSuiAtt†	0.24 ± 0.43	0.09 ± 0.29	0.23 ± 0.42	0.21 ± 0.41	0.29 ± 0.45
SSI	17.0 ± 10.6	18.1 ± 10.0	19.0 ± 7.4	15.9 ± 9.6	20.2 ± 8.7
QIDS-SR-16	12.8 ± 4.9	12.5 ± 4.5	13.6 ± 6.2	15.4 ± 7.0	13.5 ± 6.4

TABLE A.6: Additional treatment effects with 95% confidence intervals for the TRD case study. The row is u and the column is v. 1st row/column is Optimization, 2nd row/column is Augmentation, 3rd row/column is Combination, 4th row/column is Switch and 5th row/column is Watchful Waiting.

(A) $CATE_{u,v,e}$; e is `IsMale` == False

		$-0.31\,(-3.71,$	$3.70)$	$-4.42\,(-5.95,$	$1.52)$	$-2.40\,(-6.64,$	$2.44)$	$-1.75\,(-4.55,$	$4.20)$
$0.31\,(-3.70,$	$3.71)$			$-4.11\,(-5.28,$	$0.77)$	$-2.09\,(-5.78,$	$1.78)$	$-1.45\,(-3.92,$	$3.31)$
$4.42\,(-1.52,$	$5.95)$	$4.11\,(-0.77,$	$5.28)$			$2.02\,(-3.63,$	$4.14)$	$2.66\,(-1.58,$	$5.65)$
$2.40\,(-2.44,$	$6.64)$	$2.09\,(-1.78,$	$5.78)$	$-2.02\,(-4.14,$	$3.63)$			$0.65\,(-2.74,$	$6.28)$
$1.75\,(-4.20,$	$4.55)$	$1.45\,(-3.31,$	$3.92)$	$-2.66\,(-5.65,$	$1.58)$	$-0.65\,(-6.28,$	$2.74)$		

(B) $CATE_{u,v,e}$; e is `IsMale` == True

		$-1.85\,(-3.92,$	$3.75)$	$-0.74\,(-6.08,$	$1.65)$	$-0.55\,(-6.81,$	$2.67)$	$-2.02\,(-4.67,$	$4.31)$
$1.85\,(-3.75,$	$3.92)$			$1.11\,(-5.25,$	$0.85)$	$1.30\,(-5.88,$	$1.88)$	$-0.17\,(-4.07,$	$3.42)$
$0.74\,(-1.65,$	$6.08)$	$-1.11\,(-0.85,$	$5.25)$			$0.19\,(-3.91,$	$4.26)$	$-1.27\,(-1.68,$	$5.67)$
$0.55\,(-2.67,$	$6.81)$	$-1.30\,(-1.88,$	$5.88)$	$-0.19\,(-4.26,$	$3.91)$			$-1.46\,(-2.89,$	$6.50)$
$2.02\,(-4.31,$	$4.67)$	$0.17\,(-3.42,$	$4.07)$	$1.27\,(-5.67,$	$1.68)$	$1.46\,(-6.50,$	$2.89)$		

Appendix B

Appendix to Chapter 4

B.1 Interview Guide

- Background questions: What are the psychiatrist's characteristics?
  - Name, age, gender, role at the clinic, types of patients seen, number of years as psychiatrist, number of years working with TRD, number of years working at the clinic.
- Opening question: What are the challenging/important decisions that you face when treating patients suffering from TRD at the clinic?
  - Are there other decisions (operational or clinical) that are difficult to make?
  - If the time of the next appointment is not part of the answer, probe for it:
    - Do you find that the time of the next appointment is a challenging decision?
- Main topic: What factors do you account for when deciding on the time of the next appointment?
  - Probe for different areas:
    - the medical state of the patient (e.g., depression severity, suicide ideation, side effects),
    - the trend of his medical state (e.g., stable, improving, worsening),
    - his treatment (e.g., drugs necessitating a close follow-up, newly prescribed drugs, other treatment modifications),
    - other relevant characteristics of the patient,
    - the availability of the patient and physician (e.g., number of days per week the physician is working), and

        * the clinic's characteristics (e.g., available resources such as nurses and psychotherapists, size of waiting list).
    - Probe for precision on the features.
    - Probe for exceptions.

- Secondary topic: What are the typical times between consecutive appointments?
    - Probe to check if there exists a discrete set of decisions (e.g., {2 weeks, 6 weeks, 6 months}).
    - If yes, then probe for this set and the reasons when each of these decisions are used.

- Tertiary topic: What is the maximal scheduled time between consecutive appointments?
    - Probe to check if 365 days is a good upper bound. Otherwise, what is a good upper bound? This upper bound is going to be used to differentiate between follow-up appointments and appointments due to a relapse.

B.2 List of Relevant Drugs

TABLE B.1: List of the antidepressants and add-on drugs used at the clinic. [1]Drug from the TCA class. [2]Drug from the MAOI class. [3]Drug from the anticonvulsant class.

Antidepressant	**Add-on**	
Amitriptyline[1]	Alprazolam	Methotrimeprazine
Bupropion	Amphetamine aspartate	Methylphenidate
Citalopram	Aripiprazole	Modafinil
Clomipramine[1]	Atomoxetine	Naltrexone
Desipramine[1]	Bromazepam	Nitrazepam
Desvenlafaxine	Buspirone	Olanzapine
Doxepin[1]	Carbamazepine[3]	Oxazepam
Duloxetine	Chlorpromazine	Paliperidone
Escitalopram	Clonazepam	Paliperidone inj
Fluoxetine	Dextroamphetamine	Periciazine
Fluvoxamine	Diazepam	Pimozide
Imipramine[1]	Diphenhydramine	Pramipexole
Mirtazapine	Docusate	Pregabalin[3]
Moclobemide[2]	Flupentixol	Propranolol
Nortriptyline[1]	Flurazepam	Quetiapine
Paroxetine	Gabapentin[3]	Riluzole
Sertraline	Haloperidol	Risperidone
Tranylcypromine[2]	Hydroxyzine	Temazepam
Trazodone	Lamotrigine[3]	Topiramate[3]
Venlafaxine	Liothyronine	Valproic acid[3]
	Lisdexamfetamine	Ziprasidone
	Lithium carbonate	Zolpidem
	Lorazepam	Zopiclone
	Loxapine	Zuclopenthixol
	Lurasidone	

B.3 Characterization of Data Set

TABLE B.2: Descriptive statistics of the state features and action.

Variable	Mean	Std	Min	25%	50%	75%	Max
MD_A	0.22	0.42	0.00	0.0	0.00	0.00	1.0
MD_B	0.43	0.49	0.00	0.0	0.00	1.00	1.0
MD_C	0.13	0.34	0.00	0.0	0.00	0.00	1.0
MD_D	0.22	0.41	0.00	0.0	0.00	0.00	1.0
IsMale	0.36	0.48	0.00	0.0	0.00	1.00	1.0
Age	44.09	10.31	19.00	36.0	45.00	53.00	64.0
FirstAxis	0.63	0.48	0.00	0.0	1.00	1.00	1.0
SecondAxis	0.66	0.47	0.00	0.0	1.00	1.00	1.0
ADDosageIncrease_TCA	0.02	0.15	0.00	0.0	0.00	0.00	1.0
ADDosageIncrease_Any	0.18	0.38	0.00	0.0	0.00	0.00	1.0
ADAdded_TCA	0.02	0.15	0.00	0.0	0.00	0.00	1.0
ADAdded_MAOI	0.00	0.06	0.00	0.0	0.00	0.00	1.0
ADAdded_Any	0.13	0.33	0.00	0.0	0.00	0.00	1.0
AODosageIncrease_Li	0.02	0.13	0.00	0.0	0.00	0.00	1.0
AODosageIncrease_AED	0.01	0.12	0.00	0.0	0.00	0.00	1.0
AODosageIncrease_Any	0.17	0.38	0.00	0.0	0.00	0.00	1.0
AOAdded_Li	0.02	0.12	0.00	0.0	0.00	0.00	1.0
AOAdded_AED	0.02	0.13	0.00	0.0	0.00	0.00	1.0
AOAdded_Any	0.21	0.41	0.00	0.0	0.00	0.00	1.0
FIBSERScore	6.11	5.27	0.00	0.0	6.00	10.00	18.0
FIBSERTrend_Inc	0.32	0.47	0.00	0.0	0.00	1.00	1.0
FIBSERTrend_Dec	0.35	0.48	0.00	0.0	0.00	1.00	1.0
OverallFIBSERTrend_Inc	0.34	0.47	0.00	0.0	0.00	1.00	1.0
OverallFIBSERTrend_Dec	0.48	0.50	0.00	0.0	0.00	1.00	1.0
QIDSScore	13.34	6.22	0.00	9.0	13.00	18.00	27.0
QIDSTrend_Inc	0.40	0.49	0.00	0.0	0.00	1.00	1.0
QIDSTrend_Dec	0.47	0.50	0.00	0.0	0.00	1.00	1.0
OverallQIDSTrend_Inc	0.24	0.43	0.00	0.0	0.00	0.00	1.0
OverallQIDSTrend_Dec	0.68	0.47	0.00	0.0	1.00	1.00	1.0
SSIScore	4.94	7.36	0.00	0.0	2.00	5.00	36.0
SSITrend_Inc	0.28	0.45	0.00	0.0	0.00	1.00	1.0
SSITrend_Dec	0.31	0.46	0.00	0.0	0.00	1.00	1.0
OverallSSITrend_Inc	0.26	0.44	0.00	0.0	0.00	1.00	1.0
OverallSSITrend_Dec	0.44	0.50	0.00	0.0	0.00	1.00	1.0
QLDSScore	18.45	10.58	0.00	10.0	19.00	28.00	34.0
QLDSTrend_Inc	0.40	0.49	0.00	0.0	0.00	1.00	1.0
QLDSTrend_Dec	0.45	0.50	0.00	0.0	0.00	1.00	1.0
OverallQLDSTrend_Inc	0.33	0.47	0.00	0.0	0.00	1.00	1.0
OverallQLDSTrend_Dec	0.59	0.49	0.00	0.0	1.00	1.00	1.0
FollowingTime	80.25	68.93	0.14	26.0	61.14	117.29	391.0
Action	8.49	7.43	0.43	4.0	6.00	10.00	52.0

Appendix C

Appendix to Chapter 5

C.1 List of Drugs and Definition of Desired Effects

TABLE C.1: Drugs taken into account within the antidepressant classes and add-on categories. [1]Trazodone is an antidepressant but it is used as an add-on drug.

(A) Antidepressant classes and drugs

Class	**Drug**
MAOI	Moclobemide
	Tranylcypromine
SNRI	Desvenlafaxine
	Duloxetine
	Venlafaxine
SSRI	Citalopram
	Escitalopram
	Fluoxetine
	Fluvoxamine
	Paroxetine
	Sertraline
TCA	Amitriptyline
	Clomipramine
	Desipramine
	Doxepin
	Imipramine
	Nortriptyline

(B) Add-on categories and drugs

Category	**Drug**
Antipsychotic	Chlorpromazine
	Flupentixol
	Haloperidol
	Loxapine
	Lurasidone
	Methotrimeprazine
	Olanzapine
	Paliperidone
	Periciazine
	Pimozide
	Quetiapine
	Risperidone
	Ziprasidone
	Zuclopenthixol
Anxiolytic	Alprazolam
	Bromazepam
	Buspirone
	Clonazepam
	Diazepam
	Lorazepam
	Oxazepam
Hypnotic	Diphenhydramine
	Flurazepam
	Hydroxyzine
	Nitrazepam
	Temazepam
	Trazodone[1]
	Zolpidem
	Zopiclone
Stimulant	Amphetamine aspartate
	Atomoxetine
	Dextroamphetamine
	Lisdexamfetamine
	Methylphenidate
	Modafinil
	Pramipexole

TABLE C.2: Desired effects of the add-on categories and drugs.

Variable	AODE_ ADBooster	AODE_ Antipsychotic	AODE_ Anxiolytic	AODE_ Hypnotic	AODE_ Stimulant
AO_Antipsychotic		✓	✓		
AO_Anxiolytic			✓		
AO_Hypnotic				✓	
AO_Stimulant					✓
AO_Aripiprazole		✓			✓
AO_Atomoxetine					✓
AO_Liothyronine	✓				
AO_Lithium	✓				
AO_Lurasidone		✓	✓		
AO_Methylphenidate					✓
AO_Modafinil					✓
AO_Pramipexole					✓
AO_Quetiapine		✓	✓		
AO_Trazodone				✓	

C.2 Characterization of Data Set

TABLE C.3: Sample means and standard deviations of the patient features for the second treatment definition. The one-hot encoding of binary variables has been removed to improve readability. †These means and standard deviations are computed using proportions.

Variable	Mean	Std
Pt_Age	43.556	10.600
Pt_Gender_M†	0.367	0.482
Pt_FirstAxis_N†	0.364	0.481
Pt_FirstAxis_TBI†	0.047	0.211
Pt_FirstAxis_Y†	0.455	0.498
Pt_FirstAxis_nan†	0.134	0.341
Pt_SecondAxis_N†	0.296	0.456
Pt_SecondAxis_TBI†	0.238	0.426
Pt_SecondAxis_Y†	0.329	0.470
Pt_SecondAxis_nan†	0.137	0.344

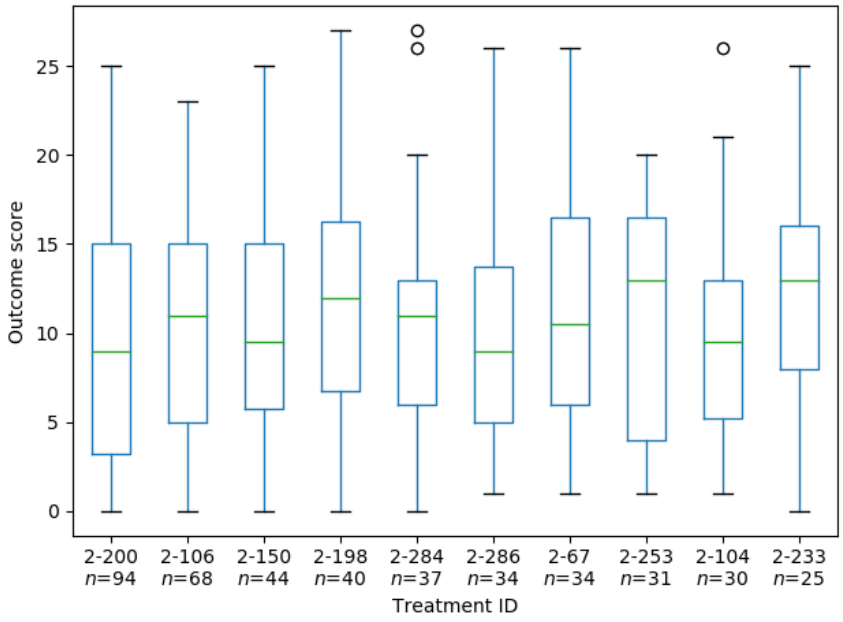

FIGURE C.1: Boxplot of outcome score for the 10 most frequent treatments under the second definition of treatment.

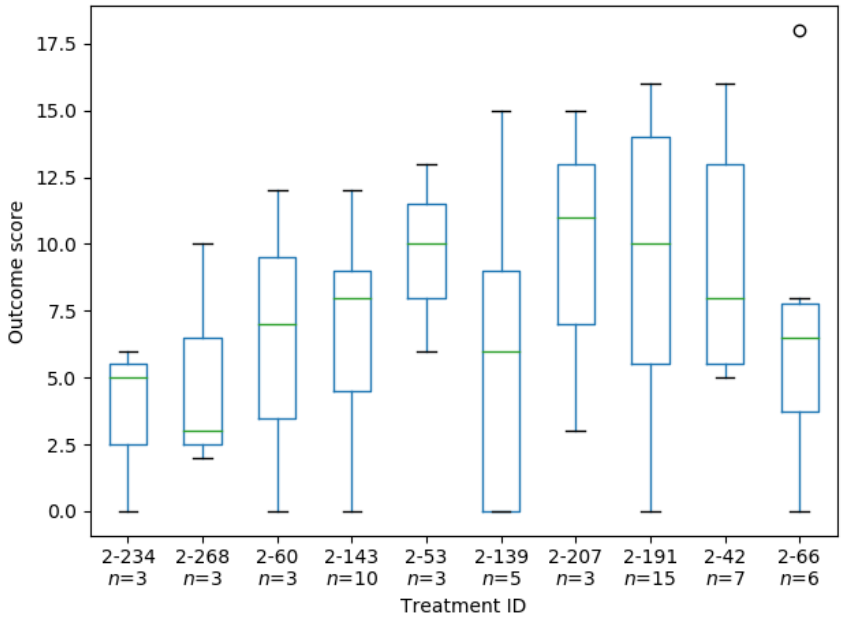

FIGURE C.2: Boxplot of outcome score for the 10 treatments with lowest 95th percentile for outcome score under the second definition of treatment.

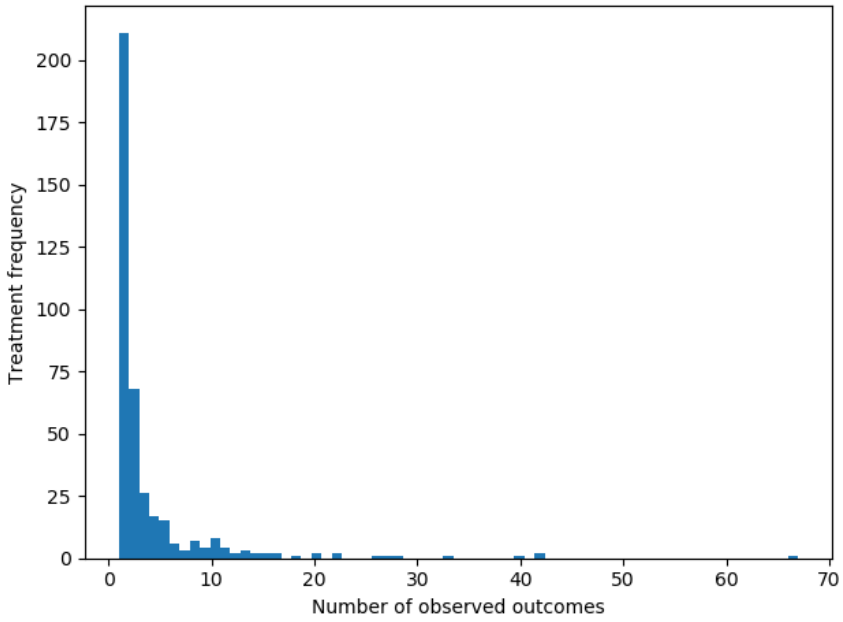

FIGURE C.3: Histogram of treatment frequency with respect to number of observed outcome scores under the first definition of treatment.

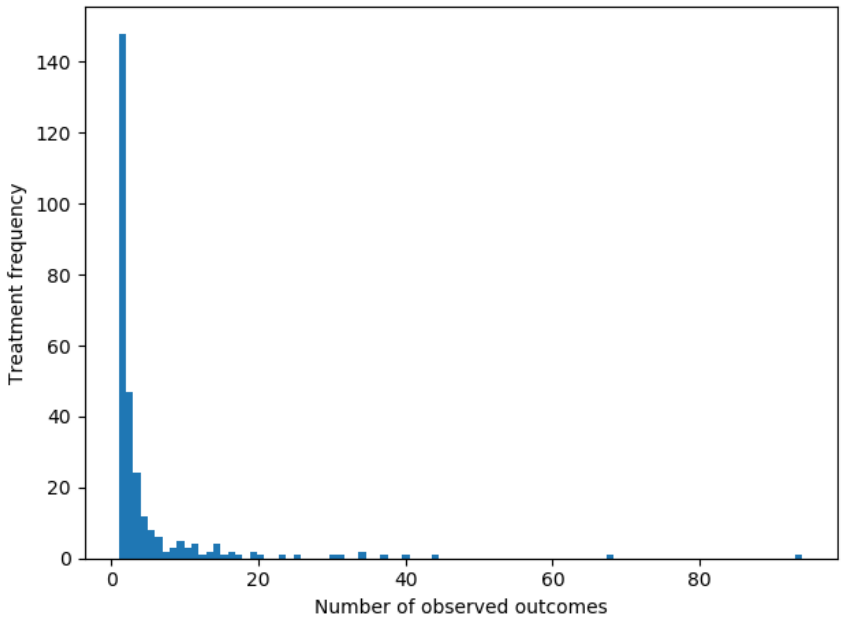

FIGURE C.4: Histogram of treatment frequency with respect to number of observed outcome scores under the second definition of treatment.

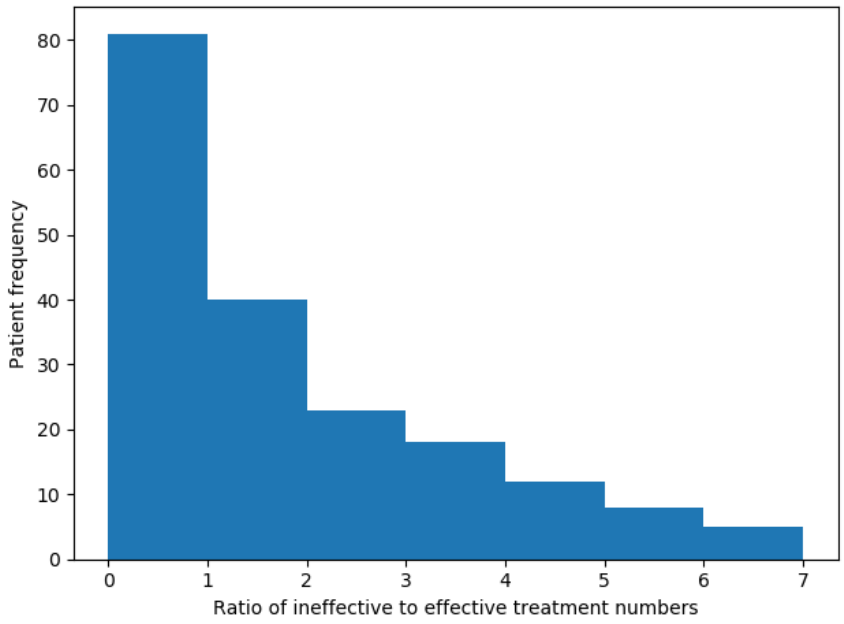

FIGURE C.5: Histogram of the ratio of ineffective to effective treatment numbers for patients with at least one remission under the second definition of treatment.

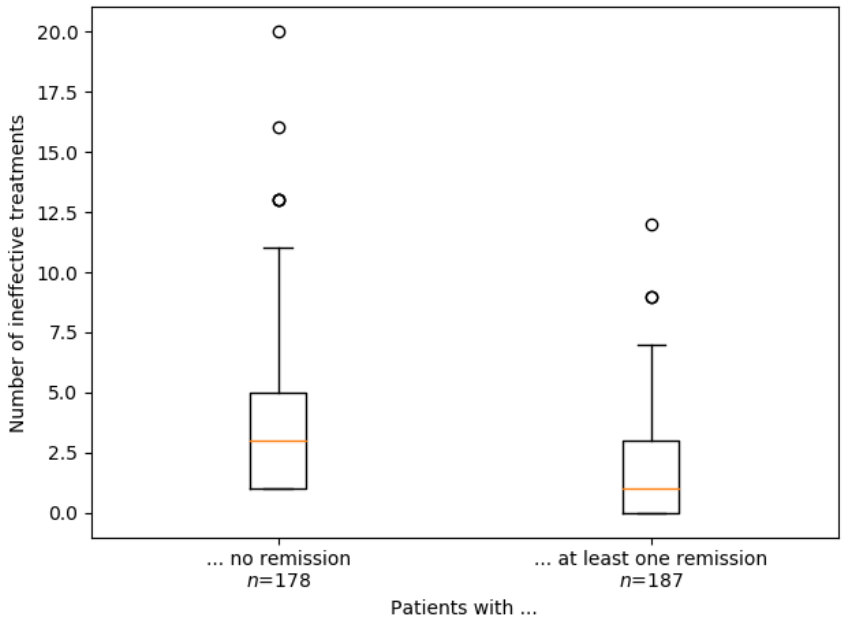

FIGURE C.6: Boxplot of the number of ineffective treatments for patients with no remission and at least one remission under the second definition of treatment.

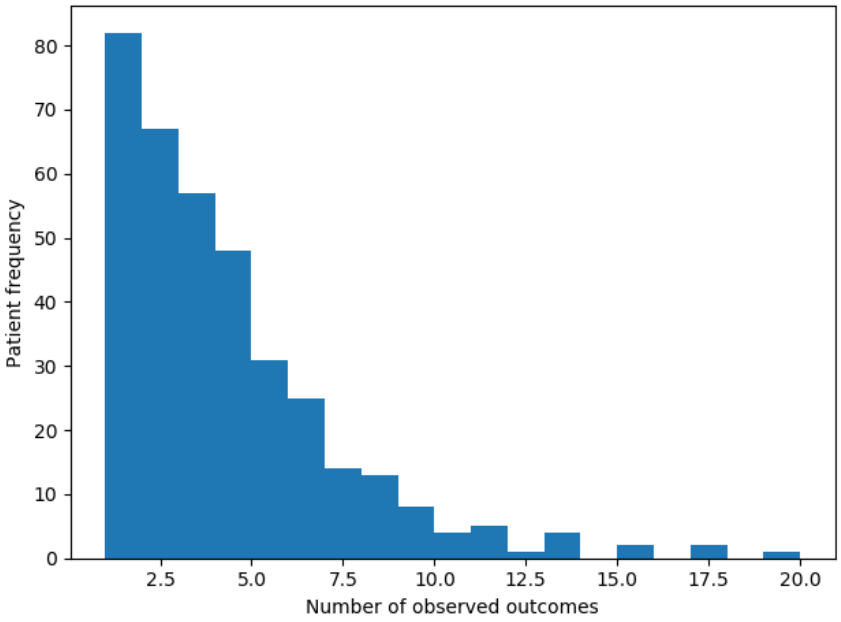

FIGURE C.7: Histogram of patient frequency with respect to number of observed outcome scores under the first definition of treatment.

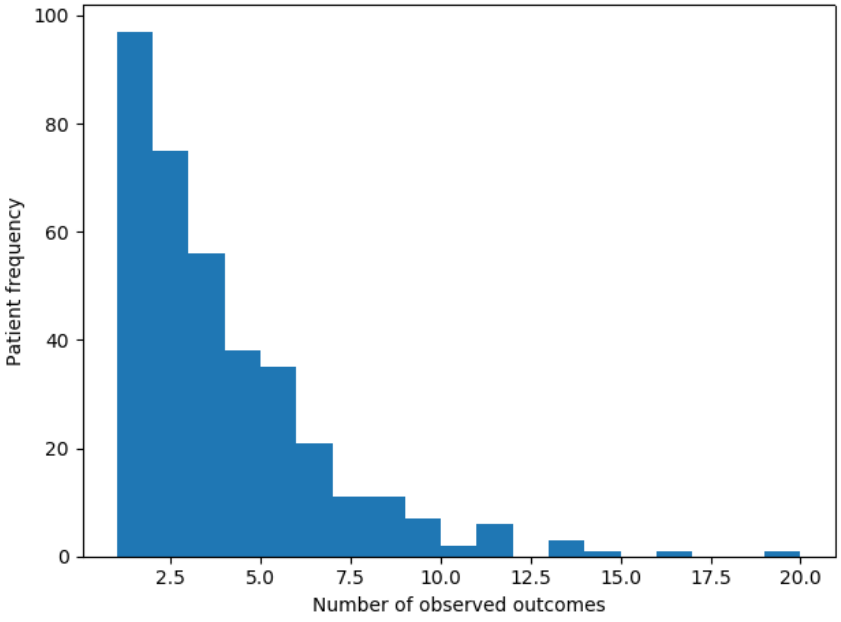

FIGURE C.8: Histogram of patient frequency with respect to number of observed outcome scores under the second definition of treatment.

C.3 Implementation Details

First, note that all the models are implemented using the code available at Cousineau (2019), my fork of the Surprise library (Hug, 2017). Thus, we will refer to each model using their class's name and parameters.

Second, the implementation of the propensity model consists in the `FM` class with the following defined and searched hyper-parameters: `rating_lst` $=$ `['userID', 'itemID']`, `user_lst` and `item_lst` defined respectively with the user and item features of Section 5.4.1, `n_factors` $\sim \mathcal{U}\{20, 300\}$, `dev_ratio` $= 0.3$, `patience` $= 20$, `n_epochs` $= 100$, `init_std` $= 0.01$, `lr` $\sim$ *log*$\mathcal{U}(0.00001, 0.1)$, `reg` $\sim$ *log*$\mathcal{U}(0.00001, 0.1)$, `refit` $=$ `True`, `binary` $=$ `True` and `random_state` $= 123$. We execute the `RandomizedSearchCV` class for 10 iterations of a 5-fold cross-validation and keep the best parameters according to the mean log loss to retrain the model on the full exposure training set.

Third, for each outcome model, the implementation and the distributions of the different hyper-parameters for the random search are provided below.

- Constant: The constant model consists in the `GlobalOnly` class. There are no hyper-parameters for this model.

- Baseline: This model consists in the `BaselineOnly` class. It is fitted by minimizing the L2-regularized RMSE between r_{ij} and $\hat{r}_{ij} = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_j$ using alternating least squares (ALS). This loss is weighted when using IPTW. For this model, the searched hyper-parameters are `reg_i` $\sim$ *log*$\mathcal{U}(0.001, 1000)$, `reg_u` $\sim$ *log*$\mathcal{U}(0.001, 1000)$ and `n_epochs` $\sim \mathcal{U}\{5, 100\}$, where $\mathcal{U}\{\cdot, \cdot\}$ corresponds to the discrete uniform distribution, $\mathcal{U}(\cdot, \cdot)$ corresponds to the continuous uniform distribution and *log*$\mathcal{U}(\cdot, \cdot)$ corresponds to the loguniform distribution, i.e., *log*$\mathcal{U}(a, b) = \exp \mathcal{U}(\log(a), \log(b))$.

- factorization machine (FM) models: μ, α and u are fitted by minimizing the L2-regularized RMSE between r_{ij} and $\hat{r}_{ij}$ (see Equation 5.4 for the definition of $\hat{r}_{ij}$) using Adam. This loss is weighted when using IPTW. If only the user or item is unknown, then the model uses a x vectors with values of zeros for the features associated respectively with the user or the item. The defined and searched hyper-parameters for all FM models are: `n_factors` $\sim \mathcal{U}\{20, 300\}$, `dev_ratio` $= 0.3$, `patience` $= 20$, `n_epochs` $= 300$, `init_std` $= 0.01$, `lr` $\sim$ *log*$\mathcal{U}(0.00001, 0.1)$, `reg` $\sim$ *log*$\mathcal{U}(0.00001, 0.1)$, `refit` $=$ `True` and `random_state` $= 123$. We now define the additional parameters of each FM model:

- FM-base: This model consists in the `FM` class with `rating_lst = ['userID','itemID']`.
- FM-features: This model consists in the `FM` class with `rating_lst = ['userID','itemID']`, and `user_lst` and `item_lst` defined respectively with the user and item features of Section 5.4.1.
- FM-outcomes: This model consists in the `FM` class with `rating_lst = ['userID','itemID','exp_u_rating']`.
- FM-full: This model consists in the `FM` class with `rating_lst = ['userID','itemID','exp_u_rating']`, and `user_lst` and `item_lst` defined respectively with the user and item features of Section 5.4.1.

Note that the predictions of all outcome models are clipped to the 0–27 range. Also note, that for unknown user and item, all outcome models return the same prediction as the Constant or IPTW-Constant models.

Finally, for each outcome model that requires randomized search, we execute the `RandomizedSearchCV` class for 500 iterations of a 5-fold cross-validation. We then keep the best parameters according to the mean RMSE to retrain the models on the full training set.

C.4 Additional Results

TABLE C.4: RMSE, MAE, FCP, NDCG and F1 metrics of the Constant, Baseline, FM-base, FM-features, FM-outcomes and FM-full models, and their IPTW variants on the random test set under the second definition of treatment.

Model	RMSE	MAE	FCP	NDCG	F1
Constant	6.92	5.88	0.000	0.931	0.000
IPTW-Constant	6.93	5.88	0.000	0.931	0.000
Baseline	5.60	4.46	0.487	0.935	0.049
IPTW-Baseline	5.55	4.44	0.475	**0.936**	0.050
FM-base	6.39	5.39	0.487	0.929	0.000
IPTW-FM-base	6.36	5.37	0.481	0.932	0.000
FM-features	5.99	4.88	0.506	0.929	0.000
IPTW-FM-features	6.04	4.96	**0.513**	0.928	0.000
FM-outcomes	6.66	5.04	0.468	0.929	0.434
IPTW-FM-outcomes	7.40	5.73	0.449	0.934	**0.483**
FM-full	5.42	4.35	0.500	0.926	0.000
IPTW-FM-full	**5.41**	**4.34**	0.487	0.924	0.000

TABLE C.5: RMSE, MAE, FCP, NDCG and F1 metrics of the Constant, Baseline, FM-base, FM-features, FM-outcomes and FM-full models, and their IPTW variants on the first intervened test set under the second definition of treatment. This first intervened test set is sampled proportionally to the inverse frequency of items in the training set.

Model	RMSE	MAE	FCP	NDCG	F1
Constant	6.97	5.90	0.000	0.942	0.000
IPTW-Constant	6.95	5.89	0.000	0.942	0.000
Baseline	5.35	4.33	0.355	0.948	0.000
IPTW-Baseline	**5.27**	**4.25**	0.231	0.946	0.000
FM-base	6.61	5.56	0.368	**0.953**	0.000
IPTW-FM-base	6.54	5.51	0.357	0.950	0.000
FM-features	5.86	4.87	0.471	0.938	0.000
IPTW-FM-features	5.83	4.74	0.460	0.936	0.000
FM-outcomes	5.98	4.52	0.319	0.941	0.400
IPTW-FM-outcomes	6.80	5.53	0.348	0.946	**0.407**
FM-full	5.33	4.34	**0.564**	0.933	0.000
IPTW-FM-full	5.30	4.31	0.531	0.935	0.000

TABLE C.6: RMSE, MAE, FCP, NDCG and F1 metrics of the Constant, Baseline, FM-base, FM-features, FM-outcomes and FM-full models, and their IPTW variants on the second intervened test set under the second definition of treatment. This second intervened test set is sampled proportionally to the inverse rating.

Model	RMSE	MAE	FCP	NDCG	F1
Constant	8.81	7.96	0.000	0.968	0.000
IPTW-Constant	8.90	8.04	0.000	0.968	0.000
Baseline	7.89	6.73	0.500	**0.979**	0.078
IPTW-Baseline	7.89	6.73	0.464	0.976	0.078
FM-base	7.96	7.12	0.476	0.975	0.000
IPTW-FM-base	7.95	7.10	**0.583**	0.978	0.000
FM-features	7.13	6.16	0.488	0.975	0.000
IPTW-FM-features	8.39	7.33	**0.583**	0.976	0.000
FM-outcomes	7.32	5.31	0.345	0.965	0.655
IPTW-FM-outcomes	7.02	**4.83**	0.476	0.973	**0.667**
FM-full	6.87	5.72	0.429	0.967	0.000
IPTW-FM-full	**6.86**	5.72	0.429	0.967	0.000

References

Abbeel, Pieter and Andrew Y. Ng (2004). "Apprenticeship learning via inverse reinforcement learning". In: *Twenty-first international conference on Machine learning - ICML '04*. New York, New York, USA: ACM Press. DOI: 10.1145/1015330.1015430.

Abrevaya, Jason, Yu-Chin Hsu, and Robert P. Lieli (2015). "Estimating Conditional Average Treatment Effects". In: *Journal of Business & Economic Statistics* 33.4, pp. 485–505. DOI: 10.1080/07350015.2014.975555.

Aggarwal, Charu C. (2016). *Recommender systems: the textbook*. Springer International Publishing.

Alagoz, Oguzhan, Turgay Ayer, and Fatih Safa Erenay (2010). "Operations Research Models for Cancer Screening". In: *Wiley Encyclopedia of Operations Research and Management Science*. Ed. by James J Cochran et al. John Wiley and Sons. DOI: 10.1002/9780470400531.eorms0597.

Alagoz, Oguzhan et al. (2004). "The Optimal Timing of Living-Donor Liver Transplantation". In: *Management Science* 50.10, pp. 1420–1430. DOI: 10.1287/mnsc.1040.0287.

— (2007a). "Choosing Among Living-Donor and Cadaveric Livers". In: *Management Science* 53.11, pp. 1702–1715. DOI: 10.1287/mnsc.1070.0726.

— (2007b). "Determining the Acceptance of Cadaveric Livers Using an Implicit Model of the Waiting List". In: *Operations Research* 55.1, pp. 24–36. DOI: 10.1287/opre.1060.0329.

Alagoz, Oguzhan et al. (2010). "Markov decision processes: a tool for sequential decision making under uncertainty." In: *Medical decision making : an international journal of the Society for Medical Decision Making* 30.4, pp. 474–483. DOI: 10.1177/0272989X09353194.

Amato, Filippo et al. (2013). "Artificial neural networks in medical diagnosis". In: *Journal of Applied Biomedicine* 11.2, pp. 47–58. DOI: 10.2478/v10136-012-0031-x.

Andersen, M. S., J. Dahl, and L. Vandenberghe (2015). *CVXOPT: A Python package for convex optimization, version 1.1.8*. URL: http://cvxopt.org/.

Ansari, Asim, Skander Essegaier, and Rajeev Kohli (2000). "Internet Recommendation Systems". In: *Journal of Marketing Research* 37.3, pp. 363–375. DOI: 10.1509/jmkr.37.3.363.18779.

Aronszajn, N. (1950). "Theory of Reproducing Kernels". In: *Transactions of the American Mathematical Society* 68.3, pp. 337–404. DOI: 10.2307/1990404.

Arthur, David and Sergei Vassilvitskii (2007). "K-Means++: the Advantages of Careful Seeding". In: *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1027–1025. DOI: 10.1145/1283383.1283494.

Ay-Woan, Pan et al. (2006). "Quality of life in depression: predictive models." In: *Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation* 15.1, pp. 39–48. DOI: 10.1007/s11136-005-0381-x.

Ayer, Turgay, Oguzhan Alagoz, and Natasha K. Stout (2012). "OR Forum—A POMDP Approach to Personalize Mammography Screening Decisions". In: *Operations Research* 60.5, pp. 1019–1034. DOI: 10.1287/opre.1110.1019.

Bar, Ariel et al. (2013). "Improving simple collaborative filtering models using ensemble methods". In: *MCS*, pp. 1–12.

Beck, Aaron T., Maria Kovacs, and Arlene Weissman (1979). "Assessment of suicidal intention: The Scale for Suicide Ideation." In: *Journal of Consulting and Clinical Psychology* 47.2, pp. 343–352. DOI: 10.1037/0022-006X.47.2.343.

Bellazzi, Riccardo and Blaz Zupan (2008). "Predictive data mining in clinical medicine: current issues and guidelines." In: *International journal of medical informatics* 77.2, pp. 81–97. DOI: 10.1016/j.ijmedinf.2006.11.006.

Bellman, R. E. (1957). *Dynamic programming*. Princeton, NJ, USA: Princeton University Press.

Bennett, Casey C and Kris Hauser (2013). "Artificial intelligence framework for simulating clinical decision-making: a Markov decision process approach." In: *Artificial intelligence in medicine* 57.1, pp. 9–19. DOI: 10.1016/j.artmed.2012.12.003.

Berlim, Marcelo T., Marcelo P. Fleck, and Gustavo Turecki (2008). "Current trends in the assessment and somatic treatment of resistant/refractory major depression: an overview." In: *Annals of medicine* 40.2, pp. 149–59. DOI: 10.1080/07853890701769728.

Berlim, Marcelo T. and Gustavo Turecki (2007). "Definition, Assessment, and Staging of Treatment-Resistant Refractory Major Depression: A Review of Current Concepts and Methods". In: *The Canadian Journal of Psychiatry* 52.1, pp. 46–54. DOI: 10.1177/070674370705200108.

Berman, Margit I. and Mark T. Hegel (2014). "Predicting depression outcome in mental health treatment: A recursive partitioning analysis". In: *Psychotherapy Research* 24.6, pp. 675–686. DOI: 10.1080/10503307.2013.874053.

Bertsekas, D. P. (2005). *Dynamic Programming and Optimal Control, Vol. I*. 3rd ed. Belmont, MA, USA: Athena Scientific.

— (2012). *Dynamic Programming and Optimal Control: Approximate Dynamic Programming, Vol. II*. 4th ed. Belmont, MA, USA: Athena Scientific.

Bertsekas, D. P. and J. Tsitsiklis (1996). *Neuro-dynamic programming*. Belmont, MA, USA: Athena Scientific.

Bertsimas, Dimitris, Vivek F. Farias, and Nikolaos Trichakis (2013). "Fairness, Efficiency, and Flexibility in Organ Allocation for Kidney Transplantation". In: *Operations Research* 61.1, pp. 73–87. DOI: 10.1287/opre.1120.1138.

Bhattacharya, S (2014). "Markov Chain Model to Explain the Dynamics of Human Depression". In: *Journal of Nonlinear Dynamics* 2014, pp. 1–9.

Bishop, Christopher M (2006). *Pattern Recognition and Machine Learning*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Blatt, D., SA Murphy, and J. Zhu (2004). *A-learning for approximate planning*. Tech. rep. Ann Arbor, MI, USA: University of Michigan.

Bogucki, Robert (2016). *Which whale is it, anyway? Face recognition for right whales using deep learning*. URL: https://deepsense.ai/deep-learning-right-whale-recognition-kaggle/ (visited on 08/02/2018).

Bortfeld, Thomas et al. (2008). "Robust Management of Motion Uncertainty in Intensity-Modulated Radiation Therapy". In: *Operations Research* 56.6, pp. 1461–1473. DOI: 10.1287/opre.1070.0484.

Boyd, Stephen P. and Lieven. Vandenberghe (2004). *Convex optimization*. New York, NY, USA: Cambridge University Press.

Bronzino, J D, R A Morelli, and J W Goethe (1989). "OVERSEER: a prototype expert system for monitoring drug treatment in the psychiatric clinic." In: *IEEE transactions on bio-medical engineering* 36.5, pp. 533–540. DOI: 10.1109/10.24255.

Brookhart, M. Alan et al. (2010). "Confounding Control in Healthcare Database Research". In: *Medical Care* 48.6-1, S114–S120. DOI: 10.1097/MLR.0b013e3181dbebe3.

Brun, Armelle, Marharyta Aleksandrova, and Anne Boyer (2014). "Can Latent Features Be Interpreted as Users in Matrix Factorization-Based Recommender Systems?" In: *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*. IEEE, pp. 226–233. DOI: 10.1109/WI-IAT.2014.102.

Burke, Robin (2002). "Hybrid Recommender Systems: Survey and Experiments". In: *User Modeling and User-Adapted Interaction* 12.4, pp. 331–370. DOI: 10.1023/A:1021240730564.

Burke, Wylie and Bruce M Psaty (2007). "Personalized medicine in the era of genomics." In: *JAMA* 298.14, pp. 1682–1684. DOI: 10.1001/jama.298.14.1682.

Cain, Lauren E et al. (2010). "When to start treatment? A systematic approach to the comparison of dynamic regimes using observational data." In: *The international journal of biostatistics* 6.2, pp. 1–24.

Carmona, Ivan Sanchez and Sebastian Riedel (2015). "Extracting interpretable models from matrix factorization models". In: *CEUR Workshop Proceedings*, pp. 1–7.

Chakraborty, Bibhas, Eric B Laber, and Yingqi Zhao (2013). "Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme." In: *Biometrics* 69.3, pp. 714–23. DOI: 10.1111/biom.12052.

Chakraborty, Bibhas and Erica E. M. Moodie (2013). *Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine*. New York, NY, USA: Springer Science+Business Media.

Chakraborty, Bibhas, Susan Murphy, and Victor Strecher (2010). "Inference for non-regular parameters in optimal dynamic treatment regimes." In: *Statistical methods in medical research* 19.3, pp. 317–343. DOI: 10.1177/0962280209105013.

Chakraborty, Bibhas and Susan A Murphy (2014). "Dynamic Treatment Regimes." In: *Annual review of statistics and its application* 1, pp. 447–464. DOI: 10.1146/annurev-statistics-022513-115553.

Chan, Carri W, Vivek F Farias, and Gabriel J Escobar (2017). "The Impact of Delays on Service Times in the Intensive Care Unit". In: *Management Science* 63.7, pp. 2049–2072. DOI: 10.1287/mnsc.2016.2441.

Chan, Timothy C. Y. et al. (2014). "Generalized Inverse Multiobjective Optimization with Application to Cancer Therapy". In: *Operations Research* 62.3, pp. 680–695. DOI: 10.1287/opre.2014.1267.

Chaovalitwongse, W. Art (2009). "Optimization and Data Mining in Epilepsy Research: A Review and Prospective". In: *Handbook of Optimization in Medicine*. Springer. Chap. 10, pp. 325–356.

Chhatwal, Jagpreet, Oguzhan Alagoz, and Elizabeth S Burnside (2010). "Optimal Breast Biopsy Decision-Making Based on Mammographic Features and Demographic Factors." In: *Operations research* 58.6, pp. 1577–1591. DOI: 10.1287/opre.1100.0877.

Chinchor, Nancy (1992). "MUC-4 Evaluation Metrics". In: *Proceedings of the Fourth Message Understanding Conference*, pp. 22–29.

Choi, Edward et al. (2017). "Generating Multi-label Discrete Patient Records using Generative Adversarial Networks". In: *Proceedings of Machine Learning for Healthcare 2017*.

Choi, J and KE Kim (2013). "Bayesian Nonparametric Feature Construction for Inverse Reinforcement Learning." In: *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence Bayesian*, pp. 1287–1293.

Choi, Jaedeug and Kee-Eung Kim (2011). "MAP Inference for Bayesian Inverse Reinforcement Learning". In: *NIPS'11 Proceedings of the 24th International Conference on Neural Information Processing Systems*, pp. 1989–1997.

Cios, Krzysztof J. and G. William Moore (2002). "Uniqueness of medical data mining". In: *Artificial Intelligence in Medicine* 26.1-2, pp. 1–24. DOI: 10.1016/S0933-3657(02)00049-0.

Cooper, K. et al. (2006). "A review of health care models for coronary heart disease interventions." In: *Health care management science* 9.4, pp. 311–324. DOI: 10.1007/s10729-006-9996-x.

Cotton, Cecilia A. and Patrick J. Heagerty (2011). "A Data Augmentation Method for Estimating the Causal Effect of Adherence to Treatment Regimens Targeting Control of an Intermediate Measure". In: *Statistics in Biosciences* 3.1, pp. 28–44. DOI: 10.1007/s12561-011-9038-1.

Cousineau, Martin (2019). *Surprise, a Python library for recommender systems*. URL: https://github.com/martincousi/Surprise.

Daume, H. and D. Marcu (2006). "Domain Adaptation for Statistical Classifiers". In: *Journal Of Artificial Intelligence Research* 26, pp. 101–126. DOI: 10.1613/jair.1872.

Dawid, A. Philip and Vanessa Didelez (2010). "Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview". In: *Statistics Surveys* 4, pp. 184–231. DOI: 10.1214/10-SS081.

De Boeck, L., J. Beliën, and W. Egyed (2014). "Dose optimization in high-dose-rate brachytherapy: A literature review of quantitative models from 1990 to 2010". In: *Operations Research for Health Care* 3.2, pp. 80–90. DOI: 10.1016/j.orhc.2013.12.004.

de Man-van Ginkel, J. M. et al. (2013). "In-Hospital Risk Prediction for Post-stroke Depression: Development and Validation of the Post-stroke Depression Prediction Scale". In: *Stroke* 44.9, pp. 2441–2445. DOI: 10.1161/STROKEAHA.111.000304.

Demic, Selver and Sen Cheng (2014). "Modeling the dynamics of disease states in depression." In: *PloS one* 9.10, pp. 1–14. DOI: 10.1371/journal.pone.0110358.

Denton, Brian T et al. (2009). "Optimizing the start time of statin therapy for patients with diabetes." In: *Medical decision making : an international journal of the Society for Medical Decision Making* 29.3, pp. 351–367. DOI: 10.1177/0272989X08329462.

Denton, Brian T. et al. (2011). "Medical decision making: open research challenges". In: *IIE Transactions on Healthcare Systems Engineering* 1.3, pp. 161–167. DOI: 10.1080/19488300.2011.619157.

Doshi-Velez, Finale and Been Kim (2017). "Towards A Rigorous Science of Interpretable Machine Learning".

Dougherty, James, Ron Kohavi, and Mehran Sahami (1995). "Supervised and Unsupervised Discretization of Continuous Features". In: *Machine Learning Proceedings 1995*, pp. 194–202. DOI: 10.1016/B978-1-55860-377-6.50032-3.

Efron, Bradley (1981). "Nonparametric standard errors and confidence intervals". In: *Canadian Journal of Statistics* 9.2, pp. 139–158. DOI: 10.2307/3314608.

Erenay, Fatih Safa, Oguzhan Alagoz, and Adnan Said (2014). "Optimizing Colonoscopy Screening for Colorectal Cancer Prevention and Surveillance". In: *Manufacturing & Service Operations Management* 16.3, pp. 381–400. DOI: 10.1287/msom.2014.0484.

Ernst, Damien, Pierre Geurts, and Louis Wehenkel (2005). "Tree-Based Batch Mode Reinforcement Learning". In: *The Journal of Machine Learning Research* 6, pp. 503–556.

Esteban, Cristóbal, Stephanie L. Hyland, and Gunnar Rätsch (2017). "Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs".

Fava, Maurizio et al. (2003). "Background and rationale for the sequenced treatment alternatives to relieve depression (STAR*D) study." In: *The Psychiatric clinics of North America* 26.2, pp. 457–494.

Feuerverger, Andrey, Yu He, and Shashi Khatri (2012). "Statistical Significance of the Netflix Challenge". In: *Statistical Science* 27.2, pp. 202–231. DOI: 10.1214/11-STS368.

First, Michael B. et al. (1997). *Structured Clinical Interview for DSM-IV Axis II Personality Disorders, (SCID-II)*. Washington, D.C., USA: American Psychiatric Press, Inc.

First, Michael B. et al. (2002). *Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Research Version, Patient Edition, (SCID-I/P)*. New York, NY, USA: Biometrics Research, New York State Psychiatric Institute.

"Analytics" (2013). In: *Encyclopedia of operations research and management science*. Ed. by Saul I Gass and Michael C Fu. Springer US, pp. 72–72.

Goldberg, Yair, Rui Song, and Michael R. Kosorok (2013). "Adaptive Q-learning". In: *From Probability to Statistics and Back: High-Dimensional Models and Processes*. Ed. by M. Banerjee et al. Beachwood, Ohio, USA: Institute of Mathematical Statistics, pp. 150–162.

Goodfellow, Ian J. et al. (2014). "Generative Adversarial Nets". In: *Proceedings of the International Conference on Neural Information Processing Systems (NIPS 2014)*, pp. 2672–2680.

Gudayol-Ferré, Esteve et al. (2012). "Prediction of remission of depression with clinical variables, neuropsychological performance, and serotonergic/dopaminergic gene polymorphisms". In: *Human Psychopharmacology: Clinical and Experimental* 27.6, pp. 577–586. DOI: 10.1002/hup.2267.

Hainmueller, J. (2012). "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies". In: *Political Analysis* 20.1, pp. 25–46. DOI: 10.1093/pan/mpr025.

Hainmueller, Jens (2014). *ebal: Entropy reweighting to create balanced samples. R package version 0.1-6.* URL: https://cran.r-project.org/package=ebal.

Hamilton, Max (1959). "The Assessment of Anxiety States by Rating". In: *British Journal of Medical Psychology* 32.1, pp. 50–55. DOI: 10.1111/j.2044-8341.1959.tb00467.x.

— (1960). "A Rating Scale for Depression". In: *J. Neurol. Neurosurg. Psychiat* 23, pp. 56–62. DOI: 10.1136/jnnp.23.1.56.

Harper, P. R. and S. K. Jones (2005). "Mathematical Models for the Early Detection and Treatment of Colorectal Cancer". In: *Health Care Management Science* 8.2, pp. 101–109. DOI: 10.1007/s10729-005-0393-7.

Hasler, Gregor (2010). "Pathophysiology of depression: do we have any solid evidence of interest to clinicians?" In: *World psychiatry : official journal of the World Psychiatric Association (WPA)* 9.3, pp. 155–161. DOI: 10.1002/j.2051-5545.2010.tb00298.x.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd ed. Springer-Verlag New York, Inc.

Hazen, G. B. (2004). "Dynamic influence diagrams: Applications to medical decision modeling". In: *Operations Research and Health Care*. Ed. by M. L. Brandeau, F. Sainfort, and W. P. Pierskalla. Boston, MA: Kluwer Academic Publishers. Chap. 24, pp. 613–638.

He, Miao, Lei Zhao, and Warren B. Powell (2010). "Optimal control of dosage decisions in controlled ovarian hyperstimulation". In: *Annals of Operations Research* 178.1, pp. 223–245. DOI: 10.1007/s10479-009-0563-y.

— (2012). "Approximate dynamic programming algorithms for optimal dosage decisions in controlled ovarian hyperstimulation". In: *European Journal of Operational Research* 222.2, pp. 328–340. DOI: 10.1016/j.ejor.2012.03.049.

Heckel, Reinhard et al. (2017). "Scalable and interpretable product recommendations via overlapping co-clustering". In: *Proceedings - International Conference on Data Engineering*, pp. 1033–1044. DOI: 10.1109/ICDE.2017.149.

Helm, Jonathan E. et al. (2015). "Dynamic Forecasting and Control Algorithms of Glaucoma Progression for Clinician Decision Support". In: *Operations Research* 63.5, pp. 979–999. DOI: 10.1287/opre.2015.1405.

Hernán, M. A. and J. M. Robins (2018). *Causal Inference*. Boca Raton: Chapman & Hall/CRC, forthcoming.

Hernán, Miguel A et al. (2006). "Comparison of dynamic treatment regimes via inverse probability weighting." In: *Basic & clinical pharmacology & toxicology* 98.3, pp. 237–242. DOI: 10.1111/j.1742-7843.2006.pto_329.x.

Hernandez-Lobato, Jose Miguel, Neil Houlsby, and Zoubin Ghahramani (2014). "Probabilistic Matrix Factorization with Non-random Missing Data". In: *Proceedings of The 31st International Conference on Machine Learning*. Vol. 32, pp. 1512–1520.

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.

Ho, Jonathan and Stefano Ermon (2016). "Generative Adversarial Imitation Learning". In: *NIPS*.

Ho, Teck-hua et al. (2017). "OM Forum—Causal Inference Models in Operations Management". In: *Manufacturing & Service Operations Management* 19.4, pp. 509–525. DOI: 10.1287/msom.2017.0659.

Holland, Paul W. (1986). "Statistics and Causal Inference". In: *Journal of the American Statistical Association* 81.396, pp. 945–960. DOI: 10.1080/01621459.1986.10478354.

Hsih, K. W. (2010). "Optimal dosing applied to glycemic control for type 2 diabetes". PhD thesis, p. 129.

Huang, J., Alexander J. Smola, and Bernhard Schölkopf (2006). *Correcting Sample Selection Bias by Unlabeled Data*. Tech. rep. Univeristy of Waterloo, p. 20.

Huang, J. et al. (2007). "Correcting Sample Selection Bias by Unlabeled Data". In: *Advances in Neural Information Processing Systems 19*, pp. 601–608.

Huang, Sandy H et al. (2014). "Toward personalizing treatment for depression: predicting diagnosis and severity". In: *Journal of the American Medical Informatics Association : JAMIA* 21.6, pp. 1069–1075. DOI: 10.1136/amiajnl-2014-002733.

Hug, Nicolas (2017). *Surprise, a Python library for recommender systems*. URL: http://surpriselib.com.

Hunter, David (2006). "First, gather the data." In: *The New England journal of medicine* 354.4, pp. 329–331. DOI: 10.1056/NEJMp058235.

Hyvönen, Saara, Pauli Miettinen, and Evimaria Terzi (2008). "Interpretable Nonnegative Matrix Decompositions Categories and Subject Descriptors". In: *KDD'08*, pp. 345–353.

Ibrahim, Rouba et al. (2016). "Designing Personalized Treatment: An Application to Anticoagulation Therapy". In: *Production and Operations Management* 25.5, pp. 902–918. DOI: 10.1111/poms.12514.

Imbens, Guido W and Jeffrey M Wooldridge (2009). "Recent Developments in the Econometrics of Program Evaluation". In: *Journal of Economic Literature* 47.1, pp. 5–86. DOI: 10.1257/jel.47.1.5.

INFORMS (2017). *Operations Research \& Analytics*. URL: https://www.informs.org/Explore/Operations-Research-Analytics (visited on 08/09/2017).

Institute of Medicine (2013). *Best Care at Lower Cost: The Path to Continously Learning Health Care in America*. Tech. rep. Washington, D.C., USA. DOI: 10.17226/13444.

Ivy, Julie Simmons (2009). "Can We Do Better? Optimization Models for Breast Cancer Screening". In: *Handbook of Optimization in Medicine*. Vol. 26. Springer Optimization and Its Applications. Boston, MA: Springer US. Chap. 2, pp. 25–52. DOI: 10.1007/b100322.

Jahrer, Michael, Andreas Töscher, and Robert Legenstein (2010). "Combining Predictions for an accurate Recommender System". In: *KDD '10 The 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Järvelin, Kalervo and Jaana Kekäläinen (2002). "Cumulated gain-based evaluation of IR techniques". In: *ACM Transactions on Information Systems* 20.4, pp. 422–446. DOI: 10.1145/582415.582418.

Jiang, Daniel R. and Warren B. Powell (2015). "An Approximate Dynamic Programming Algorithm for Monotone Value Functions". In: *Operations Research* 63.6, pp. 1489–1511. DOI: 10.1287/opre.2015.1425.

Jiang, Jing (2008). *A Literature Survey on Domain Adaptation of Statistical Classifiers*. Tech. rep. University of Illinois at Urbana-Champaign.

Johansson, Fredrik D, Uri Shalit, and David Sontag (2016). "Learning Representations for Counterfactual Inference". In: *ICML*.

Kaelbling, L. P., M. L. Littman, and A. Moore (1996). "Reinforcement learning: A survey". In: *The Journal of Artificial Intelligence Research* 4, pp. 237–385.

Kaggle Team (2015). *Taxi Trajectory Winners' Interview: 1st place, Team ?* URL: http://blog.kaggle.com/2015/07/27/taxi-trajectory-winners-interview-1st-place-team (visited on 08/02/2018).

Kahruman, Sera et al. (2010). "Scheduling the adjuvant endocrine therapy for early stage breast cancer". In: *Annals of Operations Research* 196.1, pp. 683–705. DOI: 10.1007/s10479-010-0741-y.

Kallus, N (2017). "A Framework for Optimal Matching for Causal Inference". In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*. Vol. 54, pp. 372–381.

KC, Diwas S and Christian Terwiesch (2009). "Impact of Workload on Service Time and Patient Safety: An Econometric Analysis of Hospital Operations". In: *Management Science* 55.9, pp. 1486–1498. DOI: 10.1287/mnsc.1090.1037.

KC, Diwas Singh and Christian Terwiesch (2011). "The Effects of Focus on Performance: Evidence from California Hospitals". In: *Management Science* 57.11, pp. 1897–1912. DOI: 10.1287/mnsc.1110.1401.

Kennedy, Sidney H. et al. (2016). "Canadian Network for Mood and Anxiety Treatments (CANMAT) 2016 Clinical Guidelines for the Management of Adults with Major Depressive Disorder: Section 3. Pharmacological Treatments". In: *The Canadian Journal of Psychiatry* 61.9, pp. 540–560. DOI: 10.1177/0706743716659417.

Keren, Baruch and Joseph S Pliskin (2011). "Optimal timing of joint replacement using mathematical programming and stochastic programming models." In: *Health care management science* 14.4, pp. 361–9. DOI: 10.1007/s10729-011-9172-9.

Kim, Beomjoon and Joelle Pineau (2016). "Socially Adaptive Path Planning in Human Environments Using Inverse Reinforcement Learning". In: *International Journal of Social Robotics* 8.1, pp. 51–66. DOI: 10.1007/s12369-015-0310-2.

Kim, Song-Hee et al. (2015). "ICU Admission Control: An Empirical Study of Capacity Allocation and Its Implication for Patient Outcomes". In: *Management Science* 61.1, pp. 19–38. DOI: 10.1287/mnsc.2014.2057.

Kononenko, Igor (2001). "Machine learning for medical diagnosis: history, state of the art and perspective". In: *Artificial Intelligence in Medicine* 23.1, pp. 89–109. DOI: 10.1016/S0933-3657(01)00077-X.

Koren, Yehuda (2010). "Factor in the neighbors: scalable and accurate collaborative filtering". In: *ACM Transactions on Knowledge Discovery from Data* 4.1, pp. 1–24. DOI: 10.1145/1644873.1644874.

Koren, Yehuda and Joseph Sill (2013). "Collaborative Filtering on Ordinal User Feedback". In: *Twenty-Third International Joint Conference on Artificial Intelligence*, pp. 3022–3026.

Kullback, S. (1968). *Information theory and statistics.* New York, NY: Dover Publications, Inc., p. 432.

Kurt, Murat et al. (2011). "The structure of optimal statin initiation policies for patients with Type 2 diabetes". In: *IIE Transactions on Healthcare Systems Engineering* 1.1, pp. 49–65. DOI: 10.1080/19488300.2010.550180.

Laan, Mark J van der and Maya L Petersen (2007). "Statistical learning of origin-specific statically optimal individualized treatment rules." In: *The international journal of biostatistics* 3.1, pp. 1–35.

Laber, Eric B, Daniel J Lizotte, and Bradley Ferguson (2014). "Set-valued dynamic treatment regimes for competing outcomes." In: *Biometrics* 70.1, pp. 53–61. DOI: 10.1111/biom.12132.

Laber, Eric B and Susan A Murphy (2011). "Adaptive Confidence Intervals for the Test Error in Classification." In: *Journal of the American Statistical Association* 106.495, pp. 904–913. DOI: 10.1198/jasa.2010.tm1005.

Lam, R W et al. (2016a). "Canadian Network for Mood and Anxiety Treatments (CANMAT) 2016 Clinical Guidelines for the Management of Adults with Major Depressive Disorder [Special issue]". In: *Canadian journal of psychiatry. Revue canadienne de psychiatrie* 61.9, pp. 506–603.

Lam, Raymond W et al. (2016b). "Canadian Network for Mood and Anxiety Treatments (CANMAT) 2016 Clinical Guidelines for the Management of Adults with Major Depressive Disorder: Section 1. Disease Burden and Principles of Care". In: *The Canadian Journal of Psychiatry* 61.9, pp. 510–523. DOI: 10.1177/0706743716659416.

Lavieri, Mariel S. et al. (2012). "When to treat prostate cancer patients based on their PSA dynamics". In: *IIE Transactions on Healthcare Systems Engineering* 2.1, pp. 62–77. DOI: 10.1080/19488300.2012.666631.

Lee, Brian K., Justin Lessler, and Elizabeth A. Stuart (2011). "Weight Trimming and Propensity Score Weighting". In: *PLoS ONE* 6.3. Ed. by Giuseppe Biondi-Zoccai, pp. 1–6. DOI: 10.1371/journal.pone.0018174.

Lee, Chris P., Glenn M. Chertow, and Stefanos A. Zenios (2008). "Optimal Initiation and Management of Dialysis Therapy". In: *Operations Research* 56.6, pp. 1428–1449. DOI: 10.1287/opre.1080.0613.

Lee, Eva K. and Tsung-Lin Wu (2009). "Classification and Disease Prediction Via Mathematical Programming". In: *Handbook of Optimization in Medicine*. Springer US. Chap. 12, pp. 381–430. DOI: 10.1007/978-0-387-09770-1_12.

Leshno, Moshe, Zamir Halpern, and Nadir Arber (2003). "Cost-effectiveness of colorectal cancer screening in the average risk population." In: *Health Care Management Science* 6.3, pp. 165–174. DOI: 10.1023/A:1024488007043.

Li, Zhiguo et al. (2014). "A global logrank test for adaptive treatment strategies based on observational studies." In: *Statistics in medicine* 33.5, pp. 760–771. DOI: 10.1002/sim.5987.

Liang, Dawen, Laurent Charlin, and David M. Blei (2016). "Causal Inference for Recommendation". In: *Conference on Uncertainty in Artificial Intelligence*.

Liberatore, Matthew et al. (2009). "Helping Men Decide About Scheduling a Prostate Cancer Screening Exam". In: *Interfaces* 39.3, pp. 209–217. DOI: 10.1287/inte.1080.0395.

Liberatore, Matthew J. and Wenhong Luo (2010). "The Analytics Movement: Implications for Operations Research". In: *Interfaces* 40.4, pp. 313–324. DOI: 10.2307/40793168.

Liebman, Michael N (2007). "Personalized medicine: a perspective on the patient, disease and causal diagnostics". In: *Personalized Medicine* 4.2, pp. 171–174. DOI: 10.2217/17410541.4.2.171.

Lin, Ching-Hua et al. (2011). "Early prediction of fluoxetine response for Han Chinese inpatients with major depressive disorder." In: *Journal of clinical psychopharmacology* 31.2, pp. 187–193. DOI: 10.1097/JCP.0b013e318210856f.

Liu, Shan, Margaret L Brandeau, and Jeremy D Goldhaber-Fiebert (2017). "Optimizing patient treatment decisions in an era of rapid technological advances: the case of hepatitis C treatment". In: *Health Care Management Science* 20.1, pp. 16–32. DOI: 10.1007/s10729-015-9330-6.

Liu, Yashu et al. (2014). "Sparse generalized functional linear model for predicting remission status of depression patients." In: *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pp. 364–75.

Lizotte, Daniel J, Michael Bowling, and Susan A Murphy (2012). "Linear Fitted-Q Iteration with Multiple Reward Functions." In: *Journal of machine learning research : JMLR* 13.Nov, pp. 3253–3295.

Lloyd, S. (1982). "Least squares quantization in PCM". In: *IEEE Transactions on Information Theory* 28.2, pp. 129–137. DOI: 10.1109/TIT.1982.1056489.

Maillart, Lisa M. et al. (2008). "Assessing Dynamic Breast Cancer Screening Policies". In: *Operations Research* 56.6, pp. 1411–1427. DOI: 10.1287/opre.1080.0614.

Marlin, Benjamin M. and Richard S. Zemel (2009). "Collaborative prediction and ranking with non-random missing data". In: *ACM Conference on Recommender Systems*, pp. 5–12. DOI: 10.1145/1639714.1639717.

Marostica, Eleonora et al. (2015). "Population modelling of patient responses in antidepressant studies: a stochastic approach." In: *Mathematical biosciences* 261, pp. 37–47. DOI: 10.1016/j.mbs.2014.11.007.

Mason, J.E. et al. (2014). "Optimizing the simultaneous management of blood pressure and cholesterol for type 2 diabetes patients". In: *European Journal of Operational Research* 233.3, pp. 727–738. DOI: 10.1016/j.ejor.2013.09.018.

Mason, Jennifer E et al. (2012). "Optimizing statin treatment decisions for diabetes patients in the presence of uncertain future adherence." In: *Medical decision making : an international journal of the Society for Medical Decision Making* 32.1, pp. 154–166. DOI: 10.1177/0272989X11404076.

McCaffrey, Daniel F et al. (2013). "A tutorial on propensity score estimation for multiple treatments using generalized boosted models." In: *Statistics in medicine* 32.19, pp. 3388–3414. DOI: 10.1002/sim.5753.

McSherry, Frank and Marc Najork (2008). "Computing Information Retrieval Performance Measures Efficiently in the Presence of Tied Scores". In: *30th European Conference on IR Research (ECIR)*. Springer-Verlag, pp. 414–421.

Melville, Prem, Raymond J. Mooney, and Ramadass Nagarajan (2002). "Content-boosted collaborative filtering for improved recommendations". In: *Proceedings of the 18th National Conference on Artificial Intelligence (AAAI)*, pp. 187–192. DOI: 10.1.1.16.4936.

Meyer, Mary A. and Jane M. Booker (2001). *Eliciting and Analyzing Expert Judgment*. Society for Industrial and Applied Mathematics, p. 441. DOI: 10.1137/1.9780898718485.

Miao, Yun-Qian, Ahmed K. Farahat, and Mohamed S. Kamel (2013). "Auto-Tuning Kernel Mean Matching". In: *2013 IEEE 13th International Conference on Data Mining Workshops*, pp. 560–567. DOI: 10.1109/ICDMW.2013.117.

Miller, Georges A. (1994). "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information". In: *Psychological Review* 101.2, pp. 343–352.

Moodie, Erica E. M., Nema Dean, and Yue Ru Sun (2013). "Q-Learning: Flexible Learning About Useful Utilities". In: *Statistics in Biosciences* 6.2, pp. 223–243. DOI: 10.1007/s12561-013-9103-z.

Moodie, Erica E M and Thomas S Richardson (2009). "Estimating Optimal Dynamic Regimes: Correcting Bias under the Null: [Optimal dynamic regimes: bias correction]." In: *Scandinavian journal of statistics, theory and applications* 37.1, pp. 126–146. DOI: 10.1111/j.1467-9469.2009.00661.x.

Moodie, Erica E. M., Thomas S. Richardson, and David A. Stephens (2007). "Demystifying Optimal Dynamic Treatment Regimes". In: *Biometrics* 63.2, pp. 447–455. DOI: 10.1111/j.1541-0420.2006.00686.x.

Murphy, S. A. (2003). "Optimal dynamic treatment regimes". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65.2, pp. 331–355. DOI: 10.1111/1467-9868.00389.

Murphy, S A (2005). "An experimental design for the development of adaptive treatment strategies." In: *Statistics in medicine* 24.10, pp. 1455–1481. DOI: 10.1002/sim.2022.

Murphy, Susan A et al. (2007). "Methodological Challenges in Constructing Effective Treatment Sequences for Chronic Psychiatric Disorders". In: *Neuropsychopharmacology* 32.2, pp. 257–262. DOI: 10.1038/sj.npp.1301241.

Myers, J. A. et al. (2011a). "Effects of Adjusting for Instrumental Variables on Bias and Precision of Effect Estimates". In: *American Journal of Epidemiology* 174.11, pp. 1213–1222. DOI: 10.1093/aje/kwr364.

Myers, J. A. et al. (2011b). "Myers et al. Respond to "Understanding Bias Amplification"". In: *American Journal of Epidemiology* 174.11, pp. 1228–1229. DOI: 10.1093/aje/kwr353.

National Guideline Directors (2016). *National Depression Clinical Practice Guideline*. Tech. rep. Kaiser Permanente.

Ng, Andrew Y. and Stuart Russell (2000). "Algorithms for Inverse Reinforcement Learning". In: *7th International Conf. on Machine Learning*, pp. 663–670.

Nguyen, Jennifer and Mu Zhu (2013). "Content-boosted matrix factorization techniques for recommender systems". In: *Statistical Analysis and Data Mining* 6.4, pp. 286–301. DOI: 10.1002/sam.11184.

Nikolaev, A. G. et al. (2013). "Balance Optimization Subset Selection (BOSS): An Alternative Approach for Causal Inference with Observational Data". In: *Operations Research* 61.2, pp. 398–412. DOI: 10.1287/opre.1120.1118.

Ohayon, M M (1993). "[Utilization of expert systems in psychiatry]." In: *Canadian journal of psychiatry. Revue canadienne de psychiatrie* 38.3, pp. 203–211.

Oord, Aaron van den, Nal Kalchbrenner, and Koray Kavukcuoglu (2016). "Pixel Recurrent Neural Networks". In: *ICML*.

Orellana, Liliana, Andrea Rotnitzky, and James M. Robins (2010). "Dynamic Regime Marginal Structural Mean Models for Estimation of Optimal Dynamic Treatment Regimes, Part I: Main Content". In: *The international journal of biostatistics* 6.2, Article 8. DOI: 10.2202/1557-4679.1200.

Ormoneit, D. and S. Sen (2002). "Kernel-based reinforcement learning". In: *Machine Learning* 49, pp. 161–178.

Oskooyee, KS (2011). "Predicting the severity of major depression disorder with the Markov chain model". In: *International Conference on Bioscience, Biochemistry and Bioinformatics*, pp. 30–34.

Overhage, J Marc and Lauren M Overhage (2011). "Sensible use of observational clinical data." In: *Statistical methods in medical research* 22.1, pp. 7–13. DOI: 10.1177/0962280211403598.

Paduraru, Cosmin (2013). "Off-policy Evaluation in Markov Decision Processes". PhD thesis. McGill University, p. 103.

Paltiel, A. David et al. (2004). "An Asthma Policy Model". In: *Operations Research and Health Care*. Ed. by Margaret L. Brandeau, François Sainfort, and William P. Pierskalla. Springer US. Chap. 29, pp. 659–693. DOI: 10.1007/1-4020-8066-2_26.

Pandey, Babita and R B Mishra (2009). "Knowledge and intelligent computing system in medicine." In: *Computers in biology and medicine* 39.3, pp. 215–230. DOI: 10.1016/j.compbiomed.2008.12.008.

Parker, R S, F J Doyle, and N A Peppas (2001). "The intravenous route to blood glucose control." In: *IEEE engineering in medicine and biology magazine : the quarterly magazine of the Engineering in Medicine & Biology Society* 20.1, pp. 65–73.

Patel, Meenal J et al. (2015). "Machine learning approaches for integrating clinical and imaging features in late-life depression classification and response prediction." In: *International journal of geriatric psychiatry* 30.10, pp. 1056–1067. DOI: 10.1002/gps.4262.

Patten, Scott B (2007). "An animated depiction of major depression epidemiology". In: *BMC Psychiatry* 7.23, pp. 1–6. DOI: 10.1186/1471-244X-7-23.

Patten, Scott B. et al. (2015). "Descriptive Epidemiology of Major Depressive Disorder in Canada in 2012". In: *Canadian Journal of Psychiatry* 60.1, pp. 23–30. DOI: 10.1177/070674371506000106.

Pearl, J. (2011). "Invited Commentary: Understanding Bias Amplification". In: *American Journal of Epidemiology* 174.11, pp. 1223–1227. DOI: 10.1093/aje/kwr352.

Pearl, Judea (2009a). "Causal inference in statistics: An overview". In: *Statistics Surveys* 3, pp. 96–146. DOI: 10.1214/09-SS057.

— (2009b). *Causality*. 2nd. New York, NY: Cambridge University Press, p. 484.

— (2010). "On a Class of Bias-Amplifying Variables that Endanger Effect Estimates". In: *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*

(UAI2010). Ed. by P. Grunwald and P. Spirtes. Corvallis, OR: Association for Uncertainty in Artificial Intelligence, pp. 425–432.

Pedregosa, F. et al. (2011). "Scikit-learn: Machine Learning in {P}ython". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Pfeiffer, Paul N et al. (2015). "Mobile health monitoring to characterize depression symptom trajectories in primary care." In: *Journal of affective disorders* 174, pp. 281–286. DOI: 10.1016/j.jad.2014.11.040.

Pierskalla, W. P. and D. J. Brailer (1994). "Applications of Operations Research in Health Care Delivery". In: *Handbooks in OR & MS, Vol. 6*. Ed. by S. M. Pollock, M. H. Rothkopft, and A. Barnett. North Holland: Elsevier Science B.V. Chap. 13, pp. 469–505.

Pineau, Joelle et al. (2007). "Constructing evidence-based treatment strategies using methods from computer science." In: *Drug and alcohol dependence* 88 Suppl 2, S52–60. DOI: 10.1016/j.drugalcdep.2007.01.005.

Pomerleau, Dean A. (1989). "ALVINN: An Autonomous Land Vehicle in a Neural Network". In: *NIPS*, pp. 305–313.

Powell, Warren B. (2011). *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. 2nd ed. Hoboken, NJ: John Wiley and Sons, p. 656.

Puterman, Martin L. (2005). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Hoboken, NJ: John Wiley and Sons, p. 684.

R Core Team (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ramachandran, Deepak and Eyal Amir (2007). "Bayesian inverse reinforcement learning". In: *Proceedings of the 20th international joint conference on Artifical intelligence*. Morgan Kaufmann Publishers Inc., pp. 2586–2591.

Ramdas, Kamalini et al. (2018). "Variety and Experience: Learning and Forgetting in the Use of Surgical Devices". In: *Management Science* 64.6, pp. 2590–2608. DOI: 10.1287/mnsc.2016.2721.

Ratkovic, Marc (2015a). "Balancing within the Margin: Causal Effect Estimation with Support Vector Machines".

— (2015b). *SVMMatch: Causal Effect Estimation and Diagnostics with Support Vector Machines. R package version 1.1*. URL: https://cran.r-project.org/package=SVMMatch.

Ratliff, Nathan D., J. Andrew Bagnell, and Martin A. Zinkevich (2006). "Maximum margin planning". In: *Proceedings of the 23rd international conference on Machine learning - ICML '06*. New York, New York, USA: ACM Press, pp. 729–736. DOI: 10.1145/1143844.1143936.

Rauner, Marion S. et al. (2010). "Dynamic Policy Modeling for Chronic Diseases: Metaheuristic-Based Identification of Pareto-Optimal Screening Strategies". In: *Operations Research* 58.5, pp. 1269–1286. DOI: 10.1287/opre.1100.0838.

Regnier, Eva D and Steven M Shechter (2013). "State-space size considerations for disease-progression models." In: *Statistics in medicine* 32.22, pp. 3862–3880. DOI: 10.1002/sim.5808.

Rendle, Steffen (2010). "Factorization machines". In: *Proceedings - IEEE International Conference on Data Mining, ICDM*, pp. 995–1000. DOI: 10.1109/ICDM.2010.127.

Rich, Benjamin, Erica E. M. Moodie, and David A. Stephens (2016). "Optimal individualized dosing strategies: A pharmacologic approach to developing dynamic treatment regimens for continuous-valued treatments". In: *Biometrical Journal* 58.3, pp. 502–517. DOI: 10.1002/bimj.201400244.

Robins, J M (1998). "Marginal structural models". In: *1997 Proceedings of the Section on Bayesian Statistical Science*. Alexandria, VA: American Statistical Association, pp. 1–10.

Robins, J M, M A Hernán, and B Brumback (2000). "Marginal structural models and causal inference in epidemiology." In: *Epidemiology* 11.5, pp. 550–560.

Robins, James M. (1986). "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect". In: *Mathematical Modelling* 7.9-12, pp. 1393–1512. DOI: 10.1016/0270-0255(86)90088-6.

Robins, JM (2004). "Optimal structural nested models for optimal sequential decisions". In: *Proceedings of the Second Seattle Symposium in Biostatistics*. Ed. by D.Y. Lin and P. Haegerty. New York, NY, USA: Springer-Verlag, pp. 189–326.

Romeijn, H. Edwin et al. (2006). "A New Linear Programming Approach to Radiation Therapy Treatment Planning Problems". In: *Operations Research* 54.2, pp. 201–216.

Rosenbaum, Paul R. (2005). "Observational study". In: *Encyclopedia of Statistics in Behavioral Science*. Vol. 3, pp. 1451–1462. DOI: 10.1002/0470013192.bsa454.

Rosenbaum, Paul R. and Donald B. Rubin (1983). "The central role of the propensity score in observational studies for causal effects". In: *Biometrika* 70.1, pp. 41–55. DOI: 10.1093/biomet/70.1.41.

Ross, Stéphane and J. Andrew Bagnell (2010). "Efficient Reductions for Imitation Learning". In: *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 661–668.

Ross, Stéphane, Geoffrey J Gordon, and J Andrew Bagnell (2011). "A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning". In: *14th International Conference on Artificial Intelligence and Statistics (AISTATS 2011)*.

Rosthøj, Susanne et al. (2006). "Estimation of optimal dynamic anticoagulation regimes from observational data: a regret-based approach." In: *Statistics in medicine* 25.24, pp. 4197–4215. DOI: 10.1002/sim.2694.

Rubin, Daniel L., Elizabeth S. Burnside, and Ross Shachter (2004). "A Bayesian Network to Assist Mammography Interpretation". In: *Operations Research and Health Care*. Ed. by Margaret L. Brandeau, François Sainfort, and William P. Pierskalla. Springer US. Chap. 27, pp. 695–720. DOI: 10.1007/1-4020-8066-2_27.

Rubin, Donald B. (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies". In: *Journal of Educational Psychology* 66.5, pp. 688–701.

Rush, Augustus John et al. (2003). "The 16-Item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression". In: *Biological Psychiatry* 54.5, pp. 573–583. DOI: 10.1016/S0006-3223(02)01866-8.

Rush, Augustus John et al. (2004). "Sequenced treatment alternatives to relieve depression (STAR*D): rationale and design". In: *Controlled Clinical Trials* 25.1, pp. 119–142. DOI: 10.1016/S0197-2456(03)00112-0.

Rush, Augustus John et al. (2006). "An Evaluation of the Quick Inventory of Depressive Symptomatology and the Hamilton Rating Scale for Depression: A Sequenced Treatment Alternatives to Relieve Depression Trial Report". In: *Biological Psychiatry* 59.6, pp. 493–501. DOI: 10.1016/J.BIOPSYCH.2005.08.022.

Saarela, Olli et al. (2015). "On Bayesian estimation of marginal structural models." In: *Biometrics* 71.2, pp. 279–288. DOI: 10.1111/biom.12269.

Salakhutdinov, Rusian and Andriy Mnih (2007). "Probabilistic Matrix Factorization." In: *Proc. Advances in Neural Information Processing Systems 20 (NIPS 07)*, pp. 1257–1264. DOI: 10.1145/1390156.1390267.

Sanchez, Ivan et al. (2015). "Towards Extracting Faithful and Descriptive Representations of Latent Variable Models". In: *AAAI Spring Symposium on Knowledge Representation and Reasoning*. Vol. 3, pp. 35–38.

Sauppe, Jason J., Sheldon H. Jacobson, and Edward C. Sewell (2014). "Complexity and Approximation Results for the Balance Optimization Subset Selection Model for Causal Inference in Observational Studies". In: *INFORMS Journal on Computing* 26.3, pp. 547–566. DOI: 10.1287/ijoc.2013.0583.

Schaefer, A. J. et al. (2004). "Modeling Medical Treatment using Markov Decision Processes". In: *Operations Research and Health Care*. Ed. by M. L. Brandeau, F. Sainfort, and W. P. Pierskalla. 1st ed. Springer US. Chap. 23, pp. 593–612. DOI: 10.1007/b106574.

Schmidt, Mark (2010). "Graphical Model Structure Learning with L1-Regularization". PhD thesis. University of British Columbia.

Schnabel, Tobias et al. (2016). "Recommendations as Treatments: Debiasing Learning and Evaluation". In: *Proceedings of the 33rd International Conference on Machine Learning*.

Schneeweiss, Sebastian et al. (2009). "High-dimensional Propensity Score Adjustment in Studies of Treatment Effects Using Health Care Claims Data". In: *Epidemiology* 20.4, pp. 512–522. DOI: 10.1097/EDE.0b013e3181a663cc.

Schölkopf, Bernhard et al. (2012). "On Causal and Anticausal Learning". In: *ICML*.

Schulte, Phillip J et al. (2014). "Q- and A-learning Methods for Estimating Optimal Dynamic Treatment Regimes." In: *Statistical science : a review journal of the Institute of Mathematical Statistics* 29.4, pp. 640–661. DOI: 10.1214/13-STS450.

Schwartz, Alan and George Bergus (2008). *Medical Decision Making: A Physician's Guide*. Cambridge University Press, p. 232.

Shechter, Steven M., Oguzhan Alagoz, and Mark S. Roberts (2010). "Irreversible treatment decisions under consideration of the research and development pipeline for new therapies". In: *IIE Transactions* 42.9, pp. 632–642. DOI: 10.1080/07408170903468589.

Shechter, Steven M. et al. (2008). "The Optimal Time to Initiate HIV Therapy Under Ordered Health States". In: *Operations Research* 56.1, pp. 20–33. DOI: 10.1287/opre.1070.0480.

Shi, Jinghua et al. (2011). "A survey of optimization models on cancer chemotherapy treatment planning". In: *Annals of Operations Research* 221.1, pp. 331–356. DOI: 10.1007/s10479-011-0869-4.

Shneiderman, Ben et al. (2016). *Designing the user interface : strategies for effective human-computer interaction*. 6th ed. Pearson, p. 616.

Shortreed, Susan M and Erica E M Moodie (2012). "Estimating the optimal dynamic antipsychotic treatment regime: Evidence from the sequential multiple assignment randomized CATIE Schizophrenia Study." In: *Journal of the Royal Statistical Society. Series C, Applied statistics* 61.4, pp. 577–599. DOI: 10.1111/j.1467-9876.2012.01041.x.

Shortreed, Susan M et al. (2011). "Informing sequential clinical decision-making through reinforcement learning: an empirical study." In: *Machine Learning* 84.1-2, pp. 109–136. DOI: 10.1007/s10994-010-5229-0.

Shu-Hsien Liao (2005). "Expert system methodologies and applications—a decade review from 1995 to 2004". In: *Expert Systems with Applications* 28.1, pp. 93–103. DOI: 10.1016/j.eswa.2004.08.003.

Simon, Gregory E. and Roy H. Perlis (2010). "Personalized medicine for depression: can we match patients with treatments?" In: *The American journal of psychiatry* 167.12, pp. 1445–1455. DOI: 10.1176/appi.ajp.2010.09111680.

Simon, Jay (2009). "Decision Making with Prostate Cancer: A Multiple-Objective Model with Uncertainty". In: *Interfaces* 39.3, pp. 218–227. DOI: 10.1287/inte.1080.0406.

Society for Medical Decision Making. *Definition of Medical Decision Making*. URL: http://smdm.org/hub/page/definition-of-medical-decision-making/about (visited on 11/26/2017).

Song, Hummy, Anita L Tucker, and Karen L Murrell (2015). "The Diseconomies of Queue Pooling: An Empirical Investigation of Emergency Department Length of Stay". In: *Management Science* 61.12, pp. 3032–3053. DOI: 10.1287/mnsc.2014.2118.

Song, Rui et al. (2015). "Penalized Q-Learning for Dynamic Treatment Regimes". In: *Statistica Sinica* 25.3, pp. 901–920.

Spirtes, Peter (2010). "Introduction to Causal Inference". In: *Journal of Machine Learning Research* 11, pp. 1643–1662.

Splawa-Neyman, Jerzy, D. M. Dabrowska, and T. P. Speed (1923). "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." In: *Statistical Science* 5.4, pp. 465–480.

Staats, Bradley R, Diwas S Kc, and Francesca Gino (2018). "Maintaining Beliefs in the Face of Negative News: The Moderating Role of Experience". In: *Management Science* 64.2, pp. 804–824. DOI: 10.1287/mnsc.2016.2640.

Staats, Bradley R et al. (2017). "Motivating Process Compliance Through Individual Electronic Monitoring: An Empirical Examination of Hand Hygiene in Healthcare". In: *Management Science* 63.5, pp. 1563–1585. DOI: 10.1287/mnsc.2015.2400.

Steinwart, Ingo (2002). "Support Vector Machines Are Universally Consistent". In: *JOURNAL OF COMPLEXITY* 18, pp. 768–791. DOI: 10.1006/jcom.2002.0642.

Sutton, R. S. and A. G. Barto (1998). *Reinforcement learning: An introduction*. Cambridge, MA, USA: MIT.

Syed, Umar and Robert E. Schapire (2007). "A Game-Theoretic Approach to Apprenticeship Learning". In: *Advances in Neural Information Processing Systems 20 (NIPS 2007)*, pp. 1449–1456.

— (2010). "A Reduction from Apprenticeship Learning to Classification". In: *Advances in Neural Information Processing Systems (NIPS)*, pp. 2253–2261.

Tafazzoli, Ali et al. (2009). "Probabilistic cost-effectiveness comparison of screening strategies for colorectal cancer". In: *ACM Transactions on Modeling and Computer Simulation* 19.2, pp. 1–29. DOI: 10.1145/1502787.1502789.

Taskin, Z. Caner et al. (2010). "Optimal Multileaf Collimator Leaf Sequencing in IMRT Treatment Planning". In: *Operations Research* 58.3, pp. 674–690. DOI: 10.1287/opre.1090.0759.

Trangle, M. et al. (2016). *Health Care Guideline: Adult Depression in Primary Care Guideline*. Tech. rep. Institute for Clinical Systems Improvement, p. 131.

Tunc, Sait, Oguzhan Alagoz, and Elizabeth Burnside (2014). "Opportunities for Operations Research in Medical Decision Making." In: *IEEE intelligent systems* 29.3, pp. 59–62.

Van De Klundert, Joris (2016). "Healthcare Analytics: Big Data, Little Evidence". In: *INFORMS Tutorials in Operations Research*, pp. 307–328. DOI: 10.1287/educ.2016.0158.

Vansteelandt, Stijn and Marshall Joffe (2014). "Structural Nested Models and G-estimation: The Partially Realized Promise". In: *Statistical Science* 29.4, pp. 707–731. DOI: 10.1214/14-STS493.

Vincent, Robert Durham (2014). "Reinforcement learning in models of adaptive medical treatment strategies". PhD thesis. McGill University, p. 224.

Wagholikar, Kavishwar B, Vijayraghavan Sundararajan, and Ashok W Deshpande (2012). "Modeling paradigms for medical diagnostic decision support: a survey and future directions." In: *Journal of medical systems* 36.5, pp. 3029–3049. DOI: 10.1007/s10916-011-9780-4.

Wagner, Todd H. and Jeffrey K. Jopling (2017). "Déjà Vu: Introducing Operations Research to Health Care". In: *Medical Decision Making* 37.8. DOI: 10.1177/0272989X17711909.

Wang, Jian Li et al. (2014). "Development and validation of a prediction algorithm for use by health professionals in prediction of recurrence of major depression." In: *Depression and anxiety* 31.5, pp. 451–457.

Wang, Yixin and David M. Blei (2018). "The Blessings of Multiple Causes".

Wang, Yixin et al. (2018). "The Deconfounded Recommender: A Causal Inference Approach to Recommendation".

Watkins, C. J. C. H. (1989). "Learning from delayed rewards". PhD thesis. Cambridge University.

Wong, M-L et al. (2012). "Prediction of susceptibility to major depression by a model of interactions of multiple functional genetic variants and environmental factors." In: *Molecular psychiatry* 17.6, pp. 624–633. DOI: 10.1038/mp.2012.13.

World Health Organization (2008). *The Global Burden of Disease: 2004 update*. Tech. rep. World Health Organization, p. 146. DOI: 10.1038/npp.2011.85.

Wu, Mon-Ju et al. (2015). "Prediction of pediatric unipolar depression using multiple neuromorphometric measurements: A pattern classification approach". In: *Journal of Psychiatric Research* 62, pp. 84–91. DOI: 10.1016/j.jpsychires.2015.01.015.

Xin, Yu and Harald Steck (2011). "Multi-Value Probabilistic Matrix Factorization for IP-TV Recommendations". In: *Proceedings of the fifth ACM conference on Recommender systems*, pp. 221–228. DOI: 10.1145/2043932.2043972.

Yang, Yan, Jeremy D Goldhaber-Fiebert, and Lawrence M Wein (2013). "Analyzing Screening Policies for Childhood Obesity." In: *Management science* 59.4, pp. 782–795. DOI: 10.1287/mnsc.1120.1587.

Yoo, Illhoi et al. (2012). "Data mining in healthcare and biomedicine: a survey of the literature." In: *Journal of medical systems* 36.4, pp. 2431–2448. DOI: 10.1007/s10916-011-9710-5.

Yu, Yaoliang and Csaba Szepesvari (2012). "Analysis of Kernel Mean Matching under Covariate Shift". In: *Proceedings of the 29th International Conference on Machine Learning*, pp. 1–8.

Yue, Yisong and Hoang M. Le (2018). "Imitation Learning". In: *ICML*.

Zenios, Stefanos A. (2002). "Optimal Control of a Paired-Kidney Exchange Program". In: *Management Science* 48.3, pp. 328–342. DOI: 10.1287/mnsc.48.3.328.7732.

Zhang, Baqun et al. (2012a). "A robust method for estimating optimal treatment regimes." In: *Biometrics* 68.4, pp. 1010–1008. DOI: 10.1111/j.1541-0420.2012.01763.x.

Zhang, Baqun et al. (2012b). "Estimating optimal treatment regimes from a classification perspective". In: *Stat* 1.1, pp. 103–114. DOI: 10.1002/sta.411.

Zhang, Baqun et al. (2013a). "Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions". In: *Biometrika* 100.3, pp. 681–694. DOI: 10.1093/biomet/ast014.

Zhang, Jingyu et al. (2012c). "Optimization of PSA screening policies: a comparison of the patient and societal perspectives." In: *Medical decision making : an international journal of the Society for Medical Decision Making* 32.2, pp. 337–349. DOI: 10.1177/0272989X11416513.

Zhang, Jingyu et al. (2013b). "Disease Prevention, Detection, and Treatment". In: *Encyclopedia of Operations Research and Management Science*. Ed. by S. I. Gass and M. C. Fu. 3rd ed. Boston, MA: Springer US, pp. 437–447.

Zhang, Y. and Brian T. Denton (2015). "Robust Markov Decision Processes for Medical Treatment Decisions".

Zhang, Yuanhui et al. (2014). "Second-line agents for glycemic control for type 2 diabetes: are newer agents better?" In: *Diabetes care* 37.5, pp. 1338–1345. DOI: 10.2337/dc13-1901.

Zhao, Ying-Qi and Eric B Laber (2014). "Estimation of optimal dynamic treatment regimes." In: *Clinical trials* 11.4, pp. 400–407. DOI: 10.1177/1740774514532570.

Zhao, Ying-Qi et al. (2015). "New Statistical Learning Methods for Estimating Optimal Dynamic Treatment Regimes." In: *Journal of the American Statistical Association* 110.510, pp. 583–598. DOI: 10.1080/01621459.2014.937488.

Zhao, Yingqi et al. (2012). "Estimating Individualized Treatment Rules Using Outcome Weighted Learning." In: *Journal of the American Statistical Association* 107.449, pp. 1106–1118. DOI: 10.1080/01621459.2012.695674.

Zhao, Yufan et al. (2011). "Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer." In: *Biometrics* 67.4, pp. 1422–1433. DOI: 10.1111/j.1541-0420.2011.01572.x.

Zhu, Rui et al. (2017). "Expectile Matrix Factorization for Skewed Data Analysis". In: *Proceedings of the 31th Conference on Artificial Intelligence (AAAI 2017)*, pp. 259–265.

Ziebart, Brian D et al. (2008). "Maximum Entropy Inverse Reinforcement Learning". In: *Proceeding AAAI'08 Proceedings of the 23rd national conference on Artificial intelligence*. Chicago, Illinois, pp. 1433–1438.

Zimmerman, Mark et al. (2013). "Severity classification on the Hamilton depression rating scale". In: *Journal of Affective Disorders* 150.2, pp. 384–388. DOI: 10.1016/j.jad.2013.04.028.

Zubizarreta, José R. (2015). "Stable Weights that Balance Covariates for Estimation with Incomplete Outcome Data". In: *Journal of the American Statistical Association* 110.511, pp. 910–922. DOI: 10.1080/01621459.2015.1023805.

Zubizarreta, Jose R. and Amine Allouah (2016). *sbw: Stable weights that balance covariates. R package version 0.0.2.*