Supporting Therapy Success by Developing Predictive Models in E-Mental-Health

Faculty of Business and Economics
LEUPHANA UNIVERSITY LÜNEBURG

Approved thesis as a requirement for the award of the title of

Doctor of rerum politicarum

-DR. RER. POL.-

Approved thesis by Vincent Bremer, M.A.

born 16th of June 1987 in Kiel, Germany

Submitted on:	September 8, 2020
Oral defense (disputation) on:	July 13, 2021
First Supervisor:	Prof. Dr. Burkhardt Funk
Second Supervisor:	Prof. Dr. Heleen Riper
Third Supervisor:	Prof. Dr. Peter Niemeyer

The individual articles of this cumulative dissertation have been or will be published as follows:

Bremer, V., Becker, D., Funk, B., and Lehr, D. (2017). Predicting the Individual Mood Level based on Diary Data. Proceedings of the Twenty-Fifth European Conference on Information Systems, Guimaraes, June 5-10, pp. 1161-1177.

Bremer, V., Becker, D., Genz, T., Funk, B., and Lehr, D. (2018). A two-stage approach for the prediction of mood levels based on diary data. ECML/PKDD Nectar Track.

Bremer, V., Becker, D., Kolovos, S., Funk, B., Van Breda, W., Hoogendoorn, M., and Riper, H. (2018). Predicting Therapy Success and Costs for Personalized Treatment Recommendations Using Baseline Characteristics: Data-Driven Analysis. Journal of Medical Internet Research, 20(8):e10275.

Bremer, V., Funk, B., and Riper, H. (2019). Heterogeneity matters – predicting self-esteem in online interventions based on Ecological Momentary Assessment data. Depression Research and Treatment, 2019:3481624.

Bremer, V., Chow, P., Funk, B., Thorndike, F., and Ritterband, L. (2020). Developing a Process for the Analysis of User Journeys and the Prediction of Dropout in Digital Health Interventions: Machine Learning Approach. Journal of Medical Internet Research, 22(10):e17738.

Bremer, V. (2020). Machine learning-based decision support systems in clinical practice: Potential and limitations (white paper).

Year of publication: 2021

Abstract

Mental health is an important factor in an individuals' life - more than 300 million individuals suffered from depression in 2015. Online-based interventions have been developed for the treatment of various mental disorders. These types of interventions have proven their efficacy and can lead to positive outcomes for suffering patients. During these interventions, a large amount of patient-specific data is gathered that can be utilized to increase treatment outcomes by informing decision-making processes of psychotherapists, experts in the field, and patients.

The articles included in this dissertation focus on the analysis of such data collected in digital psychological treatments by using machine learning approaches. This dissertation utilizes various machine learning methods such as Bayesian models, regularization techniques, or decision trees to predict different psychological factors, such as mood or self-esteem, dropout of patients, or treatment outcomes and costs. These models are evaluated using a variety of performance metrics, for example, receiver operating characteristics curve, root mean square error, or specialized performance metrics for Bayesian inference. These types of analyses can support decision-making for psychologists and patients, which can, in turn, lead to better recommendations and subsequently to increased outcomes for patients and simultaneously more insight about the interplay between psychological factors. The contribution of this interdisciplinary dissertation is manifold and can be classified at the intersection of Information Systems, health economics, and psychology. The analysis of user journey data has not yet been fully examined in the field of psychological research. A process for this endeavor is developed and a technical implementation is provided for the research community. The application of machine learning in this context is still in its infancy. Thus, another contribution is the exploration and application of machine learning techniques for the revelation of correlations between psychological factors or characteristics and treatment outcomes as well as their prediction. Additionally, economic factors are predicted to develop a process for treatment type recommendations. This approach can be utilized for finding the optimal treatment type for patients on an individual level considering predicted treatment outcomes and costs. By evaluating the predictive accuracy of multiple machine learning techniques based on various performance metrics, the importance of considering heterogeneity among patients' behavior and affect is highlighted in some articles. Furthermore, the potential of machine learning-based decision support systems in clinical practice has been examined from a psychotherapists' point of view.

Dedication and acknowledgement

This dissertation is dedicated to my wife, my daughters, my parents, and my late grandfathers. My wife has helped me through the whole process of this research project not only emotionally by providing emotional comfort but also professionally by providing valuable advice. We both have worked on our dissertations simultaneously; this immense burden has helped to understand each other and enabled us to provide crucial feedback for both of our research projects. Thank you, Veronica, for all of your help and understanding. Additionally, I want to dedicate this dissertation to my daughters Emilia and Luna. They have helped me to recognize what the word *importance* really means. They have pushed my work attitude toward the end of my dissertation and enabled me to see life from a different perspective. I also want to thank my parents for always believing in me and for the continuous support and trust not only throughout my childhood but also throughout my life as an adult. Without my parents, my wife, and my daughters, I would not be who I am today and where I am today.

I also thank my academic advisor Burkhardt Funk and express my deep gratitude. He has supported me since I pursued my Bachelor's degree and introduced me to the field of data analytics, which I then started to love. He offered me a scholarship, positions as a research associate, and additionally provided an excellent research environment. His critical guidance, valuable comments, and enormous support helped me tremendously and continuously pushed me forward. I have been fortunate to learn from him and collaborate with him on many research projects. Additionally, I want to thank the other members of my academic committee, Heleen Riper and Peter Niemeyer; Thank you for your very valuable comments and advice. I also thank my colleagues Dennis Becker, Martin Stange, Christoph Martin, Sebastian Mair, Felix Krieger, Ward van Breda, and Spyros Kolovos for the great discussions, exchange, and collaboration throughout the last years.

I furthermore acknowledge the EU funded project E-COMPARED. A variety of data that has been collected in this project was utilized in multiple papers presented in this dissertation. I therefore thank the EU for funding and the E-COMPARED consortium and all of the colleagues from this project for the fantastic cooperation. Finally, I want to thank the colleagues from the University of Virginia Lee Ritterband, Philip Chow, and Frances Thorndike for having me as an exchange student and developing exciting approaches together.

TABLE OF CONTENTS

List of Tables

List of Figures

CHAPTER 1

INTRODUCTION

1.1 Introduction

Mental disorders such as depression or anxiety are among the main burdens of disease in nowadays' society (Vos et al., 2015). More than 300 million individuals suffered from depression in 2015 (WHO, 2017). Mental health problems are associated with lower physical health and can lead to a substantial negative effect on a person's quality of life and mental well-being in both short and long-term (Buntrock et al., 2014; Scott et al., 2016; Smit et al., 2006). Mental health problems are widespread and do not only affect the health of individuals but can lead to an increase in financial expenses for governments (Leger, 1994; Schofield et al., 2011). Based on data from 2010, mental disorders were estimated to be 798 billion euros just in the European Union and 2.5 trillion US dollars globally (Trautmann et al., 2016). To make things worse, it is estimated that global mental health costs will increase to 6 trillion dollars a year by 2030 (Bloom et al., 2011).

In the mental health sphere, e-health is a relatively new field in this context and takes advantage of information and communication technology in order to support healthcare in terms of outcome improvement and cost reduction (Eysenbach, 2001; Riper et al., 2010). Internet-based interventions have been developed for the prevention and treatment of psychological disorders; they represent one possibility to close the gap between treatment and demand (Fassbinder et al., 2015). Online-based treatments have been shown to produce benefits for patients as well as leading to similar outcomes compared to face-to-face treatment (Carlbring et al., 2018; Sander et al., 2016; Yang et al., 2018). These psychological interventions also generate tremendous amounts of patient-specific data such as self- or observer rated retrospective questionnaire data, Ecological Momentary Assessment data, and interactions with the computerized systems (i.e.

login information and message exchanges). This data contains viable information about the patients' behavior and represents a gateway for interdisciplinary collaboration between the fields of Information Systems, health economics, and psychology.

Information Systems can contribute in this context by the development and application of machine learning techniques - a subdomain of artificial intelligence - for the analysis and prediction of generated data. Machine learning can be described as the process of programming a computer to learn patterns from historical data and, subsequently, the application of these patterns to new data in order to generalize the learned patterns and generate predictions (Alpaydin, 2009). Algorithms that can model and analyze data, predecessors of today's machine learning one might say, exist for more than 60 years (Kononenko, 2001). Since these early days, machine learning has been applied in various healthcare domains such as diabetes, radiology, cancer treatment, chronic diseases, and recently, in mental health as well (Becker et al., 2018; Belle et al., 2013; Kavakiotis et al., 2017; Lindblom et al., 2012; Riano et al., 2012). There are applications that aim to identify strokes (Lee and Yoon, 2017), look for skin lesions (Esteva et al., 2017), or predict clinical remission in depression (Chekroud et al., 2016).

The application of machine learning in the context of mental health is still in its infancy (Clifton et al., 2015) and can contribute in a twofold manner, which are often not distinguished (Yarkoni and Westfall, 2017). First, shedding light on relationships between psychological factors such as sleep or mood, which are synonymously called *psychological concepts* in later chapters, can lead to an increased understanding of the patients' behavior and can offer indicators of how psychological factors influence each other. Low self-esteem, for example, is connected to lower mental and physical health and these problems are linked to mood states and psychological well-being (Paradise and Kernis, 2002; Steiger et al., 2014; Trzesniewski et al., 2006). Mental disorders such as depression can be heterogeneous and, therefore, patients can show various symptoms (Goldberg, 2011). Thus, finding correlations between these psychological factors can be beneficial for understanding relationships among disorders. Second, predictive modeling can support treatment decisions by offering predictions about individual treatment progress or potential outcomes of treatment. These predictions could support psychologists' and patients' decision-making, for example, when it is necessary to intervene if critical states of psychological factors are reached or when patients are of high risk to drop out of treatment. Patients can also benefit from the results of predictive models in terms of increased self-management and prevention. Predictive modeling can support the treatment process by offering advanced diagnostic, prognostic, detection, prevention, and treatment selection processes (Becker et al., 2018; Dwyer et al., 2018; Shatte et al., 2019). The resulting insight can aid in the decision-making process of patients as well as professionals in the field of psychology and potentially increase treatment outcomes for patients (Clifton et al., 2015; Shatte et al., 2019; Triantafyllidis and Tsanas, 2019; Wallace et al., 2012).

At this point, this dissertation aims at exploring and providing directions on how to analyze data of Internet-based interventions using various types of machine learning techniques. By

doing so, relationships between psychological factors are analyzed in order to shed light on these correlations and on possible drivers of mental disorders. At the same time, predictions of psychological factors and intervention outcomes are generated and evaluated to support the decision-making process of psychologists in this context. For example, if being able to predict future states of psychological factors, psychologists might provide *better* treatment recommendations and might be able to intervene before increased aggravation of disorders. Thus, the purpose of this dissertation is manifold.

The goals are to support the prevention of adverse clinical outcomes, the creation of personalized treatment strategies, and eventually the enhancement of treatment efficacy by supporting decision-making processes. Therefore, the overall obstacle this dissertation tackles is the generation of increased insight about patients' behavior and subsequent prediction of various psychological factors and treatment outcomes by exploring and developing machine learning approaches in digital health interventions.

1.2 Background and state of the art

This section introduces the psychological factors and outcomes that are utilized as target variables for the application of machine learning approaches in this dissertation. Psychological research and existing studies that focus on the application of machine learning in the field of e-mental-health are demonstrated.

1.2.1 Inference and prediction of data in e-mental-health

In e-mental-health, the application of machine learning and analysis and prediction of mental health-related data is still in its infancy and comprises a variety of subdomains. The following section illustrates related literature in the subdomains this dissertation focuses on and demonstrates their importance.

Mood levels

Mood disorders affect the physical health of suffering patients as well as social communication; illustrating the importance of analyzing mood levels in research (Byrne and Byrne, 1993; Byrne, 1986). Evidence-based treatments for depression such as cognitive behavioral therapy exist (Dobson and Dobson, 2009), however, they are often not individualized to specific patients even though these patients might not be affected similarly. Various research projects attempted to predict mood levels based on different types of data. Likamwa et al. (2013), for example, utilized unobtrusive mobile phone data of 32 students in order to predict the mood level and mood fluctuations. They applied multi-linear regression models for an individualized prediction of mood. They were able to predict the mood level with up to 93% accuracy. Zulueta et al. (2018) analyzed the relationship between mood disturbances and mobile phone keyboard data in patients

suffering from bipolar disorder. Based on these data, the authors created various features such as session length, autocorrect rate, or backspace ratio. The features were then mapped to cognitive and behavioral domains, for example, social activity (increase in keyboard activity). Mixed effect linear models were used for the prediction of depressive symptoms. They found that keystroke data does predict depressive symptoms and mood disturbances. Fluctuations in mood have also been predicted based on acoustic data. Weidman et al. (2020) extracted 88 acoustic features from audio recordings of individuals and subsequently utilized three machine learning techniques, namely random forests, neural networks, and support vector machines in order to predict within-person mood fluctuations. However, the performance of either machine learning technique did not lead to a substantially better performance compared to chance. Thus, this study indicates that predictions of mood fluctuations are not yet possible by using acoustic data (Weidman et al., 2020). Additionally, various studies utilized Ecological Momentary Assessment (EMA) data to predict the mood level of individuals (Kanning and Schlicht, 2010; Mikus et al., 2018; Starr and Savila, 2013; Van Breda et al., 2016). These studies applied different machine learning techniques such as recurrent neural networks, linear regression, or multi-level models. The authors showed that the utilization of EMA data in this context can contribute to the predictive performance of statistical procedures, revelation of relationships between psychological factors, and is therefore an appropriate type of data in order to predict the mood level of individual patients. However, there are also other studies, for example by van Breda et al. (2018), who found that utilizing EMA data does not lead to a significantly improved predictive performance. It is, thus, important to analyze if a significant signal exists in EMA data that can help to predict psychological factors.

Self-esteem levels

Self-esteem is important for an individuals' general well-being and mental and physical health (Lemola et al., 2013; Steiger et al., 2014; Trzesniewski et al., 2006). Various studies utilized self-esteem as a predictor for depression (Cheng and Furnham, 2003; Orth et al., 2008; Park and Yang, 2017; Steiger et al., 2014). Steiger et al. (2014), for example, showed that low levels of self-esteem as well as changes in self-esteem influence depression states of individuals. Specifically, based on a 23-year longitudinal study, they found that self-esteem levels and changes in self-esteem are predictors for depression in adulthood. Individuals who experienced low self-esteem in adolescent years were more likely to develop symptoms of depression two decades later (Steiger et al., 2014). Another study by Orth et al. (2008) also found that low levels of self-esteem significantly predict subsequent states of depression. These studies support the vulnerability assumption, which understands low levels of self-esteem as a contributor to future depression (Manna et al., 2016). On the other hand, the scar model assumes low self-esteem to be a consequence of depression. Research does agree on the fact that low self-esteem levels and depression are related, however, the actual relationship between these factors is not yet clearly understood (Sowislo and Orth, 2013). Sowislo and Orth (2013) and Manna et al. (2016) compared the vulnerability and scar assumptions in their articles. Both found supporting evidence for the vulnerability assumption.

Sowislo and Orth (2013) found that the effect of self-esteem on depression was significantly stronger than vice versa ($\beta = -0.16$; $\beta = -0.08$) while pointing out that interventions focusing on self-esteem might reduce the risk of depression. Steiger et al. (2015) found evidence for both models, however, the results indicate a stronger effect for the vulnerability assumption as well. That might be one indicator for the fact that literature that tries to predict self-esteem is rather scarce compared to scientific articles that utilize self-esteem for the prediction of depression. Swann et al. (2007) argued for the importance of designing and delivering interventions for an improvement of self-views. They argued that individuals who have negative self-views are less able to cope with general life events. Thus, analyzing and predicting self-esteem can lead to important information about patients in this context and support their treatment progress.

Intervention outcomes and treatment recommendation

The prediction of treatment outcomes is a challenging task; if successful, however, it can reveal crucial insight and information for individualized treatment strategies and planning (McMahon, 2014). Various studies aimed at predicting treatment outcomes of an intervention in the field of mental health. These outcomes were often defined as post-test results or specific psychological questionnaire scores and were based on data that is gathered before and during the treatment phase (DeRubeis et al., 2014; Huibers et al., 2015; Proudfoot et al., 2013). A multitude of predictors were utilized to predict outcomes on a population and individual patient-level such as therapeutic relationship (Priebe et al., 2011), EMA data and usage/sensor data of mobile phones (Asselbergs et al., 2016; Becker et al., 2016; Mikus et al., 2018; van Breda et al., 2016), or log data of programs in Internet-delivered interventions (Whitton et al., 2015). Månsson et al. (2015), for example, predicted individualized long-term outcomes of patients suffering from social anxiety disorders by applying support vector machines. They were able to predict long-term outcomes one year after treatment with 92% accuracy. Pearson et al. (2018) utilized regularized regression and random forests in order to predict depressive symptoms among patients that have finished an Internet intervention. They found that the utilized predictors only had marginal contributions to the predictive performance, however, by applying an ensemble of the aforementioned machine learning techniques, they were able to reach better predictive performance compared to their baseline model (a linear model that predicted the outcome based on the outcomes' baseline information). Additionally, Saunders et al. (2016) classified 10.693 patients into specific profiles. These profiles showed the predictive capabilities of forecasting treatment outcomes in psychological routine care. van Breda et al. (2018) used random forests and general linear models in order to predict therapy success defined by the Patient Health Questionnaire-9 (Kroenke et al., 2001). They created various encoding and feature selection settings and showed that baseline data of an Internet-based intervention has predictive power in forecasting therapy outcomes by reaching AUC values between 0.55 and 0.76.

As depicted, research studies exists for the prediction of psychological outcomes. However, these predictions and generally machine learning approaches in this context are rarely converted

into tools and subsequent treatment recommendations in clinical practice (Sacchi et al., 2015). Literature in this regard is scarce but there are some studies that seek to utilize results of machine learning applications for recommendations. DeRubeis et al. (2014) developed a method, the Personalized Advantage Index (PAI), for the transformation of predictive results into recommendations of the treatment selection process for individuals in this context. This PAI essentially reveals the treatment type that is supposed to be more efficacious on an individual patient-level. For the creation of the PAI, outcomes of both treatment types are predicted for each patient. Then, the difference between the predictions is calculated, which indicates the more beneficial treatment type. However, no costs of treatment types are considered in this approach, which is often of significance when choosing between different treatment options. Nevertheless, various studies utilized this approach for recommendations of treatment types such as Lopez-Gomez et al. (2019) or Huibers et al. (2015). Lopez-Gomez et al. (2019), for example, predicted if Integrative Positive Psychological Intervention (IPPI-D) or a cognitive-behavioral therapy group intervention (CBT) was more beneficial for the treatment of depression on a patient-individual level. The authors utilized elastic net regularization for the prediction of change of depressive symptoms. They found that both treatment types were clinically effective, however, it was predicted that IPPI-D was more beneficial for 73% of the patients. Another study by Tomlinson et al. (2020) plans on developing a decision support tool targeting patients that suffer from major depressive disorder. Here, predictive models such as neural networks and support vector machines will be developed and evaluated that aim at forecasting effects of various antidepressants. The eventual results will be utilized in a computerized decision support system in order to support the decision-making of clinicians and patients on and individual patient-level. This system will consider preferences of clinicians as well as patients and will be evaluated and compared in a randomized controlled trial (Tomlinson et al., 2020).

Dropout in digital health interventions

Dropout of patients in Internet-based treatments is an important issue (Melville et al., 2010) and can be defined as when a patient prematurely discontinues the program. This definition is close to Eysenbachs description of *non-usage dropout attrition*, which leads to losing the patient from the intervention. Internet-based treatments for psychological disorders with minimal therapist contact have shown to suffer from dropout rates between 30% and 50% (Melville et al., 2010). Another study found 28% dropout rates for guided (therapist contact) and 74% dropout rates for unguided (no therapist contact) Internet-based interventions (Richards and Richardson, 2012). Convincing patients to continuously adhere to the intervention decreases dropout rates and can simultaneously improve outcomes for individuals while playing a major role for societal and economic costs (Donkin et al., 2011; Lutz et al., 2018). Predictive modeling can lead to a deeper understanding of the patients' behavior and reveal relationships between input data and adherence. Understanding these relations can support the prevention of dropout (Karyotaki et al., 2015) since these relationships can support the identification of high-risk patients and provide

indicators for psychologists on how to intervene. With this information, especially early on in the intervention, psychologists can target these individuals and develop strategies for persuading them in order to stay active. Thus, predictive modeling can inform individualized treatment recommendations in this context.

In the literature, there are studies for two different objectives: revealing relationships between dropout and other psychological factors as well as predicting dropout of treatment programs. Research projects that aimed at unraveling the relationship between psychological factors and dropout found indicators that male gender, lower education, severity of the disease, lower number of treating therapists, or lower self-esteem levels can possibly lead to a higher risk of dropping out of treatment (Christensen et al., 2009; Karyotaki et al., 2015; Kegel and Flückiger, 2015; Reneses et al., 2009). Studies that aim at predicting dropout using machine learning techniques are scarce, however, a relatively new study by Pedersen et al. (2019) aimed at predicting dropout among 2684 patients who suffer from chronic lifestyle diseases. They compared various machine learning techniques such as logistic regression and random forests. Their final model, a random forest, resulted in 89% precision in classifying dropout.

1.2.2 Acceptance of machine learning in clinical practice

As outlined above, machine learning approaches are evaluated and applied for various purposes in the context of e-mental-health research. In clinical practice or routine care, however, actual machine learning based tools are rarely utilized (Clifton et al., 2015; Sacchi et al., 2015; Triantafyllidis and Tsanas, 2019) even though they could lead to beneficial outcomes for individual patients (Triantafyllidis and Tsanas, 2019). There could be various reasons for this fact. First of all and as mentioned above, the application of machine learning in clinical practice is *in its infancy* and attention to its potential is quite recently received (Clifton et al., 2015). Second, even though a large amount of data is being collected in clinical practice, data quality plays an important role in the application of machine learning techniques. Thus, incorrect or incomplete data can impact analyses and lead to low predictive performance (Scott et al., 2019). Even though this statement holds for every field machine learning can be applied to, consequences of following *wrong* recommendations based on *uncertain* predictions in the medical domain might not only have financial consequences but most importantly negative health-related outcomes.

Another reason could be that the analysis of raw data requires collaboration between data related professionals and clinical experts. The models need to be built and implemented, but at the same time, these models often depend on expert knowledge from the clinicians in terms of features included in the analysis. For an assessment of the actual effectiveness of machine learning in a clinical setting, it might be beneficial to evaluate how these machine learning techniques and subsequent tools affect patient care as was recently done by Shimabukuro et al. (2017) for patients suffering from severe sepsis. They found that the utilization of machine learning led to improved outcomes for patients. Additionally, it might be beneficial to analyze

potential advantages and disadvantages regarding the application of machine learning in clinical practice from an experts' point of view.

1.3 Contribution

Table 1.1 illustrates the articles, the journal or conference they are published in, and the corresponding ranking of the journal or conference. The articles are included in the order they have been published or submitted. This dissertation is contributing to existing research in various ways. Before delving into the specifics of each chapter, some general contributions of this dissertation are discussed (Figure 1.1).

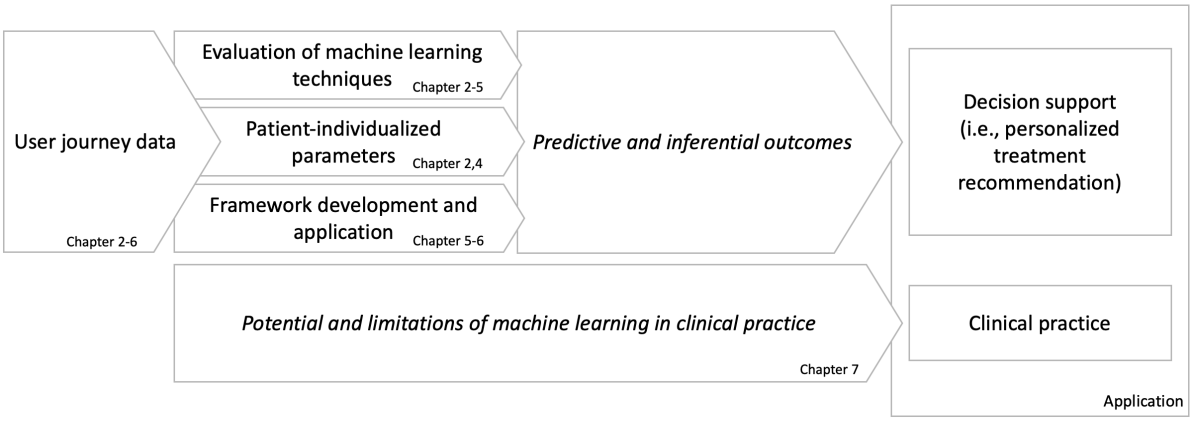


Figure 1.1: Contributions of this dissertation.

One contribution is the illustration of how user journey data can be analyzed in the context of Internet-based psychological interventions based on machine learning techniques. User journey analysis, also referred to as clickstream data analysis, is well established in the field of Online marketing (Chatterjee et al., 2003; Nottorf and Funk, 2013). However, its application in the field of psychology has not yet been examined. Various articles in this dissertation focus on the analysis of user journey data (Chapter 2,4,5,6). This type of data may be defined as a sequence of experiences or interactions an individual encounters to reach a specific goal. In the context of psychological Internet-based interventions, a user journey is the path they take to navigate through an online program. Because individuals interact with these online programs at various points in time, a user journey is here defined as observations of a varying set of features for a particular participant at a specific point in time. The creation of user journey data requires the transformation of the raw data into a wide format. Chapter 5 illustrates this process in more detail. In behavioral research, however, scientists typically utilize pretest-posttest designs when analyzing their data (Dimitrov and Rumrill, 2003). EMA data, for example, are often ignored in that process even though they perfectly suit a personalized analysis and therefore can potentially increase the chance of developing more personalized interventions. Thus, this

dissertation supports the introduction of user journey analysis to the field of psychology and offers a framework for its application (Chapter 5).

Another contribution of this dissertation is the application, exploration, and evaluation of machine learning techniques for the prediction of psychological factors and subsequent revelation of correlations between them. Throughout history, the field of psychology has focused on the explanation of causal mechanism of psychological disorders in order to understand human behavior while it is often assumed that the statistical model that explains the data best also creates the best prediction of future behavior (Yarkoni and Westfall, 2017). Thus, it is uncommon in psychological research to evaluate the predictive performance of applied models but instead to report goodness of fit (Yarkoni and Westfall, 2017), which is no indicator of predictive performance when applied to out-of-sample data. In this dissertation, there was thus a focus on the evaluation of applied machine learning techniques regarding their predictive accuracy. Additionally, complex predictive models such as Bayesian estimation techniques or ordinal logit models have not yet been common in the field of social science research (Grilli and Rampichini, 2012; Yarkoni and Westfall, 2017). However, a continuously growing body of research projects acknowledges and utilizes machine learning in this context as also outlined above (Dwyer et al., 2018; Shatte et al., 2019). Furthermore, statistical models were adapted in order to consider patient-individualized parameters. Through this approach, personalized effects can be revealed that can lead to deeper insight about the patients' behavior (i.e., one patient might be affected differently by certain executed activities compared to another patient). In Chapter 7, limitations and potentials of the application of machine learning techniques in clinical practice were analyzed from a psychotherapists' point of view - possible drivers and inhibitors of the utilization of machine learning in psychotherapeutic treatment are discussed.

Paper	Journal	Ranking
Predicting the individual mood level based on diary data (Chapter 2)	Full paper published at European Conference on Information Systems (2017) and short paper at ECML/PKDD Nectar Track (2018)	B (VHB)
Predicting therapy success and costs for personalized treatment recommendations using baseline characteristics: a data-driven analysis (Chapter 3)	Journal of Medical Internet Research (2018)[1]	4.67 (IF)
Heterogeneity matters – predicting self-esteem in online interventions based on Ecological Momentary Assessment data (Chapter 4)	Depression Research and Treatment (2019)	1.85 (IF)
Developing a process for the analysis of user journeys and the prediction of dropout in digital health interventions: Machine Learning approach (Chapter 5)	Journal of Medical Internet Research (2020) (edited version published)	5.03 (IF)
Machine learning-based decision support systems in clinical practice: Potential and limitations (Chapter 7)	White paper (2020)	-

Table 1.1: Published articles in this dissertation and the corresponding ranking of journals/conferences according to VHB Jourqual V3 (VHB) and Impact Factor (IF) at the time of submission/publication.

1.3.1 Inference and prediction of data in e-mental-health

This section highlights the contributions of each chapter included in this dissertation. These contributions can be divided into the topics outlined in Section *Background and state of the art*.

[1]This paper received the best paper award of Leuphana University Lüneburg

Mood levels

In Chapter 2, the mood level of individual patients was predicted based on free text diary data. The contribution of this article is manifold. A two-step approach was developed in which the free text of the patients was first classified into different activity categories by utilizing a bag-of-words approach and recurrent neural networks. Then, partial ordered logit models were developed. Reasons for mental problems and disorders can differ among individuals. The human mind and behavior are unique. How psychological factors influence one another on an individual patient-level therefore plays an important role in therapeutic processes. Thus, the logit models were modified in order to account for individual behavior of the patients and the parameters were estimated by using Bayesian estimation techniques. This two-step approach is a new contribution and indicates one possibility for a personalized analysis of free text data. Additionally, the individualized prediction of mood and how the mood is affected by daily executed activities can offer valuable information for psychotherapists.

Self-esteem levels

Even though it is strongly debated if depression predicts self-esteem or vice versa, researchers agree that self-views do represent an important factor for depression (Orth et al., 2008). Thus, Chapter 4 focuses on the prediction of self-esteem based on EMA data by utilizing two different models that account for the ordinal structure of the data. Continuing the work of Chapter 2, these ordinal logit models included varying degrees of heterogeneity among the patients and the parameters were estimated using Bayesian estimation techniques. These models were then evaluated based on various predictive performance measures. The results indicated that models that allow for more heterogeneous parameters lead to an increased prediction performance. Thus, one contribution is the emphasis and evaluation regarding the implementation of individual parameters into the models. Subsequently, this article demonstrated patient-individualized inferential results. Self-reported sleep quality or enjoyed activities could influence patients differently. The revelation of such differences in impact can support decision-making processes of practitioners by gaining deeper insight into patients' behavior implying that personalized models could lead to the development of personalized treatment strategies.

Intervention outcomes and treatment recommendation

Chapter 3 of this dissertation focused on two topics: the prediction of outcome and costs of treatment, and based on that, individualized treatment recommendations whenever multiple treatment types exist for the patients. Following Dwyer et al. (2018), who mention that there are no patient-individualized methods for the selection of treatment options in psychotherapeutic treatment, Chapter 3 closes this gap and offers a possible solution for this task. In this article, outcomes and costs were predicted based on baseline data from a two-arm randomized controlled trial on an individual patient-level before the onset of treatment. Various machine learning techniques were evaluated in this process. Then, the incremental cost-effectiveness ratio (ICER)

was applied to the predictions in order to find the most appropriate treatment type for an individual patient. This combination, the utilization of predictions and the ICER, for personalized treatment recommendations is a novel approach in this context and can lead to increased outcomes for patients and decreased associated costs. Even though similar to the research of DeRubeis et al. (2014) regarding the prediction of treatment outcomes, Chapter 3 also considered costs in the process of treatment recommendation and utilized the ICER as a selection tool.

Dropout in interventions

Chapter 5 contributes to existing research by providing a framework for the analysis of user journey data and application of machine learning techniques in the context of digital health interventions. A step-by-step guide for this type of analysis was introduced and discussed. Additionally, a technical implementation for the research community was provided. This framework was then applied to data from an automated web-based program in order to predict dropouts of patients at different points in time of the intervention. As it is important to identify high-risk dropout patients early in the intervention (Lutz et al., 2018), and the applied framework reached predictive performance better than chance early on, the application of the framework was a success in this regard. Because it is important to understand how variables influence dropout (Melville et al., 2010), Shapley additive explanation values (SHAP) were utilized for revealing the importance and impact of the features. This metric is a relatively new concept in the field of machine learning (Lundberg and Lee, 2017) and has not yet been utilized in the field of psychology. Because the number for participants was limited in the study, the developed framework was repeatedly applied to a larger commercial dataset in Chapter 6, which resulted in similar and as well promising results.

1.3.2 Intention to utilize machine learning in clinical practice

Research indicates that the application of machine learning in psychotherapy might be beneficial for the patients and improve care (Clifton et al., 2015; Shatte et al., 2019; Wallace et al., 2012) while utilization of such tools is uncommon in clinical practice (Sacchi et al., 2015). Regardless of how beneficial the application of machine learning techniques in psychotherapy may be, if practitioners claim no interest in adopting decision support systems based on machine learning, no such systems will be utilized. Thus, Chapter 7 explored the potentials and limitations of machine learning based decision support systems in clinical practice. For this purpose, eight semi-structured interviews with professional psychotherapists were conducted. The resulting data were qualitatively analyzed. To the best of my knowledge, no study to date has investigated potentials and limitations of machine learning in this context from a psychotherapists' point of view.

1.4 Utilized data

A variety of data was used in this dissertation for the analysis and prediction of the different aspects outlined above. The data utilized originates from two EU-funded projects E-COMPARED[2] and GET.ON[3]. Additionally, data from the *Sleep Healthy Using The Internet^{TM} (SHUTi)*[4] program has been used. The E-COMPARED project was a comparative study aiming to assess the clinical and cost-effectiveness of blended cognitive behavioral therapy treatment for major depression compared to TAU (treatment as usual). The GET.ON project developed online-based health programs in order to generally prevent diseases, promote health factors, and support patients during the process of finding a place on a treatment program. SHUTi is a personalized and interactive online application that is designed to improve the sleep of adults with insomnia.

The types of data, a brief description, and the corresponding chapter in which the type of data was used are illustrated in Table 1.2. Each project gathered baseline data. Baseline data usually cover information before the start of an intervention, include a variety of questionnaires tailored to the specific disease the intervention focuses on, and serve as a comparison for the evaluation of treatment and its effectiveness. Another type of data that was gathered in each project is EMA data. Dairies, one form of EMA data, were utilized in this dissertation. They are indicators for symptoms, behavior, and cognition close in time to the participants' experience in their natural environment and can take on the form of a rating on a scale or free text provided by the patients (Iida et al., 2012). Furthermore, log data of the Internet-based program such as patient individual logins to the system or sent emails have been used. This type of data can support the understanding of patient engagement with the intervention, indicate possible factors for individualizing treatment designs (Morrison and Doherty, 2014), and might be a predictive factor of various mental disorders. Additionally, semi-structured interviews were conducted, transcribed, and utilized for a qualitative analysis in Chapter 7.

[2] http://www.e-compared.eu/
[3] http://www.geton-training.de/
[4] https://www.myshuti.com

Type of data	Description	Chapter
Baseline data	Various questionnaires filled out before the onset of treatment	3,5,6
Ecological Momentary Assessment (EMA)	Diaries kept by the patients	2,4,5,6
Log data	Interaction with computer system	5,6
Interviews	Semi-structured interviews with professional therapists	7

Table 1.2: Types of data utilized in the articles included in this dissertation.

1.5 Conclusion

This dissertation demonstrates how machine learning techniques can be modified and applied in the context of digital health interventions. The results are promising and indicate the possibility for substantial health-related as well as financial benefits. The developed approach that supports the selection of most suitable treatment types for patients can increase treatment outcomes for individuals while at the same time saving financial resources. Including patient-specific parameters in the machine learning techniques have shown to provide deeper information about the patients, especially important in terms of patient-specific recommendations, even though these models are computationally more expensive. Additionally, inferential results have illustrated how various psychological factors are intertwined. The provided process for user journey analysis in this context could lead to an increased usage of this type of analysis based on machine learning in psychological research and eventually improve decision-making of experts in the field. Successful prediction of dropout in digital health interventions has been shown to be feasible based on this process. Acting on these predictions and developing possible micro-interventions for targeting high-risk patients is a next step for the evaluation of such predictions in clinical practice. Thus, even though this dissertation does provide valuable insight and these approaches are validated based on the utilized data, they have not been implemented in clinical practice or randomized controlled trials. The next step would be to utilize such models and evaluate their benefit for professionals and patients in the field of mental health.

A major limitation of this dissertation is the size of the utilized data. This limited number of observations leads to high uncertainty of inferential results and predictions. Consequently, the predictions might, in some cases, not be reliable enough in order to be utilized in clinical

practice. Thus, the approaches outlined in the articles would need to be evaluated based on various datasets that include a higher number of patients. In order to evaluate the process developed in Chapter 5, it was applied to a commercial and larger dataset in Chapter 6. The results were very promising as well.

EMA data is utilized in Chapter 2, 4, 5, and 6. The used EMA data might not represent the full psychological factor (i.e. self-esteem was not measured with the widely accepted Rosenberg Self-Esteem Scale (Robins et al., 2001) but only consists of one question). It is questionable if one question for the EMA measures can capture complex emotions and states such as mood or self-esteem. Additionally, self-reported data is not evaluated by a professional; thus, it can lack objectivity and can also lead to falsely reported data and social desirability bias (Logan et al., 2008; Moskowitz and Young, 2006). Another limitation most studies suffer from is the number of missing values contained in the datasets. Incomplete data can impact analyses and result in reduced predictive accuracy. Omitting or imputing observations that include missing data are viable options for handling them. However, both options can bias estimates and can potentially lead to higher uncertainty and subsequently invalid conclusions.

Nevertheless, there is room for further research. The additional utilization of unobtrusive measures through smartphones or smartwatches such as GPS or accelerometer data, text messages, or screen time data, as already introduced by various research projects, could lead to an increased predictive performance of machine learning techniques and insight about the patients' behavior. Situations in which psychotherapists could intervene when critical levels of psychological factors are reached could then be determined more accurately. In addition, chatbots that communicate with patients and enable the booking of appointments or help to diagnose diseases based on entered symptoms could increase cost-effectiveness and allow for continuous monitoring of patients and subsequent transmission of reminders. The collected data from chatbots could then be used to improve machine learning outcomes. Besides additional information that can be utilized as input for machine learning techniques, more studies are needed that evaluate the performance and benefits of machine learning in digital health interventions. In this context, deep collaboration between psychologists and data specialists is needed in order to combine machine learning techniques and experimental designs for the evaluation of such models. Without this, the integration of decision support systems based on machine learning in clinical practice seems improbable.

References

Alpaydin, E. (2009). *Introduction to Machine Learning*. MIT Press.

Asselbergs, J., Ruwaard, J., Ejdys, M., Schrader, N., and Sijbrandij, M. (2016). Mobile Phone-Based Unobtrusive Ecological Momentary Assessment of Day-to-Day Mood : An Explorative Study. *JMIR Mental Health*, 18(3):e72.

Becker, D., Breda, W. V., Funk, B., Hoogendoorn, M., and Ruwaard, J. (2018). Predictive modeling in e-mental health : A common language framework. *Internet Interventions*, 12:57–67.

Becker, D., Bremer, V., Funk, B., Asselbergs, J., Riper, H., and Ruwaard, J. (2016). How to Predict Mood? Delving into Features of Smartphone-Based Data. In *Twenty-second Americas Conference on Information Systems*, San Diego (USA).

Belle, A., Kon, M. a., and Najarian, K. (2013). Biomedical informatics for computer-aided decision support systems: a survey. *The Scientific World Journal*, 2013:769639.

Bloom, D. E., Cafiero, E., Jané-Llopis, E., Abrahams-Gessel, S., Bloom, L. R., Fathima, S., Feigl, A. B., Gaziano, T., Mowafi, M., Pandya, A., Prettner, K., Rosenberg, L., Seligman, B., Stein, A., and Weinstein, C. (2011). The Global Economic Burden of Noncommunicable Diseases. Pgda working papers, Geneva: World Economic Forum.

Buntrock, C., Ebert, D. D., Lehr, D., Cuijpers, P., Riper, H., Smit, F., and Berking, M. (2014). Evaluating the efficacy and cost-effectiveness of web-based indicated prevention of major depression: design of a randomised controlled trial. *BMC psychiatry*, 14:25–34.

Byrne, A. and Byrne, D. (1993). The effect of exercise on depression, anxiety and other mood states: A review. *Journal of Psychosomatic Research*, 37(6):565–574.

Byrne, D. (1986). Psychological factors and disease. In King, N. and Remenyi, A., editors, *Health Care: A Behavioural Approach*, pages 33–38. Grune & Stratton, Sydney.

Carlbring, P., Andersson, G., Cuijpers, P., Riper, H., and Hedman-Lagerlöf, E. (2018). Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. *Cognitive Behaviour Therapy*, 47(1):1–18.

Chatterjee, P., Hoffman, D. L., and Novak, T. P. (2003). Modeling the Clickstream: Implications for Web-Based Advertising Efforts. *Marketing Science*, 22(4):520–541.

Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H., Cannon, T. D., Krystal, J. H., and Corlett, P. R. (2016). Cross-trial prediction of treatment outcome in depression: A machine learning approach. *The Lancet Psychiatry*, 3(3):243–250.

Cheng, H. and Furnham, A. (2003). Personality, self-esteem, and demographic predictions of happiness and depression. *Pers Individ Differ*, 34(6):921–42.

Christensen, H., Griffiths, K. M., and Farrer, L. (2009). Adherence in internet interventions for anxiety and depression. *Journal of Medical Internet Research*, 11(2):e13.

Clifton, D. A., Niehaus, K. E., Charlton, P., and Colopy, G. W. (2015). Health Informatics via Machine Learning for the Clinical Management of Patients. *Yearbook of Medical Informatics*, 10(1):38–43.

DeRubeis, R. J., Cohen, Z. D., Forand, N. R., Fournier, J. C., Gelfand, L. A., and Lorenzo-Luaces, L. (2014). The personalized advantage index: Translating research on prediction into individualized treatment recommendations. A demonstration. *PLoS ONE*, 9(1):e83875.

Dimitrov, D. and Rumrill, P. (2003). Pretest-posttest designs and measurement of change. *Work*, 20(3):159–165.

Dobson, D. and Dobson, K. (2009). *Evidence-Based Practice of Cognitive-Behavioral Therapy*. Guilford Press.

Donkin, L., Christensen, H., Naismith, S. L., Hons, B. A., Neuro, D., Neal, B., Chb, M. B., Hickie, I. B., and Glozier, N. (2011). A Systematic Review of the Impact of Adherence on the Effectiveness of e-Therapies. *Journal of Medical Internet Research*, 13(3):e52.

Dwyer, D. B., Falkai, P., and Koutsouleris, N. (2018). Machine Learning Approaches for Clinical Psychology and Psychiatry. *Annu Rev Clin Psychol*, 14:91–118.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542:115–118.

Eysenbach, G. (2001). What is e-health? *Journal of Medical Internet Research*, 3(2):e20.

Fassbinder, E., Hauer, A., Schaich, A., Schweiger, U., Jacob, G. A., and Arntz, A. (2015). Integration of e-Health Tools Into Face-to-Face Psychotherapy for Borderline Personality Disorder : A Chance to Close the Gap Between Demand and Supply? *Journal of clinical psychology*, 71(8):764–777.

Goldberg, D. (2011). The heterogeneity of "major depression". *World Psychiatry*, 10(3):226–228.

Grilli, L. and Rampichini, C. (2012). Multilevel models for ordinal data. In Kenett, R. S. and Salini, S., editors, *Modern Analysis of Customer Surveys: with Applications using R*, chapter 19, pages 391–413. Wiley.

Huibers, M. J. H., Cohen, Z. D., Lemmens, L. H. J. M., and Arntz, A. (2015). Predicting Optimal Outcomes in Cognitive Therapy or Interpersonal Psychotherapy for Depressed Individuals Using the Personalized Advantage Index Approach. *PLoS One*, 10(11):e0140771.

Iida, M., Shrout, P. E., Laurenceau, J.-P., and Bolger, N. (2012). Using diary methods in psychological research. In Cooper, H., Camic, P. M., Long, D. L., Panter, A. T., Rindskopf, D., and Sher, K. J., editors, *APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics.*, pages 277–305. American Psychological Association, Washington.

Kanning, M. and Schlicht, W. (2010). Be active and become happy: an ecological momentary assessment of physical activity and mood. *Journal of sport & exercise psychology*, 32(2):253–261.

Karyotaki, E., Kleiboer, A., Smit, F., Turner, D. T., Pastor, A., Andersson, G., Berger, T., Botella, C., Breton, J., Carlbring, P., Christensen, H., de Graaf, E., Griffiths, K., Donker, T., Farrer, L., Huibers, M., Lenndin, J., Mackinnon, A., Meyer, B., Moritz, S., Riper, H., Spek, V., Vernmark, K., and Cuijpers, P. (2015). Predictors of treatment dropout in self-guided web-based interventions for depression : An ' individual patient data ' meta-analysis. *Psychological Medicine*, 45(13):2717–26.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., and Chouvarda, I. (2017). Machine Learning and Data Mining Methods in Diabetes Research Machine Learning and Data Mining Methods in Diabetes Research. *Computational and Structural Biotechnology Journal*, 15:104–116.

Kegel, A. F. and Flückiger, C. (2015). Predicting Psychotherapy Dropouts : A Multilevel Approach. *Clin Psychol Psychother*, 22(5):377–386.

Kononenko, I. (2001). Machine learning for medical diagnosis : history , state of the art and perspective. *Artif Intell Med*, 23(23):1.

Kroenke, K., Spitzer, R., and Williams, J. (2001). The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med*, 16(9):606–13.

Lee, C. H. and Yoon, H.-j. (2017). Medical big data : promise and challenges. *Kidney Res Clin Pract*, 36(1):3–11.

Leger, D. (1994). The cost of sleep-related accidents: a report for the National Commission on Sleep Disorders Research. *Sleep*, 17(1):84–93.

Lemola, S., Räikkönen, K., Gomez, V., and Allemand, M. (2013). Optimism and self-esteem are related to sleep. Results from a large community-based sample. *International Journal of Behavioral Medicine*, 20(4):567–571.

Likamwa, R., Liu, Y., Lane, N., and Zhong, L. (2013). MoodScope: building a mood sensor from smartphone usage patterns. In *MobiSys '13 Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services*, page pp.

Lindblom, K., Gregory, T., Wilson, C., Flight, I. H., and Zajac, I. (2012). The impact of computer self-efficacy, computer anxiety, and perceived usability and acceptability on the efficacy of a decision support tool for colorectal cancer screening. *J Am Med Inform Assoc*, 19(3):407–12.

Logan, D. E., Claar, R. L., and Scharff, L. (2008). Social desirability response bias and self-report of psychological distress in pediatric chronic pain patients. *Pain*, 136(3):366–372.

Lopez-Gomez, I., Lorenzo-Luaces, L., Chaves, C., Hervas, G., Derubeis, R. J., and Vazques, C. (2019). Predicting optimal interventions for clinical depression: Moderators of outcomes in a positive psychological intervention vs. cognitive-behavioral therapy. *Gen Hosp Psychiatry*, 61:104–110.

Lundberg, S. M. and Lee, S.-i. (2017). A Unified Approach to Interpreting Model Predictions. In *Neural Information Processing Systems (NIPS)*, pages 4765–4774.

Lutz, W., Schwartz, B., Hofmann, S. G., Fisher, A. J., Husen, K., and Rubel, J. A. (2018). Using network analysis for the prediction of treatment dropout in patients with mood and anxiety disorders : A methodological proof- of-concept study. *Scientific Reports*, 8(1):7819.

Manna, G., Falgares, G., Ingoglia, S., Como, M. R., and Santis, S. D. (2016). The relationship between self-esteem , depression and anxiety : Comparing vulnerability and scar model in the Italian context. *Mediterranean Journal of Clinical Psychology*, 4(3):1–17.

Månsson, K. N. T., Frick, A., Boraxbekk, C.-J., Marquand, A. F., Williams, S. C. R., Carlbring, P., Andersson, G., and Furmark, T. (2015). Predicting long-term outcome of Internet-delivered cognitive behavior therapy for social anxiety disorder using fMRI and support vector machine learning. *Translational Psychiatry*, 5(3):e530.

McMahon, F. J. (2014). Prediction of treatment outcomes in psychiatry-where do we stand? *Dialogues in Clinical Neuroscience*, 16(4):455–464.

Melville, K. M., Casey, L. M., and Kavanagh, D. J. (2010). Dropout from internet-based treatment for psychological disorders. *British Journal of Clinical Psychology*, 49(4):455–471.

Mikus, A., Hoogendoorn, M., Rocha, A., Gama, J., Ruwaard, J., and Riper, H. (2018). Predicting short term mood developments among depressed patients using adherence and ecological momentary assessment data. *Internet Interventions*, 12:105–110.

Morrison, C. and Doherty, G. (2014). Analyzing Engagement in a Web-Based Intervention Platform Through Visualizing Log-Data. *Journal of Medical Internet Research*, 16(11):e252.

Moskowitz, D. S. and Young, S. N. (2006). Ecological momentary assessment: what it is and why it is a method of the future in clinical psychopharmacology. *Journal of psychiatry & neuroscience : JPN*, 31(1):13–20.

Nottorf, F. and Funk, B. (2013). The economic value of clickstream data from an advertiser's perspective. In *Proceedings of the 21st European Conference on Information Systems*.

Orth, U., Robins, R. W., and Roberts, B. W. (2008). Low Self-Esteem Prospectively Predicts Depression in Adolescence and Young Adulthood. *J Pers Soc Psychol*, 95(3):695–708.

Paradise, A. and Kernis, M. H. (2002). Self-esteem and Psychological Well-being: Implications of Fragile Self-esteem. *Journal of Social and Clinical Psychology*, 21:345–361.

Park, K. and Yang, T.-c. (2017). HHS Public Access. *Sociol Q*, 58(3):429–446.

Pearson, R., Pisner, D., Meyer, B., Shumake, J., and Beevers, C. G. (2018). A machine learning ensemble to predict treatment outcomes following an Internet intervention for depression. *Psychol Med*, pages 1–12.

Pedersen, D., Mansourvar, M., Sortso, C., and Schmidt, T. (2019). Predicting Dropouts From an Electronic Health Platform for Lifestyle Interventions: Analysis of Methods and Predictors. *Journal of Medical Internet Research*, 21(9):e13617.

Priebe, S., Richardson, M., Cooney, M., Adedeji, O., and McCabe, R. (2011). Does the therapeutic relationship predict outcomes of psychiatric treatment in patients with psychosis? A systematic review. *Psychother Psychosom*, 80(2):70–7.

Proudfoot, J., Clarke, J., Birch, M., Whitton, A., Parker, G., Manicavasagar, V., Harrison, V., Christensen, H., and Hadzi-Pavlovic, D. (2013). Impact of a Mobile Phone and Web Program on Symptom and Functional Outcomes for People With Mild-To-Moderate Depression, Anxiety and Stress: A Randomised Controlled Trial. *BMC psychiatry*, 13(312).

Reneses, B., Munoz, E., and Lopez-Ibor, J. (2009). Factors predicting drop-out in community mental health centres. *World Psychiatry*, 8(3):173–177.

Riano, D., Real, F., Lopez-Vallverdu, J. A., Campana, F., Ercolani, S., Mecocci, P., Annicchiarico, R., and Caltagirone, C. (2012). An ontology-based personalization of health-care knowledge to support clinical decisions for chronically ill patients. *J Biomed Inform*, 45(3):429–46.

Richards, D. and Richardson, T. (2012). Computer-based psychological treatments for depression: a systematic review and meta-analysis. *Clin Psychol Rev*, 32(4):329–342.

Riper, H., Andersson, G., Christensen, H., Cuijpers, P., Lange, A., and Eysenbach, G. (2010). Theme Issue on E-Mental Health: A Growing Field in Internet Research. *Journal of Medical Internet Research*, 12(5):e74.

Robins, R., Hendin, H., and Trzesniewski, K. H. (2001). Measuring Global Self-Esteem: Construct Validation of a Single-Item Measure and the Rosenberg Self-Esteem Scale. *Pers Soc Psychol Bull*, 27:151–161.

Sacchi, L., Quaglini, S., Lanzola, G., and Viani, N. (2015). Personalization and Patient Involvement in Decision Support Systems: Current Trends. *Yearbook of medical informatics*, 10(1):106–118.

Sander, L., Rausch, L., and Baumeister, H. (2016). Effectiveness of Internet-Based Interventions for the Prevention of Mental Disorders: A Systematic Review and Meta-Analysis. *JMIR Mental Health*, 3(3):e38.

Saunders, R., Cape, J., Fearon, P., and Pilling, S. (2016). Predicting treatment outcome in psychological treatment services by identifying latent profiles of patients. *J Affect Disord*, 197:107–15.

Schofield, D., Shrestha, R., Percival, R., Passey, M., Callander, E., and Kelly, S. (2011). The personal and national costs of mental health conditions: impacts on income, taxes, government support payments due to lost labour force participation. *BMC Psychiatry*, 11(72).

Scott, I., Cook, D., Coiera, E., and Richards, B. (2019). Machine learning in clinical practice: prospects and pitfalls. *Med J Aust*, 211(5):203–205.

Scott, K. M., Lim, C., Al-Hamzawi, A., Alonso, J., Bruffaerts, R., Caldas-de Almeida, J. M., Florescu, S., de Girolamo, G., Hu, C., de Jonge, P., Kawakami, N., Medina-Mora, M. E., Moskalewicz, J., Navarro-Mateu, F., O'Neill, S., Piazza, M., Posada-Villa, J., Torres, Y., and Kessler, R. C. (2016). Association of Mental Disorders With Subsequent Chronic Physical Conditions. *JAMA Psychiatry*, 73(2):150–158.

Shatte, A. B. R., Hutchinson, D. M., and Teague, S. J. (2019). Machine learning in mental health : a scoping review of methods and applications. *Psychol Med*, 49(9):1426–1448.

Shimabukuro, D. W., Barton, C. W., Feldman, M. D., Mataraso, S. J., and Das, R. (2017). Effect of a machine learning-based severe sepsis prediction algorithm on patient survival and hospital length of stay : a randomised clinical trial. *BMJ Open Respir Res*, 4(1):e000234.

Smit, F., Cuijpers, P., Oostenbrink, J., Batelaan, N., de Graaf, R., and Beekman, A. (2006). Costs of nine common mental disorders: implications for curative and preventive psychiatry. *Journal of Mental Health Policy and Economics*, 9(4):193–200.

Sowislo, J. F. and Orth, U. (2013). Does Low Self-Esteem Predict Depression and Anxiety? A Meta-Analysis of Longitudinal Studies. *Psychological bulletin*, 139(1):213–240.

Starr, L. and Savila, J. (2013). Temporal Patterns of Anxious and Depressed Mood in Generalized Anxiety Disorder: A Daily Diary Study. *Behav Res Ther*, 50(2):131–141.

Steiger, A., Allemand, M., Robins, R., and Fend, H. (2014). Low and Decreasing Self-Esteem During Adolescence Predict Adult Depression Two Decades Later. *Journal of Personality and Social Psychology*, 106(2):325–38.

Steiger, A., Fend, H., and Allemand, M. (2015). Testing the Vulnerability and Scar Models of Self-Esteem and Depressive Symptoms from Adolescence to Middle Adulthood and Across Generations. *Developmental Psychology*, 51(2):236–247.

Swann, W., Chang-Schneider, C., and Larsen McClarty, K. (2007). Do people's self-views matter? Self-concept and self-esteem in everyday life. *American Psychologist*, 62(2):84–94.

Tomlinson, A., Furukawa, T. A., Efthimiou, O., Salanti, G., De Crescenzo, F., Singh, I., and Cipriani, A. (2020). Personalise antidepressant treatment for unipolar depression combining individual choices, risks and big data (petrushka): rationale and protocol. *Evidence-Based Mental Health*, 23(2):52–56.

Trautmann, S., Rehm, J., and Wittchen, H. (2016). The economic costs of mental disorders. *EMBO Rep*, 17(9):1245–1249.

Triantafyllidis, A. K. and Tsanas, A. (2019). Applications of Machine Learning in Real-Life Digital Health Interventions : Review of the Literature. *Journal of Medical Internet Research*, 21(4):e12286.

Trzesniewski, K. H., Donnellan, M. B., Moffitt, T. E., Robins, R. W., Poulton, R., and Caspi, A. (2006). Low self-esteem during adolescence predicts poor health, criminal behavior, and limited economic prospects during adulthood. *Developmental psychology*, 42(2):381–90.

van Breda, W., Bremer, V., Becker, D., Hoogendoorn, M., Funk, B., Ruwaard, J., and Riper, H. (2018). Predicting therapy success for treatment as usual and blended treatment in the domain of depression. *Internet Interventions*, 12:100–104.

Van Breda, W., Hoogendoorn, M., Eiben, A. E., Andersson, G., Riper, H., Ruwaard, J., and Vernmark, K. (2016). A feature representation learning method for temporal datasets. In *IEEE Symposium Series on Computational Intelligence*, pages 1–8.

van Breda, W., Pastor, J., Hoogendoorn, M., Ruwaard, J., Asselbergs, J., and Riper, H. (2016). Exploring and Comparing Machine Learning Approaches for Predicting Mood Over Time. In

Innovation in Medicine and Healthcare 2016, pages 37–47, Tenerife, Spain. Springer International Publishing.

Vos, T., Barber, R., and Bell, B. (2015). Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet*, 22(386):743–800.

Wallace, B. C., Small, K., Brodley, C. E., Lau, J., and Trikalinos, T. (2012). Deploying an interactive machine learning system in an evidence-based practice center. *Proceedings of the 2nd ACM SIGHIT symposium on International health informatics*, pages 819–824.

Weidman, A., Sun, J., Vazire, S., Quoidbach, J., Ungar, L., and Dunn, E. (2020). (Not) hearing happiness: Predicting fluctuations in happy mood from acoustic cues using machine learning. *Emotion*, 20(4):642–658.

Whitton, A. E., Proudfoot, J., Clarke, J., Birch, M.-r., Parker, G., Manicavasagar, V., and Hadzipavlovic, D. (2015). Breaking Open the Black Box : Isolating the Most Potent Features of a Web and Mobile Phone-Based Intervention for Depression , Anxiety , and Stress. *JMIR Ment Health*, 2(1):e3.

WHO (2017). Depression and Other Common Mental Disorders: Global Health Estimates. *Geneva: World Health Organization*.

Yang, D., Hur, J.-W., Kwak, Y. B., and Choi, S.-W. (2018). A Systematic Review and Meta-Analysis of Applicability of Web-Based Interventions for Individuals with Depression and Quality of Life Impairment. *Psychiatry Investig*, 15(8):759–766.

Yarkoni, T. and Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology : Lessons From Machine Learning. *Perspect Psychol Sci*, 12(6):1100–1122.

Zulueta, J., Piscitello, A., Rasic, M., Easter, R., Babu, P., and Scott, A. (2018). Predicting Mood Disturbance Severity with Mobile Phone Keystroke Metadata : A BiAffect Digital Phenotyping Study. *J Med Internet Res*, 20(7):e241.

CHAPTER 2

PREDICTING THE INDIVIDUAL MOOD LEVEL BASED ON DIARY DATA

Bremer, V., Becker, D., Funk, B., and Lehr, D. (2017). Predicting the Individual Mood Level based on Diary Data. Proceedings of the Twenty-Fifth European Conference on Information Systems, Guimaraes, June 5-10, pp. 1161-1177.

Bremer, V., Becker, D., Genz, T., Funk, B., and Lehr, D. (2018). A two-stage approach for the prediction of mood levels based on diary data. ECML/PKDD Nectar Track.

Abstract

Understanding mood changes of individuals with depressive disorders is crucial in order to guide personalized therapeutic interventions. Based on diary data, in which clients of an online depression treatment report their activities as free text, we categorize these activities and predict the mood level of clients. We apply a bag-of-words text mining approach for activity categorization and explore recurrent neural networks to support this task. Using the identified activities, we develop partial ordered logit models with varying levels of heterogeneity among clients to predict their mood. We estimate the parameters of these models by employing Markov Chain Monte Carlo techniques and compare the models regarding their predictive performance. Therefore, by combining text mining and Bayesian estimation techniques, we apply a two-stage analysis approach in order to reveal relationships between various activity categories and the individual mood level. Our findings indicate that the mood level is influenced negatively when participants report about sickness or rumination. Social activities have a positive influence on the mood. By understanding the influences of daily activities on the individual mood level, we hope to improve the efficacy of online behavior therapy, provide support in the context of clinical decision-making, and contribute to the development of personalized interventions.

2.1 Introduction

A good state of mental health is crucial for every individual as it provides general motivation in achieving life goals and a healthy environment. However, many individuals lack proper mental health and suffer from a variety of mental health disorders. Studies report a striking 7% of the European society as having suffered from major depression (Wittchen et al., 2011). Depression, being just one out of hundreds of various types of mental disorders, creates a mental, social, emotional, and financial burden that not just affects the 7% of individuals diagnosed, but also the families of those individuals while at the same time even imposing financial expenses at the government level (Gustavsson et al., 2011; Leger, 1994). In the health sphere, major depression is associated with a substantial loss of quality of life and increased mortality rates (Buntrock et al., 2014).

Since mood changes can play a crucial role regarding depression and are experienced by many individuals on a daily basis, we focus on the prediction of the mood level in this study. The changes of mood can be affected by executed activities and varying events throughout the day (Weinstein and Mermelstein, 2007). These events and subsequently mood levels have a stake in determining well-being and cognitive functions such as problem solving (Isen et al., 1987), creativity, and the performance level (Nadler et al., 2010). Because various activities from walking a dog, to volunteering, cleaning the house, or having a drink out with friends affect mood in different and complex ways (Weinstein and Mermelstein, 2007), we attempt to analyze the effects that different activities can have on the mood level of an individual. Despite the fact that the importance of daily activities for a person's happiness and well-being is well known (Tadic et al., 2013), there is no specific indicator of how different activities explicitly affect the mood level of a client. Although it is recognized that changes in mood exist - the actual origin of them and how certain activities are connected with mood changes is not yet understood completely (Weinstein and Mermelstein, 2007). In that context, we hope to provide further insight.

For our approach, we utilize diary data that is provided by participants of an online depression treatment (Buntrock et al., 2014). Diary data is often collected using Ecological Momentary Assessments (EMA) (Iida et al., 2012; Smyth and Stone, 2003). These methods and online healthcare treatments have been established in order to treat depression and other mental disorders; resulting in the collection of a new kind of data. EMA methods are one option to collect data on symptoms and behavior close in time to a clients experience - and most importantly - in their natural environment (Iida et al., 2012). These methods and Internet-based treatment in general can potentially lead to an increase of quality of treatment by providing deeper insight into the daily lives of the participants (Eysenbach, 2001; Iida et al., 2012).

The field of Information Systems (IS) can contribute to gaining insight into individual behavior and E-Mental-Health in different ways (Agarwal and Dhar, 2014). Developing statistical models and applying techniques such as text mining for the analysis of data represent powerful ways of improving the understanding of clients in therapeutic treatments. They simultaneously provide

the opportunity to reveal relationships and effects between psychological concepts (Agarwal and Dhar, 2014) and can therefore inform decision-making in the E-Health sector (Jardim, 2013).

In this study, we utilize text mining techniques in order to categorize free text diary data and use partial ordered logit models to subsequently predict the mood level. By doing so, we illustrate the importance of accounting for heterogeneity among individual clients. For parameter estimation, Markov Chain Monte Carlo (MCMC) techniques are employed. Besides studying the relationship between activities and the mood level, we contribute to the field of Information Systems by providing a mixed method approach to analyze diary data. We can further support the decision-making process in online therapy by, for example, offering important insights into *how* and *when* to intervene in online therapy.

In the following chapters, we first discuss related literature. We then present the experimental setting of our study including a brief description of the dataset, introduce the predictors, and describe our text mining approach and model development. Finally, we illustrate the results, point out some limitations, and conclude our work.

2.2 Related literature

A low mood level can potentially result in severe depression (Minden, 2000). An entire set of behavioral patterns are affected by mood changes and a low mood level. Since we know that different activities influence the mood in various ways (Weinstein and Mermelstein, 2007), it is increasingly important to study and evaluate the impact of daily events on the mood level of participants. In this chapter, we demonstrate the importance of this topic in general and illustrate the state-of-the-art regarding mood changes. We do not specify activity categories in this chapter but illustrate general relationships of psychological concepts and mood.

According to early research in the field of behavioral theories, pleasant events have great potential of improving the mood level of individuals and general well-being (Grosscup and Lewinsohn, 1980; Lewinsohn and Amenson, 1978). Researchers have also identified the existence of a relationship between specific activities such as exercise (Wang et al., 2012) or social activities (Byrne and Byrne, 1993; Clark and Watson, 1988) and the mood level of individuals. In recent years, the impact of low moods was further investigated whereas serious consequences could be identified. In the case of experiencing long-lasting and "excessive" low mood levels, severe depression can occur (Nesse, 2000). This state of enduring negative mood can subsequently lead to low self-esteem, pessimism, sadness, loss of pleasure in favorite activities, and even an increased risk for cardiovascular disease (Both et al., 2008; Nesse, 2000; Penninx et al., 2001). Besides that, Donaldson and Lam (2004) find a relationship between depression, mood, rumination, and problem solution skills. Specifically, they reveal that depressed and ruminative participants with a lower mood level are more challenged by problems and deliver less effective solutions compared to less ruminative individuals. Additionally, Reis et al. (2000) utilize hierarchical linear models

in order to analyze factors that influence emotional well-being. Their results indicate that three concepts are crucial for an individual - autonomy, competence, and relatedness. They also find that the mood level is explicitly increasing on the weekend and decreasing on Mondays. The latter finding is also consistent with the results of Becker et al. (2016) who seek to predict the mood level of 27 healthy Dutch students by utilizing smartphone-based data and a varying set of statistical methods.

Regarding our approach of categorizing free text into activity categories, we are not the first to implement text mining techniques. Balog et al. (2006), for example, provide solutions for determining mood changes and irregularities in blog posts. Their work compares corpus frequencies of terms which lead to an identification of the decisive factors regarding mood changes. Kramer et al. (2014) utilize bag-of-words techniques (Linguistic Inquiry and Word Count) and study how emotional posts spread on Facebook. They find that a reduction in positive news leads to less positive and more negative posts by users and vice versa.

Improving predictions of psychological concepts such as mood can lead to essential benefits for clients and reduce health care costs (McMahon, 2014). Since it is often difficult for therapists to predict specific outcomes for patients (Hannan et al., 2005), computerized methods can help and support the decision-making process (Garg et al., 2005). Bright et al. (2012), for example, found that clinical decision support can lead to improved preventive care services. Furthermore, various researchers utilize predictive models in the field of E-Mental-Health in order to reveal relationships between concepts or investigate the acceptance of technological systems (Chih et al., 2014; Hah and Bharadwaj, 2012; Hippisley-Cox et al., 2008; Wilson and Lankton, 2004). Therefore, it is increasingly important to develop approaches and models in order to further support the therapist's work and aim to provide efficient tools that can enhance decision processes and eventually individual outcomes. To the best of our knowledge, there is no study that seeks to categorize free text diary data from interventions and simultaneously predicts specific outcomes.

2.3 Setting, predictors, and extracting activities

The dataset utilized in our research is acquired from two separate trials of an online depression treatment that "evaluate the efficacy of a newly developed guided self-help Internet-based intervention compared to an online psychoeducation on depression" (Buntrock et al., 2014; Ebert et al., 2014). The participants are recruited from the German population via the GET.ON[1] research website. The dataset represents the responses of 440 clients who are 18 years or older, suffer from subthreshold depression, do not have a major depressive episode, and have Internet access. Participants who have a history of psychotic disorders, currently receive psychotherapy, or show a notable suicidal risk are excluded (Buntrock et al., 2014; Ebert et al., 2014). All clients gave their informed consent. Our dataset is based on an activity diary that has been kept by

[1] http://www.geton-training.de/

the participants; the data has been gathered through a secured online-based assessment system (Buntrock et al., 2014).. In this diary, the clients specify their daily activities as free text and simultaneously report their corresponding individual mood level once a day. In total, we received 9,192 diary entries. Most of the analyzed clients are female (76.2%). The majority (82.4%) of participants are employed (at least part-time) and the average age is 45 years (SD=11.5).

2.3.1 Activity categories

Work related activities

In this study, work related activities are defined as all actions that can be linked to duties on the job; examples for this category can be 'call at work' or 'office meeting'. Work related activities can have positive or negative influences on individuals based on the type of the experience. Great achievements at work, for example, can increase the mood level and *bad* experiences at work decrease the mood state. Stone (1987) and Stewart and Barling (1996) state that work related stress factors are strongly associated with negative mood and especially when not being able to detach from work, they can also be a crucial aspect for recovery processes (Cropley and Zijlstra, 2011). On the other hand, Tadic et al. (2013) find that participation in daily work related activities increases the chance of being in a momentary state of happiness. One reason for this finding could be the fact that work related activities foster cognitive abilities that in turn can result in greater achievements at work. Interested in the general effects work related activities have on mood, we examine the effects of *good* and *bad* perceived work related activities on the mood level. We hypothesize negative effects from these events because we assume that the continuous stress factor of work potentially outweighs possible momentary feelings of satisfaction that arise out of great work outcomes.

Recreational activities

Recreational activities aim to rebuild psychological resources (Rook and Zijlstra, 2006) and negative effects that result from exertion (Demerouti et al., 2012). These activities can potentially increase life satisfaction, distract from work stress, and are an important factor for the sleep quality of an individual (Sluiter et al., 2003). With work and sleep taking up a large amount of an individual's day, it is more important to find other activities that help to cope with the daily stress many individuals experience. Thus, how an individual spends alone or leisure time is important for the recovery process (Cropley and Zijlstra, 2011) and can, furthermore, support overcoming daily stress and, in turn, prevent low mood levels (Qian et al., 2014). In our data, we define recreational activities as leisure time activities. The reported text fields are only assigned to the recreational activity category if they are executed completely alone (otherwise they would be assigned to the category social activity which will be introduced below). We expect recreational activities to have a positive effect on the mood level.

Necessary Activities

We define necessary activities as the kind of action that is frequently needed in an individual's life. Examples of necessary activities are grocery shopping and household chores such as cleaning and vacuuming. These activities do not necessarily need to be perceived as negative - however, they can often be *unwanted* or tedious activities that require energy and are more likely to decrease the mood level of an individual (Bolger et al., 1989). We hypothesize negative influences on the mood level from this activity category.

Exercise

Physical activity can be defined as any movement that requires "energy expenditure". Exercise is a more structured way of physical activity and seeks to increase physical fitness (Caspersen et al., 1985). However, in this study, we use the terms interchangeably. Previous literature widely assumes that both exercise and physical activity in general can influence the psychological well-being and happiness of individuals positively and can further even benefit the individual mood state (Byrne and Byrne, 1993; Kanning and Schlicht, 2010; Wang et al., 2012). Netz and Lidor (2003) also show that clients are often "less anxious, tense, depressed, angry, and confused after exercising than before". Therefore, we hypothesize positive effects from physical activities on the mood level.

Sickness

Sickness can lead to decreasing levels of mood and previous literature already indicates that life threatening diseases such as stroke and cancer influence the mood level negatively (McCorkle and Quint-Benoliel, 1983; Robinson et al., 1984). But how does the state of "normal and every day life sickness" such as a cold or a headache influence the mood level? In an attempt to answer this question, we use this activity category as predictor to measure its influence and predict the individual mood level. We expect this category to have a negative influence on the mood.

Sleep related activities

Sleep loss can be associated with changes of mood, fatigue, and stress (Dinges et al., 1997; Rosen et al., 2006). But sleep can also be perceived as a state of relaxation and rest when experiencing *good* sleep quality and an appropriate amount of sleep. Under that condition, sleep can be used to recover and improve the mood level (Bolger et al., 1989; Dinges et al., 1997). Therefore, we are interested in how sleep affects the mood level of the participants. We are uncertain of how sleep related activities affect the mood level.

Rumination

We define rumination as a state of repetitively reflecting and thinking about upsetting situations and life in general. Rumination can possibly lead to a multitude of negative emotions (Thomsen et al., 2003). Furthermore, depressed individuals are more "self-focused" than non depressed

individuals (Ingram and Smith, 1984; Larsen and Cowan, 1988). Ruminative responses, a specific type of self-focusing, represents the state of primarily thinking about depressive symptoms and their consequences (Nolen-Hoeksema and Morrow, 1993). Constantly being reminded of those symptoms and their aftermath can have a negative effect on the mood level of individuals. On the other hand, ruminative phases might provide insight and support to overcome personal problems (Watkins and Baracaia, 2001). In our analysis, the free text fields are assigned to the rumination category whenever states of *serious thoughts* are reported. Therefore, we include positive as well as negative thoughts in our rumination category. Nevertheless, we hypothesize negative effects on the mood level from this category.

Social activities

Social activities have been shown to result in an increased "positive affect" for individuals (Weinstein and Mermelstein, 2007). Previous research finds a consistent positive relationship between social activities and a person's mood level (Clark and Watson, 1988; David et al., 1997). Moreover, social activeness can also lead to a general increased well-being and improve negative mood states (Weinstein and Mermelstein, 2007). In our analysis, we define social activities as a state of spare or leisure time where at least one person is present besides the participant. These social activities can either be *good* or *bad* experiences. We expect positive effects from social activities on the individual mood level.

2.3.2 Text Mining: Extracting activities

We seek to categorize free text diary data and apply various statistical models in order to predict the individual mood level of the clients. The aim of this chapter is to demonstrate how we categorize the free text into the above specified activity categories. For this purpose, we utilize a bag-of-words (BoW) approach and extent the results by applying recurrent neural networks (RNN) (Elman, 1990). We use the RNN extension in order to categorize free texts that have not yet been classified by the BoW technique. The outcomes of both are then compared by their predictive performance. Figure 2.1 illustrates our approach.

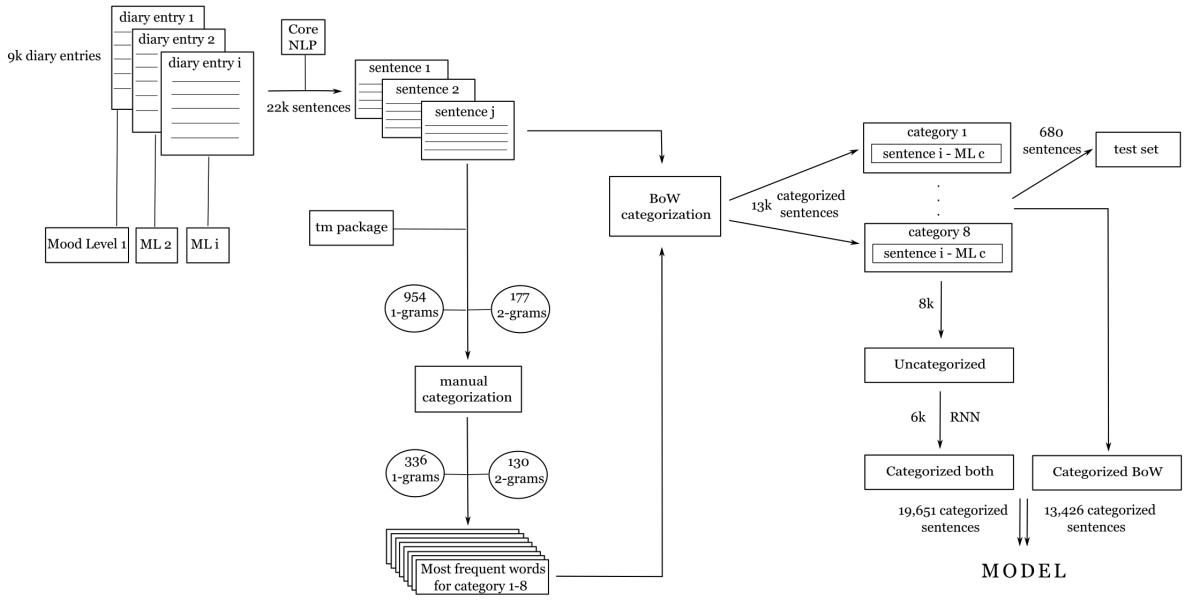


Figure 2.1: Process of the text mining approaches for the categorization of diary entries.

First, we separate the diary entries into multiple sentences by using the `NLP` package (Hornik, 2016). The separation of sentences appears necessary because the extent and format of the content in the diary entries varies tremendously. Some clients only provide keywords whereas others state short paragraphs for their daily activities. This process results in 21,598 different sentences. Afterwards, we convert the free text to lower case and remove punctuation, numbers, and stop words (words that do not have any contribution in content, i.e. here, too, nor, about, etc.) by utilizing the `tm` package (Feinerer et al., 2008). In a next step, we identify the most frequent 2-grams and 1-grams and require that they appear at least 10 times in the corpus specified above. The 177 most frequent 2-grams and the 954 most frequent 1-grams are then manually inspected. Specifically, two authors independently categorize the most frequent 1- and 2-grams into the previously defined activity categories. Only the 1- and 2-grams that are assigned identically by both authors (336 1-grams and 130 2-grams) are utilized for the BoW technique. To measure the inter-rate agreement rate, we calculate Cohen's Kappa: For the 1-grams, we achieve a value of .57. According to Landis and Koch (2008), this value can be considered to be a "Moderate" agreement. For the 2-grams, the Cohen's kappa coefficient is .75 ("Substantial") (Landis and Koch, 2008); it achieves a higher kappa coefficient because it includes more context information.

The algorithm then searches for the n-grams in the free text fields reported by the participants and assigns them to the activity categories. Whenever a sentence is connected with various categories, the sentence is assigned multiple times. This method results in 13,426 categorized sentences. Since 8,032 sentences cannot be categorized by the previously described approach (they do not contain any of the n-grams), we explore the predictive power of a recurrent neural network in this context. To do so, we use the 13,426 categorized sentences to train the recurrent

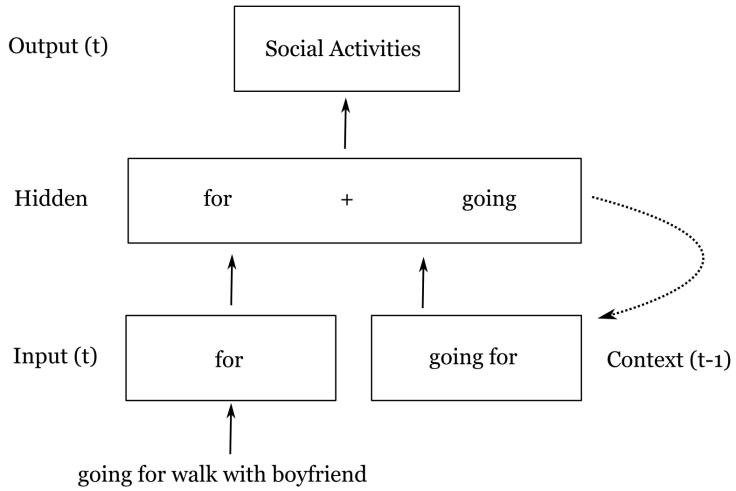


Figure 2.2: Visualization of the Elman Network.

neural network.
Why do we choose RNN? RNN architectures have been shown to produce strong results in language processing (Kombrink et al., 2011). One reason for that is the RNN's capability of word embedding, this is, each word is represented as a vector and the value of this vector is changed during learning. After the learning phase, similar words of the same category are in proximity to each other in the vector space. This fact enables RNNs to extend the available vocabulary for classifying further sentences.

The RNN is implemented as an Elman network that consists of three layers: the input, hidden, and output layer (Elman, 1990). In this network, each word is represented as a vector and presented to the input layer. In the hidden layer, the input is combined with the previous output of the hidden layer. This result is then redirected into so called "context units". These context units model a temporal memory that allows the consideration of word sequences. The word-vectors are then sequentially being presented to the neural network that subsequently tries to estimate the category of the sentence (output layer). Figure 2.2 illustrates an example of the classification process of the neural network. In this case, the first word that is presented to the network is "going". The word vector is combined with the content of the context unit in the hidden layer. For the first word, the context layer consist of zeros and has no influence because no word was previously presented. The next word, which is the word "for", is combined with the previous result of the hidden layer. This specific step is represented in Figure 2.2. After this step, the context unit consists of a vector representing both words. This process is repeated for each word of the sentence. In the end, the output layer estimates the probability for each category. The category with the highest probability is subsequently selected. The RNN classifies 6,225 sentences that are not already assigned by the BoW technique. In the end, 1,807 sentences are not determined, because these sentences consist of words that do not appear in the 13,426 sentences

used for training purposes. The results of both approaches are then merged and utilized as input for the statistical model described in the next section.

2.4 Model development

For analyzing the effects of the activity categories on the individual mood level and predicting specific outcomes, we use a partial ordered logit model and employ MCMC techniques for estimating the parameters. It is important to consider that the effects of activities on the mood level also depend on factors within the individuals and can therefore be influenced by exclusive personal and behavioral factors (Gable et al., 2000; Weinstein and Mermelstein, 2007). Therefore, heterogeneity among clients can be an important aspect in statistical analyses. By developing multiple models with varying levels of heterogeneity among participants, we not only seek to compare our models and demonstrate the importance of heterogeneity, but achieve a greater prediction performance. We hypothesize that the results become more accurate when allowing for more heterogeneity in the model. In the following, we iteratively illustrate the utilized models and their modifications which account for an increasing amount of heterogeneity.

The dependent variable in this analysis is the mood level on a scale from one to ten. Even though the scores on the scale are ranked, the "real" space between them remains unidentified and cannot be interpreted as real numbers (Norusis, 2010). Ordered logit models address this challenge and are often used in research when ordinal outcomes are involved (Liu and Koirala, 2012).

$$\theta_{ijt} = \frac{P(mood_{jt} \leq i \mid x_{jt})}{P(mood_{jt} > i \mid x_{jt})}.$$

In general, this model seeks to estimate the odds θ_{ijt} of being at or below a specific rank of the dependent variable $mood_{jt}$ given the data x_{jt} for each rank i, a specific client j, and every repeated measurement at time t. Here, $mood_{jt}$ represents the EMA response for the mood level for client j at time t. x_{jt} is a vector of length eight (for each activity category) where each element accounts for the number of executed activity for client j at time t. Since specific sentences can be assigned multiple times - also to different activity categories - and the same activity can also be executed more than once a day, multiple elements in the vector x_{jt} can exceed one.

The ordered logit model is based on the proportional odds assumption (Peterson and Harrell, 1990). This assumption represents the belief that the relationship between the independent variables and the outcome of the dependent variable do not depend on the rank (Liu and Koirala, 2012; McCullagh, 1980; Peterson and Harrell, 1990). Specifically, the independent variables (activity categories) have the same effect on the outcome variable across all ranks of the mood level. However, this assumption is often violated in real datasets which can lead to serious problems of interpreting the results (Liu and Koirala, 2012). When the proportional odds assumption is violated, models that allow for varying effects of the predictors among the outcome ranks have been shown to be a better fit compared to the ordered logit model (Liu and Koirala,

2012). Thus, we perform likelihood ratio tests of the proportional odds assumption by utilizing the `ordinal` package (Christensen, 2015) in `R`. The results indicate a violation of the proportional odds assumption for the variables social activities, work related activities, necessary activities, exercise, sickness, and rumination. Based on these results, we then develop a partial ordered logit model. In this case, only some relationships between predictors and the dependent variable do not depend on the rank. Specifically, the variables that violate the proportional odds assumption are allowed to have varying effects among the ranks of the mood level.

$$\begin{aligned}\ln(\theta_{ijt}) = \alpha_i - (&\beta_{social_i}\, x_{social_{jt}} + \beta_{work_i}\, x_{work_{jt}} + \beta_{recreational}\, x_{recreational_{jt}} \\ &+ \beta_{necessary_i}\, x_{necessary_{jt}} + \beta_{exercise_i}\, x_{exercise_{jt}} + \beta_{sickness_i}\, x_{sickness_{jt}} \\ &+ \beta_{sleep}\, x_{sleep_{jt}} + \beta_{rumination_i}\, x_{rumination_{jt}}).\end{aligned} \tag{2.1}$$

In this partial ordered logit model, α_i represents the boundaries of the categories where $i = 1, ..., 9$. $x_{[...]jt}$ stands for a specific independent variable (executed activity) of participant j at time t and $\beta_{[...]}$ are the parameters to be estimated for each predictor. The β-terms vary among the ranks for the variables that violate the proportional odds assumption which is indicated by the index i. Equation 2.1 does not account for any heterogeneity among the clients. The parameters that represent the relationships between the predictors and the dependent variable illustrate the general influences and do not consider any difference in behavior. This is the first version of the model (*Model 1*) we utilize for predicting the mood level of the participants.

Rossi et al. (2012), Farewell (1982), and Johnson (2003) discuss the aspect of "scale usage heterogeneity". Specifically, this term implies that participants often do not rank a given scale the same way but develop diverse response styles. This varying behavior can lead to a preferred usage of the scale (i.e. only using the middle part of the scale) and even to biased analyses (Rossi et al., 2012). By implementing client specific cutoffs into α_i, we seek to address the problem of a heterogeneous usage of the scale. Specifically, we sample α_i from a normal distribution. This procedure results in nine specific values that represent the cutoffs for the boundaries of the categories. We then sample user specific cutoffs based on the previously sampled values. This process is indicated by α_{ij} in the following equation:

$$\begin{aligned}\ln(\theta_{ijt}) = \alpha_{ij} - (&\beta_{social_i}\, x_{social_{jt}} + \beta_{work_i}\, x_{work_{jt}} + \beta_{recreational}\, x_{recreational_{jt}} \\ &+ \beta_{necessary_i}\, x_{necessary_{jt}} + \beta_{exercise_i}\, x_{exercise_{jt}} + \beta_{sickness_i}\, x_{sickness_{jt}} \\ &+ \beta_{sleep}\, x_{sleep_{jt}} + \beta_{rumination_i}\, x_{rumination_{jt}}).\end{aligned}$$

This model addresses "scale usage heterogeneity" (*Model 2*). We further hypothesize that not only differences in scale usage exist among the participants. Precisely, we assume varying effects of the items (activities) on different participants and we thus implement client specific β-parameter values to account for the differing influences each concept can have on an individual j. We sample these user specific values from a normal distribution as well. Therefore, *Model 3* is the modification that only accounts for the varying effects the predictors can have on an individual.

We also combine both alterations and thus include α and β heterogeneity terms (*Model 4*):

$$\begin{aligned}\ln(\theta_{ijt}) = \alpha_{ij} - (&\beta_{social_{ij}} x_{social_{jt}} + \beta_{work_{ij}} x_{work_{jt}} + \beta_{recreational_j} x_{recreational_{jt}} \\ &+ \beta_{necessary_{ij}} x_{necessary_{jt}} + \beta_{exercise_{ij}} x_{exercise_{jt}} + \beta_{sickness_{ij}} x_{sickness_{jt}} \\ &+ \beta_{sleep_j} x_{sleep_{jt}} + \beta_{rumination_{ij}} x_{rumination_{jt}}).\end{aligned}$$

For all modifications, the logits for every client j and every response in time t are calculated and then transferred into a probability. Thereupon, a specific outcome for the dependent variable (mood level) is sampled from a categorical distribution based on the individual probabilities. We realize the models in `R` and include JAGS (Just Another Gibbs Sampler) for MCMC sampling (Plummer, 2003). We implement three chains in JAGS to create three independent samples from the posterior distribution. We perform 40,000 iterations when running the MCMC algorithm and store every twentieth draw from the last 20,000 iterations for each of the three chains. In terms of convergence, all chains succeed for the reported variables.

2.4.1 Prior settings and model comparison

Based on current literature and our assumptions, we implement specific priors for the predictors. We decide to set a weak negative prior for the variables work related activities, necessary activities, sickness, and rumination. Moreover, we implement a weak positive prior for the variables social activities, recreational activities, and exercise. We also set an uninformative prior for the variable sleep related activities because we do not have further information or are deeply assured as to how this variable could influence the mood level. However, by setting a weak prior for the other predictors we allow for high variance. By doing so, we take previous knowledge, findings of related literature, and our assumptions into account but allow the data to have strong influence on the analysis.

Furthermore, we compare our developed models and attempt to obtain information about the necessary levels of heterogeneity among the participants; concurrently finding the model that has the greatest performance in predicting the test dataset. We start by estimating the parameters with the partial ordered logit model, which does not account for heterogeneity among the participants (Model 1). Consequently, we implement solely scale usage heterogeneity (Model 2; α-terms), only the influences of each psychological concept on an individual (Model 3; β-terms), and subsequently we estimate the parameters by implementing both heterogeneity terms (Model 4). We compare the models by using the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002). The DIC incorporates a measure of fit and a measure of model complexity (Berg et al., 2004). A smaller DIC value suggests a superior fit to the data. We choose the DIC for model comparison because it is especially suited for Bayesian models that are estimated by MCMC methods and it does not require additional Monte Carlo sampling (Berg et al., 2004). This method has been shown to perform adequately regarding a variety of examples (Berg et al., 2004; Spiegelhalter et al., 1998). According to Ando (2007) and Richardson (2002), however, the DIC

can be prone to select overfitted models. Thus, we predict the mood level of the individuals in the test dataset based on the varying models and each text mining approach and then utilize the Root Mean Square Error (RMSE) as well as the Mean Absolute Error (MAE) as performance indicators. In the following section, we present the results of the model comparison and of our analysis.

2.5 Results and discussion

We utilize free text diary data and categorize them into defined activity categories. First, we classify the free text by applying a BoW approach. The resulting dataset is then used as input for partial ordered logit models with different levels of heterogeneity among the participants. We then extend the BoW approach by utilizing RNN techniques that are trained on the already categorized data. This enables us to classify an additional set of free text. Consistently, we repeatedly use the resulting dataset as input for the models and compare model fit and predictive performance of the text mining procedures and statistical models. Table 2.1 illustrates the results of the DIC calculation:

Model	DIC (BoW)	DIC (RNN)
Model 1 (No Heterogeneity)	27717.15	33718.62
Model 2 (α Specific Term)	23644.13	28447.73
Model 3 (β Specific Term)	22775.33	27666.42
Model 4 (α Specific Term & β-Term)	22376.95	26950.70

Table 2.1: Model comparison with different levels of heterogeneity for each text mining approach, bag-of-words, and RNN.

The reason for an increase of the DIC of the RNN extension in comparison to the BoW approach is the number of observations in the different datasets. Therefore, we do not compare the DIC across the text mining approaches but among the varying statistical models. As we can see, the DIC value is highest for the model without any heterogeneity in both text mining approaches. This indicates a superior performance of models that account for differences among the clients and illustrates the importance of heterogeneity. Expecting every participant to behave the same way is not realistic and therefore it is important to account for differences among the individuals.

As expected, the model that includes α- and β-terms has the lowest DIC value. Even though the DIC penalizes the number of parameters in the model, the general fit of this model is better to an extent where the number of parameters are not affecting the performance of Model 4 compared to the others. Model 3 also appears to be better than Model 2. This might indicate that heterogeneity in the β-coefficients produces a better balance between model fit and model complexity. In order to verify these results and achieve an indicator for the predictive performance, we predict the individual mood values for each text mining approach and model.

For the comparison and execution of an out-of-sample test, we randomly extract mood entries and their corresponding activities from the data before training the model. We select at most one entry for each client, only from users who provide more than one observation, and - of course - only categorized activities. This random process results in 301 selected mood entries (680 sentences). We then predict the mood levels of each observation of the test set and compare the results. We also report performance measures for a so called *Mean Model*; here, we use the average mood level of the training set as predictions for the test dataset (in this case the number 6).

Measure	Model 1 BoW	Model 1 RNN	Model 2 BoW	Model 2 RNN	Model 3 BoW	Model 3 RNN	Model 4 BoW	Model 4 RNN	Mean Model
RMSE	2.32	2.33	1.98	1.98	1.87	1.91	1.81	1.86	1.91
MAE	1.78	1.82	1.48	1.49	1.41	1.41	1.37	1.37	1.53

Table 2.2: Prediction performance for each model and text mining approach.

As illustrated in Table 2.2, we can see that the *Mean Model* produces a greater predictive performance compared to the model without heterogeneity and even compared to the model including scale usage heterogeneity. However, when more heterogeneity terms are accounted for and the complexity of the model simultaneously increases, the prediction performance clearly grows. Table 2.2 also indicates that the usage of the RNN does not contribute but rather decreases the predictive performance compared to the BoW approach. This can potentially arise because the training data, which is based on the BoW approach, might not be accurate enough for the RNN to generate new knowledge and connections between the words and categories. Thus, the deep learning algorithm might only add noise to the prediction. Another reason for this finding could be that users often specify their activities as keywords - therefore, the RNN cannot contribute to the prediction. Model 4 for the BoW approach has the greatest prediction performance. Consequently, we choose this model for our analysis regarding the effects of the activity categories on the mood level.

Variable	50%	95% - CI
Work related activities	-0.15	(-0.78;0.47)
Recreational activities	0.54	(-0.57;1.61)
Necessary activities	-0.07	(-0.65;0.53)
Exercise	0.62	(-0.05;1.26)
Sickness	-4.92	**(-6.35;-3.50)**
Sleep related activities	-0.25	(-2.14;1.67)
Rumination	-5.47	**(-7.02;-3.96)**
Social activities	1.50	**(1.03;1.98)**

Table 2.3: Estimated model parameters (significant parameters in bold).

As illustrated in Table 2.3, we find that the category sickness has a strong negative and significant effect on mood. When being sick, it is a logical assumption and might even be natural to have a lower mood level. Furthermore, our analysis suggest that the category rumination

affects the mood level in a negative way. This can be due to the fact that individuals tend to think more about their problems and reflect on *bad* experiences rather than on *good* experiences. Therefore, negative events outweigh the positive in the stated ruminating activities. We further find that social activities have a significant positive effect on the mood level. This finding is consistent with previous research (Clark and Watson, 1988; David et al., 1997; Sonnentag, 2001; Weinstein and Mermelstein, 2007). Spending time and engaging with others, especially when they are of a general happy nature, might help people to cope with their problems. Another reason for this finding might be linked to the uplifting aspect of having some companionship, sharing a moment, and interacting with somebody else either in conversation or an activity; demonstrating the powerful force that comes with connecting. This can be literally giving a friend a ring and exchanging a few words, or calling friends to schedule an in-person meeting. The strong bonds of friendship can mean support and can provide feelings of enlivenment. Therefore, start browsing through your phone's contacts list - it might be time to call up your friends.

The results also indicate a tendency that physical activities influence the individual mood level positively - even though this result is barely insignificant. Previous literature in the field of psychology often reveals positive effects from exercise such as enhanced psychological well-being, reduced anger, and mood improvements in general (Byrne and Byrne, 1993; Kanning and Schlicht, 2010; Netz and Lidor, 2003; Yeung, 1996). However, we expected a stronger and significant influence from physical activities.

The rest of the predictors show insignificant results. Especially surprising are the results for the category recreational activities since we expected a strong positive and significant influence because activities that are directly chosen by an individual are beneficial; he/she would not have chosen that specific activity otherwise. The insignificance for necessary activities might be due to the fact that individuals perceive necessary activities differently. Some clients might enjoy grocery shopping whereas others do experience this activity as a chore. Furthermore, sleep related activities could be insignificant because individuals might not perceive a bad sleep experience as important enough to report. Therefore, a reason for the insignificance of the predictors is certainly related to the differences in behavior among the participants. Some individuals, for instance, might feel rewarded to a certain degree after they have finished up duties and therefore receive positive moods whereas others experience certain activities more as an ordeal. Thus, we emphasize the importance of individual preferences and heterogeneity by implementing parameters for every participant in the model and demonstrate how the performance level of the utilized models, indicated by the prediction performance, increases the more heterogeneity is implemented. Additionally, some of our findings are consistent with our hypotheses. We can confirm that daily activities, especially social contacts, rumination, and sickness, do influence the mood level of individuals, which is consistent with literature in the field of psychology (Weinstein and Mermelstein, 2007).

2.6 Limitations and conclusion

We analyze the effects of daily activities on the individual mood level, predict the mood of the participants, and simultaneously compare a BoW text mining approach including an extension of this method by coupling BoW and RNN. Furthermore, we evaluate statistical models with different levels of heterogeneity among the clients. We do so by developing varying partial ordered logit models and employing MCMC techniques for parameter estimation. Thus, we emphasize the importance of heterogeneity and seek to foreshadow how analyses and their prediction performance can be improved by considering individual behavior. Furthermore, our results support the development of treatment by focusing on factors that negatively influence the mood level, for example sickness and rumination, and concurrently emphasize and reinforce social activities. Gaining deeper insight into the relationship between certain activities and mood offers the opportunity to achieve greater therapeutic success and can additionally provide an indication for individuals who suffer from mood changes and depression. Therefore, the developed model can serve as a decision support tool for the treating therapist in order to enhance the well-being of the individual, improve the quality of therapy strategies as well as the general therapy outcome. The therapist can then make enhanced decisions of *when* and *how* to intervene. Thus, our method can potentially be utilized by researchers and practitioners to develop and extend decision support systems for therapeutic interventions in healthcare.

Besides the implications and insight our analysis provides, we also outline some limitations regarding our research approach. The developed text mining algorithm, for example, does not classify all reported text fields to the corresponding activity category correctly. Certainly, it is not simple to classify all text fields accurately because they are often hard to assign in terms of ambiguity. Besides that, the definition of our categories can be questioned. Our exercise category, for instance, represents all physical activities. We do not distinguish between type of exercise, type of sickness, or type of leisure activity. Individuals might also perceive necessary activities differently whereas some individuals might consider cooking as recreational activity and others as necessary. This fact can also lead to insignificant results. Thus, more precise definitions and therefore more categories might result in more accurate outcomes, predictions, implications, and insight. Moreover, although our analyzed dataset is comparatively large, we only possess self-reported and optional data from clients. Even though ecological momentary data is perfectly suited for gathering information on experiences, the data is not reported objectively. Developing more accurate "psychological and biological" measures can further enhance analyses and make results more representative (Cropley and Zijlstra, 2011).

Evidently, further room for improvement and more analyses exists. An implementation of additional categories and other factors such as dropout information or other psychological concepts can further improve our analyses. Developing an enhanced text mining approach can potentially increase prediction performance and at the same time provide a more accurate support system. In the future, we seek to implement such factors, create other techniques to categorize

text fields, develop more statistical models to gain deeper insight into the clients' behavior, reveal relationships between psychological concepts in order to support and help clients in need, and simultaneously make an attempt to provide guidance for more personalized interventions and an increased therapy success.

References

Agarwal, R. and Dhar, V. (2014). Big Data , Data Science , and Analytics : The Opportunity and Challenge for IS Research. *Information Systems Research*, 25(3):443–448.

Ando, T. (2007). Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika*, 94(2):443–458.

Balog, K., Mishne, G., and de Rijke, M. (2006). Why Are They Excited? Identifying and Explaining Spikes in Blog Mood Levels. In *EACL '06 Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics*, pages 207–2010, Stroudsburg, PA (USA).

Becker, D., Bremer, V., Funk, B., Asselbergs, J., Riper, H., and Ruwaard, J. (2016). How to Predict Mood? Delving into Features of Smartphone-Based Data. In *Twenty-second Americas Conference on Information Systems*, San Diego (USA).

Berg, A., Meyer, R., and Yu, J. (2004). Deviance Information Criterion for Comparing Stochastic Volatility Models. *Journal of Business & Economic Statistics*, 22(1):107–120.

Bolger, N., DeLongis, a., Kessler, R. C., and Schilling, E. a. (1989). Effects of daily stress on negative mood. *Journal of personality and social psychology*, 57(5):808–818.

Both, F., Hoogendoorn, M., Klein, M., and Treur, J. (2008). Modeling the Dynamics of Mood and Depression. In *18th European Conf. on Artificial Intelligence*, Patras (Greece).

Bright, T. J., Wong, A., Dhurjati, R., Bristow, E., Bastian, L., Coeytaux, R. R., Samsa, G., Hasselblad, V., Williams, J. W., Musty, M. D., Wing, L., Kendrick, A. S., Sanders, G. D., and Lobach, D. (2012). Effect of Clinical Decision-Support Systems: A Systematic Review. *Ann Intern Med*, 157(1):29–43.

Buntrock, C., Ebert, D. D., Lehr, D., Cuijpers, P., Riper, H., Smit, F., and Berking, M. (2014). Evaluating the efficacy and cost-effectiveness of web-based indicated prevention of major depression: design of a randomised controlled trial. *BMC psychiatry*, 14:25–34.

Byrne, A. and Byrne, D. (1993). The effect of exercise on depression, anxiety and other mood states: A review. *Journal of Psychosomatic Research*, 37(6):565–574.

Caspersen, C. J., Powell, K. E., and Christenson, G. M. (1985). Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research. *Public Health Rep*, 100(2):126–131.

Chih, M.-Y., Patton, T., McTavish, F., Isham, A., Judkins-Fisher, C., Atwood, A., and Gustafson, D. (2014). Predictive modeling of addiction lapses in a mobile health application. *Journal of Substance Abuse Treatment*, 46(1):29–35.

Christensen, R. H. B. (2015). Ordinal - Regression Models for Ordinal Data.

Clark, L. and Watson, D. (1988). Mood and the mundane: Relations between daily life events and self-reported mood. *Journal of personality and social psychology*, 54:296–308.

Cropley, M. and Zijlstra, F. (2011). Work and rumination. In Langan-Fox, J. and Cooper, C. L., editors, *Handbook of stress in the occupations*, chapter 24, pages 487–499. Edward Elgar Publishing, Cheltenham, UK.

David, J., Green, P., Martin, R., and Suls, J. (1997). Differential roles of neuroticism, extraversion, and event desirability for mood in daily life: An integrative model oftop-down and bottom-up influences. *Journal of personality and social psychology*, 73:149–159.

Demerouti, E., Bakker, A. B., Sonnentag, S., and Fullagar, C. J. (2012). Work-related flow and energy at work and at home: A study on the role of daily recovery. *Journal of Organizational Behavior*, 33(2):276–295.

Dinges, D. F., Pack, F., Williams, K., Gillen, K. a., Powell, J. W., Ott, G. E., Aptowicz, C., and Pack, a. I. (1997). Cumulative sleepiness, mood disturbance, and psychomotor vigilance performance decrements during a week of sleep restricted to 4-5 hours per night. *Sleep*, 20(4):267–277.

Donaldson, C. and Lam, D. (2004). Rumination, mood and social problem-solving in major depression. *Psychol Med*, 34(7):1309–1318.

Ebert, D. D., Lehr, D., Baumeister, H., Boß, L., Riper, H., Cuijpers, P., Reins, J. A., Buntrock, C., and Berking, M. (2014). GET.ON Mood Enhancer: efficacy of Internet-based guided self-help compared to psychoeducation for depression: an investigator-blinded randomised controlled trial. *Trials*, 15(1):39.

Elman, J. L. (1990). Finding Structure in Time. *Cognitive Science*, 14(2):179–211.

Eysenbach, G. (2001). What is e-health? *Journal of Medical Internet Research*, 3(2):e20.

Farewell, V. T. (1982). A note on regression analysis of ordinal data with variability of classification. *Biometrika*, 69(3):533–538.

Feinerer, I., Hornik, K., and Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software*, 25(5):1–54.

Gable, S. L., Reis, H. T., and Elliot, a. J. (2000). Behavioral activation and inhibition in everyday life. *Journal of personality and social psychology*, 78(6):1135–1149.

Garg, A., Adhikari, N., McDonalds, H., Rosas-Arellano, M., Devereaux, P., Beyene, J., Sam, J., and Haynes, R. (2005). Effects of Computerized Clinical Decision Support Systems on Practitioner Performance and Patient Outcomes. *Journal of the American Medical Association*, 293(10):1223–1238.

Grosscup, S. J. and Lewinsohn, P. M. (1980). Unpleasant and pleasant events, and mood. *J Clin Psychol*, 36(1):252–259.

Gustavsson, A., Svensson, M., Jacobi, F., Allgulander, C., Alonso, J., Beghi, E., Dodel, R., Ekman, M., Faravelli, C., Fratiglioni, L., Gannon, B., Jones, D. H., Jennum, P., Jordanova, A., Jönsson, L., Karampampa, K., Knapp, M., Kobelt, G., Kurth, T., Lieb, R., Linde, M., Ljungcrantz, C., Maercker, A., Melin, B., Moscarelli, M., Musayev, A., Norwood, F., Preisig, M., Pugliatti, M., Rehm, J., Salvador-Carulla, L., Schlehofer, B., Simon, R., Steinhausen, H. C., Stovner, L. J., Vallat, J. M., den Bergh, P. V., van Os, J., Vos, P., Xu, W., Wittchen, H. U., Jönsson, B., and Olesen, J. (2011). Cost of disorders of the brain in Europe 2010. *European Neuropsychopharmacology*, 21(10):718–779.

Hah, H. and Bharadwaj, A. (2012). A Multi-level Analysis of the Impact of Health Information Technology on Hospital Performance. In *Thirty Third International Conference on Information Systems*, Orlando (USA).

Hannan, C., Lambert, M. J., Harmon, C., Nielsen, S. L., Smart, D. W., Shimokawa, K., and Sutton, S. W. (2005). A lab test and algorithms for identifying clients at risk for treatment failure. *Journal of Clinical Psychology*, 61(2):155–163.

Hippisley-Cox, J., Coupland, C., Vinogradova, Y., Robson, J., Minhas, R., Sheikh, A., and Brindle, P. (2008). Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2. *BMJ*, 336(7659):1475–1482.

Hornik, K. (2016). NLP: Natural Language Processing Infrastructure.

Iida, M., Shrout, P. E., Laurenceau, J.-P., and Bolger, N. (2012). Using diary methods in psychological research. In Cooper, H., Camic, P. M., Long, D. L., Panter, A. T., Rindskopf, D., and Sher, K. J., editors, *APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics.*, pages 277–305. American Psychological Association, Washington.

Ingram, R. E. and Smith, T. W. (1984). Depression and internal versus external focus of attention. *Cognitive Therapy and Research*, 8(2):139–151.

Isen, A. M., Daubman, K. a., and Nowicki, G. P. (1987). Positive affect facilitates creative problem solving. *Journal of Personality and Social Psychology*, 52(6):1122–1131.

Jardim, S. (2013). The Electronic Health Record and its Contribution to Healthcare Information Systems Interoperability. *Procedia Technology*, 9:940–948.

Johnson, T. R. (2003). On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. *Psychometrika*, 68(4):563–583.

Kanning, M. and Schlicht, W. (2010). Be active and become happy: an ecological momentary assessment of physical activity and mood. *Journal of sport & exercise psychology*, 32(2):253–261.

Kombrink, S., Mikolov, T., Karafiát, M., and Burget, L. (2011). Recurrent Neural Network Based Language Modeling in Meeting Recognition. *12th Annual Conference of the International Speech Communication Association*, 2877-2880.

Kramer, A. D. I., Guillory, J. E., and Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 111(24):8788–8790.

Landis, J. R. and Koch, G. G. (2008). The Measurement of Observer Agreement for Categorical Data Published by : International Biometric Society Stable URL : http://www.jstor.org/stable/2529310. *Society*, 33(1):159–174.

Larsen, R. J. and Cowan, G. S. (1988). Internal focus of attention and depression: A study of daily experience. *Motivation and Emotion*, 12(3):237–249.

Leger, D. (1994). The cost of sleep-related accidents: a report for the National Commission on Sleep Disorders Research. *Sleep*, 17(1):84–93.

Lewinsohn, P. M. and Amenson, C. S. (1978). Some relations between pleasant and unpleasant mood-related events and depression. *Journal of abnormal psychology*, 87(6):644–654.

Liu, X. and Koirala, H. (2012). Ordinal regression analysis: Using generalized ordinal logistic regression models to estimate educational data. *Journal of Modern Applied Statistical Methods*, 11(1):242–254.

McCorkle, R. and Quint-Benoliel, J. (1983). Symptom distress, current concerns and mood disturbance after diagnosis of life-threatening disease. *Social science & medicine*, 17(7):431–438.

McCullagh, P. (1980). Regression Models for Ordinal Data. *Journal of the Royal Statistical Society*, 42(2):109–142.

McMahon, F. J. (2014). Prediction of treatment outcomes in psychiatry-where do we stand? *Dialogues in Clinical Neuroscience*, 16(4):455–464.

Minden, S. L. (2000). Mood disorders in multiple sclerosis: diagnosis and treatment. *Journal of neurovirology*, 6(2):160–167.

Nadler, R. T., Rabi, R., and Minda, J. P. (2010). Better mood and better performance. Learning rule-described categories is enhanced by positive mood. *Psychological science : a journal of the American Psychological Society / APS*, 21(12):1770–1776.

Nesse, R. M. (2000). Is depression an adaptation? *Archives of general psychiatry*, 57(1):14–20.

Netz, Y. and Lidor, R. (2003). Mood alterations in mindful versus aerobic exercise modes. *The Journal of psychology*, 137(5):405–419.

Nolen-Hoeksema, S. and Morrow, J. (1993). Effects of rumination and distraction on naturally occurring depressed mood. *Cognition & Emotion*, 7(6):561–570.

Norusis, M. J. (2010). Ordinal Regression. In *PASW Statistics 18.0 Advanced Statistical Procedures Companion*, chapter 4, pages 69–89. Prentice Hall.

Penninx, B., Beekman, A., Honig, A., Deeg, D., Schoevers, R., van Eijk, J., and van Tilburg, W. (2001). Depression and cardiac mortality: results from a community-based longitudinal study. *Arch Gen Psychiatry*, 58(3):221–227.

Peterson, B. and Harrell, F. E. (1990). Partial Proportional Odds Models for Ordinal Response Variables. *Journal of the Royal Statistical Society*, 39(2):205–217.

Plummer, M. (2003). {JAGS}: A program for analysis of {Bayesian} graphical models using {Gibbs} sampling. In *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*.

Qian, X., Yarnal, C., and Almeida, D. M. (2014). Does leisure time moderate or mediate the effect of daily stress on positive affect? An examination using eight-day diary data. *Journal of Leisure Research*, 46(1):106–124.

Reis, H. T., Sheldon, K. M., Gable, S. L., Roscoe, J., and Ryan, R. M. (2000). Daily Well-Being: The Role of Autonomy, Competence, and Relatedness. *Personality and Social Psychology Bulletin*, 26:419–435.

Richardson, S. (2002). Discussion of a paper by D. J. Spiegelhalter et al. *J.R. Statist. Soc.*, B(64):626–7.

Robinson, R. G., Kubos, K. L., Starr, L. B., Rao, K., and Price, T. R. (1984). Mood disorders in stroke patients. Importance of location of lesion. *Brain : a journal of neurology*, 107 (Pt 1):81–93.

Rook, J. W. and Zijlstra, F. R. H. (2006). The contribution of various types of activities to recovery. *European Journal of Work and Organizational Psychology*, 15(2):218–240.

Rosen, I. M., Gimotty, P. a., Shea, J. a., and Bellini, L. M. (2006). Evolution of sleep quantity, sleep deprivation, mood disturbances, empathy, and burnout among interns. *Academic medicine*, 81(1):82–85.

Rossi, P. E., Allenby, G. M., and McCulloch, R. (2012). *Bayesian Statistics and Marketing*. Wiley Series in Probability and Statistics. Wiley.

Sluiter, J. K., de Croon, E. M., Meijman, T. F., and Frings-Dresen, M. H. W. (2003). Need for recovery from work related fatigue and its role in the development and prediction of subjective health complaints. *Occupational and environmental medicine*, 60(Suppl 1):i62–i70.

Smyth, J. M. and Stone, A. a. (2003). Ecological momentary assessment research in behavioral medicine. *Journal of Happiness Studies*, 4(1):35–52.

Sonnentag, S. (2001). Work, recovery activities and well-being: a diary study. *J Occup Health Psychol*, 6(3):196–210.

Spiegelhalter, D., Best, N. G., and Carlin, B. P. (1998). Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. Technical Report , MRC Biostastatistics Unit, Cambridge.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 64(4):583–616.

Stewart, W. and Barling, J. (1996). Daily work stress, mood and interpersonal job performance: A mediational model. *Work & Stress*, 10(4):336–351.

Stone, A. (1987). Event content in a daily survey is differentially associated with concurrent mood. *Journal of personality and social psychology*, 52:56–58.

Tadic, M., Oerlemans, W. G. M., Bakker, A. B., and Veenhoven, R. (2013). Daily Activities and Happiness in Later Life: The Role of Work Status. *Journal of Happiness Studies*, 14(5):1507–1527.

Thomsen, D. K., Yung Mehlsen, M., Christensen, S., and Zachariae, R. (2003). Rumination—relationship with negative mood and sleep quality. *Personality and Individual Differences*, 34(7):1293–1301.

Wang, F., Orpana, H. M., Morrison, H., De Groh, M., Dai, S., and Luo, W. (2012). Long-term association between leisure-time physical activity and changes in happiness: Analysis of the prospective National Population Health Survey. *American Journal of Epidemiology*, 176(12):1095–1100.

Watkins, E. and Baracaia, S. (2001). Why do people ruminate in dysphoric moods? *Personality and Individual Differences*, 30:723–734.

Weinstein, S. M. and Mermelstein, R. (2007). Relations between daily activities and adolescent mood: the role of autonomy. *Journal of Clinical Child and Adolescent Psychology*, 36(2):182–194.

Wilson, V. and Lankton, N. (2004). Modeling Patients' Acceptance of Provider-delivered E-health. *J Am Med Inform Assoc*, 11(4):241–248.

Wittchen, H. U., Jacobi, F., Rehm, J., Gustavsson, a., Svensson, M., Jönsson, B., Olesen, J., Allgulander, C., Alonso, J., Faravelli, C., Fratiglioni, L., Jennum, P., Lieb, R., Maercker, a., van Os, J., Preisig, M., Salvador-Carulla, L., Simon, R., and Steinhausen, H. C. (2011). The size and burden of mental disorders and other disorders of the brain in Europe 2010. *European Neuropsychopharmacology*, 21(9):655–679.

Yeung, R. R. (1996). Review the Acute Effects of Exercise on Mood State. *Journal of psychosomatic research*, 40(2):123–141.

CHAPTER 3

PREDICTING THERAPY SUCCESS AND COSTS FOR PERSONALIZED TREATMENT RECOMMENDATIONS USING BASELINE CHARACTERISTICS: DATA-DRIVEN ANALYSIS

Bremer, V., Becker, D., Kolovos, S., Funk, B., Van Breda, W., Hoogendoorn, M., and Riper, H. (2018). Predicting Therapy Success and Costs for Personalized Treatment Recommendations Using Baseline Characteristics: Data-Driven Analysis. Journal of Medical Internet Research, 20(8):e10275.

Abstract

Background: Different treatment alternatives exist for psychological disorders. Both clinical and cost effectiveness of treatment are crucial aspects for policy makers, therapists, and patients and thus play major roles for healthcare decision-making. At the start of an intervention, it is often not clear which specific individuals benefit most from a particular intervention alternative or how costs will be distributed on an individual patient level.

Objective: This study aimed at predicting the individual outcome and costs for patients before the start of an internet-based intervention. Based on these predictions, individualized treatment recommendations can be provided. Thus, we expand the discussion of personalized treatment recommendation.

Methods: Outcomes and costs were predicted based on baseline data of 350 patients from a two-arm randomized controlled trial that compared treatment as usual and blended therapy for depressive disorders. For this purpose, we evaluated various machine learning techniques, compared the predictive accuracy of these techniques, and revealed features that contributed most to the prediction performance. We then combined these predictions and utilized an incremental

cost-effectiveness ratio in order to derive individual treatment recommendations before the start of treatment.

Results: Predicting clinical outcomes and costs is a challenging task that comes with high uncertainty when only utilizing baseline information. However, we were able to generate predictions that were more accurate than a predefined reference measure in the shape of mean outcome and cost values. Questionnaires that include anxiety or depression items and questions regarding the mobility of individuals and their energy levels contributed to the prediction performance. We then described how patients can be individually allocated to the most appropriate treatment type. For an incremental cost-effectiveness threshold of 25,000 €/quality-adjusted life year, we demonstrated that our recommendations would have led to slightly worse outcomes (1.98%), but with decreased cost (5.42%).

Conclusions: Our results indicate that it was feasible to provide personalized treatment recommendations at baseline and thus allocate patients to the most beneficial treatment type. This could potentially lead to improved decision-making, better outcomes for individuals, and reduced health care costs.

3.1 Introduction

In a clinical context, different forms of behavioral interventions such as face-to-face or internet-based treatments exist for patients with depressive disorders. Clinical and cost effectiveness studies provide important knowledge regarding these treatment alternatives (Ryder et al., 2009). However, questions remain as to which particular individuals prefer particular treatment types or receive an increased benefit from one specific treatment option over another, especially before the treatment begins. Therapists or other clinicians often make decisions based on personal understanding and experience, leading to high uncertainty or nonoptimal decisions (Ryder et al., 2009). This uncertainty can potentially result in worse treatment outcomes for individuals and increased health care costs. Simultaneously, policy makers and stakeholders increasingly demand cost-effectiveness evidence in order to support their conclusions and decisions (Knapp, 1999).

For supporting these admittedly difficult and complex decisions, approaches exist based on cost analysis or decision analysis (Ryder et al., 2009; Van Breda et al., 2016). The incremental cost-effectiveness ratio (ICER) is a widespread indicator for cost effectiveness (Russell et al., 1996). The goal is to support the mentioned decisions by identifying actions that, on average, maximize a specific result (Ryder et al., 2009) such as quality-adjusted life years (QALYs). The ICER is applied on a population level, which means that average values of costs and outcomes are considered for population-level decisions (Ryder et al., 2009; Sculpher, 2015). This procedure does not consider any heterogeneity among individuals regarding outcomes and costs. Individual patients, for example, respond differently to treatment and have varying mindsets regarding

risks (Ioannidis and Garber, 2011; Kravitz et al., 2004). Thus, the average outcomes and costs often do not necessarily represent the best decision for an individual (Ioannidis and Garber, 2011). Even though these aspects are well known, cost-effectiveness analyses based on average values are still widely used (Ioannidis and Garber, 2011).

Predictive analyses can provide crucial insight into aspects that influence outcomes and costs of interventions and can be beneficial for patients as well as society (Jones et al., 2007). Research that seeks to forecast outcomes for patients with depression already exists. One study, for example, predicted treatment success in the domain of depression and showed that baseline data has predictive power in this context (van Breda et al., 2018). Another study predicted treatment outcomes of treatment-resistant patients with depression and thereby revealed important predictors such as severity and suicidal risk, among others (Kautzky et al., 2018). These types of statistical procedures can ultimately result in the development of decision support systems in the context of health interventions. In the field of depression treatment, these systems often lead to positive effects and even a reduction of symptoms in various situations (Triñanes et al., 2015).

This study focused on making personalized treatment recommendations. For this purpose, we predicted the outcomes and costs for different treatment types, at baseline, on an individual patient level. We applied various machine learning techniques, evaluated them based on their predictive performance, and revealed important features that contributed to the prediction. In order to derive personalized treatment recommendations, we applied an individualized cost-effectiveness analysis based on the ICER. Unlike its traditional utilization based on the ratio of average values, we used individual predictions for each treatment type and its alternative. The predictions and their generated information can provide additional knowledge and enable practitioners, as well as researchers, to individually assign patients at baseline to their most appropriate treatment type in terms of outcomes and costs. This approach is applied to data from an internet-based two-arm randomized controlled trial in the domain of depression.

The forecast of individual outcomes and costs is one of the most important aims in clinical research (Dunlop, 2015), and personalized analyses and illustrations of cost effectiveness in this context are of increased interest and need (Ioannidis and Garber, 2011; O'Hagan and Stevens, 2002). Thus, we contribute to existing research by attempting to predict these factors at the start of treatment for each individual and by further proposing a conceptual approach for treatment recommendations, as applied to empirical data.

3.2 Methods

3.2.1 Data and preprocessing

The data we utilized originate from the European Union-funded project E-Compared in which the clinical and cost effectiveness of blended treatment (BT) for depression, where internet-based

and face-to-face treatments are combined in one integrated treatment protocol, is evaluated and compared with treatment as usual (TAU) in 9 different countries (Kleiboer et al., 2016). Participants were aged 18 years or older, met criteria for a major depressive disorder, were not of high suicidal risk, were not being treated for depression, and had access to an internet connection. Table 3.1 illustrates the different questionnaires used in the study.

Data	Description
1 Demographic data	N/A
2 Current treatment	Current treatment type, medication, provider
3 MINI International Neuropsychiatric Interview	Structured clinical interview for making diagnoses
4 Quick Inventory of Depressive Symptomatology (16-Item) (Self-Report)	Quick Inventory of Depressive Symptomatology
5 Patient Health Questionnaire-9	Questions regarding depressive symptoms
6 5-level EQ-5D	EuroQol questionnaire; measuring generic health status; for calculation of quality-adjusted life years
7 Costs Associated with Psychiatric Illness	Measurement of healthcare costs and productivity losses
8 Treatment preferences	Individual preferences for blended treatment or treatment as usual

Table 3.1: Data utilized in this study.

The data consisted of individualized information regarding depressive symptoms, medical costs, and other factors. These questionnaires are widely utilized and known and can be found elsewhere (EuroQolGroup, 1990; Hakkaart-van Roijen et al., 2002; Kleiboer et al., 2016; Kroenke et al., 2001; Rush et al., 2003). The data in the E-Compared project were collected multiple times during the trial: at baseline, 3 months, 6 months, and 12 months. Questionnaires 3, 4, 6, and 7 (according to Table 1) were also available, not only at baseline but also after other times during data acquisition. Because we were interested in recommendations before the start of the actual treatment, we solely used the baseline information as features in this study.

We used QALY as an outcome, as measured by the EuroQol questionnaire (5-level EQ-5D version). Utility weights were calculated using the Dutch tariffs (Versteegh et al., 2016). These weights are a preference-based measure of quality of life anchored at 0 (worst perceivable health) and 1 (perfect health). QALYs were calculated by multiplying the utility weights with the amount of time a participant spent in a particular health state. Transitions between the health states were linearly interpolated. The costs that we aimed to forecast were measured from the societal

perspective (including healthcare utilization and productivity losses) based on the adapted version of the Trimbos and Institute for Medical Technology Assessment questionnaires on Costs Associated with Psychiatric Illness (Hakkaart-van Roijen et al., 2002). Dutch unit costs were used to value healthcare utilization and productivity losses (Hakkaart-van Roijen et al., 2015). Costs for the online part of BT included maintenance and hosting of the treatment and costs that occurred for a therapist to provide feedback to participants. We decided to use costs from a societal perspective because they represent interests of society and all other stakeholder groups (Ryder et al., 2009). More information on the calculation of the costs can be found elsewhere (Kolovos et al., 2016). As dependent variables, we utilized QALY and costs that appear after a 6-month period. This allowed for more observations compared with the data at 12 months (350 patients vs 212 patients) because not all patients had already finished the treatment process. Because we focused on the outcome data up to 6 months, QALY could have a maximum value of 0.5 in our analysis.

During the data preprocessing phase, we merged all mentioned data from Table 3.1. This process led to 309 features that could be utilized for the prediction. We then calculated the costs and QALY for each individual. We only included patients for which both dependent variables were not missing. By splitting the dataset into groups for the different treatment types (TAU and BT), some factor levels of an item or feature can go missing. We removed 97 features that had just one level or were missing. Table 3.6 (Appendix A) lists the omitted items from the questionnaires. The resulting dataset still contained 29,568 missing values. Disregarding these values, and thus deleting them, would lead to a substantial decrease in observations. We therefore utilized two different methods for handling them in order to evaluate which method would perform better regarding the predictive performance. We first imputed the numeric values by sampling from a normal distribution based on the mean value and SD of the corresponding feature. We imputed the categorical predictors by sampling from the categorical distribution of those features. As a second approach, we imputed the missing values by the median (numeric variable) and mode (categorical variable). Finally, we ended up with a dataset of 350 observations (1 for each patient) and 212 features. In the following, we have reported only the results for the latter imputation procedure. In Table 3.7 and Table 3.8 (Appendix B-C), we have also demonstrated the final performances for the first imputation method. However, we decided to utilize the latter method because it led to the best performance in terms of prediction.

3.2.2 Approach and statistical analysis

In order to derive individual treatment recommendations, we utilized the baseline features as input for predicting individual level outcome and costs based on the treatment type, as seen in Figure 3.1. We applied various machine learning techniques to evaluate which yielded the highest prediction performance. As mentioned by several studies, it is beneficial to compare different statistical procedures in order to eventually find the most precise model, especially when

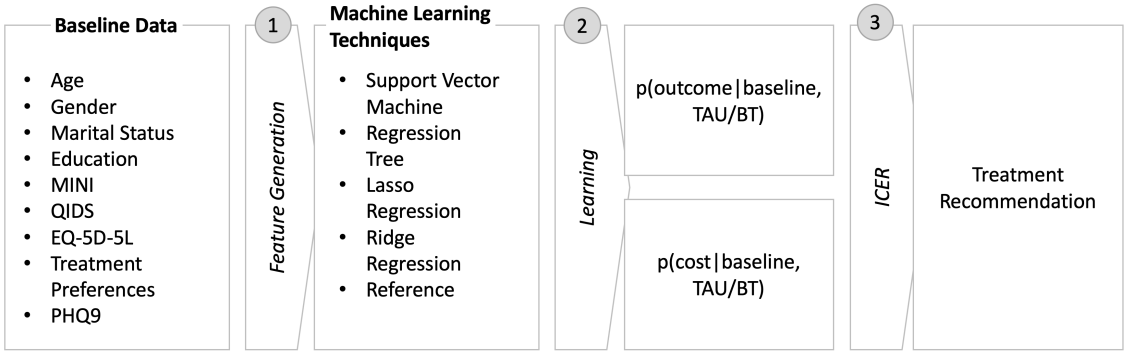


Figure 3.1: Process for deriving treatment recommendations for individuals (BT: blended treatment; ICER: incremental cost-effectiveness ratio; TAU: treatment as usual).

predicting costs due to the challenging nature of this activity (Diehr et al., 1999; Jones et al., 2007; Manning and Mullahy, 2001). Because the data consist of numerous features, we applied a feature selection method to reveal variables that contributed to the prediction performance. To demonstrate how the forecasts can be beneficial in recommending treatment types on an individual patient level, we applied the ICER to the predictions.

Specifically, we estimated the conditional probability $p(o, c \mid b, tt)$ for each treatment type, where o is the outcome, c is the costs, b reflects the baseline features, and tt is one of the two treatment types. Given the limited amount of data, we assumed that the conditional probability could be factorized as follows: $p(o, c \mid b, tt) = p(o \mid b, tt)p(c \mid b, tt)$.

For the prediction of outcome and costs, we used linear regression and support vector regression (SVR). The latter method has shown good predictive capabilities in various fields (Burges, 1998). We further utilized regression trees and ridge regression. For finding the optimal parameters, we applied a grid-based search and cross-validation. Additionally, we defined the mean of all outcomes or costs as a reference measure. If unable to achieve a better prediction performance compared with the reference measure, it is questionable if the application of more advanced statistical methods is appropriate in this context. For finding the model that achieves the highest prediction performance, we used leave-one-out cross-validation. That is, one observation is utilized as the test set and the remaining observations are used for training the model. This procedure is repeated for every single observation in the dataset. The error measures we used were root mean square error (RMSE) and mean absolute error (MAE). We have presented both error measures because debate exists as to which measure is more appropriate for the demonstration of predictive performance (Chai and Draxler, 2014; Willmott et al., 2009).

When utilizing a vast number of features, overfitting presumably occurs. Thus, we used Lasso regression to select features that contributed to the predictive performance. Lasso is a linear regression that introduces a penalty term called regularizer (Tibshirani, 1996). The error function of the regression, which is to be optimized, consists of the mean square error of the misclassified

samples and a term that penalizes the absolute value of the sum of regression coefficients. This linear penalty enforces useless coefficients to shrink toward zero in order to produce a sparse solution. The corresponding optimization problem is illustrated below, where X is the baseline feature, Y is the outcome or costs, and β is the coefficient:

$$\min_{\beta} \quad \{\|Y - X\beta\|^2 + \lambda\|\beta\|_1\}.$$

The parameter λ influences the strength of the penalty. Specifically, the higher the value of λ, the higher the penalty. A higher penalty leads to sparser solutions (more coefficients are shrunk to zero). The optimal λ's are found by utilizing cross-validation. After obtaining the specific features that appear to add to the predictive accuracy, we again predicted the outcome values and costs based on the aforementioned machine learning techniques. This time, however, we only utilized the features that were identified by the Lasso regression. Finally, we selected the algorithm that produced the smallest error and therefore performed best for the outcome and cost predictions. Based on these individual predictions, we calculated the ICER, as seen in the equation:

$$\text{ICER} = \frac{(\text{Cost}_{\text{BT}} - \text{Cost}_{\text{TAU}})}{(\text{Outcome}_{\text{BT}} - \text{Outcome}_{\text{TAU}})}.$$

The ICER was then visualized in the cost-effectiveness plane (Black, 1990). By predicting the costs and outcomes at baseline and utilizing the ICER, we could then make recommendations about individual patient allocation. We implemented the mentioned models and processes in R (R Core Team, 2018).

3.3 Results

3.3.1 Overall findings

Before we focused on the outcome and cost predictions, we illustrated the general improvements of the patients for TAU and BT. The E-Compared project hypothesized noninferiority between both treatment types (ie, BT is not less effective) (Kleiboer et al., 2016). Improvement was defined as the difference of the start and end value of the cumulated PHQ9 values. The PHQ9 questionnaire is a reliable measure for depression severity (Kroenke et al., 2001). Because we only investigated the improvements for a 6-month period, these results are not final; however, they can indicate a trend. Table 3.2 shows that the mean baseline score for PHQ9 was 15.35 for BT and 15.42 for TAU. At the 6-month measurement, the scores were 7.85 and 9.49, respectively. Furthermore, 154 patients in the BT group and 140 patients in the TAU group showed improvement. Therefore, we can see that the PHQ9 value decreased more strongly for BT and that the number of improvements for BT exceeded the outcome of TAU. Applying a t test for the comparison of the mean end values resulted in the rejection of the hypothesis that both samples had the same mean ($P = .006$).

	TAU	Blended
Start Patient Health Questionnaire-9 (mean)	15.42	15.35
End Patient Health Questionnaire-9 (mean)	9.49	7.85
Number of patients with improvement	140	154
Number of patients without improvement	38	18

Table 3.2: Mean of Patient Health Questionnaire-9 scores at baseline and end for treatment as usual and blended treatment as well as the numbers of patients in each condition that improved (N=350).

3.3.2 Outcome and cost prediction

Table 3.3 illustrates the prediction performance for all utilized machine learning techniques and all baseline features. Overall, the SVR and regression tree had the smallest errors for performance measures. The ridge regression also performed better than the reference measure. Based on a Wilcoxon test, MAEs differed significantly ($SVR: P_O = .030, P_C < .001; Tree: P_O = .001, P_C < .001; Ridge: P_O = .049, P_C < .023$). Since we had more features than observations, we did not apply ordinary least squares regression when utilizing all baseline features.

	Algorithm	MAE_O	$RMSE_O$	MAE_C	$RMSE_C$
1	SVR	0.0714	0.0997	6299.63	9360.50
2	Regression Tree	0.0698	0.0992	6573.94	9406.11
3	Ridge Regression	0.0711	0.1000	6557.69	9187.78
4	Reference Measure	0.0770	0.1017	7024.11	9539.54

Table 3.3: Results for prediction performance based on all baseline features for varying machine learning approaches (MAE: mean absolute error; RMSE: root mean square error).

We then performed Lasso regression in order to select the important features that contributed to the prediction performance. Table 3.9, Table 3.10, Table 3.11, and Table 3.12 (Appendix D-G) show the important features that were utilized and their corresponding coefficient. By applying cross-validation, we chose specific λ values that minimized the mean cross-validated error. For TAU and BT, we used all features up to a λ value of 0.01485 and 0.01479, respectively (433.83 and 651.14 for the cost prediction).

Multiple features appeared repeatedly. Various questions regarding the medication use and the amount of consultations of some kind of therapist, practitioner, or treatment program occurred most often (24 and 16 times, respectively). Furthermore, the anxiety or depression items (6 times), mobility (5 times), origin of the patient (7 times), and energy level questions (4 times) appeared to have an influence on the prediction performance. Using the selected features, we then repeatedly applied the above specified statistical methods in order to achieve a better accuracy.

	Algorithm	MAE_O	$RMSE_O$	MAE_C	$RMSE_C$
1	SVR	0.0575	0.0812	5164.22	8026.46
2	Regression	0.0590	0.0793	6436.63	15319.89
3	Regression Tree	0.0684	0.0952	6573.94	9406.11
4	Ridge Regression	0.0553	0.0747	4590.00	6607.31
5	Reference Measure	0.0770	0.1017	7024.11	9539.54

Table 3.4: Results for prediction performance based on selected baseline features for varying machine learning approaches (MAE: mean absolute error, RMSE: root mean square error).

We observed a general increase in performance (Table 3.4). All statistical methods performed better than the reference measure (except for RMSE for linear regression and cost prediction), which was again confirmed by a significant Wilcoxon test for MAEs ($SVR : P_O < .001, P_C < .001; Regression : P_O < .001, P_C < .001; Tree : P_O = .002, P_C < .001; Ridge : P_O < .001, P_C < .001$). This suggested that feature selection resulted in more accurate predictions in this context. The overall results demonstrate that some machine learning approaches are beneficial when predicting the outcomes and costs. Since ridge regression predicted the outcome and costs best, we utilized this model in the following analysis.

Figure 3.2 illustrates the predicted and observed values for each treatment type and dependent variable (QALY/costs). For estimating the ridge regression penalty term, we implemented 100 cross-validation runs and utilized the parameter that minimized the mean cross-validated error among these runs. The predictions were sorted in an ascending order. The blue markers or lines are the predictions and the black markers are the observed values where the y-axis demonstrates the value of the QALY/costs and the x-axis represents the corresponding patient. We observed that the predicted outcome and costs showed high uncertainty. The broader range of the actual observations around the blue markers for the cost predictions indicated that these were more difficult to achieve than outcome predictions in this context. Visually, however, the trend of the predictions appeared to be as expected, and as illustrated by the increased performance compared with the reference measure; this result indicates a step in the right direction.

3.3.3 Treatment recommendation

In order to derive individual treatment recommendations, we represent the differential outcomes and costs in the cost-effectiveness plane, where the y-axis is the difference between the costs of each treatment type and the x-axis is the difference between the clinical effects, as seen in Figure 3.3 (Black, 1990). Each quadrant has a different meaning. In our context, the NE quadrant represents higher costs and positive effects for BT; the SE quadrant indicates that BT is less expensive and more effective (BT dominates); the SW quadrant demonstrates the case where BT is less expensive but less effective; and the NW quadrant displays the situation where BT is more expensive and less effective (TAU dominates) (Klok and Postma, 2004). As a first step, a threshold had to be defined that specified up to which point an additional improvement was worth

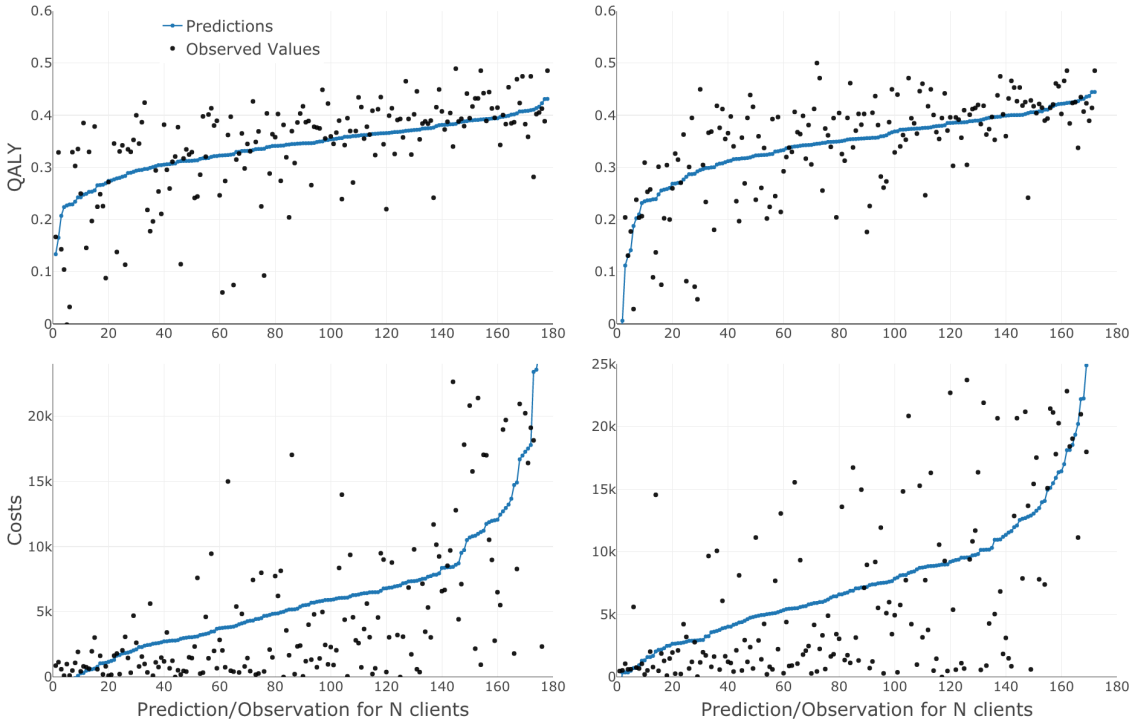


Figure 3.2: Predicted and observed values for quality-adjusted life years and costs and both treatment types (left panels for treatment as usual and right panels for blended treatment).

the costs. In the context of this study, the monetary amount or willingness to pay for gaining one QALY differed by country (Klok and Postma, 2004); the commonly used UK WTP thresholds for QALYs are between 25,000 and 35,000 €/QALY (National Institute for Health and Care Excellence (NICE), 2013). For this study, we used the conservative estimation of 25,000 €/QALY. A value above this threshold indicated that the treatment type was too expensive. Each patient represented by a green cross received the treatment type we would have recommended based on the prediction. On the contrary, each patient that had a red circle should have received the other treatment type based on the forecasts. Questionnaire items that deviate tremendously for either TAU or BT create high differences when calculating the ICER. The point for the participant at the bottom of Figure 3.3 at (-0.04, -60.420), for example, is due to the fact that this patient reported a large number of hospital admissions. Since these are very expensive, it led to very high costs for this particular patient, and thus, the difference in costs between BT and TAU was high. Following this process, it is possible to recommend the likely most beneficial treatment type, on an individual level, at baseline.

Table 3.5 is a contingency table consisting of the patients for whom we recommended a specific treatment type. Only 46.57% (163/350) of all patients were treated using the treatment type we would recommend based on our models and the particular ICER threshold. We then calculated potential outcomes and costs on a population level assuming the patients would have been

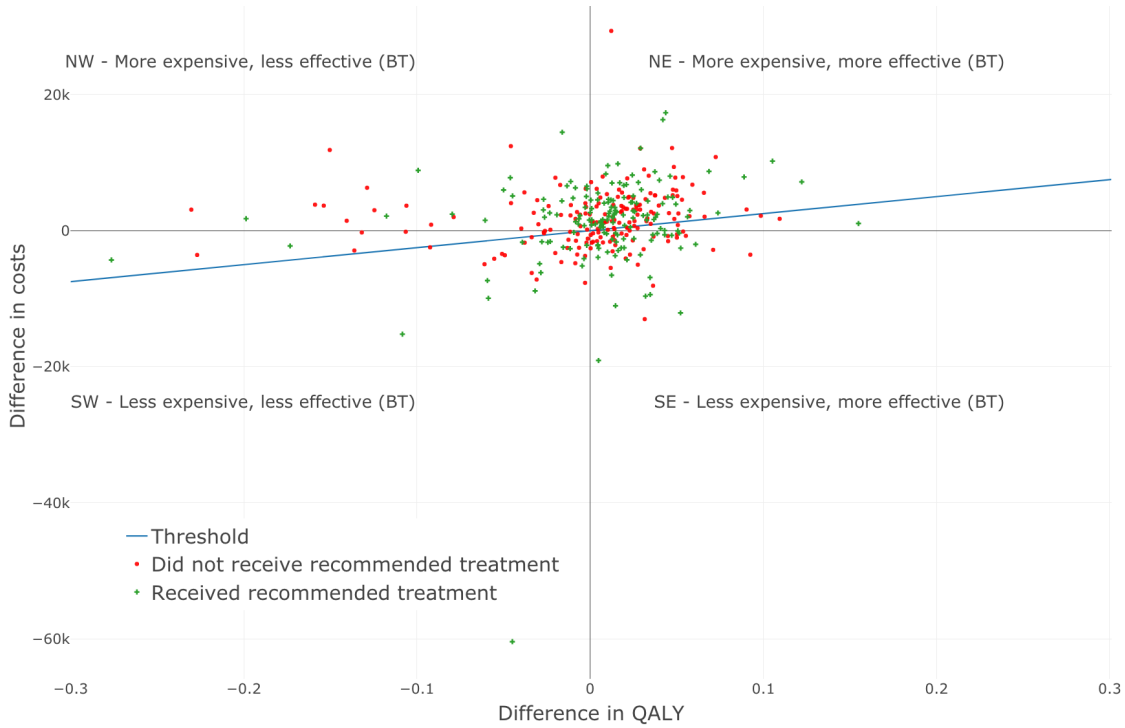


Figure 3.3: Expected improvement for all patients in relation to costs. The x-axis illustrates the difference in quality-adjusted life years (blended treatment - treatment as usual) and the y-axis the difference in costs (blended treatment - treatment as usual).

allocated according to the predictions. For patients who had already received the recommended treatment type, we utilized the observed outcomes and costs. For patients for whom the actual treatment type was not recommended, we utilized the predictions of the model. Then, QALYs would have decreased by 1.98%, while at the same time, a reduction in costs of 5.42% could have been achieved.

	Recommended BT	Recommended TAU	
Received BT	20%	29.14%	$\sum = 172$
Received TAU	24.29%	26.57%	$\sum = 178$
	$\sum = 155$	$\sum = 195$	

Table 3.5: Treatment recommendation for all patients (N=350).

3.4 Discussion

3.4.1 Principal findings

Given the growth in demand for personalized treatments and the need for a reduction in costs, predictions of outcomes and costs, in the context of mental health, are increasingly important (Van Breda et al., 2016). In this study, we proposed an approach for personalized treatment

recommendations at baseline. Here, individuals are assigned to the most beneficial treatment before treatment, which can, if desired, even be automated. We derived these recommendations by predicting patient individual QALYs and costs based on data from a European Union-funded project. We then used the ICER and the cost-effectiveness plane as an individualized treatment recommendation tool. Nowadays, decisions are often made based on the ICER; we proposed a feasible path that allows the individualization and tailoring of this process.

We illustrated that the utilization of all baseline features is not necessarily appropriate in this context. Taking advantage of feature selection techniques can increase prediction performance. As a result, we found that consultations with some kind of therapist, medication usage, anxiety or depression information (severity), mobility items (ie, "I have no problems in walking about"), and origin of the patient play an important role when predicting outcomes and costs in the context of digital health interventions. Therefore, including questionnaires that contain these factors and subsequently utilizing these features in statistical analyses when predicting outcomes and costs can be beneficial. We further illustrated that experimentation with different statistical methods benefits the final results since considerable varying performances occurred among the methods.

However, we demonstrated that prediction is a challenging task. Even though the results suggest that predictive power exists in the baseline features, our analyses indicated that the predictions, and thus the recommendations, come with uncertainty when only baseline information is available. In general, the predictive uncertainty is due to two sources. The first source is the uncertainty in the estimated parameters. With an increased amount of data, the uncertainty in parameter estimation reduces. This does not mean that we would achieve perfect predictions because the second source is related to the variance of treatments that cannot be explained by the model. More specifically, the models do not fully represent the reality and all its complexity. Hence, although the estimation of the model parameters improves with more data, the uncertainty that results from the model specifications and inability of the baseline information to precisely predict results remains. Nevertheless, we showed that we were able to predict the outcomes and costs better, compared with using the mean of the dependent variables as prediction (reference measure). Therefore, we are convinced that the baseline features do include some information regarding the forecast of outcomes and costs and can support practitioners in their decision-making process. Thus, combining these results with the ICER enabled us to provide treatment recommendations on an individual level.

As mentioned earlier, if the patients would have been allocated according to our predictions, QALYs would have decreased by 1.98% and a simultaneous reduction in costs of 5.42% could have been achieved. These results are based on a specific ICER threshold. When applying this procedure in a real-world setting, this threshold can be adjusted to values set by experts or policy makers or available budgets. These experts must make decisions regarding the monetary resources they would want to spend on a specific QALY gain. Thus, the outcome and costs can be controlled by setting this threshold. As suggested by a previous study (Lord et al., 2006), the

cost-effectiveness decision rule might be modeled in a nonlinear form. For example, the value of improvements may vary among the outcome levels. Particularly, a difference between 0.1 and 0.2 on the scale might be more important than a difference between 0.8 and 0.9, even though the absolute difference is the same. The absolute severity of the symptoms can also play an additional role in this context. It might not be justifiable to spend additional monetary effort if a specific patient already does not suffer from severe symptoms. Therefore, experts in the field need to choose appropriate values for the ICER threshold based on their experiences and knowledge and even consider a nonlinear specification.

Even though these results are preliminary, the implementation of such predictive models in clinical decision support systems for usage in interventions can be beneficial. We envision developing a system that incorporates these models and provides treatment recommendations for individuals. However, investment into other aspects is necessary for the realization of such support systems. Besides the technical implementation, the creation of information systems in this context also requires interdisciplinary collaboration among clinicians, computer scientists, and other decision makers (Sim et al., 2001). Future users, for example decision makers or therapists, need to be educated appropriately and also be involved in the design phase of the system and its requirements and development, while at the same time, the IT specialists need to be confronted with content-related issues of the user (Berg, 2001; Hartswood et al., 2000). Thus, implementation should be carefully planned and considered as organizational development (Atkinson and Peel, 1998). Furthermore, a vast amount of financial and organizational resources can be required for the implementation (Sim et al., 2001), and clinical decision makers need to understand the value and limitations of such decision support systems. Additionally, we need to be cautious with the interpretability of the results because in individual cases, recommendations might lead to suboptimal outcomes and high uncertainty depending on the particular context. Overall, these systems may be used in the future to support the decision-making process of clinicians and therapists and not to replace their treatment recommendations.

3.4.2 Limitations

This study has certain limitations. One limitation is the fact that we utilized data after a 6-month period. Usually, the preferred outcome for cost-effectiveness analysis is based on 12 months. Another limitation, which is closely associated with the previous aspect, is the size of the dataset we used. Given the complexity of the problem, it is inevitable that variations in performance occur when predicting other datasets. Thus, for achieving higher accuracy in predictions, obtaining more data is crucial. Even though our results are promising, more data and evaluations are needed in order to investigate the generalizability of these outcomes and improve the predictive accuracy of statistical techniques. Besides the size of the dataset, the data are heterogeneous in different ways. For example, the data were collected from 9 different European countries, with each having their own country-specific conditions (Kleiboer et al., 2016). This can result in

country-specific patterns in the data. Given the limited amount of observations on a national level, we have not explored this multi-level structure. Additionally, the dataset consists of a large amount of missing values that needed imputation. Making all baseline questions mandatory for the patients can lead to an increased performance of the statistical procedures and can therefore lower uncertainty.

3.4.3 Conclusions

This study investigated how patients can be allocated to different treatment types in order to increase clinical and cost effectiveness. We demonstrated how to predict outcomes and costs in this context and proposed an approach for individualized treatment recommendations by utilizing the ICER. Simultaneously, we evaluated a variety of machine learning techniques and demonstrated specific features that contribute to the prediction performance. The results are indicative of progress. We hope that policy makers increasingly understand the benefit of predictive modeling in this context and apply these types of models to make better and simultaneously more personalized treatment choices. We further hope that we can contribute to the decision-making process in this field by providing a path that allows the prediction of eventual outcomes and costs on an individual basis before the onset of treatment.

References

Atkinson, C. and Peel, V. (1998). Transforming a hospital through growing, not building, an electronic patient record system. *Methods Inf Med*, 37(3):285–93.

Berg, M. (2001). Implementing information systems in health care organizations: myths and challenges. *International Journal of Medical Informatics*, 64(2-3):143–56.

Black, W. C. (1990). The CE plane: a graphic representation of cost-effectiveness. *Medical Decision Making*, 10(3):212–4.

Burges, C. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167.

Chai, T. and Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3):1247–1250.

Diehr, P., Yanez, D., Ash, A., Hornbrook, M., and Lin, D. Y. (1999). Methods for analyzing health care utilization and costs. *Annu. Rev. Public Health*, 20:125–144.

Dunlop, B. (2015). Prediction of treatment outcomes in major depressive disorder. *Expert review of clinical pharmacology*, 8(6):669–72.

EuroQolGroup (1990). EuroQol-a new facility for the measurement of health-related quality of life. *Health policy*, 16(3):199–208.

Hakkaart-van Roijen, L., van der Linden, N., Bouwmans, C., Kanters, T., and Tan, S. (2015). *Kostenhandleiding: Methodologie van kostenonderzoek en referentieprijzen voor economische evaluaties in de gezondheidszorg*.

Hakkaart-van Roijen, L., Van Straten, A., Donker, M., and Tiemens, B. (2002). *Trimbos/iMTA Questionnaire for Costs Associated with Psychiatric Illness (TiC-P)*. Institute for Medical Technology Assessment, Erasmus University.

Hartswood, M., Procter, R., Rouncefield, M., and Sharpe, M. (2000). Being there and doing IT in the workplace: A case study of a co-development approach in healthcare. In *Proceedings of the participatory design conference*, pages 96–105.

Ioannidis, J. P. A. and Garber, A. M. (2011). Individualized cost-effectiveness analysis. *PLoS Medicine*, 8(7):e1001058.

Jones, J., Amaddeo, F., Barbui, C., and Tansella, M. (2007). Predicting costs of mental health care: a critical literature review. *Psychological Medicine*, 37(4):467–477.

Kautzky, A., Dold, M., Bartova, L., Spies, M., Vanicek, T., Souery, D., Montgomery, S., Mendlewicz, J., Zohar, J., Fabbri, C., Serretti, A., Lanzenberger, R., and Kasper, S. (2018). Refining Prediction in Treatment-Resistant Depression: Results of Machine Learning Analyses in the TRD III Sample. *J Clin Psychiatry*, 79(1):16m11385.

Kleiboer, A., Smit, J., Bosmans, J., Ruwaard, J., Andersson, G., Topooco, N., Berger, T., Krieger, T., Botella, C., Baños, R., Chevreul, K., Araya, R., Cerga-Pashoja, A., Cieślak, R., Rogala, A., Vis, C., Draisma, S., van Schaik, A., Kemmeren, L., Ebert, D., Berking, M., Funk, B., Cuijpers, P., and Riper, H. (2016). European COMPARative Effectiveness research on blended Depression treatment versus treatment-as-usual (E-COMPARED): study protocol for a randomized controlled, non-inferiority trial in eight European countries. *Trials*, 17(1):387.

Klok, R. and Postma, M. (2004). Four quadrants of the cost-effectiveness plane: Some considerations on the south-west quadrant. *Expert Review of Pharmacoeconomics and Outcomes Research*, 4(6):599–601.

Knapp, M. (1999). Economic Evaluation and Mental Health : Sparse Past . . . Fertile Future ? *The Journal of Mental Health Policy and Economics*, 2(4):163–167.

Kolovos, S., Kenter, R., Bosmans, J., Beekman, A., Cujipers, P., Kok, R., and van Straten, A. (2016). Economic evaluation of Internet-based problem-solving guided self-help treatment in comparison with enhanced usual care for depressed outpatients waiting for face-to-face treatment: A randomized controlled trial. *Journal of affective disorders*, 200:284–292.

Kravitz, R. L., Duan, N., and Braslow, J. (2004). Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *Milbank Q*, 82(4):661–87.

Kroenke, K., Spitzer, R., and Williams, J. (2001). The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med*, 16(9):606–13.

Lord, J., Laking, G., and Fischer, A. (2006). Non-linearity in the cost-effectiveness frontier. *Health Econ*, 15(6):565–77.

Manning, W. G. and Mullahy, J. (2001). Estimating log models : to transform or not to transform? *Journal of Health Economics*, 20(4):461–94.

National Institute for Health and Care Excellence (NICE) (2013). Guide to the methods of technology appraisal. *Process and Methods Guides*, 9.

O'Hagan, A. and Stevens, J. W. (2002). The probability of cost-effectiveness. *BMC Medical Research Methodology*, 2:5.

R Core Team (2018). R: A Language and Environment for Statistical Computing.

Rush, A. J., Trivedi, M. H., Ibrahim, H. M., Carmody, T. J., Arnow, B., Klein, D. N., Markowitz, J. C., Ninan, P. T., Kornstein, S., Manber, R., and Others (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. *Biological psychiatry*, 54(5):573–83.

Russell, L., Gold, M., Siegel, J., Daniels, N., and Weinstein, M. (1996). The role of cost-effectiveness analysis in health and medicine. *JAMA*, 276(4):1172–77.

Ryder, H., McDonough, C., Tosteson, A., and Lurie, J. (2009). Decision Analysis and Cost-effectiveness Analysis. *Semin Spine Surg*, 21(4):216–222.

Sculpher, M. (2015). Clinical trials provide essential evidence, but rarely offer a vehicle for cost-effectiveness analysis. *Value in Health*, 18(2):141–142.

Sim, I., Gorman, P., Greenes, R., Haynes, R., Kaplan, B., Lehmann, H., and Tang, P. (2001). Clinical decision support systems for the practice of evidence-based medicine. *Journal of the American Medical Association*, 8(6):527–534.

Tibshirani, R. (1996). Regression Selection and Shrinkage via the Lasso. *Journal of the Royal Statistical Society Series B*, 58(1):267–88.

Triñanes, Y., Atienza, G., Louro-González, A., De-las Heras-Liñero, E., Alvarez-Ariza, M., and Palao, D. J. (2015). Development and impact of computerised decision support systems for clinical management of depression: A systematic review. *Revista de Psiquiatría y Salud Mental*, 8(3):157–166.

van Breda, W., Bremer, V., Becker, D., Hoogendoorn, M., Funk, B., Ruwaard, J., and Riper, H. (2018). Predicting therapy success for treatment as usual and blended treatment in the domain of depression. *Internet Interventions*, 12:100–104.

Van Breda, W., Hoogendoorn, M., Eiben, A. E., Andersson, G., Riper, H., Ruwaard, J., and Vernmark, K. (2016). A feature representation learning method for temporal datasets. In *IEEE Symposium Series on Computational Intelligence*, pages 1–8.

Versteegh, M., M. Vermeulen, K., M. A. A. Evers, S., de Wit, G. A., Prenger, R., and A. Stolk, E. (2016). Dutch Tariff for the Five-Level Version of EQ-5D. *Value in Health*, 19(4):343–352.

Willmott, C. J., Matsuura, K., and Robeson, S. M. (2009). Ambiguities inherent in sums-of-squares-based error statistics. *Atmospheric Environment*, 43(3):749–752.

3.5 Appendix

A

Item No.	Item Description
1	Do you have access to a fast Internet connection (apref2)
2	Treatment program at other institution 1/2/3/4 (aTicp5d1/aTicp5e1/aTicp5f1/aTicp5g1)
6	Other primary care 1/2/3/4/5 (aTicp1i1/aTicp1j1/aTicp1h1/aTicp1g1/aTicp1k1)
11	How many times did you consult other primary care 4/5 (aTicp1j2/aTicp1k2)
13	Number of days a day-time treatment program (institution 2/3/4) (aTicp5e2/aTicp5f2/aTicp5g2)
16	Number of parts of days a part-time treatment program (institution 2/3/4) (aTicp5e3/aTicp5f3/aTicp5g3)
19	Other Tranquilizers or sleep medication (aTicp27/aTicp28)
21	Other mental care 1/2/3/4 (aTicp2h1/aTicp2i1/aTicp2j1/aTicp2k1)
25	Other complementary by name 1/2/3/4/5 (aTicp3f1/aTicp3g1/aTicp3h1/aTicp3i1/aTicp3j1)
30	How many times did you consult other complementary therapists 2/3/4/5 (aTicp3g2/aTicp3h2/aTicp3i2/aTicp3j2)
34	Received domestic care, number of months (aTicp9a)
35	Other provider of treatment (atreat15b)
36	Medication use (period Amitryptiline (Tryptizol) (aTicp11d1)
37	How long have you been in treatment (atreat15c)
38	Medication use (dosage Nortriptyline (Nortrilen) (aTicp16b)
39	Other type of treatment (atreat17a)
40	Medication use (Frequency Nortriptyline (Nortrilen) (aTicp16c)
41	Depressive episode current (amini1)
42	Medication use (period Nortriptyline (Nortrilen) (aTicp16d1)
43	Specify drugs taken (amini20b)
44	Antidepressants (aTicp20)
45	Specify drugs (amini20d)
46	Other antidepressants (aTicp21)
47	Psychotic disorder current (amini21)
48	Period other depressants (aTicp21d1)
49	Self care (aEQ5D5L2)
50	Pain/Discomfort (aEQ5D5L4)
51	How many times did you consult the Speech therapist (aTicp1d)
52	Times of rehabilitation clinic admissions (aTicp6c1)
53	Other medication for mental health complaints (aTicp29)

Item No.	Item Description
54	Medication use (other period Amitryptiline (Tryptizol)) (aTicp11d2)
55	Nights of rehabilitation clinic admissions (aTicp6c2)
56	Medication use (other medications for mental health complaints) (aTicp30)
57	Medication use (period Flurazepam (Dalmadorm)) (aTicp25d1)
58	How many times did you consult other mental care 4 (aTicp2k2)
59	Type of other institution at which admissions 1 (aTicp6e1)
60	Period other medication for mental health complaints (aTicp30d1)
61	How long have you been taking antipsychotic medication (atreat8)
62	Type of other institution at which admissions 2 (aTicp6f1)
63	Other Treatment (atreat2b)
64	How long have you been taking this other medication (atreat12)
65	Times of other institution 2 admissions (aTicp6f2)
66	Anti-depressants (atreatMed)
67	Who is delivering the psychotherapy (atreat14a)
68	Nights of other institution admissions (aTicp6f3)
69	Other provider of anti-depressants (atreat5a)
70	Recency of hypomanic episode (amini6b)
71	Type of other institution at which admissions 3 (aTicp6g1)
72	Who is providing the tranquillizers (atreat7)
73	Abuse non-alcohol psychoactive substance use disorder current (amini20a)
74	Times of other institution admissions (aTicp6g2)
75	Other provider of antipsychotic medication (atreat9a)
76	Abuse non-alcohol psychoactive substance current (amini20c)
77	Nights of other institution admissions (aTicp6g3)
78	Other provider of sleep medication (atreat11a)
79	Depressive episode lifetime (amini23)
80	Received nurse care, number of months, in last 3 months (aTicp7a)
81	Other provider of this other medication (atreat13a)
82	Who is delivering the other treatment (atreat15a)
83	Received daily care, number of months (aTicp8a)
84	Other provider of psychotherapy (atreat14b)
85	Received daily care, number of hours per week (aTicp8b)
86	Thoughts of Death or Suicide (aQIDS12)
87	Medication use (Other medication for mental health complaints) (aTicp30a)
88	Medication use (Nortriptyline (Nortrilen)) (aTicp16a)
89	How many times did you consult the Holistic therapist (aTicp3c)

Item No.	Item Description
90	Medication use (Amitryptiline (Tryptizol)) (aTicp11a)
91	Marital status (aMarital)
92	Hypomanic episode (amini6a)
93	How many times did you consult the Natural healer (aTicp3e)
94	How many times did you consult the Professional from a clinic for alcohol and drugs or similar institution (aTicp2f)
95	Who is providing the anti-depressants (atreat5)
96	How long have you been taking tranquillizers (atreat6)
97	Who is providing this other medication (atreat13)

Table 3.6: Omitted items.

B

	Algorithm	MAE_O	$RMSE_O$	MAE_C	$RMSE_C$
1	SVR	0.0737	0.1014	6539.97	9466.11
2	Regression Tree	0.0765	0.1040	6471.05	9346.40
3	Ridge Regression	0.0647	0.0870	6044.44	8395.56

Table 3.7: Results for prediction performance based on sampling from normal and categorical distribution and all baseline features for varying machine learning approaches (MAE: mean absolute error, RMSE: root mean square error).

C

	Algorithm	MAE_O	$RMSE_O$	MAE_C	$RMSE_C$
1	SVR	0.0598	0.0838	5195.46	8211.82
2	Regression	0.0599	0.0794	5204.89	7194.91
3	Regression Tree	0.0707	0.0958	6229.68	9278.88
4	Ridge Regression	0.0565	0.0746	5001.75	6921.86

Table 3.8: Results for prediction performance based on sampling from normal and categorical distribution and selected baseline features for varying machine learning approaches (MAE: mean absolute error, RMSE: root mean square error).

D

Feature	Parameter Coefficient
(Intercept)	3.90e−1
Cumulated PHQ value (aPHQScore)	−3.01e−3
Anxiety/Depression (I am slightly anxious or depressed) (aEQ5D5L5)	2.90e−2

Feature	Parameter Coefficient
How many times did you consult the General practitioner (aTicp1a)	−1.73e−3
Mobility (I have no problems in walking about) (aEQ5D5L)	1.50e−2
General interest (I have virtually no interest in the activities I used to enjoy) (aQIDS13)	−6.16e−3
Usual activities (I have severe problems doing my usual activities) (aEQ5D5L3)	−1.44e−2
Anxiety/Depression (I am severely anxious or depressed) (aEQ5D5L5)	−9.66e−3
How many times did you consult other primary care (aTicp1h2)	−3.27e−3
Medication use (other period other depressants 1) (aTicp20d2)	−3.18e−3
How many times: medical specialist at an outpatient clinic (aTicp4)	−2.73e−3
Medication use (dosage Venlafaxine (Efexor)) (aTicp19b)	−7.90e−5
Medication use (Frequency other depressants 1) (aTicp20c)	−1.71e−3
Agoraphobia current (yes) (amini11)	−1.26e−3
Age at baseline (aAge)	−5.20e−5

Table 3.9: Important baseline features based on Lasso regression for TAU (including single levels for each item) and QALY prediction for λ=0.01485.

E

Feature	Parameter Coefficient
(Intercept)	2.93e−1
Mobility (I have severe problems in walking about) (aEQ5D5L1)	−1.76e−1
Mobility (I have no problems in walking about) (aEQ5D5L1)	3.18e−2
Cumulated PHQ value (aPHQScore)	−2.22e−3
Energy level (I really cannot carry out most of my usual daily activities because I just do not have the energy) (aQIDS14)	−1.86e−2
Anxiety/Depression (I am severely anxious or depressed) (aEQ5D5L5)	−1.83e−2
Usual activities (I have no problems doing my usual activities) (aEQ5D5L3)	1.63e−2
Usual activities (I have severe problems doing my usual activities) (aEQ5D5L3)	−1.81e−2
How many times did you consult other mental care 3 (aTicp2j2)	−5.05e−3
Medication use (period Other medication for mental health complaints) (other) (aTicp29d1)	7.26e−2
How many times did you consult the general practitioner (aTicp1a)	−2.89e−4
Trouble concentrating on things, such as reading the newspaper or watching television (Nearly every day) (aPHQ07)	−5.28e−3
How many times did you consult the Dietician (aTicp1e)	−1.05e−2
Who is providing the sleep medication (Psychiatrist) (atreat11)	2.71e−3
Medication use (dosageFluoxetine (Prozac)) (aTicp14b)	−5.70e−5

Table 3.10: Important baseline features based on Lasso regression for BT (including single levels for each item) and QALY prediction for λ=0.01479.

F

Feature	Parameter Coefficient
(Intercept)	1.17e+6
Little interest or pleasure in doing things (Several days) (aPHQ01)	−5.68e+2
Trouble falling or staying asleep, or sleeping too much (Nearly every day) (aPHQ03)	1.40e+3
Do you have a preference for one of the treatments offered (No preference) (apref1)	−3.97e+2
Do you have a preference for one of the treatments offered (Treatment as usual not including the online treatment) (apref1)	−3.97e+2
Mobility (I have no problems in walking about) (aEQ5D5L1)	−8.39e+2

Feature	Parameter Coefficient
Mobility (I have slight problems in walking about) (aEQ5D5L)	2.15e+2
Anxiety/Depression (I am severely anxious or depressed) (aEQ5D5L5)	1.11e+3
Anxiety/Depression (I am slightly anxious or depressed) (aEQ5D5L5)	−4.17e+2
How many times did you consult the general practitioner (aTicp1a)	6.83e+1
How many times did you consult a therapist for physical therapy (aTicp1b)	2.05e+1
How many times did you consult the Dietician (aTicp1e)	4.19e+1
How many times did you consult other primary care 1 (aTicp1g2)	5.99e+2
How many times did you consult the Psychiatrist (aTicp2d)	1.57e+2
How many times did you consult other mental care 1 (aTicp2h2)	−9.30e+1
How many times did you consult the Acupuncturist (aTicp3a)	5.17e+1
How many times: medical specialist at an outpatient clinic (aTicp4)	2.46e+2
Times of other institution admissions (aTicp6e2)	−4.07e+2
Medication use (dosage Citalopram (Cipramil)) (aTicp12b)	2.42e+1
Medication use (other periodCitalopram (Cipramil)) (aTicp12d2)	−3.80e+1
Medication use (Fluoxetine (Prozac)) (aTicp14a)	−6.93e+2
Medication use (other period Nortriptyline (Nortrilen)) (aTicp16d2)	2.52e+3
Medication use (other) (aTicp17d1)	−2.36e+3
Medication use (dosage other depressants 1) (aTicp20b)	6.13e+1
Medication use (Frequency Oxazepam (Seresta)) (aTicp22c)	−9.58e+1
Medication use (period Oxazepam (Seresta)) (aTicp22d1)	−1.41e+3
Medication use (period Zopiclon (Imovane)) (aTicp26d1)	1.42e+3
Medication use (dosage Other Tranquilizers or sleep medication) (aTicp27b)	8.46e+0
Medication use (other period for Other medication for mental health complaints) (aTicp29d2)	−4.05e+2
Do you have a paid job (yes) (aTicp39)	1.05e+3
How many hours does your contract specify (aTicp40)	3.72e+1
Did health problems oblige you to call in sick from work at any time (Yes, I was off work during the full three months) (aTicp42)	4.24e+3
On which date did you call in sick from work first because of health problems (aTicp43)	−8.50e−5
On how many working days did you call in sick from work because of health problems in the past three months (aTicp45)	5.47e+1
Was your job performance adversely affected by health problems (yes) (aTicp46)	7.22e+2
Rate how well performed on days bothered by health problems (aTicp48)	−1.44e+2
What type of treatment do you receive (Medication) (aTreat2a)	−8.90e+0
How long have you been taking sleep medication (1-6 months) (atreat10)	−2.75e03

Feature	Parameter Coefficient
How long have you been taking sleep medication (More than 1 year) (atreat10)	7.83e+2
Who is providing the sleep medication (Psychiatrist) (atreat11)	−6.85e+2
What type of treatment did you receive (Medication) (atreat17)	−7.69e+2
Suicidal risk current (yes) (amini5a)	5.67e+2
Recency manic episode (Lifetime) (amini7b)	3.52e+1
Agoraphobia current (yes) (amini11)	−1.53e+3
Panic disorder without agoraphobia current (amini12)	4.79e+2
Panic disorder with agoraphobia current (yes) (amini13)	−3.19e+3
Obsessive compulsive disorder current (yes) (amini16)	−7.25e+2
Falling asleep (I take at least 30 minutes to fall asleep, some nights) (aQIDS01)	−2.66e+1
Increased Appetite (I regularly eat more often and or greater amounts of food than usual) (aQIDS07)	6.12e+2
Weightloss (I have lost 2.5 kilos or more) (aQIDS08)	−6.54e+2
Weightgain (I have gained 2.5 kilos or more) (aQIDS09)	1.15e+3
Weightgain (I have not had a change in my weight) (aQIDS09)	1.03e+1
Concentration/Decision Making (Most of the time, I struggle to focus my attention or to make decisions) (aQIDS10)	−7.03e+1
Energy level (I have to make a big effort to start or finish my usual daily activities) (aQIDS14)	7.81e+1
Energy level (I really cannot carry out most of my usual daily activities because I just do not have the energy) (aQIDS14)	1.16e+3
Country code (Germany) (cc)	4.78e+2
Country code (Poland) (cc)	−2.84e+3
Country code (Netherlands) (cc)	4.06e+3
Country code (Spain) (cc)	−7.36e+2
Country code (UK) (cc)	4.06e+3

Table 3.11: Important baseline features based on Lasso regression for TAU (including single levels for each item) and cost prediction for λ=433.83.

G

Feature	Parameter Coefficient
(Intercept)	−1.52e+6
Little interest or pleasure in doing things (Not at all) (aPHQ01)	−5.83e+2
Trouble falling or staying asleep, or sleeping too much (Several days) (aPHQ03)	−2.31e+2

Feature	Parameter Coefficient
Poor appetite or overeating (Nearly every day) (aPHQ05)	1.74e+3
Trouble concentrating on things, such as reading the newspaper or watching television (Not at all) (aPHQ07)	−5.50e+2
Willing to carry a Smartphone delivered by treatment team (no) (apref4)	−1.44e+3
Usual activities (I have no problems doing my usual activities) (aEQ5D5L3)	−2.35e+1
Anxiety/Depression (I am moderately anxious or depressed) (aEQ5D5L5)	1.56e+2
How many times did you consult the industrial physician (aTicp1f)	5.16e+2
How many times did you consult other primary care 1 (aTicp1g2)	7.37e+2
How many times did you consult other mental care 3 (aTicp2j2)	1.62e+2
How many times did you consult the Acupuncturist (aTicp3a)	9.68e+2
Nights of regular hospital admissions (aTicp6a2)	4.01e+1
Times of other institution admissions (aTicp6e2)	−2.48e+2
Medication use (Citalopram (Cipramil)) (aTicp12a)	−7.97e+2
Nights of regular hospital admissions (aTicp15d2)	−3.80e+2
Medication use (dosage Venlafaxine (Efexor)) (aTicp19b)	1.88e+1
Medication use (period Venlafaxine (Efexor)) (aTicp19d1)	1.69e+3
Medication use (dosage other depressants 1) (aTicp20b)	1.49e+1
Medication use (other antidepressant 2) (yes) (aTicp21a)	1.85e+3
Medication use (dosage other depressants 2) (aTicp21b)	2.36e+2
Medication use (other period Oxazepam (Seresta)) (aTicp22d2)	4.46e+2
Medication use (dosage Other medication for mental health complaints) (aTicp30b)	1.35e+3
Do you have a paid job (yes) (aTicp39)	2.45e+3
How many hours does your contract specify (aTicp40)	5.94e+0
Job questions: over how many days are these hours distributed (aTicp41)	1.56e+2
Did health problems oblige you to call in sick from work at any time (Yes, I was off work during the full three months) (aTicp42)	1.01e+4
On which date did you call in sick from work first because of health problems (aTicp43)	1.12e−4
On how many working days did you call in sick from work because of health problems in the past three months (aTicp45)	1.61e+2
Number of hours you had to catch up on work unable to perform (aTicp49b)	5.35e+0
How long have you been in psychotherapy (6 months-1 year) (atreat14c)	1.19e+3
How long have you been in psychotherapy (Less than one month) (atreat14c)	−1.83e+10
What type of treatment did you receive (Psychotherapy) (atreat17)	2.62e+3
Falling asleep (I take at least 30 minutes to fall asleep, some nights) (aQIDS01)	9.57e+2

Feature	Parameter Coefficient
General interest (There is no change from usual in how interested I am in other people or activities) (aQIDS13)	−1.85e+3
Energy level (I really cannot carry out most of my usual daily activities because I just do not have the energy) (aQIDS14)	2.76e+1
Feeling Restless (I do not feel restless) (aQIDS16)	9.59e+2
Country code (Germany) (cc)	2.79e+2
Country code (UK) (cc)	−2.97e+2

Table 3.12: Important baseline features based on Lasso regression for BT (including single levels for each item) and cost prediction for λ=651.14.

CHAPTER 4

HETEROGENEITY MATTERS – PREDICTING SELF-ESTEEM IN ONLINE INTERVENTIONS BASED ON ECOLOGICAL MOMENTARY ASSESSMENT DATA

Bremer, V., Funk, B., and Riper, H. (2019). Heterogeneity matters – predicting self-esteem in online interventions based on Ecological Momentary Assessment data. Depression Research and Treatment, 2019:3481624.

Abstract

Self-esteem is a crucial factor for an individual's well-being and mental health. Low self-esteem is associated with depression and anxiety. Data about self-esteem is oftentimes collected in Internet-based interventions through Ecological Momentary Assessments and is usually provided on an ordinal scale. We applied models for ordinal outcomes in order to predict the self-esteem of 130 patients based on diary data of an online depression treatment and thereby illustrated a path of how to analyze EMA data in Internet-based interventions. Specifically, we analyzed the relationship between mood, worries, sleep, enjoyed activities, social contact and the self-esteem of patients. We explored several ordinal models with varying degrees of heterogeneity and estimated them using Bayesian statistics. Thereby, we demonstrated how accounting for patient-heterogeneity influences the prediction performance of self-esteem. Our results show that models that allow for more heterogeneity performed better regarding various performance measures. We also found that higher mood levels and enjoyed activities are associated with higher self-esteem. Sleep, social contact, and worries were significant predictors for only some individuals. Patient-individual parameters enable us to better understand the relationships between the variables on a patient-individual level. The analysis of relationships between self-

esteem and other psychological factors on an individual level can therefore lead to valuable information for therapists and practitioners.

4.1 Introduction

Access to mental care is limited; by providing further access, Internet-based interventions can close the gap between treatment and demand (Karyotaki et al., 2017; Saddichha et al., 2014; Titzler et al., 2018). At the same time, Online-based interventions may lead to comparable outcomes compared to face-to-face treatment (Carlbring et al., 2018; Saddichha et al., 2014). In Internet-based interventions, data about various psychological factors, for example the self-esteem level of individuals, is often collected. Self-esteem is closely related to psychological well-being and satisfaction with life (Paradise and Kernis, 2002). Low levels of self-esteem are associated with serious mental problems such as depression, anxiety, (Sowislo and Orth, 2013) or eating disorders (Silvera et al., 1998). Trzesniewski et al. (2006), found that low self-esteem can lead to "negative real-world consequences" such as mental and physical health problems, misconduct, and worse economic outlooks. In the literature, however, there is a debate if low mood levels affect self-esteem or vice versa. Two models exist for each assumption. The vulnerability model assumes that self-esteem is a risk for depression whereas the scar model interprets self-esteem rather as an outcome or aftermath of depression (Manna et al., 2016). One study, for example, found that low self-esteem can predict depression decades later (Steiger et al., 2014). Steiger et al. (2015) found that the vulnerability and the scar model are valid over decades with weaker effects for the scar model. A reoccurring finding is that low levels of self-esteem are associated with serious mental illnesses which, in turn, are known to be associated with decreased quality of life, tremendous health care costs, as well as increased costs for individuals and governments (Gustavsson et al., 2011; Leger, 1994; Paradise and Kernis, 2002; Silvera et al., 1998; Silverman et al., 2015; Sowislo and Orth, 2013). Thus, we aimed at predicting the self-esteem level of individuals in this study and analyze its relationships with a variety of psychological factors.

Data about self-esteem and other psychological factors such as mood levels or social interactions are often assessed by Ecological Momentary Assessments (EMA). These EMA methods collect data regarding behavior, symptoms, and cognition close in time to the users' experience and in their natural environment (Iida et al., 2012; Moskowitz and Young, 2006). Diaries, which are used for the analysis in this paper, are one example of EMA methods that are often utilized (Iida et al., 2012).

Due to multiple measures per individual, this data has a nested structure (Nezlek, 2003; Waegeman et al., 2008). As is common in the social sciences (Long, 2014), self-reports of diary data can be ranked on an ordinal scale. Individuals are often prompted to rank their mood level, for instance, by providing a score between one and ten for a specific question such as *"How is*

your mood right now?". Data with this structure needs to be analyzed by utilizing appropriate statistical models that can account for the ordinality in the measurements, for example, ordinal logit models or generalized linear models. In research studies, however, this is often not the case (Forrest and Andersen, 1986; Jakobsson, 2004; LaValley and Felson, 2002). Jakobsson (2004) and LaValley and Felson (2002) analyzed a multitude of journal articles; even though ordinal scales were often used, they came to the conclusion that frequently there were no appropriate data representation techniques or data analysis methods present. They found that solely 49% (La Valley et al.: 39.4%) of the analyzed articles had proper data presentation and 57% (La Valley et al.: 63.4%) had appropriate data analysis. This is alarming since an improper handling can lead to bias and incorrect interpretation of statistical effects (Hedeker, 2015).

Each patient behaves differently, has different experiences, and can be affected by psychological factors in various ways. Repeated measurements provided by patients can therefore not be considered to be independent (Bolger et al., 2003). Considering the differences among patients by implementing patient-individual parameters might lead to a better model fit (representation of the pattern in the data) and an increased prediction performance (ability to predict unobserved values of the dependent variable). By revealing these patient-individual parameters, individual effects for the independent variables (psychological factors) can be obtained for each patient, which in turn can result in individualized decision support systems and subsequently individualized recommendations in a clinical context.

In this study, we thus combined ordinal models appropriate for the analysis of diary data - namely the ordinal logit model (Liu, 2014; McCullagh, 1980) and the less frequently utilized stereotype logit model (Anderson, 1984; Liu, 2014) - and proposed to extend the models by including patient specific parameters in order to account for heterogeneity among the participants. General mixed models are often applied when analyzing data that includes repeated measurements (Bolger et al., 2003). Hedeker (2015), for example, discussed mixed effects logistic regression models for ordinal data and illustrated a possible hierarchical structure in which the effect each patient has on the outcome value is considered. In contrast to this study, our approach considered different influences of the psychological factors on the individuals which led to individual slopes. These patient specific coefficients can potentially result in more information on how the analyzed psychological factors are related to the self-esteem of the patients on an individual level and can therefore lead to a knowledge gain for researchers and practitioners. We applied the models to self-reported diary data from an Internet-based depression treatment (Kleiboer et al., 2016) in order to predict the self-esteem of individuals. At the same time, we revealed the relationship between a variety of psychological/psychosocial factors (mood, worries, sleep, enjoyed activities, social contact) and the self-esteem level of patients. Thus, this study contributes to existing research by gaining insight into the patients' behavior and how their self-esteem is related to a variety of factors on an individual level and thus by highlighting the importance of individuality in this context.

4.2 Materials and methods

4.2.1 Data

The data we utilized for our approach is acquired from an EU funded two-arm randomized controlled trial that compared bCBT (blended cognitive behavior therapy - experiment group) and face-to-face treatment (control group) (Kleiboer et al., 2016). Participants were 18 years or older, met criteria for a major depressive disorder, were not of high suicidal risk, were not currently being treated for depression, and had access to an Internet connection. The utilized data was based on diary data that has been assessed in the study through an EMA mobile-application between February 2015 and January 2017. The diary questions were sent via email or text message depending on the therapists' choice. The mood level of the participants was collected every day at a random time between 10a.m. and 8p.m. All other factors were collected on specific days; the first and last seven days of the intervention and one random day each week in the intervention period. All factors could be ranked on a scale from one to ten. We only utilized days on which all factors were assessed, which resulted in the analysis of 130 patients and their 2326 observations including all psychological factors that will be introduced in the following.

Self-Esteem | The dependent variable in our analysis was the self-esteem of the patients. It was assessed through the question "How do you feel about yourself right now?". This question is closely related to an item of the state self-esteem scale (Heatherton and Polivy, 1991) and can represent a person's self-image (Graham, 2009). The same question has also been utilized in another study that measured self-esteem for individuals and has shown to be correlated with the Rosenberg self-esteem scale (Clasen et al., 2015; Robins et al., 2001; Rosenberg, 1965). In this study, we defined this question as the self-esteem level.

Mood | Mood is an important factor for an individual's well-being, physical health, and behavioral patterns (Cohen and Rodriguez, 1995; Minden, 2000). We analyzed the relationship between these factors and hypothesized that the mood level is positively related to self-esteem. This predictor was assessed by the question "How is your mood right now?".

Worry | Worries are connected to anxiety disorders (Hoyer et al., 2002) and depression (Diefenbach et al., 2001). Since the act of worrying can potentially create feelings and thoughts that impact self-respect or cause individuals to underestimate themselves, it could be linked to self-esteem. We hypothesized that this factor is negatively related to the self-esteem of the patients. Worries were assessed by asking the patients "How much do you worry at the moment?".

Sleep | Sleep supports various functions of the human body such as repair and restorative processes (Curcio et al., 2006) and is a crucial aspect for the well-being of an individual (Gray and Watson, 2002). Prior research found that low levels of sleep can lead to lower self-esteem (Lemola et al., 2013). We hypothesized that "good" self-reported sleep levels can lead to higher levels of self-esteem. Sleep was assessed through the question "How well did you sleep last night?".

Enjoyed activities | This concept relates to any action that has been executed by the

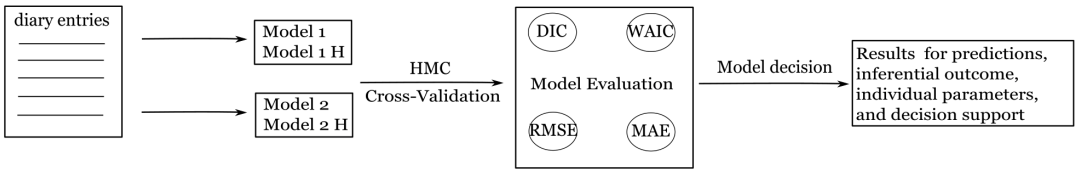


Figure 4.1: Graphic visualization of approach.

participant that day. It describes to what degree the patient has relished a specific day by the performed activities. Since we assumed that joy - that in turn can trigger happiness - can potentially boost the self-esteem of individuals, we hypothesized that enjoyed activities are positively linked to self-esteem. The predictor enjoyed activities was assessed by the question "How much did you enjoy activities today?".

Social contact | Social contact can provide important emotional support; and the lack thereof can be linked to depression (Frasure-Smith et al., 2000). We hypothesized a positive relationship between social contact and self-esteem. Social contact was assessed by asking the individuals "How much were you involved in social interaction today?".

4.2.2 Statistical analysis

4.2.2.1 Approach

We applied two different models for predicting the self-esteem at time t based on the aforementioned predictors and their scores at time t – the ordered logit and stereotype logit model. Both approaches account for the ordinality in the measurements. Four models were eventually used because we modified each method by implementing patient-specific parameters in order to consider how they are individually affected by the psychological factors (Figure 4.1). We used Hamiltonian Monte Carlo techniques (HMC) for parameter estimation (Carpenter et al., 2017), applied cross-validation, and evaluated the models by comparing their outcomes based on various performance measures. We then utilized the model that performed best for illustrating the concrete predictions, the inferential outcomes (relationship between psychological factors and self-esteem), and the patient-individual parameters.

4.2.2.2 Ordinal logistic regression model

One method that was utilized is the frequently used proportional odds or ordered logit model (OLM) that was initially proposed by McCullagh (1980). This model estimates the odds of observing a specific rank or less of self-esteem (score on the scale) for patient j at time step t for $rank_{jt} = 1, \cdots, C$, where C is the number of ranks or the highest category on a scale (ten in our

analysis since self-esteem is rated on a scale from one to ten) (Norusis, 2010):

$$\theta_{cjt} = \frac{P(rank_{jt} \leq c \mid x_{jt})}{P(rank_{jt} > c \mid x_{jt})}.$$

The estimation then follows Equation 4.1. The parameters α_c are the boundaries of the categories or threshold; also called cutpoints where $c = 1, \cdots, C-1$. This parameter has therefore nine distinct values. Furthermore, the cutpoints are following the constraint $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_{C-1}$. x_{jt} is a vector of length five that represents the observations of the psychological concepts for each client j at each time step t. The β parameters are the weights to be estimated that reveal relationships between the concepts and are utilized for the self-esteem prediction. This model is based on the proportional odds assumption. This means that the OLM assumes all β terms and their effects to be equal among all the levels of the dependent variable. As we can see, the β parameters do not vary among the ordinal levels or in any other fashion. Therefore, no individual effects are captured. The fixed β - coefficient for all patients in the data leads to the unrealistic assumption that all individuals are similarly related to the psychological factors.

$$ln(\theta_{cjt}) = \alpha_c - (x_{jt}\beta) \tag{4.1}$$

However, humans possess very unique and intricate qualities - each person has a different personality, opinion, thinking structure, and behavior; this can in turn lead to patient-individual effects from the predictors (Gable et al., 2000; Weinstein and Mermelstein, 2007). We further assumed that including patient-individual parameters could lead to a greater prediction performance because more variance can potentially be explained. However, this process comes with the sacrifice of an increased model complexity. Nevertheless, we modified the model by introducing an additional index j into the β parameters which accounts for the varying effect a predictor can have on an individual. The OLM then yields the following form:

$$ln(\theta_{cjt}) = \alpha_c - (x_{jt}\beta_j).$$

4.2.2.3 Stereotype ordinal logit model

Another model that is less frequently used in research, presumably due to the rare existence of already implemented software packages (Ahn et al., 2012; Liu and Koirala, 2012), is the stereotype ordinal logit model. This model was created by Anderson (1984) in order to tackle the restrictive nature of the OLM due to its proportional odds assumption that is often violated in real datasets (Kohavi, 1995). It can be seen as an extension of the multinomial logistic regression - with the distinction that less parameters have to be estimated (Ahn et al., 2012). We additionally applied this model in order to compare the performance of both techniques and to demonstrate that heterogeneous parameters are not only beneficial when utilizing the OLM, but also in other statistical procedures. As in the OLM, θ_{cjt} is estimated; this is the odds of observing a specific

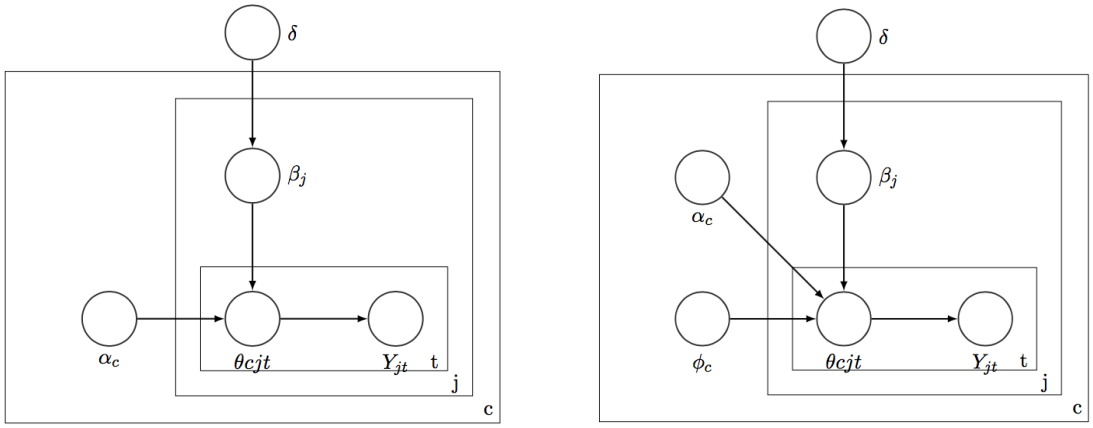


Figure 4.2: Graphic visualization for both models as plate notation (OLM left, SOLM right).

rank of self-esteem in comparison to a baseline category (in our case the last category ten) for patient j at time t.

$$\theta_{cjt} = \frac{P(rank_{jt} = c \mid x_{jt})}{P(rank_{jt} = C \mid x_{jt})}$$

The procedure of the stereotype logit model for the estimation is illustrated in Equation 4.2 for $c = 1, \cdots, C$. As we can see by the index j, the β parameters already consider individual effects. The original model does not include this index. The α_c's are the intercepts and the ϕ_c parameters are a score for the different levels of the outcome variable. Ordinality is only given as long as the constraint $0 = \phi_1 \leq \phi_2 \leq \cdots \leq \phi_C = 1$ is considered. Specifically, for a four point scale, two ϕ's are to be estimated. For a ten point scale, eight ϕ's are to be estimated.

$$P(rank = c|x_{jt}) = \frac{exp(\alpha_c + \phi_c x_{jt}\beta_j)}{\sum_{c=1}^{C} exp(\alpha_c + \phi_c x_{jt}\beta_j)} \tag{4.2}$$

4.2.2.4 Parameter setting

Enabled by the Bayesian approach, we set different priors based on assumptions and already exsiting literature mentioned above. In this context, priors are beliefs in terms of probability distributions about the effects of the predictors that can be set before the actual data is considered. We set weak positive priors for the predictors mood, sleep, enjoyed activities, and social contact. For the variable worry, we set a weak negative prior. Implementing weak priors means sampling the corresponding parameter with high variance. Thereby, prior knowledge from related literature is taken into account while at the same time, the data strongly affects the analyses. Figure 4.2 illustrates the hierarchical structure of both models including heterogeneity parameters as a plate notation.

The parameters are distributed as shown in Equation 4.3 where $\sigma^2 = 100$ (high variance). The expected value for the hyper-parameter δ is either -1 or 1 depending on our definition as either a weak negative or positive prior. The parameters δ and $\alpha_{1..C}$ are sampled from a normal

distribution. The heterogeneous parameters for each client, $\beta_{1..J}$, are also sampled from a normal distribution, however, they are based on the vector δ. We decided to sample from a normal distribution because this allowed the parameters to evenly take on a positive or negative value. This means that we assumed that patients exist for whom a specific coefficient is positive whereas other patients are negatively affected. The results for δ indicate the effects each predictor has on the self-esteem on a population level. We utilized this parameter for prediction for the models that do not consider heterogeneity. The β parameters for each patient were used for the prediction of the individual models and illustration of the individual parameters.

$$\begin{aligned} \delta &\sim \mathcal{N}(\mu \in \{-1, 1\}, \sigma^2) \\ \alpha_c &\sim \mathcal{N}(0, \sigma^2) \\ \gamma_c &\sim Dir(\mathrm{A}) \\ \beta_j &\sim \mathcal{N}(\delta, \sigma^2) \\ Y_{jt} &\sim Cat(\theta_{cjt}) \end{aligned} \tag{4.3}$$

Solely in the stereotype model, as we can see in Figure 4.2, θ_{cjt} also depends on ϕ_c. This parameter is the cumulative sum of γ_c which follows a $Dirichlet(\mathrm{A}_1, \cdots, \mathrm{A}_C)$ distribution where $\mathrm{A}_{1...C} = 1$. Since the stereotype model requires ϕ_c to be constrained, for example, steadily increasing, initialized with 0, and limited to 1, sampling γ_c from a Dirichlet distribution is an appropriate procedure to meet this constraint (Ahn et al., 2012). As a final step, the actual predicted self-esteem level for each individual at each point in time (Y_{jt}) is sampled from a categorical distribution based on θ_{cjt}. For each model, we performed 60,000 iterations on four chains when running the Hamiltonian Monte Carlo algorithm and stored every twentieth draw from the last 30,000 iterations. We implemented the models in Python[1] and utilized STAN (Carpenter et al., 2017) for Monte Carlo procedures.

4.2.2.5 Performance measures

We implemented 10-fold stratified cross-validation in order to determine the model that achieves the best prediction performance. In 10-fold cross-validation, the dataset is divided into ten equally sized chunks (in our case each patient has observations in the training as well as the test dataset). Then, the models are trained on nine chunks and the tenth is predicted. This process is repeated ten times until every chunk is utilized as test data. 10-fold cross-validation is widely used and has also been shown to be suited for real-world datasets (Kohavi, 1995; McLachlan et al., 2005).

We utilized the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002) as indicator for measure of fit and model complexity (Berg et al., 2004). The DIC is often used for model comparison and selection; especially in a Bayesian context (Gelman et al., 2014). The performance of a model is evaluated by the trade-off between how well the model fits the data and the

[1] https://www.python.org/

complexity of the model. The model fit is expressed by the deviance (the lower the value, the better the fit), which is essentially the difference between a saturated model (a model that explains all variance in the responses) and the actual model. A penalty term is added to the model fit that is increasing with a rise in number of parameters (Spiegelhalter et al., 2002). Thus, models are preferred that have a smaller number of parameters. We chose the DIC as an indicator for model selection and comparison because it has been performing sufficiently regarding a variety of examples (Berg et al., 2004; Spiegelhalter et al., 1998).

According to Ando (2007) and Richards and Richardson (2012), however, the DIC can tend to prefer overfitted models and is only based on a point estimate (Plummer, 2008; Vehtari et al., 2016). Thus, we also utilized the widely applicable or Watanabe-Akaike information criterion (WAIC) (Watanabe, 2010). The WAIC is infrequently used in research and practice because of its additional computational effort (Vehtari et al., 2016). According to Vehtari et al. (2016), the WAIC represents an improvement of the DIC. Since the calculation for the number of parameters is based on each data point of the log likelihood, which is not the case for the DIC, the outcome is more stable and reliable. The WAIC (as well as the DIC) suggests a superior performance the smaller the value. For reasons of comparison and because of the mentioned issues regarding the DIC, we utilized both measures in our analyses. For readers interested in the exact derivations and steps regarding the calculation of the DIC and WAIC, we refer to the papers of Spiegelhalter et al. (2002) and Vehtari et al. (2016) respectively.

We further used the root-mean-square error (RMSE) and mean absolute error (MAE) as performance indicators. There is a debate about the selection of choosing either one of these measures. Willmott and Matsuura (2005) and Willmott et al. (2009), for example, criticized the usage of the RMSE and came to the conclusion that it is not a good indicator for the average model performance. They emphasized to only utilize the MAE since it is more natural compared to the RMSE. However, Chai and Draxler (2014) showed that the RMSE can be a better indicator for model performance. Since there is no specific agreement in the literature as to which measure is more reliable, we decided to report both measures in our analysis.

Additionally, we defined a *mean model*. This model uses the arithmetic mean of the self-esteem value among the whole training set as prediction for each self-esteem value in the test data. Since we included heterogeneous parameters, we also used a *mean individual model* that utilizes the arithmetic mean of the training set on an individual patient level as predictions. We used these measures for comparison and as a baseline model – if we would not achieve a higher prediction performance than the *mean models*, it is questionable if the creation of such complex models is even worth the effort.

4.3 Results and discussion

4.3.1 Principal results

We can see that the *mean individual model* clearly performed better compared to the *mean model* (Table 4.1). It is also indicated that all created models performed better than the *mean models* regarding the RMSE and MAE (the other performance measures are not generatable for the *mean models*). Indicated by a Wilcoxon-Test, the errors differed significantly ($P < .05$). Therefore, creating such models is beneficial in regard to predictive performance in this context. The results further indicate that the implementation of patient-individual parameters was advantageous; both models performed better regarding each of the performance measures when accounting for individual effects even though the complexity of the models (number of parameters) increased (indicated by DIC as well as WAIC). This result highlights the importance of accounting for individual parameters. We can further see that the stereotype logit model benefit more from heterogeneity. Thus, we decided to utilize this model for further demonstration and analysis.

Model	RMSE	MAE	DIC	WAIC
No Heterogeneity Ordered logit	1.20	0.81	6204.88	6205.87
Heterogeneity Ordered logit	1.18	0.80	5914.91	6143.56
No Heterogeneity Stereo	1.21	0.82	6364.25	6369.72
Heterogeneity Stereo	1.08	0.73	5871.31	5772.14
Mean model	1.90	1.48	-	-
Mean individual model	1.38	0.98	-	-

Table 4.1: Results - performance for each model based on performance measures.

Figure 4.3 illustrates the predictive performance of this model in more detail. Specifically, it shows the observed values of self-esteem in the test data as a line and the predictions of the test data as crosses. The values are sorted in ascending order according to the observed values. Oftentimes, the predictions were the exact observed self-esteem value. Only once, the prediction was four categories off, however, it was frequently falsely predicted with a distance of two ranks. Since the predictions were close to the observed value most of the time, also indicated by the performance measures, we consider this a good result.

Table 4.2 demonstrates the effects of the psychological factors on the self-esteem. Here, the analysis was executed based on all data without withholding observations for evaluation of the models. The results indicate that the mood level of the patients is significantly related to the self-esteem.

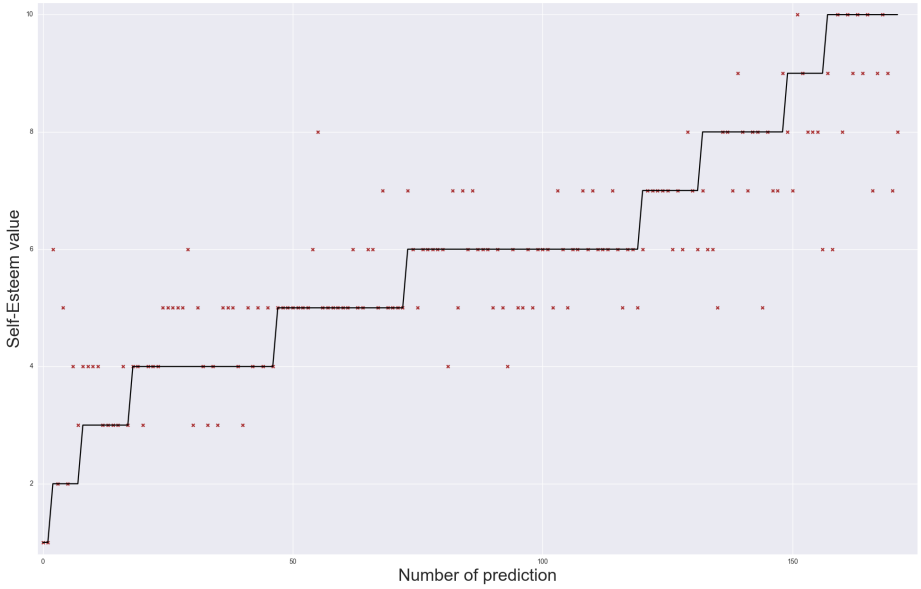


Figure 4.3: Graphic visualization of predicted and observed values.

Variables	Median	2.5% HDI	97.5% HDI
Mood	**16.82**	**14.25**	**19.55**
Worry	-1.05	-2.95	0.69
Sleep	1.50	-0.52	3.42
Enjoyed activities	**4.26**	**2.37**	**6.24**
Social contact	0.81	-1.34	2.82

Table 4.2: Results - estimated model parameters including High Density Interval (significant parameters in bold).

Since recent literature found that low self-esteem is linked to depressive moods (Martyn-Nemeth et al., 2009) and mood changes can modify self-concepts (DeSteno and Salovey, 1997), this finding is plausible. As already indicated by Pressmann et al. (2009), who found that enjoyable leisure activities are related to factors for well-being, we show that enjoyed activities significantly increased the self-esteem. When individuals experience certain activities as fun and pleasure, they might be involved in actions that can boost their confidence, be of avail, and foster feelings of happiness that can in turn increase the sense of self-worth. Therefore, joy and doing well in a specific activity can potentially lead to feelings of reward and satisfaction and thus to an increased self-esteem.

The other predictors were not significant. However, for some of the patients, these predictors might be significantly related to the self-esteem. Figure 4.4 illustrates the distributions of the individual β parameters for each patient and each predictor. The values in this Figure cannot

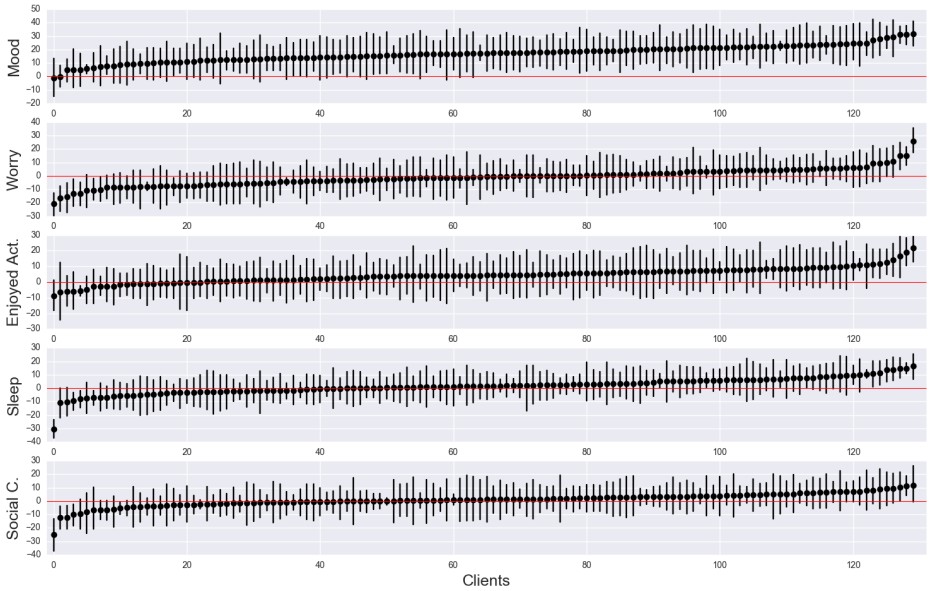


Figure 4.4: Graphic visualization of parameter distribution for each patient.

be read horizontally for each patient among the predictors; this means that the first patient for one predictor is not the same as the first patient for another predictor because the values are sorted in ascending order according to the individual mean value of the corresponding distribution. The horizontal line represents the zero value for the parameter and is an indicator for significance. The parameters varied tremendously, which again indicates the importance of considering heterogeneity. Even though the overall result for the variable worry, for instance, was insignificant, individuals exist for whom the outcome, the negative effect, was significantly true and vice versa. This finding occurs for every predictor except the mood level. Mood seemed to not be negatively related to self-esteem for any patient. Thus, the overall parameter for this predictor was highly significant. This individualized information can potentially help therapists to make refined and improved decisions on an individual level. Some patients were affected negatively by certain factors and some positively; with this procedure, it is possible to detect those specific patients. The gained information can lead to an increased understanding of patient-individual behavior and improved decision-making which can in turn result in personalized interventions and potentially better treatment outcomes.

4.3.2 Limitations

Besides the implications this study provides, we also depict some limitations and directions for further improvement and research opportunities. One limitation is the usage of diary data. Self-reported data is not inspected personally by a professional; even though this fact enables

researchers to collect data in their natural environment, it lacks objectivity and can also lead to falsely reported data and social desirability bias (Logan et al., 2008; Moskowitz and Young, 2006). Furthermore, we measured self-esteem only based on one question. Even though this question is related to one item of the state self-esteem scale (Heatherton and Polivy, 1991), it might not represent the whole complexity of self-esteem. We also obtained data for only 130 patients and 2326 observations. We believe that applying the modified models on other datasets in order to confirm the results can lead to an increased representativity. More data could improve the accuracy of gained information and especially enhance prediction performance. Therefore, more research in this context is necessary for a verification of the results.

Another aspect that can be viewed critically is the attempt of predicting a self-esteem value of a *new* patient that has not been seen before by the model. Unfortunately, even though we would have access to varying parameters for the individuals, we would not have any information on the *new* patient; therefore, we would predict this patients' self-esteem based on the overall parameter δ. In fact, we would not perform less accurate compared to models that do not account for heterogeneous influences, however, we would also not benefit from the modified models. Nevertheless, after obtaining some information about the new patient and a recalculation of the models, we could obtain individual parameters for this patient. Thus, the utilization of the modified models is initially not beneficial for new patients, but after an initial data collection period, valuable results can be generated.

Another important aspect is the question of the exact impact of more accurate predictions. How can the illustrated improvement be translated into practical benefits? If a therapist is able to provide more refined recommendations, how are the individuals affected, how can this be converted into higher outcomes, and what role do costs play in this question? We seek to tackle challenges in this context in further research.

4.4 Conclusion

In this study, we predicted the self-esteem level of participants based on collected EMA data from a two-arm randomized controlled trial. We modified two statistical models by including heterogeneous slopes for each patient and employed Hamiltonian Monte Carlo techniques for parameter estimation. Therefore, one purpose of this study was to highlight the importance of individuality in such analyses. We illustrated a path of how individual parameters can be considered in an ordinal context and demonstrated how the prediction performance of different models is influenced by doing so. Individual parameters did not only increase the performance of these models but also allow practitioners to investigate differences among patients; possibly leading to knowledge gain and deeper insight about the patients. We further emphasized the importance of self-esteem in this context and investigated its relationships with other psychological factors. We found that the self-esteem level of patients was positively related to mood and when individuals

experienced joyful activities. We further found that worries can be negatively linked to self-esteem whereas better sleep and social contact can be positively related to self-esteem. These latter results were not significant overall, however, we demonstrated that for some individuals these effects are significant. With our approach, we hope we can provide valuable information in the mental health sphere and support the decision-making process in personalized interventions.

References

Ahn, J., Mukherjee, B., Banerjee, M., and Cooney, K. A. (2012). Bayesian Inference for the Stereotype Regression Model: Application to a Case-Control Study of Prostate Cancer. *Statistics in medicine*, 29(25):997–1003.

Anderson, J. A. (1984). Regression and Ordered Categorical Variables. *Journal of Royal Statistical Society*, 46:1–30.

Ando, T. (2007). Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika*, 94(2):443–458.

Berg, A., Meyer, R., and Yu, J. (2004). Deviance Information Criterion for Comparing Stochastic Volatility Models. *Journal of Business & Economic Statistics*, 22(1):107–120.

Bolger, N., Davis, A., and Rafaeli, E. (2003). Diary methods: capturing life as it is lived. *Annual review of psychology*, 54:579–616.

Carlbring, P., Andersson, G., Cuijpers, P., Riper, H., and Hedman-Lagerlöf, E. (2018). Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. *Cognitive Behaviour Therapy*, 47(1):1–18.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, 76(1):1–32.

Chai, T. and Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3):1247–1250.

Clasen, P. C., Fisher, A. J., and Beevers, C. G. (2015). Mood-reactive self-esteem and depression vulnerability: Person-specific symptom dynamics via smart phone assessment. *PLoS ONE*, 10(7):1–16.

Cohen, S. and Rodriguez, M. S. (1995). Pathways linking affective disturbances and physical disorders. *Health Psychology*, 14:374–380.

Curcio, G., Ferrara, M., and De Gennaro, L. (2006). Sleep loss, learning capacity and academic performance. *Sleep Medicine Reviews*, 10(5):323–337.

DeSteno, D. A. and Salovey, P. (1997). The effects of mood on the structure of the self-concept. *Cognition & Emotion*, 11(4):351–372.

Diefenbach, G. J., Mccarthy-larzelere, M. E., Williamson, D. A., Mathews, A., Manguno-mire, G. M., and Bentz, B. G. (2001). Anxiety, Depression, and the Content of Worries. *Depression and Anxiety*, 14:247–250.

Forrest, M. and Andersen, B. (1986). Ordinal scale and statistics in medical research. *Bmj*, 292(6519):537–538.

Frasure-Smith, N., Lespérance, F., Gravel, G., Masson, A., Juneau, M., Talajic, M., and Bourassa, M. G. (2000). Social support, depression, and mortality during the first year after myocardial infarction. *Circulation*, 101(16):1919–24.

Gable, S. L., Reis, H. T., and Elliot, a. J. (2000). Behavioral activation and inhibition in everyday life. *Journal of personality and social psychology*, 78(6):1135–1149.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016.

Graham, Z. (2009). *Better Than a Stick in the Eye: A Method for Resolving Conflicts and Bringing about Changes in Marriages, Families, and the Workplace*. AuthorHouse.

Gray, E. K. and Watson, D. (2002). General and specific traits of personality and their relation to sleep and academic performance. *Journal of Personality*, 70(2):177–206.

Gustavsson, A., Svensson, M., Jacobi, F., Allgulander, C., Alonso, J., Beghi, E., Dodel, R., Ekman, M., Faravelli, C., Fratiglioni, L., Gannon, B., Jones, D. H., Jennum, P., Jordanova, A., Jönsson, L., Karampampa, K., Knapp, M., Kobelt, G., Kurth, T., Lieb, R., Linde, M., Ljungcrantz, C., Maercker, A., Melin, B., Moscarelli, M., Musayev, A., Norwood, F., Preisig, M., Pugliatti, M., Rehm, J., Salvador-Carulla, L., Schlehofer, B., Simon, R., Steinhausen, H. C., Stovner, L. J., Vallat, J. M., den Bergh, P. V., van Os, J., Vos, P., Xu, W., Wittchen, H. U., Jönsson, B., and Olesen, J. (2011). Cost of disorders of the brain in Europe 2010. *European Neuropsychopharmacology*, 21(10):718–779.

Heatherton, T. F. and Polivy, J. (1991). Development and validation of a scale for measuring state self-esteem. *Journal of Personality and Social Psychology*, 60(6):895–910.

Hedeker, D. (2015). Methods for Multilevel Ordinal Data in Prevention Research. *Prev Sci*, 16(7):997–1006.

Hoyer, J., Becker, E. S., and Margraf, J. (2002). Generalized anxiety disorder and clinical worry episodes in young women. *Psychological Medicine*, 32(7):1227–1237.

Iida, M., Shrout, P. E., Laurenceau, J.-P., and Bolger, N. (2012). Using diary methods in psychological research. In Cooper, H., Camic, P. M., Long, D. L., Panter, A. T., Rindskopf, D., and Sher, K. J., editors, *APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics.*, pages 277–305. American Psychological Association, Washington.

Jakobsson, U. (2004). Statistical presentation and analysis of ordinal data in nursing research. *Scandinavian Journal of Caring Sciences*, 18(4):437–440.

Karyotaki, E., Riper, H., Twisk, J., Hoogendoorn, A., Kleiboer, A., Mira, A., MacKinnon, A., Meyer, B., Botella, C., Littlewood, E., Andersson, G., Christensen, H., Klein, J. P., Schröder, J., Bretón-López, J., Scheider, J., Griffiths, K., Farrer, L., Huibers, M. J., Phillips, R., Gilbody, S., Moritz, S., Berger, T., Pop, V., Spek, V., and Cuijpers, P. (2017). Efficacy of self-guided internet-based cognitive behavioral therapy in the treatment of depressive symptoms a meta-analysis of individual participant data. *JAMA Psychiatry*, 74(4):351–359.

Kleiboer, A., Smit, J., Bosmans, J., Ruwaard, J., Andersson, G., Topooco, N., Berger, T., Krieger, T., Botella, C., Baños, R., Chevreul, K., Araya, R., Cerga-Pashoja, A., Cieślak, R., Rogala, A., Vis, C., Draisma, S., van Schaik, A., Kemmeren, L., Ebert, D., Berking, M., Funk, B., Cuijpers, P., and Riper, H. (2016). European COMPARative Effectiveness research on blended Depression treatment versus treatment-as-usual (E-COMPARED): study protocol for a randomized controlled, non-inferiority trial in eight European countries. *Trials*, 17(1):387.

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. *International Joint Conference on Artificial Intelligence*, 14(12):1137–1143.

LaValley, M. P. and Felson, D. T. (2002). Statistical presentation and analysis of ordered categorical outcome data in rheumatology journals. *Arthritis and rheumatism*, 47(3):255–259.

Leger, D. (1994). The cost of sleep-related accidents: a report for the National Commission on Sleep Disorders Research. *Sleep*, 17(1):84–93.

Lemola, S., Räikkönen, K., Gomez, V., and Allemand, M. (2013). Optimism and self-esteem are related to sleep. Results from a large community-based sample. *International Journal of Behavioral Medicine*, 20(4):567–571.

Liu, X. (2014). Fitting Stereotype Logistic Regression Models for Ordinal Response Variables in Educational Research (Stata). *Journal of Modern Applied Statistical Methods*, 13(2):528–543.

Liu, X. and Koirala, H. (2012). Ordinal regression analysis: Using generalized ordinal logistic regression models to estimate educational data. *Journal of Modern Applied Statistical Methods*, 11(1):242–254.

Logan, D. E., Claar, R. L., and Scharff, L. (2008). Social desirability response bias and self-report of psychological distress in pediatric chronic pain patients. *Pain*, 136(3):366–372.

Long, J. S. (2014). Regression models for nominal and ordinal outcomes. In Best, H. and Wolf, C., editors, *The SAGE handbook of regression analysis and causal inference*, chapter 9, pages 173–204. SAGE Publications, London.

Manna, G., Falgares, G., Ingoglia, S., Como, M. R., and Santis, S. D. (2016). The relationship between self-esteem , depression and anxiety : Comparing vulnerability and scar model in the Italian context. *Mediterranean Journal of Clinical Psychology*, 4(3):1–17.

Martyn-Nemeth, P., Penckofer, S., Gulanick, M., Velsor-Friedrich, B., and Bryant, F. B. (2009). The relationships among self-esteem, stress, coping, eating behavior, and depressive mood in adolescents. *Research in Nursing and Health*, 32(1):96–109.

McCullagh, P. (1980). Regression Models for Ordinal Data. *Journal of the Royal Statistical Society*, 42(2):109–142.

McLachlan, G., Do, K. A., and Ambroise, C. (2005). *Analyzing Microarray Gene Expression Data*. Wiley Series in Probability and Statistics. Wiley.

Minden, S. L. (2000). Mood disorders in multiple sclerosis: diagnosis and treatment. *Journal of neurovirology*, 6(2):160–167.

Moskowitz, D. S. and Young, S. N. (2006). Ecological momentary assessment: what it is and why it is a method of the future in clinical psychopharmacology. *Journal of psychiatry & neuroscience : JPN*, 31(1):13–20.

Nezlek, J. B. (2003). Using Multilevel Random Coefficient Modeling to Analyze Social Interaction Diary Data. *Journal of Social and Personal Relationships*, 20(4):437–469.

Norusis, M. J. (2010). Ordinal Regression. In *PASW Statistics 18.0 Advanced Statistical Procedures Companion*, chapter 4, pages 69–89. Prentice Hall.

Paradise, A. and Kernis, M. H. (2002). Self-esteem and Psychological Well-being: Implications of Fragile Self-esteem. *Journal of Social and Clinical Psychology*, 21:345–361.

Plummer, M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics*, 9(3):523–539.

Pressmann, S., Matthews, K., Cohen, S., Martire, L., Scheier, M., Baum, A., and Schulz, R. (2009). Association of Enjoyable Leisure Activities With Psychological and Physical Well-Being. *Psychosom Med*, 71(7):725–732.

Richards, D. and Richardson, T. (2012). Computer-based psychological treatments for depression: a systematic review and meta-analysis. *Clinical psychology review*, 32(4):329–342.

Robins, R., Hendin, H., and Trzesniewski, K. H. (2001). Measuring Global Self-Esteem: Construct Validation of a Single-Item Measure and the Rosenberg Self-Esteem Scale. *Pers Soc Psychol Bull*, 27:151–161.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton University Press, Princeton, NJ.

Saddichha, S., Al-Desouki, M., Lamia, A., Linden, I. A., and Krausz, M. (2014). Online interventions for depression and anxiety - a systematic review. *Health psychology and behavioral medicine*, 2(1):841–881.

Silvera, D. H., Bergersen, T. D., Bjørgum, L., Perry, J. A., Rosenvinge, J. H., and Holte, A. (1998). Analyzing the relation between self-esteem and eating disorders: differential effects of self-liking and self-competence. *Eating and weight disorders : EWD*, 3(2):95–99.

Silverman, B. G., Hanrahan, N., Bharathy, G., Gordon, K., and Johnson, D. (2015). A systems approach to healthcare: agent-based modeling, community mental health, and population well-being. *Artificial Intelligence in Medicine*, 63(2):61–71.

Sowislo, J. F. and Orth, U. (2013). Does Low Self-Esteem Predict Depression and Anxiety? A Meta-Analysis of Longitudinal Studies. *Psychological bulletin*, 139(1):213–240.

Spiegelhalter, D., Best, N. G., and Carlin, B. P. (1998). Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. Technical Report , MRC Biostastatistics Unit, Cambridge.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 64(4):583–616.

Steiger, A., Allemand, M., Robins, R., and Fend, H. (2014). Low and Decreasing Self-Esteem During Adolescence Predict Adult Depression Two Decades Later. *Journal of Personality and Social Psychology*, 106(2):325–38.

Steiger, A., Fend, H., and Allemand, M. (2015). Testing the Vulnerability and Scar Models of Self-Esteem and Depressive Symptoms from Adolescence to Middle Adulthood and Across Generations. *Developmental Psychology*, 51(2):236–247.

Titzler, I., Saruhanjan, K., Berking, M., Riper, H., and Ebert, D. D. (2018). Barriers and facilitators for the implementation of blended psychotherapy for depression: A qualitative pilot study of therapists' perspective. *Internet Interventions*.

Trzesniewski, K. H., Donnellan, M. B., Moffitt, T. E., Robins, R. W., Poulton, R., and Caspi, A. (2006). Low self-esteem during adolescence predicts poor health, criminal behavior, and limited economic prospects during adulthood. *Developmental psychology*, 42(2):381–90.

Vehtari, A., Gelman, A., and Gabry, J. (2016). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, (July):1–20.

Waegeman, W., De Baets, B., and Boullart, L. (2008). ROC analysis in ordinal regression learning. *Pattern Recognition Letters*, 29(1):1–9.

Watanabe, S. (2010). Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. *Journal of Machine Learning Research*, 11:3571–3594.

Weinstein, S. M. and Mermelstein, R. (2007). Relations between daily activities and adolescent mood: the role of autonomy. *Journal of Clinical Child and Adolescent Psychology*, 36(2):182–194.

Willmott, C. J. and Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30(1):79–82.

Willmott, C. J., Matsuura, K., and Robeson, S. M. (2009). Ambiguities inherent in sums-of-squares-based error statistics. *Atmospheric Environment*, 43(3):749–752.

CHAPTER 5

DEVELOPING A PROCESS FOR THE ANALYSIS OF USER JOURNEYS AND THE PREDICTION OF DROPOUT IN DIGITAL HEALTH INTERVENTIONS: MACHINE LEARNING APPROACH

Bremer, V., Chow, P., Funk, B., Thorndike, F., and Ritterband, L. (2020). Developing a Process for the Analysis of User Journeys and the Prediction of Dropout in Digital Health Interventions: Machine Learning Approach. Journal of Medical Internet Research, 22(10):e17738.

Abstract

Background: User dropout is a widespread concern in the delivery and evaluation of digital health (i.e., web- and mobile application) interventions. Researchers have yet to fully realize the potential of the large amount of data generated by these technology-based programs. Of particular interest is the ability to predict who will drop out of an intervention. This may be possible through the analysis of user journey data – self-reported as well as system generated data produced by the path (or journey) an individual takes to navigate through a digital health intervention.

Objective: The purpose of this study is to provide a step-by-step process for the analysis of user journey data and eventually to predict dropout in the context of digital health interventions. The process is applied to data of an Internet-based intervention for insomnia as a way to illustrate its use. The completion of the program is contingent upon completing 7 sequential cores, which includes an initial tutorial core. Dropout is defined as not completing the 7th core.

Methods: Steps of user journey analysis, including data transformation, feature engineering, and statistical model analysis and evaluation, are presented. Dropouts were predicted based

on data of 151 participants from a fully automated web-based program (SHUTi) that delivers cognitive behavioral therapy for insomnia. Logistic regression with L1 and L2 regularization, support vector machines, and boosted decision trees were used and evaluated based on their predictive performance. Relevant features from the data are reported that predict user dropout.

Results: Accuracy of predicting dropout (AUC values) varied depending on the program core and the machine learning technique. After model evaluation, boosted decision trees achieved AUC values ranging between 0.6-0.9. Additional handcrafted features, including time to complete certain steps of the intervention, time to get out of bed, and days since last interaction with the system, contributed to the prediction performance.

Conclusions: The results support the feasibility and potential of analyzing user journey data in order to predict dropout. Theory driven handcrafted features increased prediction performance. The ability to predict dropout on an individual level could be used to enhance decision-making for researchers and clinicians as well as inform dynamic intervention regimens.

5.1 Introduction

The efficacy of digital (i.e., Internet, web, mobile) behavioral interventions to improve a range of health-related outcomes has been well documented (Carlbring et al., 2018; Erbe et al., 2017; Saddichha et al., 2014). However, adherence to these interventions is a significant issue (Melville et al., 2010). Intervention dropout, defined as when a participant prematurely discontinues a program, to Internet-based treatments for psychological disorders typically vary between 30-50% (Horsch et al., 2015; Melville et al., 2010; Torous et al., 2020). However, the reason for such high dropout rates is still unclear (Torous et al., 2020), whereas longer treatment duration and user engagement appear to be associated with improved treatment outcomes and greater effectiveness of the digital intervention (Alkhaldi et al., 2015; Funk et al., 2010; Vandelanotte et al., 2007; Wickwire, 2019). Furthermore, in a research setting, high dropout rates and, consequently, low exposure to the digital content, might affect the reported effects of a digital intervention and the validity of the results (Brouwer et al., 2011; Geraghty et al., 2010). While researchers have highlighted the need for a science of user attrition (Eysenbach, 2005), there have been few advances in predicting dropout through advanced quantitative approaches in eHealth interventions (Pedersen et al., 2019). In particular, prior work has identified hypothetical factors influencing attrition in eHealth programs, such as ease of leaving the intervention, unrealistic expectations on behalf of users, usability and interface issues, and amount of workload required to benefit from an intervention (Eysenbach, 2005).

Such factors are likely to impact how a user ultimately engages with a program and could provide indicators for predictive factors but do little to advance predictive modeling of dropout when not applied in data-driven studies. Research suggests that an increased completion of modules in digital therapeutics increases treatment outcomes (Donkin et al., 2011). Identifying

those patients that are likely to drop out of treatment and addressing the related issues can, thus, improve treatment outcomes and can be the basis of the development of micro-interventions that target these high-risk participants in order to reengage them to complete the program (Fernandez-Alvarez et al., 2017). Thus, predicting dropout on a participant-level supports the decision-making of experts in the target field and consequently leads to more personalized treatment strategies. In addition, inferential results can increase insight into the causes of attrition by revealing data-driven indicators. Participant-specific factors can help to identify individuals that benefit more from digital therapies compared to individuals for whom face-to-face treatment might be a better approach. To evaluate the possibility of predicting dropout in digital interventions and to shed light on some indicators of dropout, the aim of the current investigation is to propose a process for user journey analysis to predict dropout from a digital intervention.

A wealth of data can be collected through the use of digital interventions. They often feature content that is administered over time as users complete tasks or components of the intervention, typically over several weeks or months (Christensen et al., 2016; Murray et al., 2016; Ritterband et al., 2009a,b). Digital interventions also track and log different types of user interactions (e.g., frequency of logins). These data provide a nuanced understanding of a participants' usage behavior over the course of an intervention (Iida et al., 2012). Combined with self-reported data, passively collected user data could be captured and used to provide deeper insight about how likely users are to drop out of an intervention on an individual level and lead to increased prediction performance.

A user journey is a sequence of interactions as an individual uses a digital intervention (i.e., the path an individual takes to navigate through a program). While user journeys are well known and established in the field of online marketing, its direct application to digital health interventions, to the best of our knowledge, has not yet been examined. Online marketers leverage user journeys to collect information about an individuals' behavior (Nottorf et al., 2012), often referred to as clickstream data analysis (Chatterjee et al., 2003; Stange and Funk, 2015). This increases the understanding of users' behavior by recognizing patterns in their sequence of actions. Thus, user journey analysis can reveal insight into an individuals' behavior by enabling an analysis of data (e.g., EMA or log data) that is not frequently used in the e-Health sphere (van Breda et al., 2016).

There are likely several reasons why analysis of user journeys has not achieved prominence in digital health interventions. One obstacle lies in the analysis of large amounts of raw data. Analysis of user journeys often requires transformation of raw data, feature engineering, and the application of machine learning techniques, which can be a burdensome process (Sen et al., 2006) and is not a typical skill set of eHealth behavior researchers. While user journeys have been used to predict different psychological factors such as mood, stress levels, or treatment outcomes and costs, (Becker et al., 2016; Bremer et al., 2018; Jaques et al., 2017; van Breda et al., 2018; Van Breda et al., 2016; van Breda et al., 2016) , to our knowledge, no work has provided steps to be

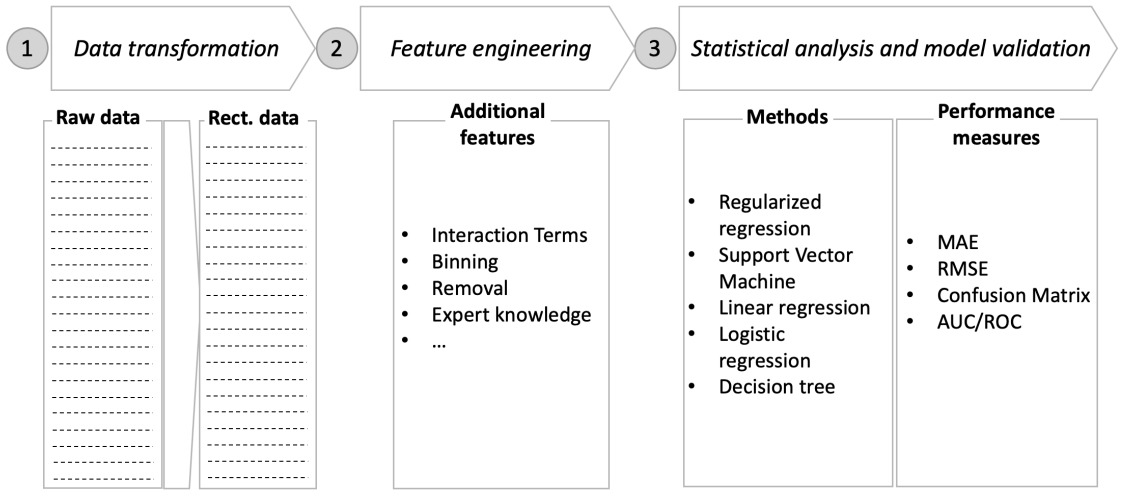


Figure 5.1: Process of analysis.

taken to analyze raw user journey data and, at the same time, predict user dropout of a digital health intervention.

The overarching goal of the current study is to establish and provide a step-by-step process describing how to leverage user journeys to predict various behaviors (e.g., dropout). This process involves several steps, including creating the basic data structure for handling user journeys, creating features that can add additional information to the existing raw data, and ultimately providing a framework for the statistical analysis. A technical implementation (R package) (Bremer, 2018; R Core Team, 2018) of this process is provided for the research community. To demonstrate the application and potential utility of this process, we use it to predict user dropout in a randomized controlled trial of a fully automated cognitive behavior therapy intervention for insomnia (Sleep Healthy Using the Internet [SHUTi]) (Gosling et al., 2014).

5.2 Methods

5.2.1 User journey process

The overarching steps of the user journey process are outlined in Figure 5.1. This process applies machine learning algorithms, specifically supervised learning, which is used when both input (e.g., logins, mood symptoms) and output data (e.g., dropout status) exist in the dataset (Kotsiantis, 2007).

It is important for researchers to clearly define the outcome variable of interest. As dependent variables can take on different measurement scales (e.g. discrete or continuous), defining the target variable has consequences for the choice of statistical models. When predicting discrete outcomes (i.e., consisting of at least two discrete categories or labels), classification is often the

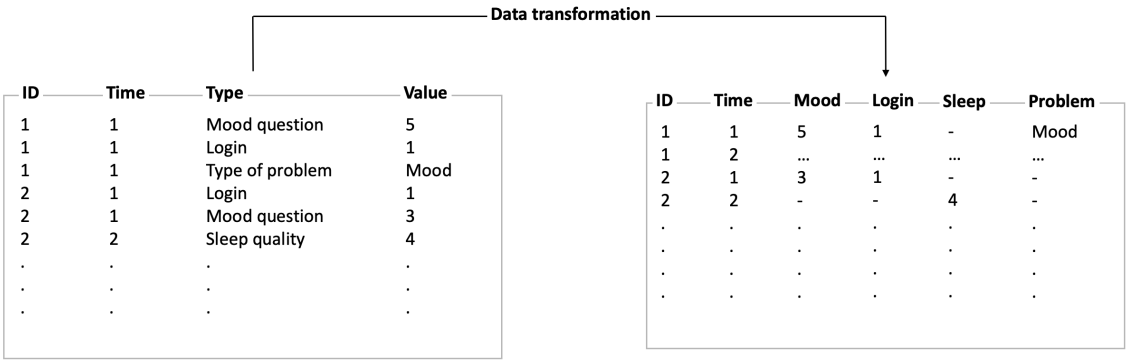


Figure 5.2: Example of data transformation in the context of digital health interventions.

appropriate approach. On the other hand, when predicting continuous outcome variables, the learning task is regression.

5.2.1.1 Step one: Data transformation

The first step to analyzing user journey data is to transform the raw data into a wide format, as can be seen in Figure 5.2. Thus, the transformed data are structured such that each row corresponds to a unique observation in "Time" for a particular user ('ID').

When transforming the raw data, it is important to specify the time window defining the time interval for which individual touchpoints are aggregated. The choice of the time window depends on the density of the observations in the raw data. For example, if a raw dataset is composed of a few touchpoints over the course of a day, choosing a time window on a scale of days avoids sparseness of the transformed data matrix. In contrast, when predicting purchases in online marketing, for example, a large number of observations exists for each user on short timescales. Here, choosing a small window (e.g., an hour) could be beneficial since the resulting matrix will not be sparse and information loss is minimal. In an Internet-based intervention, however, it is not unusual for self-reported data to be collected as little as once a day, with a user only logging into the system perhaps a few times a day. In this case, it would not make sense to choose an hour-long window because the resulting matrix would be very sparse. Thus, choosing a time window on a scale of days would be a better choice.

If multiple observations of the same type occur within a time window, one must decide how to aggregate these values. For some variables, such as diary entries, taking an average may be desirable; for other variables, such as logins, the sum is a more appropriate aggregation. The provided technical framework supports the procedure of data transformation. In addition, missing values often exist in the data. There are various procedures that can handle missing values. One might remove all rows that include missing values, however, this can lead to a reduction in observations. Other possibilities are imputation procedures such as using aggregated values of

these features or developing statistical models that predict the missing values based on other features. For more information on missing values, we refer to Batista and Monard (2003).

5.2.1.2 Step two: Feature engineering

Feature engineering can be described as the process of including additional variables into the data with the intention to achieve increased predictive performance. As statistical learning relies heavily on the input data, this step is important for improving accuracy of prediction (Domingos, 2012). There are two approaches to feature engineering: handcrafted or automated. Handcrafted feature engineering is a challenging task and requires human effort and domain knowledge. It is, therefore, appropriate for researchers with expertise in the domain that is represented by the data (e.g., sleep) to be highly involved in the process (Kanter and Veeramachaneni, 2015; Khurana et al., 2016; Lam et al., 2017). A clear understanding of the problem to be solved is necessary in order to derive meaningful features (Lam et al., 2017). Handcrafted feature engineering often involves a trial-and-error phase to experiment with different features (Domingos, 2012). Automated feature engineering involves the generation of candidate features that are evaluated based on their predictive performance. Tools exist for the application of automated feature engineering in different domains, such as natural language processing or machine vision (Cheng et al., 2011; Kanter and Veeramachaneni, 2015; Lu et al., 2014).

Interaction terms, i.e. the product of two original features, can lead to additional knowledge about their relationships and increased predictive accuracy. The provided technical framework supports generating them. In case of a large number of original features, however, including interaction terms results in many additional features.

Additionally, time-window based aggregation methods can be beneficial in terms of predictive performance in the context of digital health interventions (Van Breda et al., 2016). Here, based on a user specified time window w, various types of aggregations are performed on the original features. Figure 5.3 represents the process of this task through the exemplification of self-reported EMA data. The "Mood" level is reported by an individual at different points in time ("Time steps"). For the creation of the aggregated features, a time-window of w=3 is specified in this example. Various statistical measures, such as the sum (`"Mood_sum"`), mean (`"Mood_mean"`), minimum, maximum, and standard deviation (not shown in figure) are calculated for three consecutive measurements of the mood level (w=3) and included as additional features in the dataset. It should be noted that the creation of features can limit one's ability to reproduce study results if the feature engineering process is not well documented or if the dataset changes over time. For the case study in this paper, we created various theory driven features based on expert knowledge, which will be introduced in a later section.

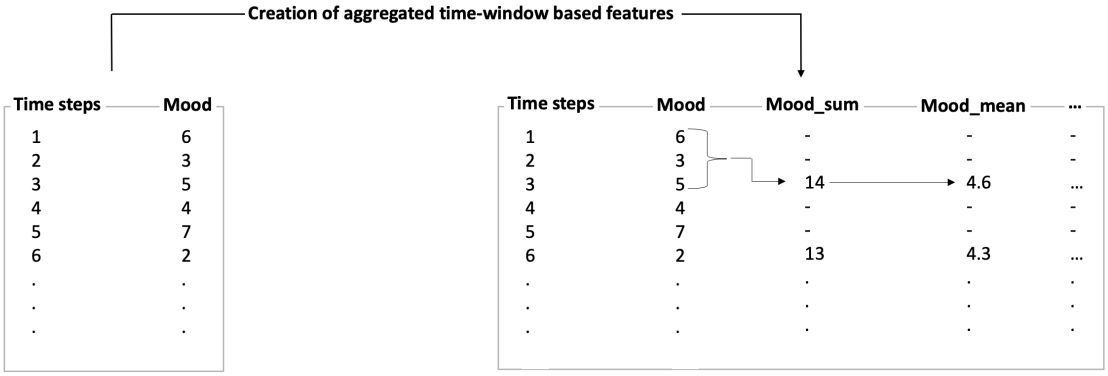


Figure 5.3: Example of creating aggregated time-window based features for w=3.

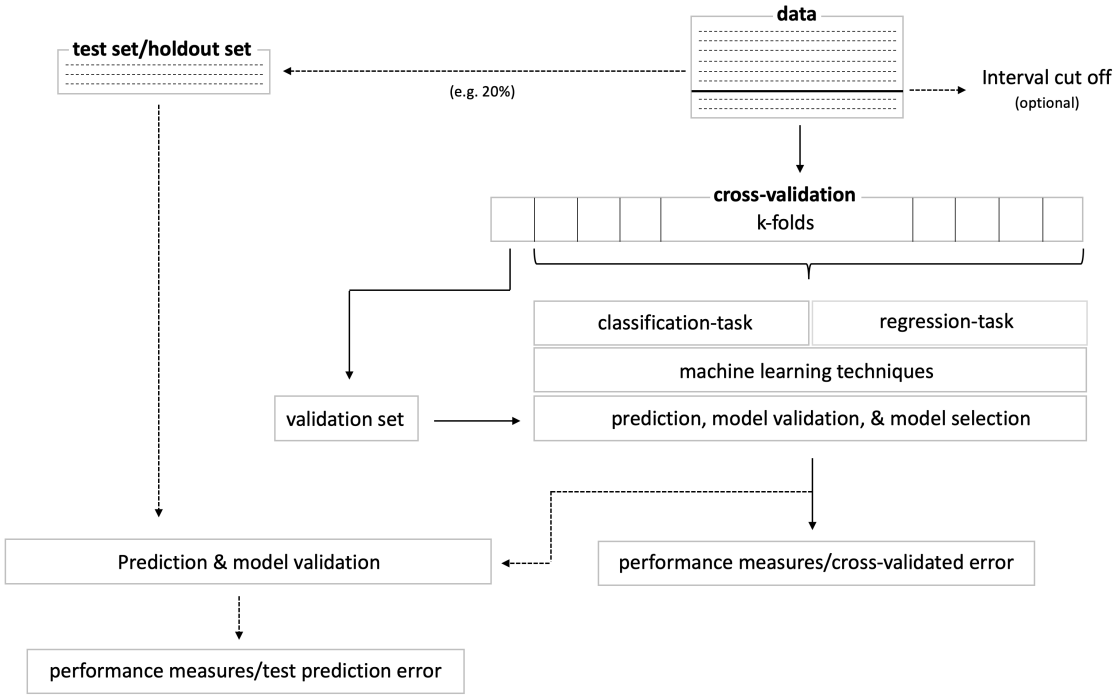


Figure 5.4: Procedure of statistical analysis.

5.2.1.3 Step three: Statistical analysis and model validation

The next step of analyzing user journey data is the application of machine learning techniques in order to predict the outcome variable. Figure 5.4 depicts this procedure. First, the dataset can be split into a training set for fitting the data and learning patterns and a test (or holdout) set. This test set is usually created if sufficient data are available. It is subsequently used for testing the final model performance of the selected algorithm. It is difficult, however, to quantify "sufficient data" as it depends strongly on the field of research, applied models, and structure of the data.

Depending on the task to be analyzed, the data can be further split based on particular points

in time. If the aim of the analysis, for example, is the prediction of the outcome of an intervention, it might be useful to evaluate at what point in time the predictive accuracy is at its peak. The longer the time-window, the higher the predictive accuracy can be assumed because more data is available. Thus, using time windows and basing the amount of usable data on these windows ("Interval cut off") can be useful in evaluating the feasibility of prediction.

There is a large number of machine learning techniques that can be applied to user journey data; some models can be applied to both learning tasks (classification or regression), such as support vector machines or decision trees, whereas others fit better for a specific task (i.e. logistic regression for classification). Researchers may wish to compare their predictive performance to justify the model selection. To gauge the predictive performance of a specified model, cross-validation is often applied. Here, the data are divided into k chunks where k-1 chunks are used for training the machine learning techniques and the remaining data chunk is used for predicting the target variable. This procedure is repeated k times until each chunk has been used as a validation set. Ultimately, the model with the best performance is selected for the specified learning task. If a holdout set was maintained, the specified model is then trained based on all data. The target variable in the holdout set is then predicted and evaluated, which leads to the test prediction error.

Model validation checks the ability of a particular model to either fit the data or predict the outcome variable (Marcus and Elias, 1998). Eventually, the one with the best performance is selected. Non-validation can lead to inaccurate predictions and thus overconfidence in the developed model (Arboretti and Salmaso, 2003). Model validation should generally be executed on the validation set for each iteration of the cross-validation procedure (cross-validated prediction error) in order to select the best model, and, subsequently, on an independent test set that was set aside earlier (test prediction error). In some cases, especially when not enough data is available, no independent test set is put aside and only the cross-validated error is reported, which can lead to an optimistic estimation of the error (Arboretti and Salmaso, 2003).

Deciding on the method of model validation also depends on the learning task. For regression, criteria such as the root-mean-square error (RMSE) or mean absolute error (MAE) are often appropriate. For the classification task, confusion matrices and receiver operating characteristics (ROC) graphs are often used as performance indicators. More information about these validation procedures and their application can be found elsewhere (Fawcett, 2006).

In the provided technical framework, logistic regression, linear regression, support vector machines, boosted decision trees, and regularization techniques are implemented. Since over-fitting can occur when utilizing a large number of features (Domingos, 2012), and some types of statistical procedures (e.g., linear regression) cannot be applied when the number of features is greater than the number of observations, alternative techniques such as regularization and feature selection may need to be used (Tibshirani, 1996). A thorough review of these techniques is outside the scope of this paper, and readers are strongly encouraged to learn more about each

of these techniques and how they pertain to their data and aims.

5.2.2 Case study

To illustrate the user journey analysis process, data were extracted from a trial of an online program (SHUTi) that is based on cognitive behavioral therapy for insomnia (CBT-I) (Thorndike et al., 2008). SHUTi is a fully automated web-delivered program that is tailored to individual users (Thorndike et al., 2008) and informed by the Model for Internet Interventions (Ritterband et al., 2009b). SHUTi is based on the primary principles of face-to-face CBT-I, including sleep restriction, stimulus control, cognitive restructuring, sleep hygiene, and relapse prevention. SHUTi contains 7 "cores" that are dispensed over time, the first core being a tutorial on how to use the program, with new cores becoming available seven days after completion of a previous core. This format was meant to mirror traditional CBT-I delivery procedures using a weekly session format. SHUTi has been found to be more efficacious than online patient education in changing primary sleep outcomes (insomnia severity, sleep onset latency, wake after sleep onset), with the majority of SHUTi users achieving insomnia remission status one year later (Ritterband et al., 2017). Thus, the efficacy of SHUTi is well established. However, similar to other digital interventions, predicting user dropout is an important yet unaddressed issue. Thus, the primary aim of this case study is to demonstrate the feasibility of predicting user dropout from data generated by a digital health intervention.

The current sample was drawn from a trial consisting of 303 participants (72% female) between the ages of 21 and 65 (Mean=43.3, SD=11.6). They were 84% White, 7% Black, 4% Asian, and 5% "other." Participants were randomly assigned (using a random number generator) to receive SHUTi or online patient education (control condition). The study was approved by the local university Institutional Review Board and the project is registered on clinicaltrials.gov (NCT01438697). Inclusionary and exclusionary criteria as well as outcomes are reported in detail elsewhere (Ritterband et al., 2017).

Data from 151 participants who were assigned to SHUTi were utilized in this paper. Both self-reported and system generated types of data are available. Participants completed a battery of self-report measures at baseline and post-intervention. A list and detailed description of the measures have been published previously (Ritterband et al., 2017) and can also be found on the clinicaltrials.gov registration page. Sleep diaries were also collected throughout the intervention period, and capture information about bedtime, length of sleep onset, number and duration of awakenings, perceived sleep quality, and arising time. Data was collected prospectively for 10 days (during a 2-week period) at each of the four assessment periods (pre- and post-intervention, and 6 and 12 month follow-ups). Sleep diary questions mirrored those from the Consensus Sleep Diary (Carney et al., 2012). Values for sleep onset latency (SOL) and wake after sleep onset (WASO) were averaged across the 10 days of diary collection at each assessment period. System generated data included individual logins and automated emails sent by the system as well as

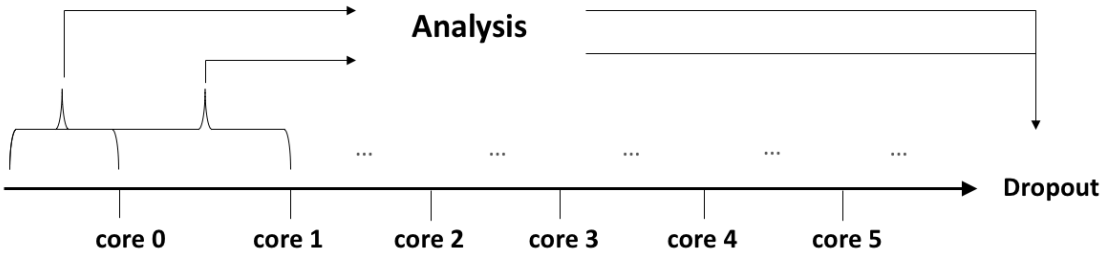


Figure 5.5: Setup of analysis for dropout prediction.

trigger events logged in the system. All data was utilized to predict user dropout, defined as not completing all 7 SHUTi cores (Core 0 through Core 6). Thus, users were classified as having dropped out or not. As noted elsewhere (Ritterband et al., 2017), 60.3% (91 of 151) participants completed all 7 cores in the SHUTi program.

5.3 Results

The primary aim was to predict whether users prematurely dropped out of SHUTi (dropped out by Core 6/completed Core 6). Therefore, the learning problem is a binary classification (drop out/did not drop out). In order to verify at what point in time of the intervention the machine learning techniques were capable of predicting dropout, separate analyses were executed after the completion of each core (Figure 5.5) and only included data up to the core in question. The number of participants included in each analysis was 146, 141, 133, 116, 102, and 101, for cores 0-5, respectively.

5.3.1 Data transformation

As a first step, the raw data was transformed into a rectangular data matrix (wide format), which led to 981 basic features. Basic features are those features that were already included in the raw data. As an example, see column "Type" in Figure 5.2. Additionally, twenty-five handcrafted and theory driven features that were derived from the raw data were implemented. These features are introduced in the next section. In total, 1006 features were used for the analyses. Whenever the same question (i.e. in the case of diary data) was administered multiple times a day, the mean of the reported values was chosen for numeric data and the mode for categorical data. To reduce the sparseness of the resulting data matrix, reported values for questionnaires such as the Insomnia Severity Index were repeated for each participant until the next occurrence of the questionnaire (this questionnaire was administered before each core). In order to address the issue of missing data, features were deleted based on their quantity of missing data. To evaluate how the deletion affects the predictive performance of the models, features were deleted that contained more than 5% (i), 10% (ii), 15% (iii), and 20% (iv) of missing values. This procedure

reduced the number of features tremendously. Additionally, categorical variables that only had one level or category were removed. Less data is available for the analysis at time point Core 0 compared to time point Core 5. Thus, the number of features for each level of missing data was 83 (i), 263 (ii), 299 (iii), and 401 (iv) features.

Because the aim of this study was to predict dropout at Core 6, each participant only had exactly one outcome value – they could either complete Core 6 or not. Users that dropped out between Cores 1-5 would be classified as having dropped out at Core 6. Therefore, the user journey data needed to be aggregated for each user. For most of the variables, the mean and mode were used as aggregation method. However, for some variables, such as login information or number of days since last contact, the sum is more appropriate. Table 5.1 illustrates the different aggregation procedures and the corresponding features. Features that are not listed were aggregated by mean and mode. The rest of the missing data were imputed using the median for numeric variables and mode for categorical features. Additionally, an imputation based on the k-nearest neighbor (KNN) algorithm was applied (k=5). Both approaches were used in order to reveal which led to a better prediction performance.

5.3.2 Feature engineering

There were twenty-five theory–driven features implemented for this case study. Some of these features, shown in Table 5.1, were handcrafted and some were already existing in the dataset. Specifically, the handcrafted features were computed from the raw data and were deemed as useful for model prediction. Few of these features are study-specific (e.g., *if the participant finished homework in Core 2*) whereas others could be used in any type of digital intervention (e.g., *if the participant logged in*). Because the number of features generated from the study data was already large, none of the generic feature generation methods were used. These twenty-five features were not deleted based on the missing value ratio (see above) because there was a clinical or theory-driven rationale that they would influence prediction performance.

5.3.3 Statistical analysis and model validation

For the learning task, a set of machine learning techniques were used in order to select the model with the best prediction performance. Specifically, support vector machines, boosted decision trees, and logistic regression with L1 and L2 regularization were applied. The optimal parameters were found by a grid-based search and cross-validation. Additionally, stratified 10-fold cross-validation was used for each analysis. In order to choose an appropriate statistical model, a heat map was created to illustrate the average area under the curve (AUC) across all core analyses for each model, imputation procedure, and threshold for percentage of missing values (Figure 5.6). As can be seen, the method of imputing the missing values did not have a strong influence on the performance of the applied statistical model. Increasing the percentage threshold influenced L1 regularization and SVM negatively whereas L2 regularization and boosted decision trees seemed

Feature aggregation method	Handcrafted features	Existing clinically important features
Sum: The sum of all observations of a specific feature for an individual	- Days since last contact (any interaction) - If sleeping duration is decreasing from Core to Core - If sleep window duration is 5 or 8 hours	- If the participant had an alcoholic drink that day - If the participant took a nap - If the system recorded a triggered event that day - If the participant logged in that day - If the system sent an email that day
Last: The last observation of a specific feature for an individual	- Difference between preferred arising time in Core 2 and Core 3 - If preferred arising time is greater than 8am in Core 2 - Average time in days to complete a Core among all Cores that have been available - Time needed in days to complete a Core in days (6 features for Core 0-5)	- If the participant finished homework in Core 2 - Number of days where no diaries have been completed in the time period of analysis - Precipitating factor includes "major life event" or "Health/Psychological"
Mean: Mean of the observations of a specific feature for an individual	- Difference between awake and arise time - Difference between preferred arise time and actual arise time (am/pm) - Difference between preferred arise time and actual arise time (minutes) - Difference between preferred bed time and actual bed time	- Naptime in minutes

Table 5.1: Aggregation of theory determined features.

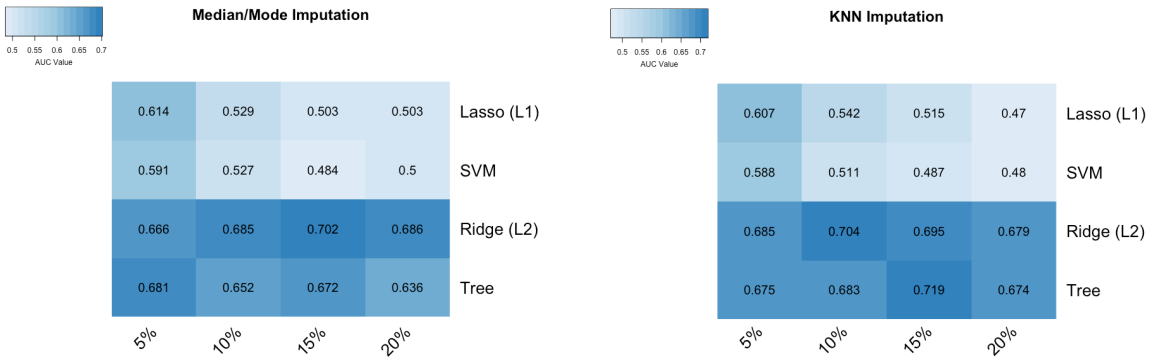


Figure 5.6: Heat map of averaged AUC values across core analyses for each model, imputation procedure, and threshold for percentage of missing values.

to not be influenced tremendously. The best average AUC value (.719) was achieved by applying boosted decision trees, deleting each feature that contained more than 15% of missing values, and imputing the rest of the missing values by KNN.

Figure 5.7 illustrates the ROC curves for each core analysis using the specified parameters. With the exception of Core 4, the AUC values increased with each analysis. For each Core, the predictions were better than random indicated by AUC values above 0.5. Generally, the AUC values ranged between 0.6 and 0.9. Importantly, the prediction of dropout appears feasible early in the intervention period (i.e., Cores 1 and Core 2). Additionally, the area under the precision-recall curve was computed (PRAUC). Across all core analyses, a PRAUC of 0.48 was observed while chance had an average of 0.24. Thus, the model performs better than chance.

Boosted decision trees were used to identify important features. Here, Shapley additive explanation values (SHAP) were used (Lundberg and Lee, 2017). SHAP values are a relatively new concept in the field of machine learning and essentially represent the importance of each feature and their contribution to the prediction by comparing the prediction of the model with and without a specified feature value depending on the order of their introduction to the model. In addition to the importance of each feature, SHAP values quantify how features contribute to the prediction of the model.

Figure 5.8 includes the five most important features according to the boosted decision trees for each core analysis. In each graph, the x-axis represents the values for each feature and the y-axis represents the SHAP values (i.e., the effect each feature has on predicting the completion of Core 6 of the intervention). In the Core 0 analysis, for example, finishing Core 0 within three days (x-Axis) has a positive influence on dropout as can be seen on the y-Axis above zero. However, taking more time to complete Core 0 (where x-Axis is greater 3) influences dropout prediction negatively as the graph approaches values under zero.

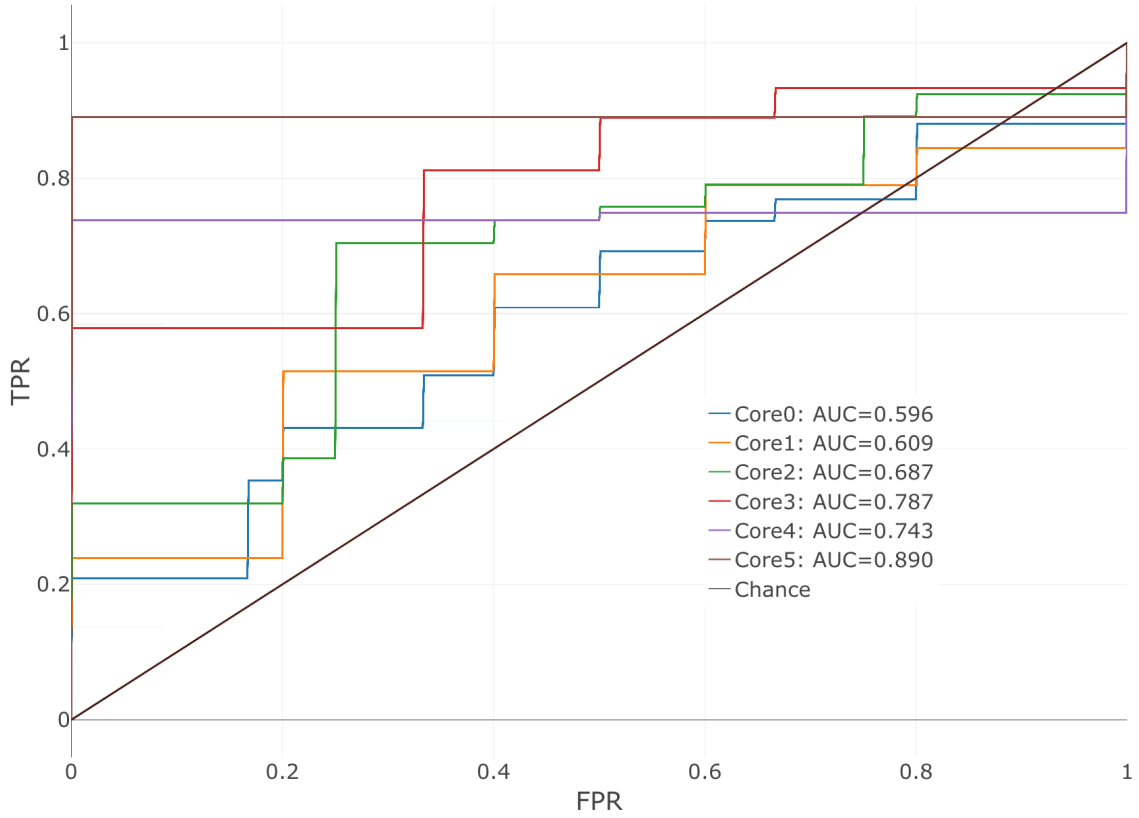
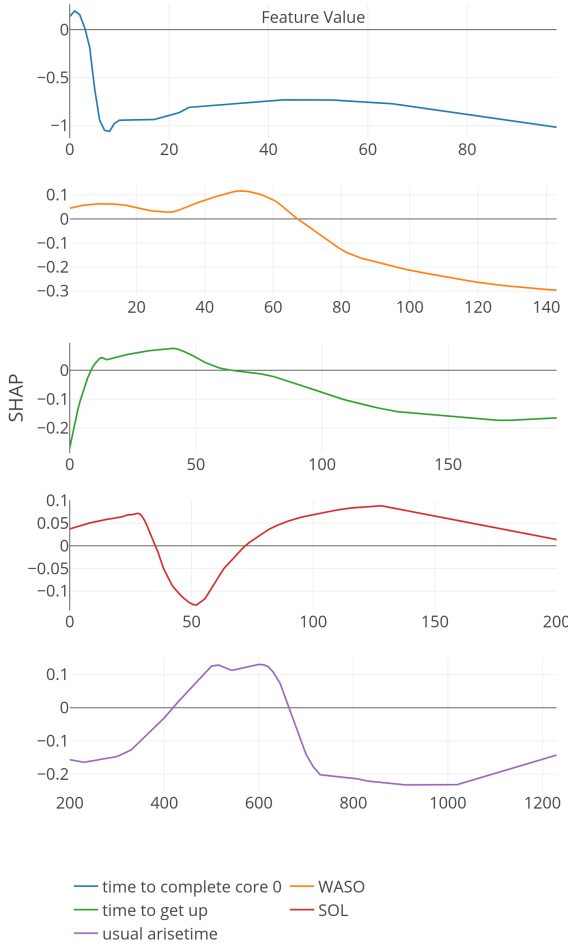
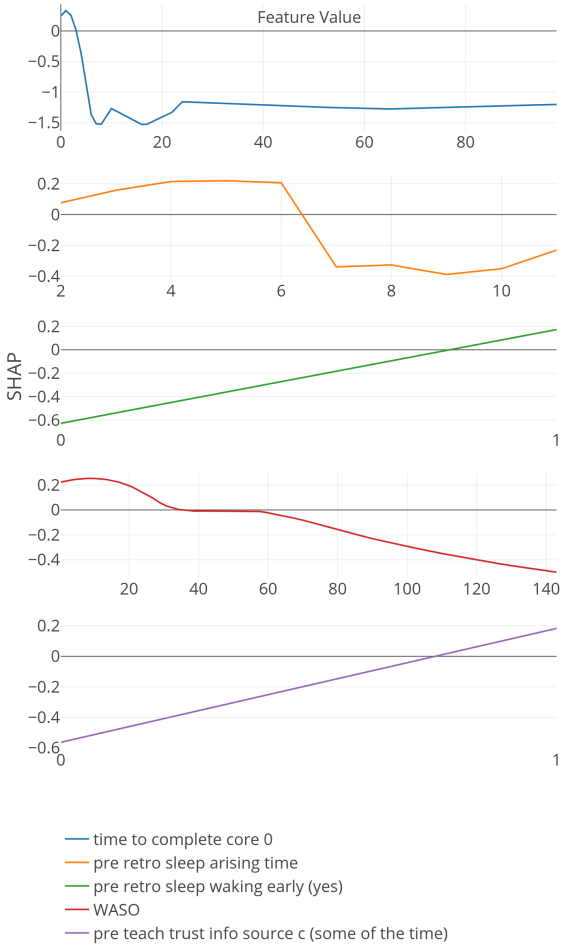


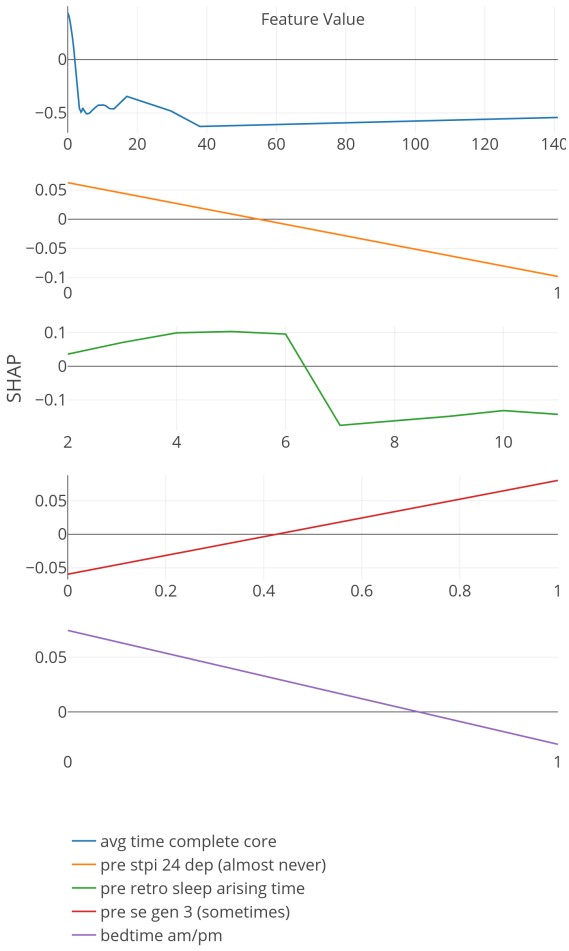
Figure 5.7: ROC for each core analysis based on boosted decision trees (15% missing value deletion, KNN imputation).



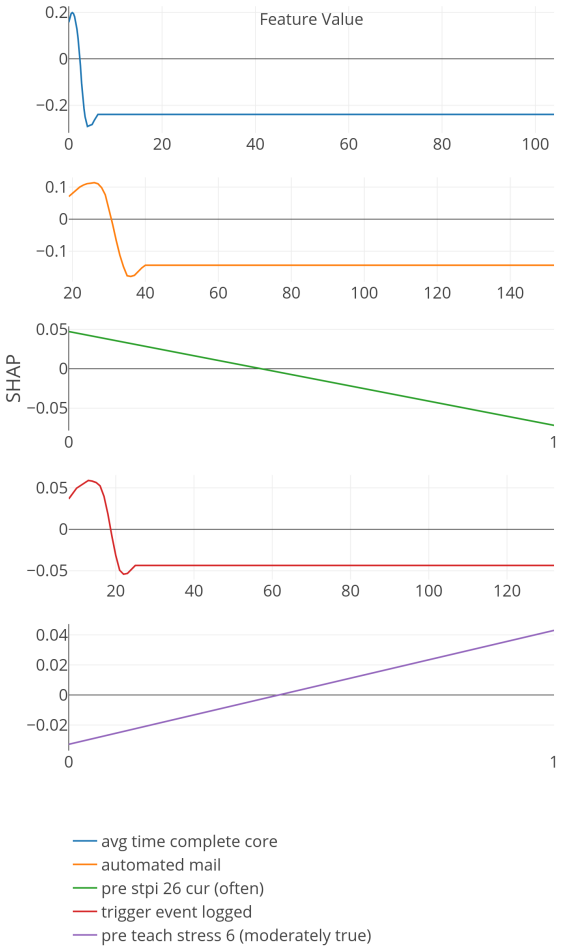
Core 0 Analysis



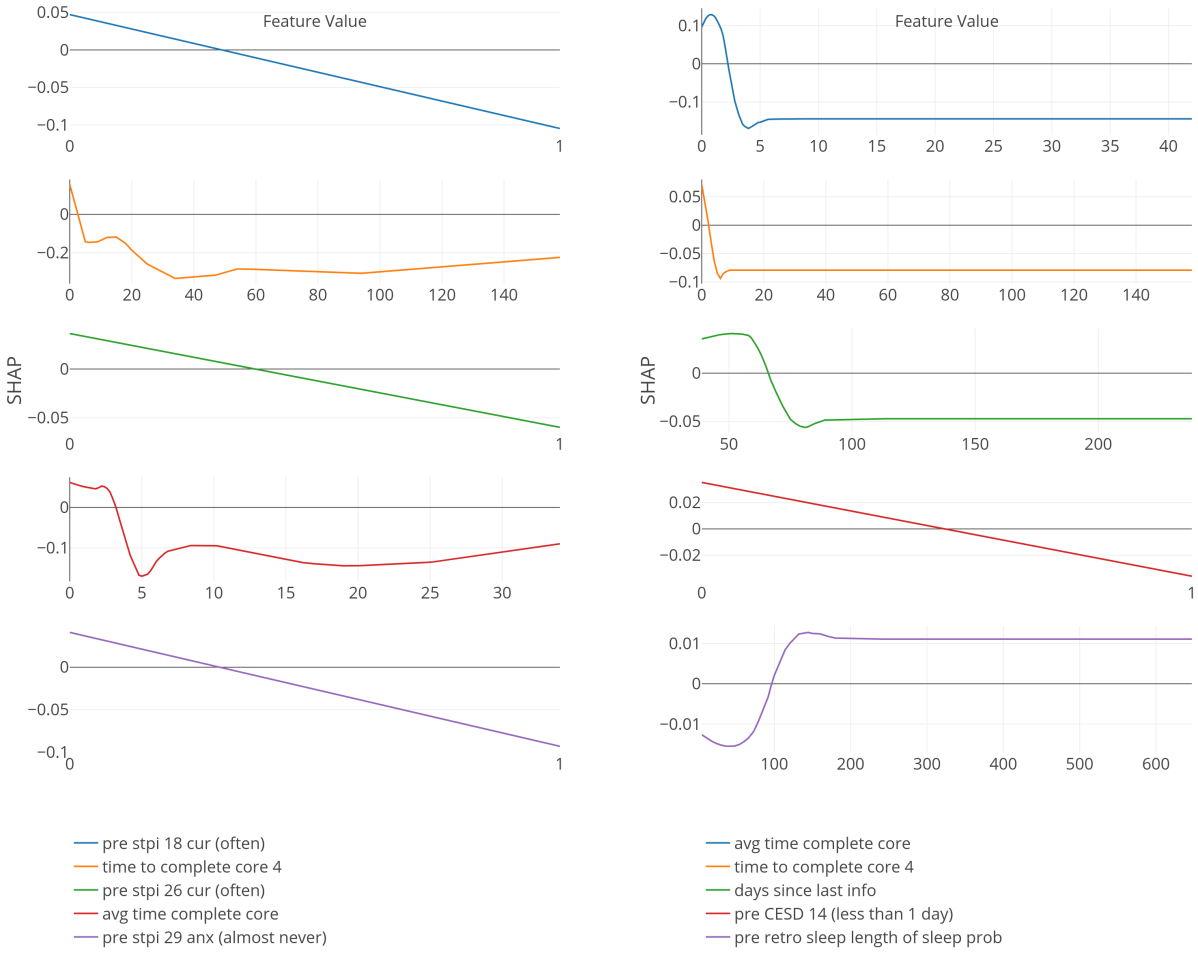
Core 1 Analysis



Core 2 Analysis



Core 3 Analysis



Core 4 Analysis

Core 5 Analysis

Figure 5.8: Five most important features for each core analysis according to boosted decision trees (15% deletion of missing values, and KNN imputation). The x-axis represents the values for each feature and the y-axis represents the SHAP values.

In general, seven out of the strongest 22 features were handcrafted and theory driven. Table 5.2 summarizes all features. Taking more time to complete the cores appeared to influence dropout. The time to complete Core 0 predicted whether a participant eventually dropped out (Core 0 and Core 1 analysis). Additionally, usual arise-time and the time needed to get out of bed (from awake to arise) affected the prediction of dropout early on in the intervention. Participants who got up earlier than 4:30am and later than 6:45am, and participants that needed less time than 9 minutes or more time than 66 minutes to get up, negatively influenced the prediction of completing Core 6 of the intervention (x-Axis of feature usual arise-time and time to get up for Core 0 respectively). Furthermore, a greater wake time after sleep onset (WASO) also appeared to influence the prediction of dropout status. These variables could, therefore, be an early indicator of dropout in this particular intervention.

In addition, if triggers were logged on more than 18 days or participants received emails on more than 30 days, dropping out was more likely (Core 3 analysis). Furthermore, if there was no interaction between the system and the participants for more than 67 days, the individuals were more likely to drop out.

5.4 Discussion

5.4.1 Principal findings

Considering the increasing usage of digital health interventions and the tremendous amount of data gathered in such interventions, a variety of methods can be used for the analysis of various data types and structures. In this study, a process for the analysis of user journey data in this context was proposed and a step-by-step guide and technical framework for the analysis as an R package was provided. Challenges of data analysis based on user journeys, such as data transformation, feature engineering, and statistical model application and evaluation were discussed. The analysis of user journeys can be a powerful tool for the prediction of various factors on an individual participant-level. Here, it has been applied to real-world data in order to predict dropout of an Internet-based intervention.

The application of the proposed process and evaluation of statistical models indicated the feasibility of dropout prediction by using this process. AUC values ranged between 0.6 and 0.9 for the selected machine learning algorithm (boosted decision trees). Most importantly, it was shown that the prediction of user dropout was possible early in the intervention, which could be helpful to clinicians and policy makers as treatment decisions are made and adjusted. Additionally, this study indicated the importance of expert knowledge and subsequent implementation of handcrafted features. Not all existing statistical models necessarily need handcrafted features because automated feature engineering can already provide crucial insight; however, handcrafted features can increase prediction performance and lead to increased interpretability. In this study, handcrafted features appeared to be among the most important features according to the boosted decision trees perhaps given the more nuanced understanding necessary for treating insomnia. It is important to keep in mind, though, that the analysis presented here was meant as a demonstration of the power of this approach. A much larger dataset is needed in order to draw more firm and generalizable conclusions.

With this caveat, a number of interesting results emerged related to features and impact on dropout prediction. For example, as participants took longer to complete earlier steps of the intervention, they were less likely to complete the final step of the intervention. Thus, a discussion about how users can be motivated to complete early steps in the intervention may be very beneficial. In addition, findings suggest that the time participants get out of bed in the morning and how much time they actually needed to get up might be an important factor for completing the sleep intervention. Participants who get out of bed between 4:30am and 6:45am

Predictors		Analysis at Each Point in Time					
Feature	Description	Core0	Core1	Core2	Core3	Core4	Core5
Core 0 Completion Date – Intervention Start Date*	Time to complete Core 0 in days	+	+				
Arise Time - Awake Time*	Difference between time awake and getting out of bed in minutes (time to get up)	+					
Usual arise time	Retrospective report specified from baseline data	+					
WASO (Wake After Sleep Onset)	Minutes awake in the middle of the night from sleep diaries	+	+				
SOL (Sleep Onset Latency)	Minutes to fall asleep from sleep diaries	+					
Baseline Arise Time (pre arising time)	Time the user specified that they got out of bed from baseline data		+	+			
Pre retro sleep waking early	User indicates having problems waking too early in the morning		+				
Pre teach trust info source c	How much user trusts health information from various sources		+				
Avg time complete core*	Average time to complete a core among all cores that have been available			+	+	+	+
Pre stpi 24 dep	How low the user feels at baseline			+			
Pre se gen 3	How well the user feels things have been going			+			
Bedtime	If participant went to bed in the am or pm			+			
Email sent*	If the system sent an email that day				+		
Pre stpi 26 cur	How stimulated the user feels at baseline				+	+	
Trigger event logged*	If the system logged a trigger event that day				+		
Pre teach stress 6	User feels s/he can solve most problems if necessary effort is put in				+		
Pre stpi 18 cur	How eager the user feels at baseline					+	
Core 4 Completion Date – Core 4 Start Date *	Time to complete Core 4 in days					+	+
Pre stpi 29 anx	How much self-confidence the user feels at baseline					+	
Days since last info*	Days since last contact (any interaction)						+
Pre CESD 14	How lonely the user feels at baseline						+
Pre retro sleep length of sleep prob	Number of months user reports having had sleep difficulties at baseline						+

Table 5.2: Summary of the unique top 5 most important features across analyses; * indicates handcrafted/theory driven features.

and do not need more time to get out of bed than 66 minutes were more likely to complete the final step of the intervention. Additionally, trigger events might only have a positive effect in the short-term, as the appearance of triggers more often than 18 days appeared to increase likelihood of dropping out. However, it could be possible that this finding only accounts for participants that would not have completed the final step of the intervention regardless. Assuming this, these participants were, therefore, not influenced by trigger events. It is also important to emphasize that these results are based on a bottom-up, data-driven learning approach. Therefore, it is up to researchers to interpret the results and cross-validate them in other samples. Predictions in this context based on user journey data and the resulting knowledge about factors that influence these predictions, especially on an individual level, could lead to the implementation of strategies that seek to improve the utilization and efficacy of digital health interventions.

5.4.2 Limitations

There are a number of limitations of this study that should be considered in interpreting the results. One limitation is the relatively limited number of participants included in the analysis and large feature space. The predictive performance of the applied models is satisfactory, especially early on in the intervention. The process and models described in the current study are technically feasible though the reliability of the ensuing results may be impacted by limitations to sample size. Due to the limited number of participants, the results of the current study should be replicated in a larger sample. Furthermore, the amount of missing values impacts the analyses and can lead to bias. Obtaining more complete data can further increase interpretability and predictive accuracy of the models. Other than time-window based features and time-dependent variables, the demonstrated steps and current analysis do not include time-dependent feature engineering such as the relation between features and observations across time. Researchers should examine the dataset they are planning to analyze to determine whether time-dynamic features could be used in their projects. Another limitation is the fact that the data are heterogeneous on an individual participant-level; thus, the application of models that consider heterogeneous parameters might provide deeper and more individualized information about the participants. However, considering the number of participants in the data, heterogeneous models have not yet been investigated. The results are, nevertheless, promising and can lead to increased knowledge about users and how dropout of digital health interventions is affected by various factors. Studies using larger datasets are necessary in order to improve model performance and confirm findings.

5.4.3 Conclusion

This study proposes a step-by-step process for the analysis of user journey data in the context of digital health interventions and provides a technical framework. Furthermore, the proposed framework was applied to data of an Internet-based intervention for insomnia to predict dropout of participants. These participants needed to complete 7 cores in order to finish the program.

Importantly, our process was able to predict user dropout at each core better than chance. The predictive performance also varied by core; while the AUC was roughly 0.6 for cores 0 and 1, it was noticeably higher for the latter cores. This indicates that the user journey process can be used to predict dropout early in the intervention and prediction accuracy increases over the course of the intervention. This may allow researchers to preemptively address dropout before it occurs by providing support to users that may be struggling to engage. Among the machine learning techniques we evaluated, boosted decision trees provided the greatest accuracy while deleting features that contained more than 15% missing values. Additionally, a varying set of features were revealed that contributed to the prediction performance of dropout in this context. Replicating the current results in a larger sample is needed to further validate the process outlined in this paper. Researchers may also wish to develop methods that predict the likelihood of user dropout over the duration of an intervention, which could enable researchers to devote resources to those at highest risk of dropping out.

References

Alkhaldi, G., Hamilton, F.L., Lau, R., Webster, R., Michie, S., Murray, E. (2015). The Effectiveness of Technology-Based Strategies to Promote Engagement With Digital Interventions: A Systematic Review Protocol. *JMIR Res Protocol*, 4(2):e47.

Arboretti, R. and Salmaso, L. (2003). Model performance analysis and model validation in logistic regression. *Statistica*, 63(2):375–396.

Batista, G. E. A. P. A. and Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, 17(5-6):519–533.

Becker, D., Bremer, V., Funk, B., Asselbergs, J., Riper, H., and Ruwaard, J. (2016). How to Predict Mood? Delving into Features of Smartphone-Based Data. In *Twenty-second Americas Conference on Information Systems*, San Diego (USA).

Bremer, V. (2018). UJ-Analysis.

Bremer, V., Becker, D., Kolovos, S., Funk, B., van Breda, W., Hoogendoorn, M., and Riper, H. (2018). Predicting Therapy Success and Costs for Personalized Treatment Recommendations Using Baseline Characteristics : Data-Driven Analysis. *Journal of Medical Internet Research*, 20(8):e10275.

Brouwer, W., Kroeze, W., Crutzen, R., de Nooijer, J., de Vries, N. K., Brug, J., and Oenema, A. (2011). Which Intervention Characteristics are Related to More Exposure to Internet-Delivered Healthy Lifestyle Promotion Interventions? A Systematic Review. *J Med Internet Res*, 13(1):e2.

Carlbring, P., Andersson, G., Cuijpers, P., Riper, H., and Hedman-Lagerlöf, E. (2018). Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. *Cognitive Behaviour Therapy*, 47(1):1–18.

Carney, C. E., Buysse, D. J., Ancoli-Israel, S., Edinger, J. D., Krystal, A. D., Lichstein, K. L., and Morin, C. M. (2012). The Consensus Sleep Diary: Standardizing Prospective Sleep Self-Monitoring. *Sleep*, 35(2):287–302.

Chatterjee, P., Hoffman, D. L., and Novak, T. P. (2003). Modeling the Clickstream: Implications for Web-Based Advertising Efforts. *Marketing Science*, 22(4):520–541.

Cheng, W., Kasneci, G., Graepel, T., Stern, D., and Herbrich, R. (2011). Automated feature generation from structured knowledge. In *Proceedings of the 20th ACM international conference on Information and knowledge management (CIKM 2011)*, pages 1395–1404.

Christensen, H., Batterham, P., Gosling, J., Ritterband, L., Griffiths, K. M., Thorndike, F., Glozier, N., O'Dea, B., Hickie, I., and Mackinnon, A. (2016). Effectiveness of an online insomnia program (SHUTi) for prevention of depressive episodes (the GoodNight Study): a randomised controlled trial. *The Lancet Psychiatry*, 3(4):333–41.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78.

Donkin, L., Christensen, H., Naismith, S. L., Hons, B. A., Neuro, D., Neal, B., Chb, M. B., Hickie, I. B., and Glozier, N. (2011). A Systematic Review of the Impact of Adherence on the Effectiveness of e-Therapies. *Journal of Medical Internet Research*, 13(3):e52.

Erbe, D., Eichert, H.-C., Riper, H., and Ebert, D. D. (2017). Blending Face-to-Face and Internet-Based Interventions for the Treatment of Mental Disorders in Adults: Systematic Review. *Journal of Medical Internet Research*, 19(9):e306.

Eysenbach, G. (2005). The law of attrition. *Journal of Medical Internet Research*, 7(1):1–9.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.

Fernandez-Alvarez, F., Diaz-Garcia, A., Gonzales-Robles, A., Banos, R., Garcia-Palacios, A., and Botella, C. (2017). Dropping out of a transdiagnostic online intervention: A qualitative analysis of client's experiences. *Internet Interventions*, 10:29–38.

Funk, K. L., Stevens, V. J., Appel, L. J., Bauck, A., Brantley, P. J., Champagne, C. M., Coughlin, J., Dalcin, A. T., Harvey-Berino, J., Hollis, J. F., Jerome, G. J., Kennedy, B. M., Lien, L. F., Myers, V. H., Samuel-Hodge, C., Svetkey, L. P., and Vollmer, W. M. (2010). Associations of internet website use with weight change in a long-term weight loss maintenance program. *Journal of medical Internet research*, 12(3):e29.

Geraghty, A. W. A., Wood, A. M., and Hyland, M. E. (2010). Attrition from self-directed interventions: investigating the relationship between psychological predictors, intervention content and dropout from a body dissatisfaction intervention. *Social science & medicine (1982)*, 71(1):30–37.

Gosling, J. A., Glozier, N., Griffiths, K., Ritterband, L., Thorndike, F., Mackinnon, A., Hehir, K. K., Bennett, A., Bennett, K., and Christensen, H. (2014). The GoodNight study-online CBT for insomnia for the indicated prevention of depression: Study protocol for a randomised controlled trial. *Trials*, 15(1):1–8.

Horsch, C., Lancee, J., Beun, R., Neerincx, M., and Brinkman, W. (2015). Adherence to Technology-Mediated Insomnia Treatment: A Meta-Analysis, Interviews, and Focus Groups. *Journal of Medical Internet Research*, 17(9):e214.

Iida, M., Shrout, P. E., Laurenceau, J.-P., and Bolger, N. (2012). Using diary methods in psychological research. In Cooper, H., Camic, P. M., Long, D. L., Panter, A. T., Rindskopf, D., and Sher, K. J., editors, *APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics.*, pages 277–305. American Psychological Association, Washington.

Jaques, N., Rudovic, O. O., Taylor, S., Sano, A., and Picard, R. (2017). Predicting Tomorrow ' s Mood , Health , and Stress Level using Personalized Multitask Learning and Domain Adaptation. *Journal of Machine Learning Research*, 66:17–33.

Kanter, J. M. and Veeramachaneni, K. (2015). Deep Feature Synthesis : Towards Automating Data Science Endeavors. In *Data Science and Advanced Analytics*, pages 1–10.

Khurana, U., Nargesian, F., Samulowitz, H., Khalil, E., and Turaga, D. (2016). Automating Feature Engineering. In *30th Conference on Neural Information Processing Systems*.

Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Informatica*, 31:249–268.

Lam, H. T., Thiebaut, J.-M., Sinn, M., Chen, B., Mai, T., and Alkan, O. (2017). One button machine for automating feature engineering in relational databases. In *CoRR abs/1706.00327*.

Lu, X., Lin, Z., Jin, H., Yang, J., and Wang, J. Z. (2014). RAPID: Rating Pictorial Aesthetics using Deep Learning. In *Proceedings of the ACM International Conference on Multimedia - MM '14*, pages 457–466.

Lundberg, S. M. and Lee, S.-i. (2017). A Unified Approach to Interpreting Model Predictions. In *Neural Information Processing Systems (NIPS)*, pages 4765–4774.

Marcus, A. H. and Elias, R. W. (1998). Some useful statistical methods for model validation. *Environmental Health Perspectives*, 106(SUPPL. 6):1541–1550.

Melville, K. M., Casey, L. M., and Kavanagh, D. J. (2010). Dropout from internet-based treatment for psychological disorders. *British Journal of Clinical Psychology*, 49(4):455–471.

Murray, E., Edin, F., Hekler, E. B., Andersson, G., Collins, L. M., Doherty, A., Hollis, C., Rivera, D. E., West, R., and Wyatt, J. C. (2016). Evaluating Digital Health Interventions. *American Journal of Preventive Medicine*, 51(5):843–851.

Nottorf, F., Mastel, A., and Funk, B. (2012). The User-journey in Online Search - An Empirical Study of the Generic-to-Branded Spillover Effect based on User-level Data. In *DCNET/ICE-B/OPTICS*, pages 145–154.

Pedersen, D., Mansourvar, M., Sortso, C., and Schmidt, T. (2019). Predicting Dropouts From an Electronic Health Platform for Lifestyle Interventions: Analysis of Methods and Predictors. *Journal of Medical Internet Research*, 21(9):e13617.

R Core Team (2018). R: A Language and Environment for Statistical Computing.

Ritterband, L., Thorndike, F., Gonder-Frederick, L., Magee, J., Bailey, E., Saylor, D., and Morin, C. (2009a). Efficacy of an Internet-based behavioral intervention for adults with insomnia. *Archives of general psychiatry*, 66(7):692–8.

Ritterband, L. M., Thorndike, F. P., Cox, D. J., Kovatchev, B. P., and Gonder-Frederick, L. A. (2009b). A behavior change model for internet interventions. *Annals of Behavioral Medicine*, 38(1):18–27.

Ritterband, L. M., Thorndike, F. P., Ingersoll, K. S., Lord, H. R., Gonder-Frederick, L., Frederick, C., Quigg, M. S., Cohn, W. F., and Morin, C. M. (2017). Effect of a web-based cognitive behavior therapy for insomnia intervention with 1-year follow-up: A randomized clinical trial. *JAMA Psychiatry*, 74(1):68–75.

Saddichha, S., Al-Desouki, M., Lamia, A., Linden, I. A., and Krausz, M. (2014). Online interventions for depression and anxiety - a systematic review. *Health psychology and behavioral medicine*, 2(1):841–881.

Sen, A., Dacin, P., and Pattichis, C. (2006). Current trends in web data analysis. *Communications of the ACM*, 49(11):85–91.

Stange, M. and Funk, B. (2015). How Much Tracking Is Necessary - The Learning Curve in Bayesian User Journey Analysis. In *European Conference on Information Systems*.

Thorndike, F. P., Saylor, D. K., Bailey, E., Gonder-Frederick, L., Morin, C., and Ritterband, L. (2008). Development and Perceived Utility and Impact of an Internet Intervention for Insomnia. *E-journal of Applied Psychology*, 4(2):32–42.

Tibshirani, R. (1996). Regression Selection and Shrinkage via the Lasso. *Journal of the Royal Statistical Society Series B*, 58(1):267–88.

Torous, J., Lipschitz, J., Ng, M., and Firth, J. (2020). Dropout rates in clinical trials of smartphone apps for depressive symptoms: A systematic review and meta-analysis. *Journal of Affective Disorders*, 263:413–19.

van Breda, W., Bremer, V., Becker, D., Hoogendoorn, M., Funk, B., Ruwaard, J., and Riper, H. (2018). Predicting therapy success for treatment as usual and blended treatment in the domain of depression. *Internet Interventions*, 12:100–104.

Van Breda, W., Hoogendoorn, M., Eiben, A. E., Andersson, G., Riper, H., Ruwaard, J., and Vernmark, K. (2016). A feature representation learning method for temporal datasets. In *IEEE Symposium Series on Computational Intelligence*, pages 1–8.

van Breda, W., Pastor, J., Hoogendoorn, M., Ruwaard, J., Asselbergs, J., and Riper, H. (2016). Exploring and Comparing Machine Learning Approaches for Predicting Mood Over Time. In *Innovation in Medicine and Healthcare 2016*, pages 37–47, Tenerife, Spain. Springer International Publishing.

Vandelanotte, C., Spathonis, K., Eakin, E., and Owen, N. (2007). Website-delivered physical activity interventions a review of the literature. *Am J Prev Med*, 33(1):54–64.

Wickwire, E. (2019). The Value of Digital Insomnia Therapeutics: What We Know and What We Need To Know. *J Clin Sleep Med*, 15(1):11–13.

CHAPTER 6

APPLICATION OF THE DEVELOPED PROCESS TO COMMERCIAL DATA

6.1 Introduction

As mentioned in Chapter 5, the limited number of participants included in the dataset used for the application of the developed framework might lead to uncertainty of the results. Therefore, the framework was applied to a larger dataset obtained from a company that develops and delivers clinically validated digital therapeutics. This chapter describes the aim, data, and preprocessing (setup), which is similar to the one described in Chapter 5, the results of this analysis, and a conceptual approach to utilize predictions in this context.

6.2 Setup

As in Chapter 5, applying the framework aims at predicting dropout of a digital health intervention at different points in time of the program (Figure 5.5). Another focus lies on the evaluation of the predictive performance of the developed process and how the generated predictions could be utilized in order to support decision-making in clinical practice. The commercial real-world dataset was derived from the SHUTi program, and thus, the data structure is similar to the one described in Chapter 5. However, no baseline information was available for most patients and instead of data from 115 participants, this dataset includes 6948 patients after data preprocessing (patients that have less than 5 days of provided data were excluded). As in Chapter 5, seven cores needed to be conducted by the patients in order to finish the treatment. Thus, at each point in time (Core 0-5) it was predicted if a specific patient will finish Core 6 of the intervention. Because of the large number of participants, a holdout dataset could be set aside that included data from 1389 (20%) patients. The number of patients included in each analysis was 5559, 5547, 5052, 4416, 3960, and 3510, for Cores 0-5, respectively. In total, 165 features were used; nine

of them were handcrafted. Specifically, the handcrafted features were the time since the last information was obtained by a patient (*diffDate*), the time between awakening and arising of a patient (*getUp*), the difference between preferred and actual arising time (*diffPref*), and how much time a patient needed to finish a specific core (six features for Core 0-5 (*diffCore*)). However, a large number of missing values exist in the data. This problem was addressed by deleting features based on the amount of missing data. To evaluate how the deletion affects the predictive performance of the models, features were deleted that contained more than 10%, 20%, 30%, and 40% of missing values. Furthermore, categorical variables that only had one level or category were removed. The average number of features for each threshold of deleted missing data across all core analyses was 39, 42, 45, and 50 features. Each patient has exactly one outcome: dropping out or not dropping out of treatment. Data for each patient was therefore, as implemented in the technical framework, aggregated by mean for numeric and mode for categorical data. For some features, however, using the mean and mode is not necessarily meaningful. The feature *diffDate* was aggregated by sum and number of logins and emails were counted. The last observation for an individual patient was utilized for all *diffCore* features. After patient-specific aggregation, the rest of the missing data were imputed by median for numeric data and mode for categorical data. Additionally, the k-nearest neighbor (KNN) algorithm was applied for imputation in order to compare both approaches in terms of predictive accuracy (k=5).

Regularized logistic regression (L1 and L2 regularization) and boosted decision trees were applied in order to predict dropout of the patients. For finding the optimal parameters for these machine learning techniques, a grid-based search and cross-validation was conducted. Table 6.1 demonstrates the hyper parameter space for the boosted tree model. Since one aim was the evaluation of the developed framework, the parameter space is equal to the implemented range in the technical framework. This space could be increased to obtain a better predictive performance. The L1 and L2 regularization models utilized 10-fold cross validation to find the best λ. Additionally, 10-fold stratified cross-validation was applied for obtaining a cross-validated error and selecting the best model. As performance indicator, the area under the receiver operating characteristics curve (ROC) was used (AUC) (Fawcett, 2006).

Hyper parameter	Tuning approach	Parameter space
Iterations	Fixed value	{1000}
Sample ratio of training instances	Fixed value	{1}
Learning rate	Grid search	$\{0.1, 0.01, 0.0001\}$
Depth of tree	Grid search	$\{2, 6, 10\}$
Gamma	Grid search	$\{0, 1\}$
Features supplied to tree	Grid search	$\{0.4, 0.8\}$
Minimum number of samples for splitting	Grid search	$\{1, 20\}$

Table 6.1: Hyper parameter space for boosted trees.

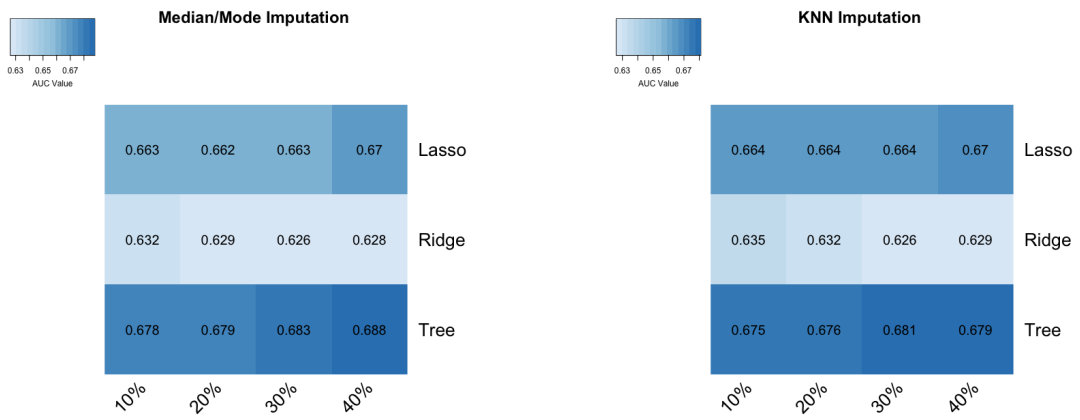


Figure 6.1: Heat map of averaged AUC values across core analyses for each model, imputation procedure, and threshold for percentage of missing values.

6.3 Results

Figure 6.1 illustrates the averaged AUC scores across all six analyses (Core 0-5) for each machine learning technique, threshold of missing values, and imputation procedure. As can be seen, the method of imputing missing values did not influence the results strongly (as in Chapter 5). The same observation accounts for the threshold of deleted missing values. Boosted decision trees lead to the best average result when features were deleted that have more than 40% of missing values while imputing the rest by median and mode.

After selecting the above mentioned setting and machine learning technique, Figure 6.2 illustrates the ROC curves for each core analysis and their corresponding AUC. The ROC curves on the left hand-side are the cross-validated results from 10-fold-cross-validation and the results on the right side depict the outcome for the holdout dataset. The AUC values for both predictions (cross-validated and holdout data) are very similar; additionally, the AUC values continuously increase across the core analyses except for Core 5 in the cross-validated ROC (left side). Thus, the predictive accuracy increases as patients progress through the program. It can also be observed that a prediction of dropout already appears feasible early on in the intervention (Core 1-2).

Because the analyzed problem is an imbalanced classification problem, the area under the precision recall curve (PRAUC) is demonstrated in Table 6.2. The precision recall curve is less sensitive to imbalanced data because it focuses on the prediction of the minority class. The threshold for a random guess or chance differs because it depends on the actual distribution of the positive and negative cases. Here, as well as the AUC indicates, it can be observed that dropout is constantly predicted better than random. These results are not only *good* for the evaluation of the framework but also for predicting dropout and can support decision-making in therapeutic processes.

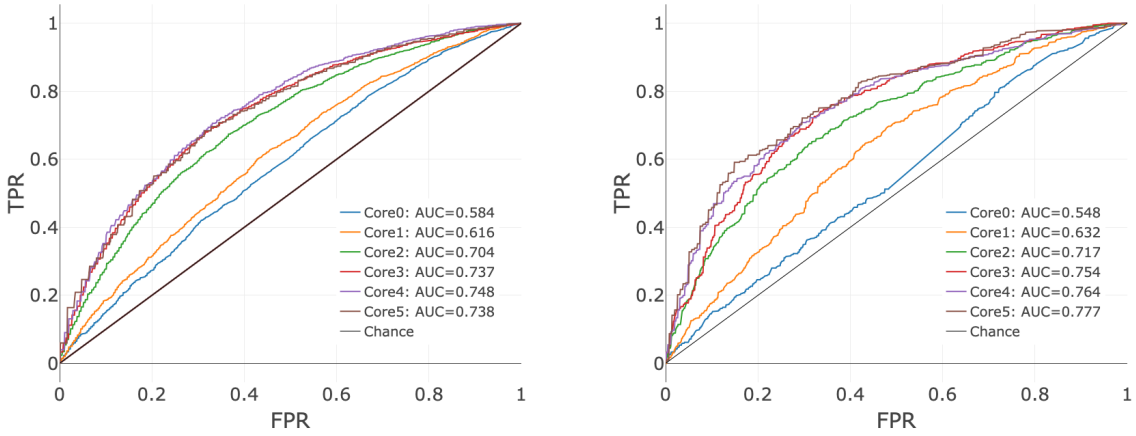


Figure 6.2: Cross-validated ROC (left) and ROC for holdout data (right).

Core	Data	PRAUC	Chance
0	Cross-validated	0.575	0.487
1	Cross-validated	0.616	0.482
2	Cross-validated	0.704	0.433
3	Cross-validated	0.737	0.360
4	Cross-validated	0.748	0.276
5	Cross-validated	0.738	0.183
0	Holdout	0.540	0.476
1	Holdout	0.621	0.488
2	Holdout	0.646	0.432
3	Holdout	0.593	0.316
4	Holdout	0.524	0.276
5	Holdout	0.434	0.186

Table 6.2: Area under the precision recall curve for different core analyses and cross-validated/holdout data.

6.4 Conceptual approach for the utilization of predictions

Since the results of the analyses are very promising, the question arises how these predictive outcomes can be utilized for therapeutic decisions in practice. Figure 6.3 demonstrates an approach for the evaluation of the predictions in terms of clinical and economic impact. As input for the machine learning models, patient-specific data is required. The output of these models are the generated predictions. These outputs basically specify the likelihood of a participant dropping out of the intervention. If the model generalizes well beyond the training set, it can be assumed that with increased data, the dropout predictions approach the actual dropout rate of the population. The next step is the development of possible micro-interventions that can potentially improve intervention adherence for specific patients. These interventions need to be developed with support of experts in the target field of the intervention. Possible interventions

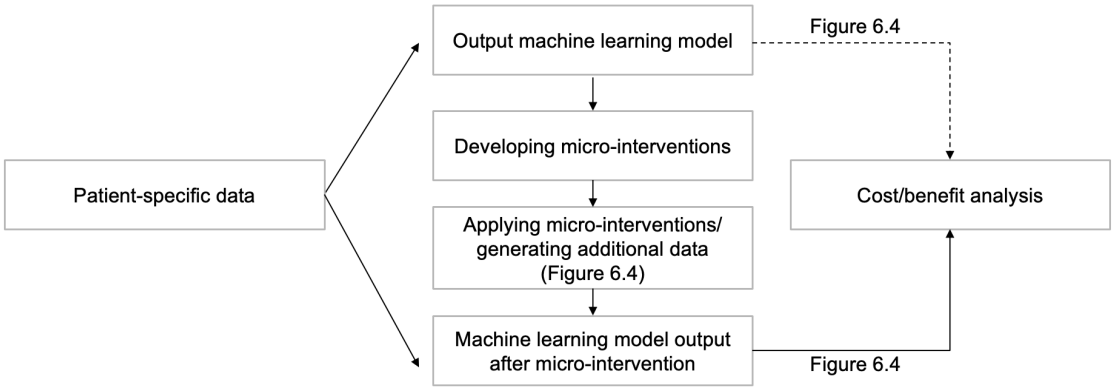


Figure 6.3: General approach for clinical and economic evaluation of predictions.

with different levels of engagement could be to send automated or personalized messages, have coaches or clinicians make personalized phone calls, or even make a personal appointment with a clinician to review the current status of treatment, moving up the hierarchy of stepped support as needed. Another general option to decrease dropout rates would be to increase motivation by offering certain rewards for completing steps of the program. These rewards could either be of financial nature or in the form of badges inspired by the concept of gamification (Cugelman, 2013; Hamari, 2017). Specific achievements could be unlocked whenever patients finish parts of the program such as completing a core or filling out seven consecutive days of sleep diaries. As mentioned earlier, these types of interventions need to be carefully developed by clinicians in the field, as interventions could target both engagement with the therapeutic and/or adherence to therapeutics assignments.

After developing such micro-interventions, they need to be utilized and evaluated for patients that are identified as high-risk. Here, the process of Figure 6.4 can be followed (will be explained in more detail below). Afterwards, more data is generated from patients that have received the micro-interventions. In order to estimate the actual impact of the micro-interventions, dropout could then be predicted repeatedly by applying machine learning models. This could lead to decreased dropout rates for individuals that are affected by the micro-interventions. Furthermore, these *new* models could reveal that the impact of different micro-interventions varies among the patients. For example, individuals that are under the age of 20 might be less affected by phone calls whereas appointments or automated messages might work better. A cost/benefit analysis can then estimate the clinical and economic impact of the predictions as well as micro-interventions.

Here, however, no data is available for this estimation because these predictions have not been utilized in practice and no new micro-interventions have been developed. Thus, an approach is demonstrated in which the cost/benefit analysis can be estimated before the micro-interventions are applied. For this task, the process of Figure 6.4 is used. In this example, the numbers are based on the predictions of the holdout data and Core 3 analysis. The average predicted dropout

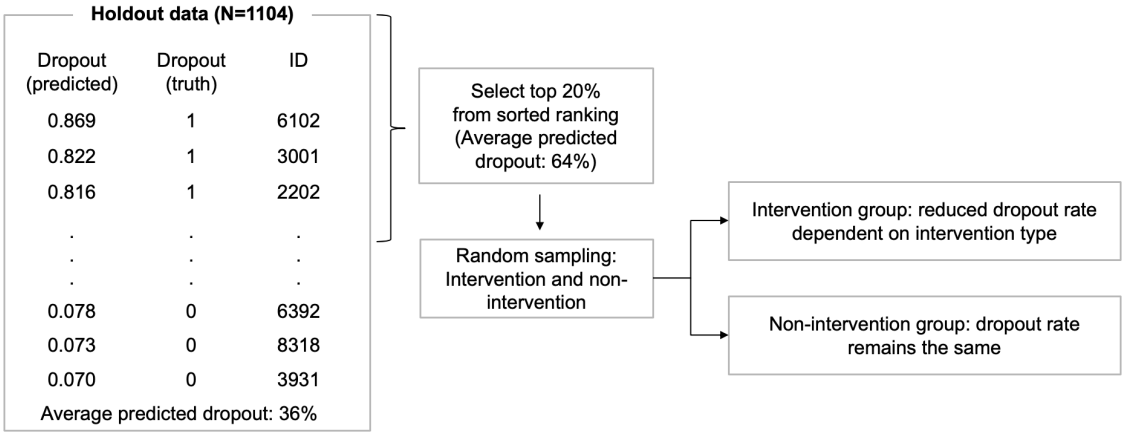


Figure 6.4: Process for utilization and evaluation of predictions based on Core 3 analysis and holdout data.

at this point in time is 36% among all patients while considering the data between Core 0 and Core 3. The first step is to sort the predictions for the individual patients in descending order. The top 20% of patients with the highest probability of dropping out of the treatment are then selected. These top 20% of patients have an average predicted dropout of 64%. Then, half of the 20% of patients with the highest probability of dropping out of the treatment would be randomly sampled and assigned to a micro-intervention whereas the others receive no micro-intervention. This process could lead to the observation that the dropout rate is reduced for the intervention group and stays the same for the non-intervention group.

Selecting the top 20% is not obligatory. Other thresholds could be utilized. For demonstrative purpose, Figure 6.5 illustrates the predicted dropout probability and the actual observed dropout rate among different thresholds. It can be seen that the trend of the predictions, as already suggested by the AUC values in Figure 6.2, follows the trend of the observed dropout rate. As the observed dropout rate decreases with a greater number of individuals, the average predicted dropout probability also decreases. It can also be observed that the average predicted dropout is always slightly higher compared to the actual dropout.

In a next step, the impact of applying the micro-interventions needs to be estimated. For this approach, however, some aspects still need to be considered. The costs of a micro-intervention type, the corresponding success rate, and the benefits of keeping an individual in the treatment play a major role. Hypotheses need to be created for these aspects and specific numbers assumed. The benefits for keeping an individual in the treatment program, for example, are not only of financial nature but also have a societal impact. This aspect could be disregarded for an example and only a financial benefit could be assumed. However, this financial benefit also depends on the point in time (i.e., Core 2 vs Core 4) and can be assumed to be decreasing with time.

Furthermore, the new probability for dropout after receiving a specific micro-intervention

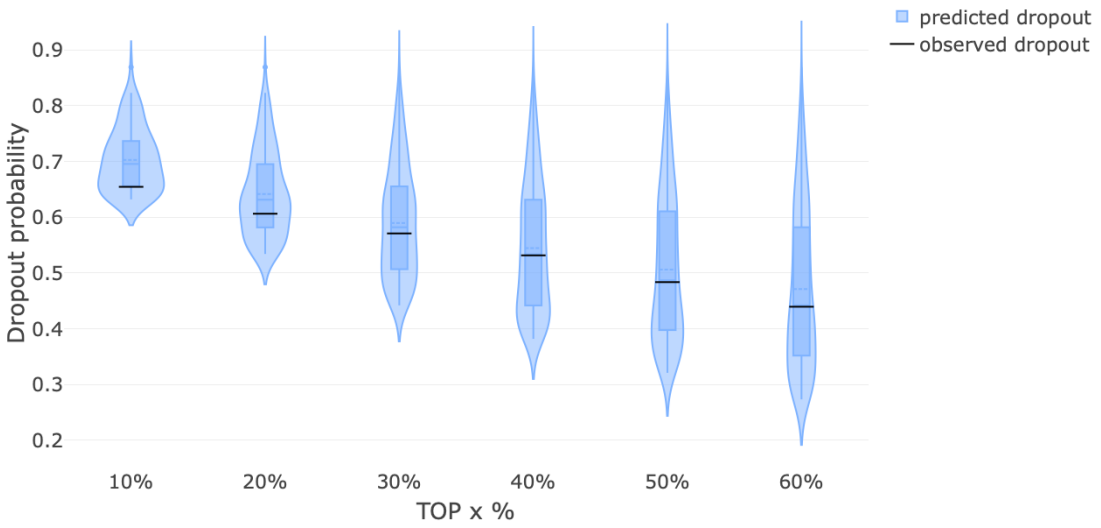


Figure 6.5: Predicted probabilities and observed dropout rate for different thresholds of selected patients based on Core 3 analysis and holdout data.

needs to be assumed. There are multiple options for doing this data transformation. Figure 6.6 demonstrates two possible approaches. One way could be the utilization of a rescaled probability density function of the beta distribution (left side of Figure 6.6). The rescaled result has the interval $[0, \max_i]$, where $\max_i$ is the strongest effect or reduction in probability each micro-intervention i can have. The new probability $\hat{p}$ after a specific micro-intervention is then the actual prediction p minus the assigned probability reduction. Individual patients are not homogeneously influenced by the micro-interventions, some might benefit more than others from being in the micro-intervention group. Thus, in this approach it is assumed that patients who have a higher chance of dropping out according to the predictions have a lower reduction in probability compared to patients that already have a low probability of dropping out. Another option would be to use a rescaled cumulative distribution function of the beta distribution for data transformation (right side of Figure 6.6). The rescaled result also has the interval $[0, \max_i]$, where $\max_i$ is the highest new possible probability for each micro-intervention i. The difference is that the new probability $\hat{p}$ is directly visible on the y-axis. For both approaches, the parameter $\max_i$ needs to be defined beforehand for each micro-intervention.

The final profit for a specific micro-intervention i can then be calculated as follows:

$$\text{profit}_i = \text{benefit} \cdot \left(N_{\text{patient}} \cdot \text{avg}_p - N_{\text{patient}} \cdot \text{avg}_{\hat{p}i}\right) - N_{\text{patient}} \cdot \text{cost}_i,$$

where N_{patient} is the number of patients in the intervention group, avg_p is their average dropout probability, $\text{avg}_{\hat{p}i}$ is the average adjusted probability after reduction due to each micro-intervention i, and cost_i are the costs for each micro intervention i. The benefit is the assumed financial benefit for keeping an individual in the treatment program. It needs to be noted

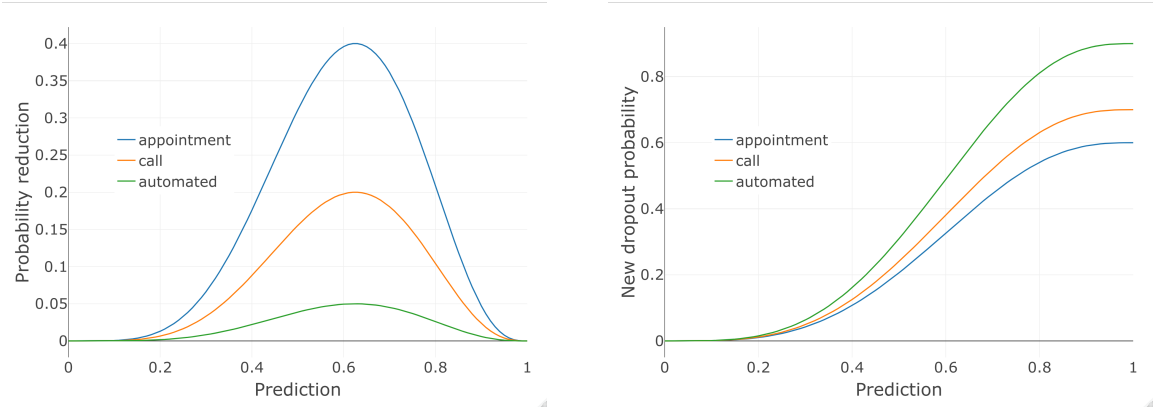


Figure 6.6: Two theoretical approaches for transformation of predictions according to different micro-interventions.

that the utilization of other underlying distributions would be possible for transforming the predictions such as the truncated normal distribution. Furthermore, this approach only approximates the generated profit for utilizing the predictions and applying the micro-interventions based on hypotheses made for various factors such as costs, benefit, and success rate of the micro-interventions. In order to evaluate the actual value of such predictions, the developed micro-interventions need to be applied and evaluated in clinical practice.

References

Cugelman, B. (2013). Gamification: What It Is and Why It Matters to Digital Health Behavior Change Developers. *JMIR Serious Games*, 1(1):e3.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.

Hamari, J. (2017). Do badges increase user activity? A field experiment on the effects of gamification. *Computers in Human Behavior*, 71:469–478.

CHAPTER 7

MACHINE LEARNING-BASED DECISION SUPPORT SYSTEMS IN CLINICAL PRACTICE: POTENTIAL AND LIMITATIONS

Bremer, V. (2020). Machine learning-based decision support systems in clinical practice: Potential and limitations (white paper).

Abstract

Computerized clinical decision support systems exist in various medical fields in order to enhance decision-making by offering patient-individualized information and recommendations. In the field of mental health, machine learning-based approaches can contribute by providing predictions of psychological factors and subsequent recommendations. However, actual machine learning-based tools are rarely utilized in clinical practice. Thus, in this study, eight semi-structured interviews with professional psychotherapists were conducted in order to shed light on the capabilities and hurdles of machine learning-based computerized clinical decision support systems in psychotherapeutic clinical practice. Thematic analysis was applied, which resulted in three main themes: Potentials, limitations, and requirements of these types of systems.

7.1 Introduction

Computerized clinical decision support systems (CCDSS) are systems that are created to enhance decision-making in a clinical context by providing patient-individualized information and recommendations (Haynes and Wilczynski, 2010). Already in the 1970s, the first system was developed and nowadays various CCDSSs exist in different medical fields (Belle et al., 2013; Sacchi et al., 2015) such as cancer treatment (Lindblom et al., 2012), chronic

diseases (Riano et al., 2012), or mental health (Stein et al., 2013). Since it can be difficult to predict negative outcomes of therapeutic interventions for therapists regularly based on experience in the context of mental health (Hannan et al., 2005), decision support systems can support decision-making and help to predict psychological factors such as mood or sleep levels, which are important for an individuals' life. Decision support systems can either be used to provide recommendations in a clinical setting or for the illustration of crucial information for the decision-making process (Sacchi et al., 2015). Often, these systems are also used for drug prescription or "ordering of medical procedures" (Sacchi et al., 2015).

The field of machine learning can contribute in regard to CCDSSs. Machine learning is a subset of artificial intelligence and often seeks to reveal relationships or patterns (Triantafyllidis and Tsanas, 2019) between input and output data in order to predict particular variables based on new observations; it is utilized in a variety of fields such as online marketing, recommender systems, or web search (Domingos, 2012). Since not only the amount and availability of data is steadily increasing in medical settings but also types of data vary from wearables and self-reported Ecological Momentary Assessment data to log data of Internet-based interventions, valuable information can be revealed from large medical data sources by using data mining and machine learning techniques. Additionally, patient-individualized analyses of these data offer the opportunity to tailor treatment to the individual and provide personalized treatment recommendations (Clifton et al., 2015). Thus, the field of machine learning "holds substantial promise for the future of medicine" (Clifton et al., 2015).

As mentioned by Belle and colleagues in their survey, systems based on machine learning have already been applied in some medical fields such as Radiology (computerized diagnosis) or cancer treatment (detection and treatment) (Belle et al., 2013). In mental healthcare, predictive analytics can have a positive influence on treatment success. Predicting patient-individual outcomes and revealing relationships between psychological factors might contribute to the understanding of the patients' behavior and support decision-making for therapists. These types of analyses can reveal crucial information regarding risk factors and symptom development (Tiemens and Kloos, 2016). In mental health research, predictive analytics based on machine learning is in its infancy but already exists and often seeks to support the diagnosis of mental health conditions, predicts treatment progression and outcomes, or aims to forecast costs of psychological interventions (Becker et al., 2016; Bremer et al., 2018; Shatte et al., 2019; van Breda et al., 2018). Tiemens and Kloos, for example, predicted treatment outcome in generalized mental healthcare and classified symptom improvement (Tiemens and Kloos, 2016). Another example is the study of Jaques and colleagues, who analyzed unobtrusive data from smartphones and wearables for the prediction of individual mood levels (Jaques et al., 2017). They found that this goal is a difficult task, however, their results indicate a considerable increase in performance compared to a non-individual approach.

In general, research indicates that a demand exists for the development and utilization of

CCDSSs (Belle et al., 2013) in medical settings and that systems based on machine learning might have great potential to support and improve care (Clifton et al., 2015; Shatte et al., 2019; Triantafyllidis and Tsanas, 2019). However, even though machine learning-based approaches are utilized in research projects (Shatte et al., 2019; Triantafyllidis and Tsanas, 2019), researchers found that machine learning-based approaches "are rarely used in the tools currently exploited in clinical practice" (Clifton et al., 2015; Kelly et al., 2019; Sacchi et al., 2015; Triantafyllidis and Tsanas, 2019). Thus, it is crucial to investigate and understand the reasons for the difficulties in implementing CCDSSs applications in a clinical context (Kaplan, 2001). The establishment of such systems is also dependent on the acceptance, usage, and perceived value of these decision support systems by the clinicians (Eichner and Das, 2010; Sacchi et al., 2015).

Therefore, the objective of this study is to understand the reason for CCDSSs not being utilized more often in clinical practice of mental health. Thus, this paper investigates the opportunities and hurdles of machine learning-based computerized clinical decision support systems (MLCCDSS) in the context of mental health from a psychotherapists' point of view. Semi-structured interviews were conducted with professional psychotherapists. Based on these data, thematic analysis was utilized for the identification of recurrent themes. Understanding the psychotherapists' view on MLCCDSSs can aid to the development of such systems and might subsequently enhance the experience of receiving treatment for patients.

7.2 Method

7.2.1 Data and participants

Semi-structured interviews were conducted with eight professional psychotherapists (50% male and 50% female) who have experience in providing psychotherapeutic advice - most of them have an office in Northern Germany. The participants were initially either contacted by phone, e-mail, or by already interviewed participants. The principal investigator consecutively visited the participants individually at their offices between November 2017 and August 2018. The initial part of the interview consisted of an introduction of the researcher and a statement of the anonymity of the psychotherapist. The participants were briefly introduced to the concept of machine learning with a focus on predictive analysis and how decision support systems can be created based on machine learning in the context of mental health.

These decision support systems were defined as software or tool that predicts future patient-individual outcomes specified by the therapist such as mood levels or sleep quality, presents inferential results, and delivers this information either to the therapist or patient by a computer screen or mobile application. Patients can then adjust their behavior according to recommendation messages or therapists can adjust their treatment or intervene when critical levels of psychological factors are indicated. In order to develop a better understanding of these factors, an example of such decision support was provided. Here, a support system was suggested that

predicts mood levels based on Ecological Momentary Assessment data/diary data. This setting was chosen because this kind of data is gathered in a variety of psychological fields and disorders (Iida et al., 2012). The first part of the interview focused on the therapists' experience with tools that support decision-making and, in particular, with tools that are based on machine learning and how they perceive these types of decision support systems. A second part consisted of the therapists' trust in statistical methods and also focused on questions regarding the predictive accuracy of such tools. Then, therapists were asked about the benefits and risks of such systems and where they believe machine learning could contribute in clinical settings. In a third part, therapists were questioned about the influence of such decision support systems and how they could affect treatment decisions. On average, the interviews lasted 30 minutes and were designed mostly open-ended in order to capture the most important aspects for each psychotherapist. The interviews were recorded and transcribed after each interview session by the principal investigator while initial thoughts were summed up as this is an important stage of initial analysis (Braun and Clarke, 2006; Fielden et al., 2011; Riessman, 1993).

7.2.2 Data analysis

Semantic thematic analysis was utilized for preparation and analysis of the interviews and identifying recurrent themes. In this context, semantic analysis means that the analysis was data-driven and focused on what the participants said and not too far beyond that (Braun and Clarke, 2006). Braun and Clarke define thematic analysis as "a method for identifying, analyzing and reporting patterns (themes) within data. It minimally organizes and describes your dataset in (rich) detail". Thematic analysis captures the participants' "point of view and descriptions of experiences, beliefs, and perceptions" (Butcher et al., 2001). This method is closely related to content analysis. However, where content analysis is often based on frequencies of word patterns and is therefore often seen as a mix between qualitative and quantitative approach (Joffe and Yardley, 2004), thematic analysis "is more involved and nuanced [...] and focuses on identifying and describing both implicit and explicit ideas" (Namey et al., 2008). Since this method is appropriate when more general aspects are analyzed based on collected data which are subject to interpretation (Alhojailan, 2012), and when the sample size of data is too small to execute statistical analyses (Joffe and Yardley, 2004), thematic analysis was utilized in this study. Since many research projects apply thematic analysis but do not explicitly state the specific undertaken steps (Attride-Jerling, 2001), it is especially crucial to clearly outline the applied procedures (Braun and Clarke, 2006). In this study, the 6-phase guide provided by Braun and Clark was followed, illustrated in Table 7.1.

Phase	Description
Data familiarization	Transcription, reading and re-reading, summarizing ideas
Initial codes	Coding interesting features systematically
Theme search	Fusing codes into potential themes
Theme review	Verifying themes and create thematic map
Definition of themes	Define final themes and create story
Report creation	Extract examples, verify results, create report

Table 7.1: The six phases of thematic analysis by Braun and Clarke.

These phases were not followed linearly; instead, phases were repeatedly moved back and forth during the analysis stage. According to the guide, the interviews were initially read and re-read to familiarize with the data and gather initial ideas. Then, parts of the text were assigned codes by generalizing relevant sections to similar statements. Coding procedures of thematic analysis can be conducted based on already existing theories in the research field (deductive coding), or exclusively from the data itself (inductive coding) (Joffe and Yardley, 2004). Inductive coding was used in this study. While coding, it was allowed that the same part of the text could be assigned to multiple codes if logically reasonable. These codes were assigned in English language and texts used for the illustration of later results were translated to make the references understandable for the international community. From the codes, themes were derived. A theme includes crucial information and meaning about the data in the context of the research question (Braun and Clarke, 2006). The analysis was conducted by the support of the computer software MAXQDA (Verbi Software, 2016).

7.3 Results

In general, some participants already gave the topic of machine learning in their respective clinical context some thought whereas others have never considered this topic before. However, only one of the eight psychotherapists already had experience with types of predictive modeling and none of the participants currently use either descriptive analyses or predictive computerized modeling techniques. Nonetheless, two of the participants have utilized computerized methods for treatment evaluation or general data exploration during their education. This might indicate that machine learning has not yet affected therapy in practice.

The thematic analysis resulted in three main themes. Potential of MLCCDSS describes the possibilities machine learning-based decision support systems can portrait and briefly outlines scenarios, which could add value to the therapeutic process. Limitations for MLCCDSS represents aspects that inhibit the realization of such systems in today's practice and Requirements for MLCCDSS represents the psychotherapists' view on required factors for a successful application of these types of systems in a clinical setting. Limitations and requirements of MLCCDSSs are not clearly separated from each other, however, limitations rather focus on the hurdles and barriers that exist in the mindset of psychotherapists while the latter has its focus on important

aspects for an implementation of MLCCDSSs in clinical practice. In the following subsections, the main themes that were derived are presented and discussed including transcribed and translated sample statements of the psychotherapists.

7.3.1 Potential of MLCCDSS

The first theme that emerged focuses on the potential of MLCCDSSs in clinical practice. The application of MLCCDSSs can generally be of a descriptive or predictive nature. Descriptive analysis can, for example, be the visualization of diary data. Since it is difficult to remember historical diaries of patients to a certain degree, illustrating the progression of treatment in terms of descriptive analyses can support to visualize the trend of therapeutic success on an individual level. Already prepared data further saves valuable time for content-related discussion during the therapeutic session. Inferential results of machine learning-based approaches can additionally enable patients to understand the relationships between behavioral factors and the disorder in terms of psychoeducation. Even though these relationships are often already known by the patients and psychotherapists, inferential results can help to deepen this understanding and can act as a proof for the connection; then, nevertheless, they do not support behavioral control.

Predictive analysis, the core of MLCCDSSs, can be utilized for the forecast of different psychological factors such as mood or sleep levels in order to provide recommendations for the psychotherapist or the patient. Predicting trends of psychological factors can help patients to regain a better feeling about themselves if they receive proof for existing highs and lows of the disorder. Reminders, that are sent to the patient when critical values of psychological factors are reached, could lead to increased self-management, better self-evaluation, and self-prevention. This could in turn lead to higher motivation of the patients for the treatment process, which was an important factor according to the participants. Specifically, the participants were interested in the prediction of risk patients that have a greater chance of dropping out of treatment or identifying patients that do not benefit from the therapeutic process. This includes recommendations of frequency and length of therapy sessions on an individual level. Furthermore, most of the participants agreed that, even though currently difficult to achieve, a prediction of suicidal behavior would be very beneficial in the context of mental health.

"Suicide - of course. Hints for this would always be great. [...] Questions about that wouldn't be asked frequently in the normal therapy sessions. So, if an algorithm can detect that, that would be great."

Suicide was the only psychological factor that the interviewed psychotherapists would act on outside of the scheduled appointments with the patients. In the case of critical levels of other psychological factors such as mood or sleep, the psychotherapists would await the next individual session instead of contacting the patient immediately. Other interesting clinical settings for MLCCDSSs according to the participants appeared to be the prediction of outcomes in the field

of addiction and relapse prevention. Recommendations for the type of treatment that might lead to an increased benefit for the patient also appeared to be important. Specifically, which approach of psychotherapy can lead to a greater benefit (comparison of treatment types), how can the treatment process be adjusted in order to provide better treatment, and when is the prescription of medication beneficial. Here, statistical results on an individual level can lead to an additional evaluation of individual patients and their treatment progress in order to adjust treatment approaches if indicated that the current approach might not work properly.

Two participants mentioned that an application of MLCCDSSs is not useful in an ambulatory setting but rather in a stationary setting. In a stationary setting, due to a larger number of patients and less fluctuation, more data can be gathered about the individuals.

"If I imagine these types of procedures in big clinics, where there is a greater number of patients, I can rather imagine this."

Thus, utilizing MLCCDSSs in large clinics can be a starting point for the evaluation and definition of necessary amounts of data in order to decrease the uncertainty of predictions. However, the application of MLCCDSSs does not fit in every psychologists' toolbox. Specifically, it was assumed that behavioral therapists rather tend to see value in such systems compared to depth psychologists.

"When I hear this, I can imagine that the general acceptance of behavioral therapists could be greater than that of analysts or depth psychologists like myself. [...] They also say how low-minded you feel today on a scale of 1-10. Analysts or people with a theoretical background like mine would never do that."

This statement could not be verified based on the interviewed participants. The sample included three depth and five behavioral psychologists and, thus, the sample size is too small in order to evaluate this hypothesis. Since behavioral psychologists often follow a structured approach to treatment (Fenn and Byrne, 2013), this statement could carry some truth and is an interesting aspect to verify in future studies.

7.3.2 Limitations for MLCCDSS

The second theme emerging from the interviews deals with the barriers MLCCDSSs face in clinical practice from a psychotherapists' point of view. Psychotherapy lives through human interaction and the relationship between psychotherapist and patient. MLCCDSSs could endanger and operationalize this relationship. It could lead to a dehumanization of the therapeutic process. Some participants hypothesized that the background of the patients is so individual that only this relationship can lead to treatment benefits. Thus, humans and their social background and environment are too complex and consist of too many variables for the analysis of behavior.

"[...] Psychotherapy lives through human interaction and from the relationship between psychotherapist and patient. Thus, I also see danger from this. It could lead to a manualization of the process. I do not like that."

In accordance with this, the aforementioned reminders could not only have positive effects. They also represent an artificial part in the therapy dynamics, which can lead to a pseudo-relationship between patient and psychotherapist. They might only be beneficial in the short term - when, i.e., reminding the patient of physical activities. In the long run, however, these reminders might damage the aim of therapy: developing an understanding of when specific types of activities might make the patient feel better without the need for a reminder. Patients need to reflect on their behavior themselves, learn mindfulness, make their own decisions, and should not be dependent on a machine or controlled by a monitoring process. Observation of patients' behavior might activate the restriction of freedom and flexibility. Thus, MLCCDSSs could lead to decreased responsibility of the patients for their own behavior and could lead to a focus on the symptoms of the disorder, which can result in a neglect of human interaction. MLCCDSSs and their subsequent recommendations could lead to a change of perception for the psychotherapists as well, which in turn could alter their decision-making process negatively. Specifically, some psychotherapists might take recommendations, for example predicted diagnoses, as a fact and rely on these predictive results more than on their clinical expertise. This newly obtained information might be a distraction and might limit the detection range for specific disorders. Summarized by one participant, MLCCDSSs represent a "[...] disempowerment of humankind."

Another limitation for MLCCDSSs could be the psychotherapists' fear of substitution and the corresponding mindset professionals might have regarding these types of systems. Most of the participants did, however, not fear to be substituted because of treatment quality - they believe that traditional treatment as usual is not substitutable from a qualitative perspective - there was rather a concern that financial aspects could lead to a substitution.

"I do think so (fear of substitution). However, I believe not because of quality but because of costs - as a society we just think economically."

Thus, insurance companies, for example, could be interested in applying such systems to save financial resources and consequently reduce therapy sessions. MLCCDSSs could then also lead to a perception change of the patients. Specifically, patients could develop an assumption that the adherence to therapy sessions would not be necessary due to the MLCCDSSs and their recommendations. Therefore, the interviewed participants hesitate to test and evaluate such systems at the expense of the patients.

One aspect that appeared to be crucial for the participants is the additional workload that is associated with the utilization of MLCCDSSs. Psychotherapy is a business and needs to be economically profitable. Thus, most participants do not want any additional effort for the application of such systems and no adjusted workflow. If realized in clinical practice, MLCCDSSs must therefore be implemented in the most flexible way and should not produce additional work for the psychotherapists.

Assuming an ongoing application of MLCCDSSs, prediction of psychological factors and the patients' behavior might also lead to paranoid patients and questions such as "can the

psychotherapist look into my head" might arise. Data security and access play an important role that can, if not secured and communicated appropriately, lead to skepticism from the patients. The results MLCCDSSs can provide should only be available to parties that aim at increasing the individualized treatment outcome and no other parties such as "the superior at work or the insurance company" should have access. One participant stated:

"[...] I always hear from patients how they control their partners' smartphones. [...] What happens if there is a text message from a therapist or a therapy tool - that can also raise questions. This drastically decreases my enthusiasm regarding these tools."

In accordance with this, MLCCDSSs could create ethical issues not only regarding data security but also when being used against the therapist. If, for example, suicidal behavior was predicted by a MLCCDSS and the psychotherapist does not react immediately while the patient actually commits suicide, the method and practice of the psychotherapist could be challenged in hindsight. Thus, a concern exists that these systems could lead to accusations based on historical cases.

Another important barrier is the feasibility of statistical procedures machine learning is based on and the interpretation of the results. Outcomes do come with uncertainty and this fact cannot be disregarded. Therefore, the accuracy of applied machine learning techniques plays a major role. They need to be evaluated properly and reach performances that provide an actual benefit. Most participants required performances that are substantially better than random guessing, however, there was one participant that would not even use MLCCDSSs even if they reached 100% accuracy due to the limitations outlined above.

7.3.3 Requirements for MLCCDSS

As already expected, the application depends on the usage experience and expected value of decision support systems. An important aspect is the communication between IT specialists and psychologists. Finding a common ground of articulation and information exchange for, i.e., requirements of MLCCDSSs seems to be crucial. Without collaboration, no systems can be created that can provide value for psychotherapists. And if the practitioners do not realize the potential and value of such systems, the patients will, thus, not benefit. As one psychotherapist made it very clear, specifying the structure of MLCCDSSs and especially their functions must be the result of deep cooperation:

"In the implementation itself, the communication between IT and psychotherapist - what does the psychotherapist want and what is possible from an IT point of view - is very important. We have to speak the same language and understand exactly what we want to achieve."

How to communicate these tools to the psychotherapists is important. Pointing out that these systems are utilized as a support mechanism and not as a decision take-over is crucial. They need to understand that the application of machine learning-based systems does not aim for

replacement but as support for their expertise. As one therapist said, the communication of such systems needs to be realized by creating no fear but excitement for the newly obtained support:

"But fears [...] are irrational [...] and that's why, if you can communicate it in a certain way, you can create those fears, and then you will not [...] win those people over."

However, it is not only important how to communicate these tools to the psychotherapists but also to the patients who will eventually be the focus of the application and provide data for such systems. Most important appears to be that patients understand the tool not as a substitution - they are not receiving a secondary treatment type but the aim is an enhancement of decisions that aid in treating a specific disorder. Two psychotherapists believed that it would be enough to explain the tool to their patients and because of the relationship that exists between patient and therapist, most of the patients would agree to participate. Explaining the function of these MLCCDSSs, data security issues, and elaborating on other barriers mentioned above, thus, being completely open about the advantages and disadvantages MLCCDSSs represent is an important aspect to consider when realizing such tools in practice.

Furthermore, these systems not only need to be communicated by their actual information they can provide but also how they function. The participants' opinion regarding how profound they need to understand the mathematical equations these machine learning algorithms are based on varied. 50% of the participants required to know the mathematical basics behind the approaches whereas the other 50% did not require to understand them fully themselves but need to be able to interpret them. For most psychotherapists, the main aspect appeared to be the origin of these algorithms and who supports and finances their development. If, for example, an insurance company developed particular algorithms for the usage in decision support systems, one participant would not trust their outcomes, which resembles an ethical issue and is of utter importance when developing MLCCDSSs in this context (Char et al., 2018):

"If you get an algorithm from a health insurance company you are not so positive about it [...], there are very specific interests behind it and these are not my interests and maybe not the interests of the patient."

Characterization and definition of scenarios for possible applications of MLCCDSSs appear to be difficult for psychotherapists. However, it is a prerequisite for both psychotherapists as well as computer scientists in order to develop systems that can provide actual value in the field. This, again, highlights the importance of proper communication between psychotherapists and computer scientists for the realization of MLCCDSSs in a clinical context. Table 7.2 illustrates a summary of the results.

Theme	Topics
Potential of MLCCDSS	
	- descriptive analysis for diary data analysis/visualization of trends
	- predictive analysis in the field of addiction, relapse prevention, suicide, or dropout
	- increased psychoeducation
	- increased self-management
	- increased self-evaluation
	- increased self-prevention
	- increased motivation
	- increased treatment adjustment
Barriers for MLCCDSS	
	- less dynamic and flexibility
	- pseudo relationship
	- decreases mindfulness
	- restriction of freedom
	- decreased responsibility for the patients' own behavior
	- change of psychotherapists' perception
	- decreased attendance of patients to therapy sessions
	- additional effort for psychotherapists
	- paranoia of patients
Requirements & challenges	
	- usage experience
	- expected value of decision support
	- communication between psychotherapist and computer science
	- communication of tool to psychotherapists
	- communication of tool to patient
	- communication of function

Table 7.2: Summary of themes and topics.

7.4 Discussion

This study provides an overview of the potentials and limitations of machine learning-based computerized clinical decision support systems from a psychotherapists' point of view. Simultaneously, challenges for their applications were highlighted. For this purpose, interviews were conducted with eight psychotherapists and a thematic analysis was applied. In general, the findings illustrate various factors for the resistance of psychotherapists regarding machine learning-based approaches. Dynamics and flexibility of treatment might be compromised and patients might experience decreased mindfulness and responsibility for their own actions. Such systems could further lead to decreased attendance of patients to therapy sessions and therefore a partial substitution - even though not qualitatively - of therapeutic treatment. Furthermore, additional workload and altered workflow appeared to be one factor that decreases the enthusiasm of the psychotherapist for a clinical application, which has also been found by (Trivedi et al., 2009). However, the design, implementation, and usage most certainly will create an increased workload. Data about the participants, which are mandatory for the application, needs to be collected and this process needs to be digitized. This data gathering process alone will create an effort for psychotherapists. Thus, there might be a hesitancy of technology adoption and the change of traditional workflows (Zheng et al., 2005). Data and information security barriers go along with the data gathering process and could decrease the patients' willingness to utilize such systems. Furthermore, major concerns exist regarding the relationship between psychotherapist and patient if MLCCDSSs are utilized in practice due to the artificial essence of such decision support systems. These concerns regarding the relationship between therapist and patient and how machine learning-based approaches could change the relationship into a connection between patient and health care system have also been discussed in the literature (Char et al., 2018).

However, potentials for the application of MLCCDSSs exist. Participants were interested in descriptive analysis of diary data in order to visualize trends and historical data. These types of analyses could increase the patients' understanding of their disorder and increase psychoeducation. The participants stated possible applications based on predictive analytics that could benefit their therapeutic process, which can lead to increased self-management, motivation, and valuable recommendations regarding the adjustment of treatment. Predictive approaches appeared to be valuable for the comparison of different treatment types and recommendations regarding treatment adjustment, point in time of beneficial drug prescription, and generally in the fields of addiction, relapse prevention, suicide, and dropout of therapeutic treatment.

Approaches that aim to predict rates of improvement over the course of therapy or intervention outcomes based on particular treatment types exist (Bremer et al., 2018; Lutz et al., 2005). Using such approaches can visualize the course of treatment or treatment success based on applied treatment types. However, these systems do currently not provide recommendations for the change of treatment type in order to reach greater treatment success. These types of recommendations would be difficult to realize since psychotherapy is not always a structured

process but relies on human interaction and interpretation. Machine learning systems that provide drug prescriptions also exist (Silva et al., 2018). However, to the best of my knowledge, there are no systems that predict if the usage of medication can lead to a decrease of symptoms in a particular disorder. And even if so, such systems should be used with caution and solely under an experts' supervision.

Processes for relapse prediction also exist in various studies for depression, eating disorders, alcohol abuse, or psychosis (Becker et al., 2018; Sullivan et al., 2017). These types of systems could represent a foundation for the development of MLCCDSSs in this field. Thus, machine learning-based tools for various mentioned fields of application by the participants have already been developed in studies. However, they have not been realized in clinical practice. Thus, what are the requirements for a successful development and application of MLCCDSSs?

The expected value of MLCCDSSs and their usage experience plays a major role as already indicated in the literature (Eichner and Das, 2010; Sacchi et al., 2015). The most important aspect might be the collaboration between psychotherapists and computer scientists. Without deep collaboration and interdisciplinary communication, no MLCCDSSs can be developed appropriately and can successfully be utilized in practice. Psychotherapists and patients should understand that the aim of MLCCDSSs is not the substitution of the psychotherapists' expertise but support of their decision-making process based on statistical analysis in order to increase treatment outcomes for the patient. The requirements and possibilities of MLCCDSSs should be defined by computer scientists and psychologists. Thus, the whole development cycle should be assisted and monitored by both parties. The system must be user-friendly and the utilized algorithms need to be comprehensible for the users (Trivedi et al., 2009).

The origin of the computerized decision support system is another important factor. Being an independent or perhaps public organization, communicating the bare interest of treatment success, and providing thorough documentation of the decision support system could be supporting arguments in order to persuade psychotherapists in using these systems in practice. On the other hand, this might not be enough. Educating therapists in the field of machine learning, their basic methods and structure, and highlighting their benefits as well as their limitations might be necessary (Char et al., 2018) to fully enable the utilization of such systems. Either way, proper evaluation of designed systems is necessary before dissemination in order to reach positive outcomes in treatment (Haynes and Wilczynski, 2010). Ethical commissions are necessary that can enable data security. Since participants stated that they would hesitate to tests MLCCDSSs at the patients' expense, research studies should be conducted that rigorously evaluate possible MLCCDSSs in interventions and if successful, highlight their actual benefit for psychotherapists as well as the patients.

Besides the insight this study provides, it also comes with limitations. Since thematic analysis relies on the interpretation of the researcher, the reliability of this method is of "concern" compared to content analysis (Namey et al., 2008). Interpretations may vary among investigators.

Additionally, the clinicians' mindset and attitude toward MLCCDSSs is an important factor. Their understanding of machine learning might differ even though introduced in the beginning of the interview. A more thorough approach could be the execution of a workshop that introduces the concept of machine learning to psychotherapists including the presentation of possible MLCCDSSs and a subsequent group discussion. Afterward, interviews could be conducted separately. This approach would lead to consistent knowledge about machine learning and deeper information among psychotherapists. Additionally, it is important to mention that the findings of this study cannot be generalized due to the small sample size and the restricted location of the interviews to Northern Germany. Conducting more interviews with psychotherapists from various countries could lead to different information. Nevertheless, the findings highlight important potentials, limitations as well as requirements for MLCCDSSs from a psychotherapists' perspective.

7.5 Conclusion

The findings of this study shed light on the current state of machine learning in clinical practice and potentials, limitations, and requirements of machine learning-based computerized clinical decision support systems. Various factors illustrated the hesitance of psychotherapists regarding these types of systems such as additional effort for the psychotherapist, problems regarding the relationship between therapist and patient, or less flexibility in the therapeutic process. Potentials were identified for descriptive as well as predictive analyses. Communication between IT and psychotherapist seems to be a crucial determinant for successful implementation of machine learning-based computerized clinical decision support systems in clinical practice. However, more research in this context is necessary to verify the results.

References

Alhojailan, M. (2012). Thematic Analysis: A critical review of its process and evaluation. *West East Journal of Social Sciences*, 1(1):39–47.

Attride-Jerling, J. (2001). Thematic networks: An analytical tool for qualitative research. *Qualitative Research*, 1(3):385–405.

Becker, D., Breda, W. V., Funk, B., Hoogendoorn, M., and Ruwaard, J. (2018). Predictive modeling in e-mental health : A common language framework. *Internet Interventions*, 12:57–67.

Becker, D., Bremer, V., Funk, B., Asselbergs, J., Riper, H., and Ruwaard, J. (2016). How to Predict Mood? Delving into Features of Smartphone-Based Data. In *Twenty-second Americas Conference on Information Systems*, San Diego (USA).

Belle, A., Kon, M. a., and Najarian, K. (2013). Biomedical informatics for computer-aided decision support systems: a survey. *The Scientific World Journal*, 2013:769639.

Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Resreach in Psychology*, 3(2):77–101.

Bremer, V., Becker, D., Kolovos, S., Funk, B., van Breda, W., Hoogendoorn, M., and Riper, H. (2018). Predicting Therapy Success and Costs for Personalized Treatment Recommendations Using Baseline Characteristics : Data-Driven Analysis. *Journal of Medical Internet Research*, 20(8):e10275.

Butcher, H., Holkup, P., Park, M., and Maas, M. (2001). Thematic Analysis of the Experience of Making a Decision to Place a Family Member With Alzheimer's Disease in a Special Care Unit. *Research in Nursing & Health*, 24:470–480.

Char, D., Shah, N., and Magnus, D. (2018). Implementing Machine Learning in Health Care - Addressing Ethical Challenges. *N Engl J Med*, 378(11):981–983.

Clifton, D. A., Niehaus, K. E., Charlton, P., and Colopy, G. W. (2015). Health Informatics via Machine Learning for the Clinical Management of Patients. *Yearbook of Medical Informatics*, 10(1):38–43.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78.

Eichner, J. and Das, M. (2010). Challenges and Barriers to Clinical Decision Support (CDS) Design and Implementation Experienced in the Agency for Healthcare Research and Quality CDS Demonstrations. Technical Report 10.

Fenn, K. and Byrne, M. (2013). The key principles of cognitive behavioural therapy. *InnovAiT*, 6(9):579–585.

Fielden, A. L., Sillence, E., and Little, L. (2011). Children's understandings' of obesity, a thematic analysis. *International Journal of Qualitative Studies on Health and Well-being*, 6(3):1–14.

Hannan, C., Lambert, M. J., Harmon, C., Nielsen, S. L., Smart, D. W., Shimokawa, K., and Sutton, S. W. (2005). A lab test and algorithms for identifying clients at risk for treatment failure. *Journal of Clinical Psychology*, 61(2):155–163.

Haynes, R. B. and Wilczynski, N. L. (2010). Effects of Computerized Clinical Decision Support Systems on Practitioner Performance and Patient Outcomes: Methods of a Decision-Maker-Researcher Partnership Systematic Review. *Implementation Science*, 5(12).

Iida, M., Shrout, P. E., Laurenceau, J.-P., and Bolger, N. (2012). Using diary methods in psychological research. In Cooper, H., Camic, P. M., Long, D. L., Panter, A. T., Rindskopf, D., and Sher, K. J., editors, *APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics.*, pages 277–305. American Psychological Association, Washington.

Jaques, N., Rudovic, O. O., Taylor, S., Sano, A., and Picard, R. (2017). Predicting Tomorrow ' s Mood , Health , and Stress Level using Personalized Multitask Learning and Domain Adaptation. *Journal of Machine Learning Research*, 66:17–33.

Joffe, H. and Yardley, L. (2004). Content and Thematic Analysis. In *Research methods for clinical and health psychology*. SAGE Publications.

Kaplan, B. (2001). Evaluating informatics applications - Some alternative approaches: Theory, social interactionism, and call for methodological pluralism. *International Journal of Medical Informatics*, 64(1):39–56.

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., and King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(195).

Lindblom, K., Gregory, T., Wilson, C., Flight, I. H., and Zajac, I. (2012). The impact of computer self-efficacy, computer anxiety, and perceived usability and acceptability on the efficacy of a decision support tool for colorectal cancer screening. *J Am Med Inform Assoc*, 19(3):407–12.

Lutz, W., Leach, C., Barkham, M., Lucock, M., Stiles, W. B., Evans, C., Noble, R., and Iveson, S. (2005). Predicting Change for Individual Psychotherapy Clients on the Basis of Their Nearest Neighbors. *Journal of Consulting and Clinical Psychology*, 73(5):904–913.

Namey, E., Guest, G., Thairu, L., and Johnson, L. (2008). Data Reduction Techniques for Large Qualitative Data Sets. In *Handbook for team-based qualitative research*. AltaMira Press.

Riano, D., Real, F., Lopez-Vallverdu, J. A., Campana, F., Ercolani, S., Mecocci, P., Annicchiarico, R., and Caltagirone, C. (2012). An ontology-based personalization of health-care knowledge to support clinical decisions for chronically ill patients. *J Biomed Inform*, 45(3):429–46.

Riessman, C. (1993). *Narrative Analysis*. Sage, London.

Sacchi, L., Quaglini, S., Lanzola, G., and Viani, N. (2015). Personalization and Patient Involvement in Decision Support Systems: Current Trends. *Yearbook of medical informatics*, 10(1):106–118.

Shatte, A. B. R., Hutchinson, D. M., and Teague, S. J. (2019). Machine learning in mental health : a scoping review of methods and applications. *Psychol Med*, 49(9):1426–1448.

Silva, P., Rivolli, A., Rocha, P., Correia, F., and Soares, C. (2018). Machine Learning for Drugs Prescription. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 548–555.

Stein, B. D., Kogan, J. N., Mihalyo, M. J., Schuster, J., Deegan, P. E., Sorbero, M. J., and Drake, R. E. (2013). Use of a computerized medication shared decision making tool in community mental health settings: impact on psychotropic medication adherence. *Community Ment Health J*, 49(2):185–92.

Sullivan, S., Northstone, K., Gadd, C., Walker, J., Margelyte, R., Richards, A., and Whiting, P. (2017). Models to predict relapse in psychosis : A systematic review. *PLoS ONE*, 12(9):e0183998.

Tiemens, B. and Kloos, M. W. (2016). Prediction of treatment outcome in daily generalized mental healthcare practice : first steps towards personalized treatment by clinical decision support. *European Journal for Person Centered Healthcare*, 4(1).

Triantafyllidis, A. K. and Tsanas, A. (2019). Applications of Machine Learning in Real-Life Digital Health Interventions : Review of the Literature. *Journal of Medical Internet Research*, 21(4):e12286.

Trivedi, M. H., Daly, E. J., Kern, J. K., Grannemann, B. D., Sunderajan, P., and Claassen, C. A. (2009). Barriers to implementation of a computerized decision support system for depression: an observational report on lessons learned in "real world" clinical settings. *BMC medical informatics and decision making*, 9(6).

van Breda, W., Bremer, V., Becker, D., Hoogendoorn, M., Funk, B., Ruwaard, J., and Riper, H. (2018). Predicting therapy success for treatment as usual and blended treatment in the domain of depression. *Internet Interventions*, 12:100–104.

Verbi Software (2016). MAXQDA Analytics Pro.

Zheng, K., Padman, R., Johnson, M. P., and Diamond, H. S. (2005). Understanding technology adoption in clinical care: Clinician adoption behavior of a point-of-care reminder system. *International Journal of Medical Informatics*, 74(7-8):535–543.