LEUPHANA
UNIVERSITÄT LÜNEBURG

Computerized assistance in online mental health treatment

Von der Fakultät Wirtschaftswissenschaften
der Leuphana Universität Lüneburg

zur Erlangung des Grades
Doktor der Naturwissenschaften
— Dr. rer. nat —

genehmigte Dissertation von
Dennis Becker

geboren am 17. März 1984 in Bremen

Eingereicht am: 12.12.2019
Mündliche Verteidigung (Disputation): 18.12.2020

Erstbetreuer und Erstgutachter: Prof. Dr. Burkhardt Funk
Zweitgutachter: Prof. Dr. Mark Hoogendoorn
Drittgutachter: Prof. Dr. Peter Niemeyer

Die einzelnen Beiträge des kumulativen Dissertationsvorhabens sind oder werden wie folgt veröffentlicht:

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

D. Becker, "Acceptance of Mobile Mental Health Treatment Applications", *Procedia - Procedia Computer Science* **98**, 220-227 (2016).

D. Becker, V. Bremer, F. Burkhardt, J. Asselbergs, H. Riper, and J. Ruwaard, "How to Predict Mood? Delving into Features of Smartphone-Based Data", *Proceedings of the 22nd Americas Conference on Information Systems* **22**, 1-10 (2016).

V. Bremer, D. Becker, B. Funk, and D. Lehr, "Predicting the individual mood level based on diary data", *Proceedings of the Twenty-Fifth Conference on Information Systems* **25**, 1161-1177 (2017).

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

D. Becker, V. Bremer, F. Burkhardt, M. Hoogendoorn, A. Rocha, and H. Riper, "Evaluation of a temporal causal model for predicting the mood of clients in an online therapy ", *Evidence-Based Mental Health* **23**, 27–33 (2020).

D. Becker, "Analysis of the Histogram Intersection Kernel for Use in Bayesian Optimization", *International Journal of Modeling and Optimization* **6**, (2017).

Veröffentlichungsjahr: 2020

Veröffentlicht im Onlineangebot der Universitätsbibliothek unter der URL:
http://www.leuphana.de/ub

LEUPHANA UNIVERSITÄT LÜNEBURG

Abstract

Institute of Information Systems

Doctor of Natural Sciences

Computerized assistance in online mental health treatment

by Dennis Becker

The wide accessibility of the Internet and web-based programs enable an increased volume of online interventions for mental health treatment. In contrast to traditional face-to-face therapy, online treatment has the potential to overcome some of the barriers such as improved geographical accessibility, individual time planning, and reduced costs. The availability of clients' treatment data fuels research to analyze the collected data to obtain a better understanding of the relationship among symptoms in mental disorders and derive outcome and symptom predictions. This research leads to predictive models that can be integrated into the online treatment process to assist clinicians and clients.

This dissertation discusses different aspects of the development of predictive modeling in online treatment: Categorization of predictive models, data analyses for predictive purposes, and model evaluation. Specifically, the categorization of predictive models and barriers against the uptake of mental health treatment are discussed in the first part of this dissertation. Data analysis and predictive modeling are emphasized in the second part by presenting methods for inference and prediction of mood as well as the prediction of treatment outcome and costs. Prediction of future and current mood can be beneficial in many aspects. Inference of users' mood levels based on unobtrusive measures or diary data can provide crucial information for intervention scheduling. Prediction of future mood can be used to assess clients' response to the treatment and expected treatment outcome. Prediction of the expected treatment costs and outcomes for different treatment types allows simultaneous optimization of these objectives and to increase the cost-effectiveness of the treatment. In the third part, a systematic predictive model evaluation incorporating simulation analyses is demonstrated and a method for model parameter estimation for computationally limited devices is presented.

This dissertation aims to overcome the current challenges of predictive model development and its use in online treatment. The development of predictive models for varies data collected in online treatment is demonstrated and how these models can be applied in practice. The derived results contribute to computer science and mental health research with client individual data analysis, the development of predictive models, and their statistical evaluation.

Acknowledgements

On my journey, I met many people that helped, encouraged, and never doubted me, which I can only restitute with my deepest gratitude.

First and foremost, my gratitude goes to my supervisors, Prof. Dr. Burkhardt Funk for his encouragement, patience, and guidance. My deepest gratitude also belongs to my second supervisor Prof. Dr. Mark Hoogendoorn for sparing neither trouble nor expense to supervise my dissertation as an external examiner. Likewise, my gratitude belongs to my third supervisor Prof. Dr. Peter Niemeyer for his advice and assistance and for keeping my progress on schedule.

I wish to express my gratitude to the European Comparative Effectiveness Research on Internet-based Depression Treatment (E-COMPARED) project for funding and thank all project participants for their support. Especially, my gratitude belongs to all my co-authors of the published research papers. I am deeply grateful for their support and contribution. I am grateful to the Leuphana Graduate School for financial support, which allowed me to present the results of my thesis at several international conferences.

Special thanks to the Leuphana Writing Center, in particular, Micha Edlich for his support and help in scientific writing. I would also like to thank my fellow Ph.D. students Vincent Bremer, Christoph Martin, Sebastian Mair, Martin Stange for their relentless support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Internet-based cognitive-behavioral therapy (ICBT) is treatment provided through a computer or a mobile device (De Graaf et al., 2009; Lambert, 2012). It is effective for the treatment of various psychiatric disorders such as mild to moderate depression (Giosan et al., 2017; Andrews et al., 2010), obsessive-compulsive disorder (Kyrios et al., 2014), anxiety, and panic disorder (Oromendia et al., 2016; Williams et al., 2016). Due to its wide range of application, it has become a fast-growing intervention channel in comparison to conventional psychotherapy (Kumar et al., 2017; Cuijpers et al., 2008). In contrast to conventional psychotherapy, online treatment helps in closing the gap between the demand and available treatment spots (Tzhak et al., 2004), provides treatment to people in rural areas (Sinclair et al., 2013), and is more cost-effective than traditional treatment (Myhr and Payne, 2006; Romero-Sanchiz et al., 2017; Lenhard et al., 2017).

Technological advances and their integration into healthcare allow increasing computerization of mental health treatment (Perna et al., 2018). The availability of mobile devices enables inquiring fine-grained measures by prompting questions in an individual's everyday life (Wichers et al., 2011; Trull and Ebner-Priemer, 2013). These measures allow new insights into the progression of clients during the treatment and are of benefit for therapists and researchers. Furthermore, mobile devices are equipped with a variety of sensors (e.g., proximity, accelerometer, temperature, humidity) that permit gathering information about current activities and environment (Dey, 2001).

With the increasing digitalization and clients' treatment tracking in recent years, the research in the field of online treatment of metal disorders flourishes (Riper et al., 2010). The collected data are a gateway to understand the psychological dynamics of mental disorders (Fisher and Boswell, 2016; Bak et al., 2016), discover knowledge (Ritterband et al., 2009), and develop support systems that can understand users' behavior and social context (Soares Teles et al., 2017). For the development of such systems, predictive models are required that combine scientific knowledge about humans and their mental and social functioning with context-aware measures. The aspiration behind these predictive models is to support therapists or clients in their day-to-day activities and help to improve wellbeing and quality of life. However, there is uncertainty regarding the challenges and barriers of these technological innovations and online treatment (Musiat, Goldstone, and Tarrier, 2014; Christensen and Hickie, 2010). There is a lack of knowledge in utilizing the collected data for knowledge discovery and the factors that influence treatment outcomes (Mcgrath et al., 2017).

Given the lack of knowledge and the importance of this field, it is necessary to research the use of treatment data to derive predictive models and their integration into online mental health treatment. For addressing these challenges, the research in this dissertation is grouped into three different parts: challenges that consider the development of online treatment programs, predictive model development, and model evaluation and parameter estimation.

To improve the model and treatment development process, this thesis examines the available literature of machine learning methods used in the field of online treatment. As a

result, a framework is derived that allows classification of the utilized method and guides the development of online treatments with computerized support. The introduction of new technologies requires recognition and acceptance of their users to reach their full potential (Taherdoost, 2018). For accessing the acceptance of mobile mental health treatment applications, a study was conducted that provides insights into inhibiting factors and barriers for their up-taking which contributes to the field of information systems.

The research of predictive models focuses on client data analysis to discover knowledge about symptom interaction and derive predictive models that can support the treatment process. For this purpose, a variety of data such as mobile phone measures, self-reported measures, and free-text diary data are utilized to derive predictions of clients' future mood and depression scores. Furthermore, screening questionnaires of clients form a European-wide depression study are utilized to demonstrate the use of this data to optimize the cost-effectiveness by simultaneously predicting the expected treatment outcome and costs for two different treatment types. This research contributes to e-mental health and computer science by providing models for client improvement and outcome prediction, cost-effectiveness optimization, and demonstrates the use of varies methods and data sources to predict symptom trajectories.

For statistical model evaluation, typically, cross-validation on study data and comparison to a reference model is used. To provide a more thorough evaluation, a novel evaluation method for models that predict multiple objectives is presented that combines model evaluation on study data with a simulation analysis. The simulation enables to estimate the expected performance on the study data and to analyze reasons for low model performance. Client-specific model parameters are required for individual symptom trajectory predictions. To estimate individual model parameters on computationally limited devices quickly, we provide a new method for model parameter estimation. The presented studies, contribute to computer science by demonstrating the use of a simulation analysis for model evaluation and a method for model parameter estimation for practical use in online treatments.

The introduction of this dissertation is structured as follows: First, research gaps in the area of online treatment are discussed. Second, the research objectives are connected to the conducted research. Third, the predictive model development process is illustrated and the development of a method that was utilized for mood prediction in two of the here presented articles is described in detail. Fourth, the main contributions are summarized and the strengths and weaknesses of each article are evaluated. Finally, the introduction concludes by discussing the results and possibilities for further research.

1.1 Research objectives in the field of online treatment

The last two decades have seen a proliferation of online-based mental health interventions (Arnberg et al., 2014). These interventions range from unsupported or unguided interventions, which provide fully automated self-help services (Leykin et al., 2014), to supported or guided interventions with remote therapist support (Adelman et al., 2014). The data collected in these interventions sparks an interest in machine learning, which is steadily integrated into online mental health treatment (Shatte, Hutchinson, and Teague, 2019; Rutledge, Chekroud, and Huys, 2019). For example, the use of intervention recommendations in self-help platforms (D'Alfonso et al., 2017) and the emergence of chatbots that provide fully automated mental health support (Kretzschmar et al., 2019). Such methods are trained on data that was collected from previous participants of the interventions. Yet, the development of such methods for use in online treatment begins with identifying treatment relevant objectives and assessing the perception of future users to identify barriers early on. For deriving models, data is analyzed and the effectiveness of different approaches is estimated.

Afterward, these models have to be systematically tested to ensure their statistical validity and provide evidence that their use in online treatment can provide benefits. Consequently, this cumulative dissertation discusses these three parts.

1.1.1 Decision support and perception of online mental health treatment

Online treatment gains increasing importance for the delivery of mental health interventions (Christensen and Griffiths, 2002; Cuijpers et al., 2008). Besides providing more people with the opportunity for treatment, online treatment promises to be more cost-effective than traditional face-to-face therapy (Tate et al., 2009; De Graaf et al., 2009; Proudfoot et al., 2003). For ensuring this cost-effectiveness, medical decision support systems can help to increase adherence and improve treatment outcomes (Vogenberg, 2009). In clinical psychology, the use of decision support systems has been evaluated (Triñanes et al., 2015) and applied to a wide range of tasks such as automated screening (Rollman et al., 2001), diagnosis support (Thomas et al., 2004), and treatment recommendations (Cannon and Allen, 2000; Kurian et al., 2009). Several studies have shown that decision support systems can improve the quality of provided care, prevent errors, reduce financial costs, and save human resources (Kawamoto et al., 2005; Steyerberg, 2009).

The integration of decision support systems into online treatment can provide similar advantages. However, a clear structure and definition of the models that could be utilized in online treatment are missing. Therefore, this challenge is addressed in this dissertation by describing the different phases of online interventions, the collected data, and the prediction targets. As part of this research, a framework is developed that allows to classify predictive models according to the treatment phase and its prediction target. Such a framework can be beneficial for the design of online interventions that aim to include predictive models to facilitate the online treatment process.

Mobile devices, such as cellphones and tablets, become increasingly important for the delivery of mobile intervention (Berry, Bucci, and Lobban, 2017). They provide sophisticated functions and user-friendly interfaces, which have become an important alternative to deliver treatment in daily practice (Varshney, 2014; Adibi, 2015). Mobile applications intend to provide services in any place and at any time, which significantly lowers geographical, temporal and organizational barriers (Silva et al., 2014). However, a client-centered design is an important challenge for the development of these applications. Tailoring these programs towards clients' needs and preferences is a key component for solving future healthcare challenges (Committee on Quality Health Care in America, 2011). Despite research on increasing the acceptance of online-based mental health treatment (Ebert et al., 2015) and acceptance of various mobile phone-based services (Zarmpou et al., 2012; Rao and Troshani, 2007), there is no adequate research on the perception of mental health treatment provided through a mobile phone. Insights obtained in the analysis help to identify inhibiting factors and challenges in the design of mobile treatment applications.

1.1.2 Data analysis and predictive models in online treatment

Mobile devices are an increasingly common approach to assess clients' mental health, behavior, and activities (Aung, Matthews, and Choudhury, 2017; Mohr, Zhang, and Schueller, 2017). They have brought forth significant advances in assessment techniques and replace traditional pen-and-paper questions because they allow a more easy and reliable assessment (Gibbons, 2017; Stone, Shiffman, and DeVries, 1999). Ecological momentary assessment (EMA) is the term used to describe the assessment of clients throughout the day in their natural environment (Robinson and Clore, 2002; Stone and Shiffman, 1994; Aan het

Rot, Hogenelst, and Schoevers, 2012). EMA can encompass a diversity of data such as diaries, open-text, and questions regarding the clients' behaviors, symptoms, and experiences with Likert-type responses (Gibbons, 2017). Besides, mobile devices enable the collection of unobtrusive data with the equipped senors (Aung, Matthews, and Choudhury, 2017). Such data provides the possibility to infer clients' location (Saeb et al., 2015b), activity (Lester, Choudhury, and Borriello, 2006), and stress level (Lu et al., 2012; Chang, Fisher, and Canny, 2011).

EMA data exhibit a strong correlation with weekly assessed depression questionnaires, which provides a link to clinical measures of depression (Saeb et al., 2015a). These daily symptom measures provide the foundation to model symptom interaction and help to understand the psychological dynamics connected with the illness (Fisher and Boswell, 2016; Bak et al., 2016). Following, these results can be used to develop predictive models for symptom prediction and personalization of online interventions (Hilvert-Bruce et al., 2012). Personalized treatment intends to customize healthcare to individuals based on their specific illness. Illness symptoms and severity can vary among individuals which subsequently can affect treatment decisions to optimize outcome (Arean, 2012). Although treatment personalization has become the center of development and research, there are no readily available tools that can be applied in practice (Perna et al., 2018).

Accordingly, this dissertation contributes to ongoing research on utilizing client data for deriving individual treatment and outcome predictions. This encompass various data types such as mobile phone measures, EMA assessments, free-text diaries, and questionnaires. Empirically collected data utilizing questionnaires often have missing values (Little and Rubin, 1989) and data collected during the treatment process from interventions and diaries can be unstructured (Barak and Grohol, 2011). This poses challenges from a computer science perspective, which leads to assumptions for data imputation of missing values and structuring of the data. Subsequently, predictive models require tailoring towards these varying data sources.

1.1.3 Validity of predictive models and personalized treatment

Machine learning and predictive models can improve clinical mental health treatment (Schnyer et al., 2017; Patel, Khalaf, and Aizenstein, 2016; Vogenberg, 2009) and its use is explored for online mental health treatment (D'Alfonso et al., 2017). Despite the impressive results of machine learning for diagnostic and predictive propose, there is a need for the evaluation of such models (Peyrou, Vignaux, and André, 2018). There always remains a risk that despite careful considerations the model does not provide accurate forecasts. Even for non-regulated applications, the technical soundness of the algorithms needs to be confirmed, especially in situations where health and life are at risk. Accordingly, for the validation of predictive models, empirical reproducibility, computational reproducibility, and statistical reproducibility are necessary to consider (Stodden, 2015). Moreover, the models developed become increasingly complex and allow prediction of trajectories for multiple objectives, which are barely covered by existing evaluation guidelines. This requires the development of new methods for model evaluation.

The integration of personalized treatment gains an increased interest and becomes a focus of research (Pritchard et al., 2017; Wittink et al., 2013). With the growing use of mobile devices, a wider assessment of client behavior is possible and ultimately individual mental health treatment (Aung, Matthews, and Choudhury, 2017). New types of models that consider individual context, which is assessed by continuous monitoring, open the path for personalized treatment (Sandstrom et al., 2016; Ben-Zeev et al., 2015). Personalized treatment may help to increase the number of positive clinical outcomes (Aung, Matthews, and Choudhury, 2017). In particular, models for individual treatment need adaptation to

individual users, which can be computationally demanding. Since these models are often provided on mobile devices and require fast parameter estimation, methods for individual parameter estimation on computationally limited devices are required.

1.2 Research objectives and included articles

This section connects the previously introduced research objectives with the included articles. It describes the structure of psychotherapeutic online interventions and emphasizes the advantages that computational models provide.

1.2.1 Predictive models in online treatment and acceptance of mobile treatment

Online therapy has been proven to be effective for the treatment of a variety of mental disorders (Andrews et al., 2010; Dölemeyer et al., 2013; Postel et al., 2010). Typically, online treatments utilize computerized cognitive behavioral therapy in the form of brief therapy (Furmark et al., 2009; McCrone et al., 2004). Brief therapies are usually shorter than traditional therapy. They are a focused application of therapeutic techniques to specifically target a symptom or behavior (Barry, 1999). In general, computerized treatment applications encompass screening and treatment functionality. Using a computer or mobile device, the interventions are processed by clients and screening questionnaires are inquired.

The inquired data can be utilized for the development of predictive models to identify treatment courses and forecast outcomes. Such tools could provide valuable information to therapists and allow them to allocate their time more efficiently among supervised clients. For a clinical environment, numerous studies have investigated biomarkers and client traits and their association with certain outcomes (Moons et al., 2009b) and defined guidelines for the development of predictive models in a clinical context (Lee, Bang, and Kim, 2016; Hemingway et al., 2013). Similar literature for online treatment is missing, therefore, Chapter 2 provides a framework that defines the typical treatment phases in an online intervention program, the various types of collected data, and categorizes the different types of predictive models that are applicable in online treatment. The presented framework is based on the available literature in the field of online treatment and implements a categorization of the models used in the literature. The article presents a structured overview of the online treatment process and provides guidance for the design of future online treatment that includes predictive models and enables researchers to categorize their models when describing them.

Given the persisting treatment gap in mental healthcare (Board, 2012), providing effective treatment via the Internet can assist in closing the gap between the demand and available treatment spots (Tzhak et al., 2004; Cameron and Thompson, 2005). Web-based treatment can be applied to a variety of mental health problems concerning mood (Cuijpers et al., 2014; Arnberg et al., 2014), anxiety (Olthuis et al., 2016; Peñate and Fumero, 2016), substance abuse (Gainsbury and Blaszczynski, 2011), eating disorders (Beintner, Jacobi, and Taylor, 2012), and allows to provide counseling, therapy, and aftercare (Lal and Adair, 2013; Riper et al., 2010). Despite the promising research findings, a poor client engagement (Deen, Fortney, and Schroeder, 2013; Apolinario-Hagen and Tasseit, 2015) and slow integration of e-mental health into clinical practice (Musiat, Goldstone, and Tarrier, 2014) are noticed, which requires the identification of barriers of acceptance and uptake of online treatment (Musiat and Tarrier, 2014; Apolinário-Hagen, Kemper, and Stürmer, 2017). Especially, for the rapid growth of mobile phone application that provide mental health care for all ages (Donker et al., 2013).

Therefore, Chapter 3 analyzes the acceptance of mobile mental health applications utilizing the technology acceptance models (TAM) (Davis, 1985). The TAM provides an empirically grounded framework to understand the facilitators and barriers of individuals' intention to

use and adopt new technologies (Venkatesh and Davis, 2000a). The results identify inhibiting factors and reveal the perception of mobile treatment applications, which are crucial for the development and advertisement of new Internet-based treatment.

1.2.2 Predictive models in online treatment

Self-reported measures are an important link to clinical health consequences. For example, self-reported mood measures are tied to scores of clinical depression (Cheng and Furnham, 2003) and strongly associated with longevity (Veenhoven, 2008). Furthermore, perceived stress is linked to susceptibility for infection and illness (Cohen, Tyrrell, and Smith, 1991), and self-reported health measures are even predictive for mortality risk (Aichele, Rabbitt, and Ghisletta, 2016). This strong association between self-reported measures and clinical relevance demonstrates the importance of data analysis and exploring their potential for predictive models. Predictive models are applicable for early diagnosis and disease stratification, selection between drug treatments, treatment adjustment, and individual prognosis (Zebley et al., 2016). These models can thus be readily applied to single clients to immediately derive useful clinical treatment objectives (Huys, Maia, and Frank, 2016; Bzdok and Meyer-Lindenberg, 2018).

These models are based on two major categories of machine learning: Supervised and unsupervised learning (Murphy, 2013). For supervised learning, labeled training data (the outcome is known) is used in either a classification or regression task. Classification only considers a binary outcome such as the presence or absence of disease. A regression task predicts a continues number such as depression severity. Unsupervised learning does not consider the outcomes and can be used for clustering similar clients according to their characteristics to derive groups of clients with similar attributes or for dimension reduction. Dimension reduction reduces the complexity of the problem by removing client attributes that show little variation or do not contribute to predicting the target label. For example, religion and ethnicity might not be relevant for the prediction of treatment outcome for a particular disease.

These machine learning methods are then applied to data to discover patterns that are predictive for the targeted variable such as the outcome of a treatment. For example, the prediction of future mood is beneficial in the case of depression because daily mood scores exhibit a significant relation to clinical depression scores (Aguilera, Schueller, and Leykin, 2015). However, there is still a need to analyze the large amount of data that is collected in online interventions to identify informative features and to provide a better understanding of the data (Mohr et al., 2013). To increase our knowledge, many of the articles included in this dissertation analyze data originating from the E-COMPARED (Kleiboer et al., 2016) depression study. In this European research study, the clinical effectiveness of blended therapy, which combines face-to-face meetings with online-based intervention, was compared to traditional clinical depression treatment. During the study period, valuable symptom questionnaires and EMA data were collected to enable detailed insight into clients' progress during the treatment.

Chapter 4 compares various machine learning methods on unobtrusively collected smartphone and EMA data for the inference of client individual mood levels. It estimates the predictive accuracy and aims to determine the relevance of a multitude of smart-phone measures. The use of unobtrusive measures provides the potential for client monitoring, reduced workload regarding symptom assessment, and can support clients with momentary intervention in previously stress-inducing situations (Heron and Smyth, 2010; Runyan et al., 2013). In Chapter 5 EMA data originating from the E-COMPARED study is utilized to shed more light on the topic of future mood and depression score prediction. Similarly to the analyzed smart-phone data, the predictive performance of different machine learning methods, which have been reported to provide the best results in other studies, are compared. Additionally, the importance of the EMA measures on future mood prediction is estimated and

EMA data is utilized to infer clients' depression scores. These depression scores can provide information about clients' progress during the treatment (Löwe et al., 2004). Client diaries provide self-reported measures that grant insight into clients' daily life and habits, which also provide information about their mood (Weinstein and Mermelstein, 2008). Accordingly, Chapter 6 analyzes the effect of daily activities on the mood level using activity diaries. Text mining is used to extract daily activities and a Bayesian model is applied to infer clients' daily mood levels from these activities. These analyzes contribute to ongoing research by exploring how heterogeneous data sources can be utilized to derive predictive models that can be used in practice to enable personalized treatment (Iniesta, Stahl, and McGuffin, 2016). The information provided by these models can assist therapists to provide clients with adequate interventions, prevent drop-out from treatment (Melville, Casey, and Kavanagh, 2010), and understand symptom improvement during treatment (Warmerdam et al., 2010).

Likewise, clients' demographics and questionnaire data can be utilized to predict treatment outcomes and expected treatment costs. Cost-effectiveness studies can provide great knowledge regarding the expected cost and outcome of treatment alternatives (Ryder et al., 2009). However, they do not provide an answer to individual preferences or the benefit of one specific treatment option over another. Typically, clients have to be assigned to a therapy type before treatment. Therapists or other clinicians often make these decisions based on personal understanding and experience, which involves high uncertainty and can result in non-optimal decisions (Ryder et al., 2009). This uncertainty can potentially result in worse treatment outcomes for individuals and increased healthcare costs. Simultaneously, policymakers and stakeholders increasingly demand cost-effectiveness evidence to support their conclusions and decisions (Knapp, 1999). A choice among different treatment types, such as in the E-COMPARED study, leads to the question for the better treatment type. Naturally, the answer depends on the individual. Therefore, Chapter 7 analyzes the E-COMPARED study data to derive that decision based on client individual baseline data. The article presents a set of tools that allow to optimize future treatments with respect to the expected costs and outcomes.

1.2.3 Model evaluation and parameter estimation

With the increasing development of predictive models for both diagnostic and prognostic predictions, there is an intensified interest in the methodology on model evaluation (Moons et al., 2009b; Ivanescu et al., 2016). Predictive models can inherit the bias presented in the training data set (Caliskan, Bryson, and Narayanan, 2017), and model development can be thought of in parallel to drug development (Rutledge, Chekroud, and Huys, 2019). This leads to the definition of guidelines for model development that encompass the definitions of outcomes and predictors, model evaluation and reporting (Steyerberg and Vergouwe, 2014). It has been shown, that suboptimal adherence to evaluation guidelines can limit the reliability and applicability of predictive models (Bouwmeester et al., 2012). As part of the E-COMPARED project, a predictive model was developed that simulates client individual trajectories for multiple symptoms at once (Altaf Hussain Abro, 2016). This model is based on psychological theories regarding the relationship between social integration and mood development. It utilizes various psychological concepts, for example, mood, social interaction, and enjoyed activities to simulate and predict these factors throughout the treatment. The available guidelines do not cover the statistical evaluation of such multi-objective models. Therefore, Chapter 8 demonstrates a model evaluation and proposes a more thorough model evaluation by combining study data and simulation analysis. Simulation is conducted by generating data using the model and enables to obtain empirical results about the model in specific scenarios, which are otherwise difficult to obtain. These simulations are necessary to improve the understanding and interpretation of the results (Morris, White, and Crowther, 2019).

For the use of such models in clinical practice, client individual parameters are crucial for deriving individual model predictions. Jaques et al. (2017) show in their study that client individual models significantly outperform one-fits-all models for future mood prediction because generic models do not consider client individual differences and therefore provide lower results in comparison. Similarly, Constantinides et al. (2018) found that personalized models provide more accuracy for mood prediction of bipolar clients and that generic models only perform close to baseline measures in their study. Given this strong evidence and need for client individual parameters, Chapter 9 analyzes Bayesian optimization using the histogram intersection kernel (HIK) for parameter estimation. Bayesian optimization is a black-box optimization algorithm designed for problems where the objective function is expensive to evaluate, not available, or without gradient information. It is a frequently used algorithm for solving challenging optimization tasks and has wide application for automatic tuning of machine learning algorithms (Hutter, Hoos, and Leyton-Brown, 2011; Snoek, Larochelle, and Adams, 2012; Wang, Mohamed, and Freitas, 2013). Usually, exponential kernels are utilized which scale cubically in computational demand and quadratically in memory usage with an increasing training data size. By utilizing the HIK, this computational burden can be reduced to subquadratic and linear memory requirements. Previously, the HIK has only been used in computer vision applications due to its fast training properties (Wu, Tan, and Rehg, 2011; Maji, Berg, and Malik, 2008). However, the HIK only provides a piecewise linear approximation, therefore, the use of this kernel in Bayesian optimization requires analysis.

1.3 Methods

This section focuses on data analysis and predictive model development. It describes the associated challenges in model development and connects the articles with the utilized machine learning methods. Then the development and implementation of a predictive model are discussed.

1.3.1 Data analysis and predictive model development

The articles presented here make use of a variety of methods from the field of machine learning. To obtain a deeper understanding of the processes that are involved in the development of predictive models, Figure 1.1 illustrates the main development steps.

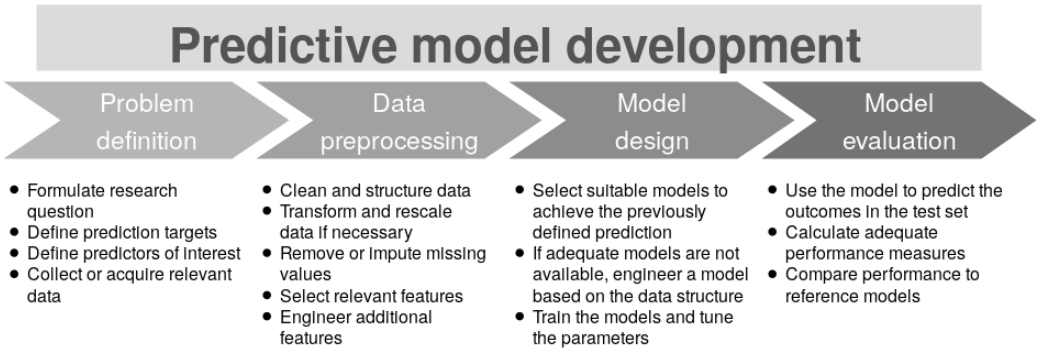


FIGURE 1.1: Main steps in predictive model development

Problem definition accounts for the identification and definition of the prediction target. This, for example, can be the treatment outcome and costs, as analyzed in Chapter 7, or future mood measures, which are examined in Chapter 4 and 5. There are, however, many other prediction targets such as diagnosis or treatment drop-out. An overview of potential

prediction targets during online treatment is provided in Chapter 2 (see Figure 2.2). With the definition of the prediction target possible predictors have to be identified. Such predictors can be recognized by investigation of the literature or the available data. If data is not available, it is collected in research studies or acquired from already conducted trials.

Data preprocessing encompass all necessary steps to process the data before a model can be trained. Study data is typically not in a format that can be utilized by a predictive model. It can have a different structure, contain irrelevant information for the prediction, and missing values. Data transformation can be rescaling of the data but might also imply that features have to be derived before use. For example, the study conducted in Chapter 6 extracts conducted activities from diary data, which are used afterward for mood prediction. Missing data are a problem in most studies that containing multi-item instrument questionnaires (Eekhout et al., 2012) and need to be accounted for by removing the cases or imputation. Imputation is a method to replace missing data with substituted values derived from the data of the available cases. Chapter 7 utilizes imputation by sampling from a distribution generated from the available cases. Only utilizing the relevant variables can often improve the prediction performance. To select contributing variables, methods such as Lasso regression can be used (Tibshirani, 1996). Lasso regression penalizes the parameters which can lead to parameters set to zero of non-contributing or highly correlating variables (Chapter 4 and 7).

Model design considers the selection or development of an appropriate model to predict the target utilizing the preprocessed data. The model to choose depends on the data and prediction target. For binary predictions, such as therapy success or failure a classification model is required. Prediction of a discrete questionnaire score favors an ordinal regression model and a continues mood level suggests a regression model. Furthermore, for the analysis of treatment data, hierarchical models are beneficial because they account for the hierarchical structure of the data. A predictive model can also consist of multiple methods, where the output of one model is used as input for the following model. In Chapter 6 a recurrent neural network is used to infer conducted activities in free-text and these activities are then used in a Bayesian model to predict the clients' mood level. After a model has been selected, it is trained on the study data except for a smaller validation and test set which are set aside for model parameter tuning and model evaluation. For model parameter tuning, random search, grid-based search, cross-validation (Chapter 7) or Bayesian optimization (Chapter 9) can be used.

Model evaluation indicates the model performance in praxis. The trained model is used to predict the data in the test set, which has not been used for model training or parameter tuning. Cross-validation is a method for model evaluation, where the model is trained and the prediction performance is estimated on changing subsets of the study data. Depending on the targeted prediction various performance measures are available. For binary predictions, the receiver operating characteristic (Metz, 1978) can be utilized for performance estimation and visualization. For regression targets, the squared and absolute error can be calculated. Additionally, it is beneficial to compare the obtained results to a reference model. This model provides a naive or state of the art prediction method and provides an objective comparison of the results. Furthermore, simulation analyses can be used for model evaluation, which enables an in-depth analysis of the model behavior. The benefits of an additional simulation analysis are demonstrated in Chapter 8.

1.3.2 Model development

Despite the availability of many machine learning methods, the algorithm that represents the data best can not be available. In these cases, the model needs to be designed and implemented. Bayesian methods allow to design models for varying types of data such as EMA or questionnaire data and have a long history of use in medical diagnosis. They provide

a tool to graphically represent knowledge regarding the data and to quantify variables as a probability distribution (Pearl, 1988). These models can handle incomplete data, allow regularization to avoid overfitting, and it is further possible to include expert knowledge (Inza, Larrañaga, and Sierra, 2001). For example, in Chapter 6 a Bayesian model for the prediction of clients' mood levels based on the daily activities is developed that accounts for the hierarchical structure of the data, the categorical mood ratings, and utilizes prior knowledge about the interaction between mood and activities derived from the literature.

For the analysis of EMA data in the articles in Chapter 4 and 5, a hierarchical Bayesian model for client individual mood prediction is developed. The employed model provides an estimate of the importance of the individual variables and utilizes variational approximation for fast parameter estimation. A fast model parameter estimation is beneficial for deriving predictions quickly and for deployment on mobile devices. To provide a deeper inside of how hierarchical models for EMA prediction can be developed, the following section describes the design, derivation, and implementation of the utilized model.

1.3.3 Bayesian models

Bayesian models or Bayesian networks are a graphical representation of a joint distribution of random variables. These random variables form an acyclic directed graph (DAG), where each node corresponds to a random variable and edges represent dependencies among these variables. Formally, the graph can be described as the tuple DAG = $(\mathbf{V}, \mathbf{E})$ which consists of a set of nodes $\mathbf{V}$ and a set of edges $\mathbf{E}$. For the graph to be directed, an edge between an ordered pair of two nodes $(A, B : A, B \in \mathbf{V} \wedge A \neq B)$ describes the binary relationship $A \rightarrow B$. The nodes (A, B) represent discrete or continues variables, and the edge indicates that B is dependent on A. Given a complete Bayesian model, which describes the relationship among the variables, unobserved variables can be inferred.

However, the exact inference of large Bayesian networks is intractable (Cooper, 1990) which makes approximate inference methods essential. For the estimation of such models usually, Markov Chain Monte Carlo (MCMC) methods are used, which approximate the model by sampling from a Markov chain with the stationary distribution of the posterior distribution (Hastings, 1970; Geman and Geman, 1970). Although these methods provide guarantees about the samples that are taken from the targeted density (Robert and Casella, 2005), they are computationally expensive even for small data sets. Thus, variational inference can be suited for larger data sets and scenarios where the model or a variety of models have to be estimated quickly (Beal, 2003; Jordan et al., 1999). It, however, only provides an approximation of the posterior distribution, does not provide the same guarantees such as MCMC methods, and underestimates the variance of the posterior distribution (Jordan et al., 1999).

1.3.4 Variational inference

For model development, we assume that the observed variables $\mathbf{X} = \{x_i, ..., x_n\}$ have been generated by the influence of several latent variables $\mathbf{Z}$. In probabilistic models, both of these are represented by a distribution that accounts for uncertainty. Under these latent variables, we summarize all parameters that are used to model the relations in the observed data. The derived probabilistic model is specified by the joint density $p(\mathbf{X}, \mathbf{Z})$. To derive an estimate of the latent variables given the observed data, the probability $p(\mathbf{Z}|\mathbf{X})$ needs to be calculated, which is called the posterior distribution. Using Bayes' rule the posterior distribution can be stated as:

$$p(\mathbf{Z}|\mathbf{X}) = \frac{p(\mathbf{X}, \mathbf{Z})}{p(\mathbf{X})} = \frac{p(\mathbf{X}|\mathbf{Z})p(\mathbf{Z})}{p(\mathbf{X})} = \frac{p(\mathbf{X}|\mathbf{Z})p(\mathbf{Z})}{\int p(\mathbf{X}, \mathbf{Z})d\mathbf{Z}}.$$

The denominator is called the marginal likelihood or model evidence which states the likelihood of the model given the observed data and can be utilized for model selection. Except for simple models, the calculation of the evidence is unavailable in closed-form. The inference of these intractable models is often based on Markov Chain Monte-Carlo methods (Hastings, 1970; Geman and Geman, 1970). Alternatives are variational methods that aim to approximate the posterior distribution p with a distribution q from a set of tractable distribution Q which results in a approximated solution for $q(\mathbf{Z}) \approx p(\mathbf{Z}|\mathbf{X})$. For measuring the similarity between q and p, variational inference uses the Kullback-Leibler (KL) (Kullback and Leibler, 1951) divergence. Essentially, the KL divergence measures the difference in information between two distributions as the expectation of the log difference between the probability of the data in the original distribution and the approximating distribution:

$$KL(q||p) = \int_{-\infty}^{\infty} q(x) \log \frac{q(x)}{p(x)} \, dx. \tag{1.1}$$

The difference between both distributions is always greater or equal to zero $KL(q||p) \geq 0$ for all q,p, and only equal to zero if $q = p$. This measure, however, is not a distance because it is not symmetrical. For deriving an expression for the variational approximation, the KL divergence of the posterior distribution $p(\mathbf{Z}|\mathbf{X})$ with respect to $q(\mathbf{Z})$ is estimated:

$$\begin{aligned} KL(q(\mathbf{Z})||p(\mathbf{Z}|\mathbf{X})) &= \int_{-\infty}^{\infty} q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X})} \, d\mathbf{Z}, \\ &= \int_{-\infty}^{\infty} q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{X}, \mathbf{Z})} \, d\mathbf{Z} + \log p(\mathbf{X}) \int_{-\infty}^{\infty} q(\mathbf{Z}) \, d\mathbf{Z}, \\ &= KL(q(\mathbf{Z})||p(\mathbf{X}, \mathbf{Z})) + \log p(\mathbf{X}). \end{aligned}$$

The result is obtained by separating the integrals, recognizing the new term as the KL divergence of the unnormalized posterior distribution, and the integral over $q(\mathbf{Z})$ is just the normalizing constant and integrates to 1. By maximizing the unnormalized KL divergence, one minimizes the above-defined KL divergence. With respect to the variational distribution $q(\mathbf{Z})$, the log model evidence $\log p(\mathbf{X})$ is constant. Since $KL(q(\mathbf{Z})||p(\mathbf{Z}|\mathbf{X})) \geq 0$, by rearranging the equation we get: $\log p(\mathbf{X}) = KL(q(\mathbf{Z})||p(\mathbf{Z}|\mathbf{X})) - KL(q(\mathbf{Z})||p(\mathbf{X}, \mathbf{Z}))$. One can notice that the log model evidence $\log p(\mathbf{X})$ is equal to the difference between the normalized and unnormalized KL divergence. Furthermore, the KL divergence of the unnormalized posterior distribution is a lower bound on the model evidence. Due to this property, the negative $KL(q(\mathbf{Z})||p(\mathbf{X}, \mathbf{Z}))$ is called the variational lower bound or the evidence lower bound (ELBO). Minimizing the lower bound amounts to maximizing a lower bound on the model evidence. To complete the specification of the optimization problem, the variational family Q needs to be defined. A widely used class of distributions is the mean-field approximation, which assumes an independent factorization:

$$q(\mathbf{Z}) = \prod_{m=1}^{M} q_m(Z_m).$$

Each latent variables Z_m is governed by its variational factor $q_m(Z_m)$, which renders the individual factors mutually independent. For the mean-field choice of Q, the problem can be optimized using coordinate descent, which iteratively optimizes each factor independently. This requires to iterate over the variational factors $q_m(Z_m)$ and for each m the evidence lower bound over q_m is optimized while keeping the other variational factors $q_j(Z_j)$ constant. This results in an optimal solution for each factor:

$$\ln q_m(Z_m) = \mathbb{E}_{j \neq m} \Big[\ln p(\mathbf{X}, \mathbf{Z}) \Big] + \mathrm{C}. \tag{1.2}$$

This procedure allows to fit the fully-factored $q(\mathbf{Z}) = q_1(Z_1) q_2(Z_2) \cdots q_m(Z_m)$ approximation of $p(\mathbf{Z}|\mathbf{X})$.

1.3.5 Hierarchical variational Bayesian regression

The target variable $\mathbf{y}_i \in \mathbb{R}$ is the client individual mood level that depends on independent variables $\mathbf{x}_i \in \mathbb{R}^D$ for each client. Since these clients all undergo the same treatment, they share similarities regarding the influence of the independent variables on the mood level, which can be captured with a hierarchical structure. For simplification, a linear relationship between the mood level and the independent variables is assumed which results in a hierarchical linear regression. The derived model is similar to the Bayesian linear regression model derived by Drugowitsch (2013) and Bishop (2006, Chapter 10). The here presented model utilizes an additional hierarchical prior to capture similarities among clients. The complete model is shown in plate notation in Figure 1.2.

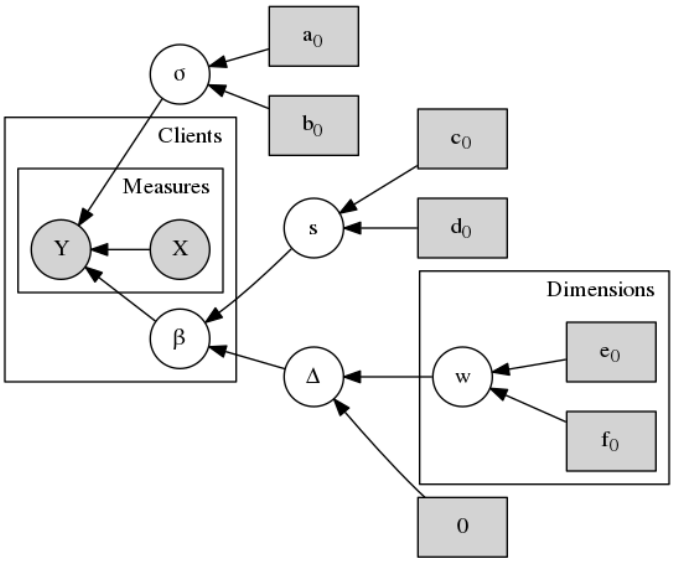


FIGURE 1.2: Plate notation of the hierarchical linear regression model

The filled nodes are observed or known and circles typically represent distributions, whereas boxes represent constants. The rectangles with names in them are plates, which indicate that their elements are observed multiple times. The model definition starts from the left of the illustration with the observed data $\mathbf{Y}$, where continues measures of mood are to be predicted, which represents a regression model. In this case, the normal distribution $\mathcal{N}(y \mid \mu, \sigma)$ is used, which has two parameters the mean value μ and a variance parameter σ. To develop this model, these parameters have both to be specified by either a distribution or a constant value. By continuing in this manner, these models appear to naturally grow in one direction. The surrounding plates illustrate that there are multiple clients with repeated observations of the targeted mood variable. The uncertainty regarding the mood measures is represented by the σ parameter. This parameter is specified by the conjugate prior of the precision a gamma distribution $\text{Gamma}(\sigma|a_0, b_0)$. The Gamma distribution has two parameters that require specification. For both of these parameters small constant values are utilized, which represents an uninformative prior. By specifying the prior as a distribution, the uncertainty will be estimated from the data. The μ parameter of the target variable is a matrix multiplication of the observed data $\mathbf{x}_{i,c}$ and a client-specific weight or parameter vector $\boldsymbol{\beta}_i$. For $\boldsymbol{\beta}_i$, the conjugate prior of the mean value a normal distribution is chosen. Utilizing conjugate priors will result in a closed-form solution when deriving the update equations later. If non-conjugate priors are used, one has to employ further methods such as Laplace approximation (Wang and Blei, 2013). This completes the description of the

observed data and allows to state the factors:

$$p(\mathbf{Y} \mid \mathbf{x}\boldsymbol{\beta}, \sigma) = \prod_{i=1}^{C} \prod_{c=1}^{M} \mathcal{N}(y_{i,c} \mid \mathbf{x}_{i,c}^{T} \boldsymbol{\beta}_i, \sigma^{-1}),$$
$$p(\sigma) = \text{Gamma}(\sigma \mid a_0, b_0).$$

This results in new client individual parameters $\boldsymbol{\beta}_i$ that need to be specified. The prior of the mean value for the client individual parameters, which is named Δ, represents the hierarchical prior. The precision s accounts for the uncertainty in the hierarchical prior. The client-specific factors are defined as follows:

$$p(\boldsymbol{\beta} \mid \Delta, s) = \prod_{i=1}^{C} \mathcal{N}\left(\boldsymbol{\beta}_i \mid \Delta, s^{-1}\right),$$
$$p(s) = \text{Gamma}(s \mid c_0, d_0).$$

For the prior on the mean of the hierarchical prior Δ, the constant 0 is used, which is a quadratic penalty on the parameter values. This penalty shrinks the parameters towards 0 similar to a quadratic regularization term in ridge regression (Bishop, 2006, Chapter 3). The chosen Gamma distribution for the prior on the precision has the role of the penalty factor similar to ride regression. For each observed variable (or dimension) an individual Gamma distribution is used. The reasoning is that during optimization of the model irrelevant parameters will shrink automatically. This process is known as automatic relevance determination (ARD) (MacKay, 1992; Tipping, 2000; Wipf and Nagarajan, 1992). Specifically, the Gamma distribution penalizes a deviation from zero heavier for non-contributing variables and less for contributing variables. This concludes the model construction with the final factors:

$$p(\Delta \mid 0, \mathbf{w}) = \mathcal{N}(\Delta \mid 0,\ \mathbf{w}^{-1}),$$
$$p(\mathbf{w}) = \prod_{d=1}^{D} \text{Gamma}(\mathbf{w}_d \mid e_0, f_0).$$

After the models is completely defined, the model's joint probability can be stated as: $p(\mathbf{Y}, \mathbf{x}, \boldsymbol{\beta}, \sigma, \Delta, s, \mathbf{w}) = p(\mathbf{Y} \mid \mathbf{x}\boldsymbol{\beta}, \sigma)p(\boldsymbol{\beta} \mid \Delta, s)p(\Delta \mid 0, \mathbf{w})p(\sigma)p(s)p(\mathbf{w})$. To derive the approximation, the posterior distribution is replaced by the factored variational posterior distribution $p(\boldsymbol{\beta}, \Delta, \sigma, s, \mathbf{w} \mid \mathbf{D}) \approx q(\boldsymbol{\beta})q(\Delta)q(\sigma)q(s)q(\mathbf{w})$. The model parameters can be estimated by using the coordinate descent algorithm stated in Equation 1.2. For implementing the coordinate descent algorithm, the update equations for the variational factors have to be derived.

Derivation of q(β_i)

To derive the update equations, we select the factors that depend on $\boldsymbol{\beta}_i$ from the joint probability, while the remaining factors are constant. The factors are replaced with their distributions and the terms are multiplied out to separate the terms. The terms that do not depend on $\boldsymbol{\beta}_i$ are constant and absorbed into the additional constant term. The remaining terms are rearranged and grouped until a recognizable distribution is received. This is the case because conjugate priors have been used, which will result in a tractable solution.

$$
\begin{aligned}
\ln q^*_{\boldsymbol{\beta}_i}(\boldsymbol{\beta}_i) &= \mathrm{E}_{\Delta,\sigma,s}\left[\ln p(\mathbf{y}_i \mid \mathbf{x}_i^T\boldsymbol{\beta}_i, \sigma) + \ln p(\boldsymbol{\beta}_i \mid \Delta, s)\right] + \mathrm{C} \\
&= \mathrm{E}_{\sigma}\left[-\frac{\sigma}{2}(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_i)^T(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_i)\right] + \mathrm{E}_{\Delta,s}\left[-\frac{s}{2}(\Delta - \boldsymbol{\beta}_i)^T(\Delta - \boldsymbol{\beta}_i)\right] + \mathrm{C} \\
&= \mathrm{E}_{\sigma}\left[\sigma\boldsymbol{\beta}_i^T\mathbf{x}_i^T\mathbf{y}_i - \frac{\sigma\boldsymbol{\beta}_i^T\mathbf{x}_i^T\mathbf{x}_i\boldsymbol{\beta}_i}{2}\right] + \mathrm{E}_{\Delta,s}\left[s\Delta^T\boldsymbol{\beta}_i - \frac{s\boldsymbol{\beta}_i^T\boldsymbol{\beta}_i}{2}\right] + \mathrm{C} \\
&= \boldsymbol{\beta}_i^T \underbrace{\left(\mathrm{E}_{\sigma}\left[\sigma\right]\mathbf{x}_i^T\mathbf{y}_i + \mathrm{E}_s\left[s\right]\mathrm{E}_{\Delta}\left[\Delta\right]\right)}_{\text{Mean}} - \frac{1}{2}\boldsymbol{\beta}_i^T \underbrace{\left(\mathrm{E}_{\sigma}\left[\sigma\right]\mathbf{x}_i^T\mathbf{x}_i + \mathrm{E}_s\left[s\right]\right)}_{\text{Precision}}\boldsymbol{\beta}_i + \mathrm{C}
\end{aligned}
$$

By completing the square over $\boldsymbol{\beta}$, the parameters of a normal distribution are derived. The expectations of the variables with respect to their variational distribution have to be evaluated. The variational distribution of the terms $\mathrm{E}_{\sigma}\left[\sigma\right]$ and $\mathrm{E}_s\left[s\right]$ are a Gamma distributions and the expected value for a Gamma distribution is $\mathrm{E}\left[\mathrm{Gamma}(a,b)\right] = \frac{a}{b}$. Consequently, the expected values are $\mathrm{E}_{\sigma}\left[\sigma\right] = \frac{a_n}{b_n}$ and $\mathrm{E}_s\left[s\right] = \frac{c_n}{d_n}$ respectively. The expectation for Δ is with respect to a normal distribution, the expected value of a normal distribution is the mean, which results into the expected value of $\mathrm{E}_{\Delta}\left[\Delta\right] = \Delta$. Replacing the expectations provides the following equations for $\boldsymbol{\beta}_i$:

$$\lambda_{\boldsymbol{\beta}_i} = \frac{a_n}{b_n}\mathbf{x}_i^T\mathbf{x}_i + \frac{c_n}{d_n}, \tag{1.3}$$

$$\boldsymbol{\beta}_i = \frac{\frac{a_n}{b_n}\mathbf{x}_i^T\mathbf{y}_i + \frac{c_n}{d_n}\Delta}{\lambda_{\boldsymbol{\beta}_i}}. \tag{1.4}$$

The update equations show that the client individual parameters dependent on their data and the hierarchical prior weighted by σ and s respectively.

Derivation of q(Δ)

Next, the update equations for the variational posterior distribution q(Δ) is derived. The factors are simplified and the terms that depend on Δ are selected while the remaining terms are constant. By rearranging the terms one can recognize a resulting normal distribution.

$$
\begin{aligned}
\ln q^*_{\Delta}(\Delta) &= \mathrm{E}_{\boldsymbol{\beta}_i,s}\left[\ln p(\boldsymbol{\beta}_i \mid \Delta, s) + \ln p(\Delta \mid \mu_0, \mathbf{w})\right] + \mathrm{C} \\
&= \mathrm{E}_{\boldsymbol{\beta}_i,s}\left[\sum_{i=1}^{C} -\frac{s}{2}(\Delta - \boldsymbol{\beta}_i)^T(\Delta - \boldsymbol{\beta}_i)\right] + \mathrm{E}_{\mathbf{w}}\left[-\frac{\mathbf{w}}{2}(\Delta - \mu_0)^T(\Delta - \mu_0)\right] + \mathrm{C} \\
&= \mathrm{E}_{\boldsymbol{\beta}_i,s}\left[-\frac{Cs\Delta^T\Delta}{2} + s\Delta^T\sum_{i=1}^{C}\boldsymbol{\beta}_i\right] + \mathrm{E}_{\mathbf{w}}\left[\mathbf{w}\Delta^T\mu_0 - \frac{\mathbf{w}\Delta^T\Delta}{2}\right] + \mathrm{C} \\
&= \mathrm{E}_{\boldsymbol{\beta}_i,s}\left[-\frac{1}{2}\Delta^T(Cs + \mathbf{w})\Delta\right] + \mathrm{E}_{\mathbf{w},s,\boldsymbol{\beta}_i}\left[\Delta^T(s\sum_{i}^{C}\boldsymbol{\beta}_i + \mathbf{w}\mu_0)\right] + \mathrm{C}
\end{aligned}
$$

Similar to the previous derivation, expected values for $\mathrm{E}_{\mathbf{w}}\left[\mathbf{w}\right] = \frac{e_n}{\mathbf{f}_n}$, and $\mathrm{E}_{\boldsymbol{\beta}_i}\left[\boldsymbol{\beta}_i\right] = \boldsymbol{\beta}_i$ need to be replaced. The expected value of s has already been stated in the update of the previous factor. This leads to the following update equations for Δ:

$$\lambda_\Delta = C\frac{c_n}{d_n} + \frac{e_n}{\mathbf{f}_n}, \tag{1.5}$$

$$\Delta = \frac{\frac{c_n}{d_n}\sum_i^C \boldsymbol{\beta}_i + \frac{e_n}{\mathbf{f}_n}\mu_0}{\lambda_\Delta}. \tag{1.6}$$

$\mathbf{f_n}$ is a vector with dimensions entries whereas e_n is a single value. Each variable has an individual Gamma distribution, but we assume the prior value e_0 for all parameters to be the same in this model. This representation is further utilized in the update for $\mathbf{w}$. The update equations show that the hierarchical prior is similar to an average parameter. It consists of the sum of all client individual parameters which is scaled by s, which measures the variance in parameters among the clients.

Derivation of q(σ)

For the definition of σ, the conjugate prior is used which will result in a Gamma distribution.

$$\begin{aligned}
\ln q^*_\sigma(\sigma) &= \mathrm{E}_{\boldsymbol{\beta}_i}\left[\ln p(\mathbf{Y} \mid \mathbf{x}\boldsymbol{\beta}, \sigma) + \ln p(\sigma \mid a_0, b_0)\right] + \mathrm{C} \\
&= \mathrm{E}_{\boldsymbol{\beta}_i}\left[\sum_{i=1}^{C}\sum_{c=1}^{M}\frac{1}{2}\ln\sigma - \frac{\sigma}{2}(y_{i,c} - \mathbf{x}_{i,c}^T\boldsymbol{\beta}_i)^2\right] + \left[(a_0-1)\ln\sigma - b_0\sigma\right] + \mathrm{C} \\
&= \mathrm{E}_{\boldsymbol{\beta}_i}\left[\sum_{i=1}^{C}\frac{M_i}{2}\ln\sigma - \frac{\sigma}{2}(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_i)^T(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_i)\right] + \left[(a_0-1)\ln\sigma - b_0\sigma\right] + \mathrm{C} \\
&= \mathrm{E}_{\boldsymbol{\beta}_i}\left[-\frac{\sigma}{2}\sum_{i=1}^{C}(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_i)^T(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_i)\right] + (a_0-1)\ln\sigma - b_0\sigma + \frac{\sum_{i=1}^{C}M_i}{2}\ln\sigma + \mathrm{C} \\
&= \mathrm{E}_{\boldsymbol{\beta}_i}\left[-\frac{\sigma}{2}\sum_{i=1}^{C}\mathbf{y}_i^T\mathbf{y}_i - 2\boldsymbol{\beta}_i^T\mathbf{x}_i^T\mathbf{y}_i + \boldsymbol{\beta}_i^T\mathbf{x}_i^T\mathbf{x}_i\boldsymbol{\beta}_i\right] + \left((a_0-1) + \frac{\sum_{i=1}^{C}M_i}{2}\right)\ln\sigma - b_0\sigma + \mathrm{C} \\
&= \mathrm{E}_{\boldsymbol{\beta}_i}\left[\underbrace{-\left(\left(\sum_{i=1}^{C}\boldsymbol{\beta}_i^T\mathbf{x}_i^T\mathbf{x}_i\boldsymbol{\beta}_i - 2\sum_{i=1}^{C}\boldsymbol{\beta}_i^T\mathbf{x}_i^T\mathbf{y}_i + \sum_{i=1}^{C}\mathbf{y}_i^T\mathbf{y}_i\right)\frac{1}{2} + b_0\right)\sigma}_{\text{Rate parameter b}}\right] \\
&\quad + \underbrace{\left((a_0-1) + \frac{\sum_{i=1}^{C}M_i}{2}\right)}_{\text{Shape parameter a}}\ln\sigma + \mathrm{C}
\end{aligned}$$

The result is a log Gamma distribution. To derive the update equation, the expected value of β^2 has to be estimated. Note that $\mathrm{E}\left[\mathbf{X}^2\right] = \mathrm{E}\left[\mathbf{X}\right]^2 + \mathrm{Var}(\mathbf{X})$. This leads to $\mathrm{E}_{\boldsymbol{\beta}_i}\left[\boldsymbol{\beta}_i^2\right] = \boldsymbol{\beta}_i^2 + \lambda_{\boldsymbol{\beta}_i}$ for the expectation of $\boldsymbol{\beta}_i^2$. The variable M_i represents the number of samples for the specific client, and their sum is the number of all observed samples. By substitution of the expectations, the following update equations are derived:

$$N = \sum_{i=1}^{C} M_i,$$

$$a_n = a_0 + \frac{N}{2}, \tag{1.7}$$

$$b_n = b_0 + \frac{1}{2}\left(\sum_{i=1}^{C} (\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_i)^T(\mathbf{y}_i - \mathbf{x}_i\boldsymbol{\beta}_i) + \sum_{i}^{C} \mathbf{x}_i^T \lambda_{\boldsymbol{\beta}_i}\mathbf{x}_i\right). \tag{1.8}$$

Considering the update equation for b_n, the first term is the sum of residual errors over all clients and the second term is the sum of the standard errors. Assuming that the residual error is high in comparison to the number of samples, the expected value $\frac{a_n}{b_n}$ of the precision would be low, which suggests a strong uncertainty in the measures. Similarly, if there is a strong uncertainty in the estimated clients' parameters, the measures are expected to be less reliable.

Derivation of q(s)

The result for factor s, which regulates the influence of the hierarchical prior on the client individual weights, will also be a Gamma distribution.

$$\begin{aligned}
\ln q_s^*(s) &= \mathrm{E}_{\boldsymbol{\beta}_i,\Delta}\left[\ln p(\boldsymbol{\beta}_i \mid \Delta, s) + \ln p(s \mid c_0, d_0)\right] + \mathrm{C} \\
&= \mathrm{E}_{\boldsymbol{\beta}_i,\Delta}\left[\sum_{i=1}^{C}\sum_{d=1}^{D} \frac{1}{2}\ln s - \frac{s}{2}(\boldsymbol{\beta}_{i,d} - \Delta_d)^2\right] + \left[(c-1)\ln s - d_0 s\right] + \mathrm{C} \\
&= \mathrm{E}_{\boldsymbol{\beta}_i,\Delta}\left[\sum_{i=1}^{C}\sum_{d=1}^{D} -\frac{s}{2}(\boldsymbol{\beta}_{i,d} - \Delta_d)^2\right] + \left((c_0 - 1) + \frac{\sum_{i=1}^{C}\sum_{d=1}^{D} 1}{2}\right)\ln s - d_0 s + \mathrm{C} \\
&= \mathrm{E}_{\boldsymbol{\beta}_i,\Delta}\left[-\left((\sum_{i=1}^{C}\sum_{d=1}^{D}\boldsymbol{\beta}_{i,d}^2 - 2\sum_{i=1}^{C}\sum_{d=1}^{D}\boldsymbol{\beta}_{i,d}\Delta_d + \sum_{d=1}^{D}\Delta_d^2)\frac{1}{2} + d_0\right)s\right] \\
&\quad + \left((c_0 - 1) + \frac{CD}{2}\right)\ln s + \mathrm{C}
\end{aligned}$$

The result contains the squared expectations of $\boldsymbol{\beta}$ and Δ, where the expected value for Δ is $\mathrm{E}_\Delta\left[\Delta^2\right] = \Delta^2 + \lambda_\Delta$. After substitution and simplification of the terms, the following update equations are derived:

$$c_n = c_0 + \frac{DC}{2}, \tag{1.9}$$

$$d_n = d_0 + \frac{1}{2}\left(\sum_{i=1}^{C}\sum_{d=1}^{D}(\boldsymbol{\beta}_{i,d} - \Delta_d)^2 + \mathrm{Spur}(\lambda_\Delta) + \sum_{i}^{C}\mathrm{Spur}(\lambda_{\boldsymbol{\beta}_i})\right). \tag{1.10}$$

The update equations show that the influence of the hierarchical prior depends on the squared differences between the individual parameters and the hierarchical prior, the variance of the hierarchical prior, and the variance of the client individual parameters. If the difference between the hierarchical prior and the clients' parameters is large, the hierarchical prior has less influence on the individual clients' parameters. Similarly, the variance of the hierarchical prior and the variance of the client individual parameters reduce the influence of the prior.

Derivation of $q(\mathbf{w})$

Finally, the update equation for $\mathbf{w}$ is derived. For simplicity, only the derivation for one dimension is shown.

$$\begin{aligned}
\ln q^*_{\mathbf{w}_d}(\mathbf{w}_d) &= \mathrm{E}_\Delta \left[\ln p(\Delta_d \mid 0, \mathbf{w}_d) + \ln p(\mathbf{w}_d \mid e_0, f_0)\right] + \mathrm{C} \\
&= \mathrm{E}_\Delta \left[\frac{1}{2}\ln \mathbf{w}_d - \frac{\mathbf{w}}{2}(\Delta_d)^2\right] + \left[(e_0 - 1)\ln \mathbf{w}_d - f_0 \mathbf{w}_d\right] + \mathrm{C} \\
&= \mathrm{E}_\Delta \left[-\frac{\mathbf{w}_d}{2}(\Delta_d)^2\right] + ((e_0 - 1) + \frac{1}{2})\ln \mathbf{w}_d - f_0 \mathbf{w}_d + \mathrm{C} \\
&= \mathrm{E}_\Delta \left[-\mathbf{w}_d(\frac{1}{2}(\Delta_d)^2 + f_0)\right] + ((e_0 - 1) + \frac{1}{2})\ln \mathbf{w}_d + \mathrm{C}
\end{aligned}$$

The result is a log Gamma distribution. The update equation is stated with $\mathbf{f}$ as a vector:

$$e_n = e_0 + \frac{1}{2}, \tag{1.11}$$

$$\mathbf{f}_n = f_0 + \frac{1}{2}(\Delta^2 + \mathrm{trace}(\lambda_\Delta)). \tag{1.12}$$

Inspecting the update equation for $\mathbf{w}$ reveals how this factor penalizes the parameters for each dimension. The first term is a quadratic penalty of the parameter values and the second term adds a penalty for the variance in the individual parameters. If a particular parameter varies among the clients, it is stronger penalized than when it shows less variation among the clients.

Variational lower bound

After all update equations have been derived, an iterative update of the variational factors will maximize the variational lower bound. To determine the convergence of the algorithm, the change in the variational lower bound can be used. This allows to monitor the optimization process and provides an approximation of the evidence $\ln p(\mathbf{x})$. Therefore, we estimate the negative *KL* divergence (Equation 1.1) of the model's joint distribution with respect to the variational distribution. The evidence lower bound is defined as the negative divergence which results in the exchange of the factors in the log fraction. The variational lower bound $\mathcal{L}(q)$ is then given by:

$$\begin{aligned}
\mathcal{L}(q) &= \int\int\int\int\int q(\boldsymbol{\beta}, \Delta, \sigma, s, \mathbf{w}) \ln\left\{\frac{p(\boldsymbol{\beta}, \Delta, \sigma, s, \mathbf{w} \mid \mathbf{D})}{q(\boldsymbol{\beta}, \Delta, \sigma, s, \mathbf{w})}\right\} d\boldsymbol{\beta}\, d\Delta\, d\sigma\, ds\, d\mathbf{w}, \\
&= \mathbb{E}_{\boldsymbol{\beta},\Delta,\sigma,s,\mathbf{w}}\Big[\ln p(\boldsymbol{\beta}, \Delta, \sigma, s, \mathbf{w} \mid \mathbf{D})\Big] - \mathbb{E}_{\boldsymbol{\beta},\Delta,\sigma,s,\mathbf{w}}\Big[q(\boldsymbol{\beta}, \Delta, \sigma, s, \mathbf{w})\Big], \\
&= \mathbb{E}_{\boldsymbol{\beta},\sigma}\Big[\ln p(\mathbf{Y} \mid \mathbf{x}\boldsymbol{\beta}, \sigma)\Big] + \mathbb{E}_{\boldsymbol{\beta},\Delta,s}\Big[\ln p(\boldsymbol{\beta} \mid \Delta, s)\Big] + \mathbb{E}_{\mathbf{w}}\Big[\ln p(\Delta \mid 0, \mathbf{w})\Big] \\
&\quad + \mathbb{E}_{\sigma}\Big[\ln p(\sigma)\Big] + \mathbb{E}_{s}\Big[\ln p(s)\Big] + \mathbb{E}_{\mathbf{w}}\Big[\ln p(\mathbf{w})\Big] \\
&\quad - \mathbb{E}_{\boldsymbol{\beta}}\Big[\ln q(\boldsymbol{\beta})\Big] - \mathbb{E}_{\Delta}\Big[\ln q(\Delta)\Big] - \mathbb{E}_{\sigma}\Big[\ln q(\sigma)\Big] - \mathbb{E}_{\mathbf{w}}\Big[\ln q(\mathbf{w})\Big].
\end{aligned} \tag{1.13}$$

The terms that involve expectations of the variational distributions $\log q(\cdot)$ are the entropies $\mathbb{H}(\cdot)$ of that distribution. The various terms are given in the following:

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\beta},\sigma}\Big[\ln p(\mathbf{Y} \mid \mathbf{x}^T\boldsymbol{\beta}, \sigma)\Big] &= \frac{N}{2}(\psi(a_n) - \ln b_n) - \frac{\sigma}{2}\left(\sum_{i=1}^{C}(\mathbf{x}_i\boldsymbol{\beta}_i - \mathbf{y}_i)^T(\mathbf{x}_i\boldsymbol{\beta}_i - \mathbf{y}_i) + \sum_{i=1}^{C}\mathbf{x}_i^T\lambda_{\boldsymbol{\beta}_i}\mathbf{x}_i\right),\\
\mathbb{E}_{\beta,\Delta,\sigma}\Big[\ln p(\boldsymbol{\beta} \mid \Delta, s)\Big] &= \frac{C}{2}(\psi(c_n) - \ln d_n) - \frac{s}{2}\left(\sum_{i=1}^{C}(\boldsymbol{\beta}_i - \Delta)^2 + \mathrm{Spur}(\lambda_\Delta) + \sum_{i=1}^{C}\mathrm{Spur}(\lambda_{\boldsymbol{\beta}_i})\right),\\
\mathbb{E}_{\Delta,\sigma}\Big[\ln p(\Delta \mid 0, \mathbf{w})\Big] &= \sum_{d=1}^{D}\frac{D}{2}\psi(e_n) - \ln f_{n_d} - \sum_{d=1}^{D}\frac{\mathbf{w}_d}{2}(\Delta_d^2 + \lambda_{\Delta_{d,d}}),\\
\mathbb{E}_{\sigma}\Big[\ln p(\sigma)\Big] &= (a_0 - 1)(\psi(a_n) - \ln b_n) - b_0\sigma,\\
\mathbb{E}_{s}\Big[\ln p(s)\Big] &= (c_0 - 1)(\psi(c_n) - \ln d_n) - d_0 s,\\
\mathbb{E}_{\mathbf{w}}\Big[\ln p(\mathbf{w})\Big] &= \sum_{d=1}^{D}(e_0 - 1)(\psi(e_n) - \ln \mathbf{f}_{n_d}) - e_0\mathbf{w}_d,\\
\mathbb{E}_{\boldsymbol{\beta}}\Big[\ln q(\boldsymbol{\beta})\Big] &= \frac{1}{2}\sum_{i=1}^{C}\ln\det(\lambda_{\boldsymbol{\beta}_i}),\\
\mathbb{E}_{\Delta}\Big[\ln q(\Delta)\Big] &= \frac{1}{2}\ln\det(\lambda_\Delta),\\
\mathbb{E}_{\sigma}\Big[\ln q(\sigma)\Big] &= a_n - \ln b_n + \ln\Gamma(a_n) + (1 - a_n)\psi(a_n),\\
\mathbb{E}_{s}\Big[\ln q(s)\Big] &= c_n - \ln d_n + \ln\Gamma(c_n) + (1 - c_n)\psi(c_n),\\
\mathbb{E}_{\mathbf{w}}\Big[\ln q(\mathbf{w})\Big] &= \sum_{d=1}^{D} e_n - \ln \mathbf{f}_{n_d} + \ln\Gamma(e_n) + (1 - e_n)\psi(e_n),
\end{aligned}
$$

where $\psi(\cdot)$ is the digamma function. During the optimization, the bound is maximized and the optimization stopped when it reaches a plateau $|\mathcal{L}(q_n) - \mathcal{L}(q_{n+1})| < \epsilon$. Generally, the ELBO is not a convex objective function. Therefore, based on the initial values this process will converge to a local optimum.

Predictive density

The approximation of the posterior distribution can be used to predict the target variable y_* based on a new observation $\mathbf{x}_*$. For this purpose, the posterior predictive distribution is required to derive a prediction and confidence intervals. The posterior predictive distribution for the new target variable is calculated by marginalizing the distribution of the target variable given the observation and parameters over the posterior distribution of the parameters.

$$
\begin{aligned}
p(y_*|\mathbf{x}_*, \mathbf{D}) &= \int\int p(y_* \mid \mathbf{x}_*, \boldsymbol{\beta}, \sigma)p(\boldsymbol{\beta}, \sigma \mid \mathbf{D})\; d\boldsymbol{\beta}\; d\sigma\\
&\approx \int\int p(y_* \mid \mathbf{x}_*, \boldsymbol{\beta}, \sigma)q(\boldsymbol{\beta}, \sigma)\; d\boldsymbol{\beta}\; d\sigma\\
&= \int\int \mathcal{N}\left(y_* \mid \mathbf{x}_*^T\boldsymbol{\beta}, \sigma^{-1}\right)\mathcal{N}\left(\boldsymbol{\beta} \mid \boldsymbol{\beta}_i, \lambda_{\boldsymbol{\beta}_i}\right)\mathrm{Gamma}(\sigma \mid a_n, b_n)\; d\boldsymbol{\beta}\; d\sigma\\
&= \int \mathcal{N}\left(y_* \mid \mathbf{x}_*^T\boldsymbol{\beta}, \sigma^{-1} + \mathbf{x}_*^T\lambda_{\boldsymbol{\beta}}\mathbf{x}_*\right)\mathrm{Gamma}(\sigma \mid a_n, b_n)\; d\sigma\\
&= St\left(y_*, \frac{a_n}{b_n} + \mathbf{x}_*^T\lambda_{\boldsymbol{\beta}_i}\mathbf{x}_*, 2a_n\right)
\end{aligned}
\tag{1.14}
$$

To obtain the result, the posterior distribution is substituted with the variational posterior and results for the convolution of normal and Gamma distributions (Bishop, 2006; Murphy, 2013) are used. First, $\boldsymbol{\beta}$ is integrated out by the convolution of two normal distributions. The marginal distribution of the resulting normal and the Gamma distribution is a Student's t-distribution with mean $\mathbf{x}_*^T \boldsymbol{\beta}_i$, precision $\frac{a_n}{b_n} + \mathbf{x}_*^T \lambda_{\boldsymbol{\beta}_i} \mathbf{x}_*$, and $2a_n$ degrees of freedom. The obtained result shows that the predicted uncertainty is the sum of the noise σ^{-1} and the variance in client individual parameters $\boldsymbol{\beta}_i$. The number of degrees of freedom is approximately the number of observed samples. Interestingly, these samples do not have to originate from an individual since it is the number of the overall observed samples. It is further possible to derive predictions for clients that have not been observed by using the hierarchical prior Δ and the variance s.

Implementation of the hierarchical Bayesian regression

The derived model can be implemented in any programming language. The model parameters are estimated using client data and the coordinate descent algorithm with the derived update equations. The implementation of the training procedure is shown in pseudo-code in Algorithm 1.

Algorithm 1 Training of the hierarchical Bayesian regression

procedure TRAINMODEL($\mathbf{X}$, $\mathbf{Y}$) ▷ Where $\mathbf{X}$ and $\mathbf{Y}$ are the training data
  $a_n = Equation\ 1.7$ ▷ Estimate a_n
  $c_n = Equation\ 1.9$ ▷ Estimate c_n
  $e_n = Equation\ 1.11$ ▷ Estimate e_n
  $last\_bound = -\infty$ ▷ Set initial bound
  for n in 1:MAX_ITERATIONS **do** ▷ Limit the maximal number of iterations
    for i in 1:N_CLIENTS **do** ▷ Update $\boldsymbol{\beta}$ for each client
      $\boldsymbol{\beta}_i = Equation\ 1.4$ ▷ Update $\boldsymbol{\beta}$ for client i
    $\Delta = Equation\ 1.6$ ▷ Update hierarchical prior Δ
    $d_n = Equation\ 1.10$ ▷ Update s
    $b_n = Equation\ 1.8$ ▷ Update σ
    $\mathbf{f}_n = Equation\ 1.12$ ▷ Update $\mathbf{w}$
    $bound = Equation\ 1.13$ ▷ Calculate lower bound
    if abs($last\_bound - bound$) $< \epsilon$ **then**
      break ▷ Algorithm is converged
    $last\_bound = bound$

After the model has been trained, client individual predictions can be derived using Equation 1.14.

1.4 Discussion

This section reviews and summarizes the implications and contributions of the eight articles included in this dissertation. In the following, the strength and weaknesses of the articles, as well as current challenges are evaluated.

1.4.1 Review of the contributions

The presented articles investigated open research questions in the field of online mental health treatment. The different types of predictive models utilized in research and applicable

to online treatment have been assessed as well as the acceptance of mobile mental health treatment. Various types of data have been analyzed to provide a deeper understanding of symptom interaction and prediction of future health states. Specifically, the use of mobile phone measures and self-reported ecological assessment was analyzed for future mood prediction. Furthermore, it was shown that ratings of mood are linked to depression and enable the prediction of depression scores. Also, the prediction of mood utilizing free-text diaries has been demonstrated. For the analysis of these various data types, the performance of existing machine learning methods has been assessed and new predictive models have been developed. For the statistical evaluation of multi-objective models, a new approach that utilizes simulation analysis has been presented and a method for model parameter estimation. The developed models can play a pivoting role in online treatment to improve and provide personalized treatment. Figure 1.3 illustrates the necessary steps to integrate predictive models into online treatment and shows the belonging of each article.

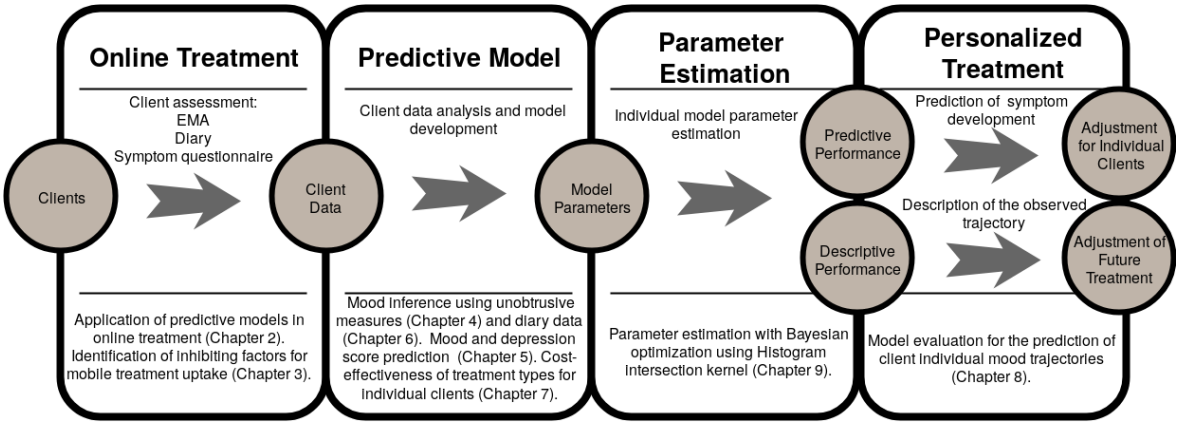


FIGURE 1.3: Process chain for integration of predictive models in online treatment

The first part of this dissertation illustrates the different types of predictive models in online treatment and analyzes the acceptance of online treatment. During the online treatment process, a variety of data such as EMA and diary are collected. Even before treatment begins, questionnaires for diagnostic purposes are inquired. The article in Chapter 2 reviews the use and types of predictions of such models in an online setting. The review provides a framework that allows classification of models that range from diagnostic purpose to required aftercare prediction. The framework can be used by researches to categorize their developed models and by software engineers for the development of online interventions that include decision support.

The second article researches the necessary steps for increasing awareness and openness towards mobile mental health treatment. For assessing the acceptance and tendency of using a mobile online treatment program, a structural equation model that is based on the technology acceptance model is proposed and evaluated in Chapter 3. This model allows to measure the acceptance and identify inhibiting factors for the uptake of mobile treatment. The derived insides allow to design mobile mental health applications more effectively by addressing users' concerns.

In the second part of this dissertation, a variety of data is analyzed utilizing multiple methods from the field of machine learning. For the treatment of depression, clients' mood and outcome prediction can enable clinicians to infer treatment decisions. These treatment decisions enable adjustment of the ongoing treatment to the individual client and can improve treatment results. Accordingly, Chapter 4-6 researches inference and prediction of

clients' mood. Specifically, Chapter 4 analyzes the use of unobtrusive measures and Chapter 6 diary data for mood inference. Chapter 5 analyzes the use of EMA data for prediction of future mood and depression scores. These models provide insights on client improvement during treatment and can be utilized in online treatment to improve decision making and provide personalized treatment.

The cost-effectiveness of traditional and blended treatment on a client individual level is analyzed in Chapter 7. Based on pre-treatment questionnaires the assigned treatment type can be adjusted utilizing a client individual prediction of the expected outcome and the costs. Although the article provides a method for the selection of an appropriate treatment utilizing baseline data, it could also be considered as treatment individualization which would be the last element of the chain in Figure 1.3. More specifically, because the presented model is based on knowledge derived from previous clients and allows to improve future iterations of the treatment. The presented methods provide a tool to improve the cost-effectiveness of depression treatment.

The third part of this dissertation considers model evaluation and parameter estimation. Personalization of predictive models provides opportunities for improving ongoing and future iterations of the treatment program. For achieving the integration of predictive models in online treatment, model evaluation is required to verify their predictive performance and increase trust in their validity. Consequently, Chapter 8 discusses the necessary steps and demonstrates a thorough model evaluation. For providing clients with personalized treatment, individual model parameters need to be estimated. These client-specific model parameters allow to derive client individual predictions. For this purpose, Chapter 9 presents a method for model parameter estimation suited for mobile devices due to its lesser memory requirement and lower computational demand.

A summary of the presented articles is shown in Table 1.1.

Title	Scientific contribution	Journal
Predictive modeling in e-mental health: A common language framework (Chapter 2)	The article introduces a framework that provides a classification of the different types of predictive models and their use during treatment. It introduces the different treatment phases and illustrates prediction types that can be obtained during treatment. Furthermore, the derived framework is used to classify predictive models that are found in the literature. The introduced framework may help in the classification of utilized models and the design of new intervention programs.	Internet Interventions (ranking: no ranking)
Acceptance of mobile mental health treatment applications (Chapter 3)	This chapter provides an analysis of the acceptance and intention to use mobile mental health applications. The utilized model is based on the technology acceptance model with additional concepts that are described in the literature for measuring the acceptance of mobile services. The results identify reasons against the up-take of mobile online interventions and can help to improve future treatment programs and advertisements.	Procedia Computer Science (ranking: no ranking)
How to predict mood? Delving into features of smartphone-based data (Chapter 4)	This chapter analyzes the use of unobtrusive mobile phone measures for the inference of users' mood levels. It contributes to the research of unobtrusive EMA measures by analyzing the use of general classifiers in comparison to client individual models. Additionally, the importance of the unobtrusive measures on mood prediction is analyzed which can provide suggestions and insights for future studies regarding unobtrusive measures. The developed model utilizes variational inference and can derive predictions more quickly than sampling-based methods and might, therefore, be used in online treatment to predict clients' mood levels.	Americas Conference on Information Systems (2016) (ranking: D)

Title	Scientific contribution	Journal
The predictive power of EMA data for mood and depression score prediction (Chapter 5)	The article contributes to the research of EMA data evaluation by analyzing the EMA data originating from the E-COMPARED depression study. The prediction performance of a variety of machine learning methods for mood prediction is evaluated. The analyses show that past mood and self-esteem measures contribute to the prediction of future mood. Furthermore, measures of mood can be utilized for the inference of clinical depression scores. These insights allow to utilize predictive models to supervise clients' improvement during the treatment.	Working paper
Predicting the individual mood level based on diary data (Chapter 6)	This article describes the analysis of free-text diary data that was collected during an online stress prevention treatment. Specifically, the influence of daily activities on clients' mood level is analyzed. For the extraction of predefined activities, text mining is utilized and a Bayesian model for effect estimation. The presented model allows inferring the mood level based on diary data, which can be a valuable tool for supervising therapists. This article further contributes to the analysis of free-text diary data and inference of clients' mood levels.	European Conference on Information Systems (2017) (ranking: B)
Predicting therapy success and costs for personalized treatment recommendations using baseline characteristics: Data-driven analysis (Chapter 7)	This article analyzes data from the E-COMPARED depression study and aims to predict clients' outcomes and treatment costs for two different treatment types. The predictions are then utilized for calculating the incremental cost-effectiveness ratio. This allows for adjusting the cost-effectiveness by selecting an individual treatment type for each client. A variety of machine learning methods are compared and it is exemplary demonstrated how to adjust the cost-effectiveness given the predicted outcome and costs. This research contributes to the optimization of online interventions with respect to the expected treatment costs and outcomes.	Journal of Medical Internet Research (ranking: A)

Title	Scientific contribution	Journal
Evaluation of a temporal causal model for predicting the mood of clients in an online therapy (Chapter 8)	Evaluation of predictive models is a mandatory task to ensure the models' integrity. The article proposes the use of simulation analysis and evaluation on study data. The use of the simulation analysis allows estimation of the theoretical performance and enables to investigate reasons for low model performances. Further, the article demonstrates a model evaluation combining both analyses. The article shows that current evaluation methods are not sufficient and contributes to the research of predictive model evaluation.	Evidence-Based Mental Health (ranking: no ranking)
Analysis of the histogram intersection kernel for use in Bayesian Optimization (Chapter 9)	The chapter provides an analysis of Bayesian optimization utilizing a linear surrogate function. The surrogate function is used instead of the unknown error function to locate a minimum of the objective function. For the reduction of the required computation time, the use of a linear kernel instead of an exponential kernel is analyzed. Previously, the linear kernel has been used in computer vision due to its fast learning properties. This method has been designed with the anticipation of fast model parameter estimation on computationally limited devices. The article researches new methods for model parameter estimation and the use of a linear surrogate function for error function approximation.	International Journal of Modeling and Optimization (ranking: no ranking)

Table 1.1: Included articles with their scientific contributions and ranking of journals according to VHB Jourqual V3

1.4.2 Evaluation of the contributions

In the following, the contributions of the articles are reconsidered and their limitations are discussed.

Acceptance and predictive modeling in online treatment

The article presented in Chapter 3 provides a structural equation model for assessing the acceptance of mobile mental health treatment applications. The model encompasses concepts of trust, social influence, and self-efficacy. These concepts have previously been shown to be connected with the uptake of online and self-learning applications. During the time of the model evaluation, the findings suggest that knowledge about the existence of these applications was considerably low and that trust in the safety of personal information was a major concern as well as the clinical effectiveness. The perception of technological innovations changes over time, which requires reevaluation to capture such changes. Furthermore, the evaluation focused on the young German population. Although social influence was not identified as a major influence on the uptake of online treatment, this could highly differ in other countries or cultures. For example, African culture has traditional beliefs that often lead people to faith healers and has a high stigma against mental disorders (Alahmed, Anjum, and Masuadi, 2018; Idoga et al., 2019). The presented model was based on the technology acceptance model, however, one could also consider a different type of model such as the elaboration likelihood model (Petty and Cacioppo, 1986) that describes how attitudes form and change, or the health belief model that aims to explain and predict health-related behaviors (Janz and Becker, 1984).

The article presented in Chapter 2 provides a framework that proposes three distinct phases during online treatment where predictive modeling can aid in supporting treatment decisions. It further lists the collected data in these phases and what type of models are suited for a specific task. According to the presented framework, existing literature was reviewed and the methods that are used in these articles were classified. However, it is not a systematic literature review and therefore based on multiple searches to explore and discover more relevant literature. A structured review would help in describing the use of predictive models in online treatment at a particular point in time. Furthermore, online treatment changes and improves over time, therefore, the presented framework might also change to include the newest technologies.

Predictive models in online treatment

The use of unobtrusive mobile phone measures and EMA data for inference of the current mood level is analyzed in Chapter 4. The analysis of the importance of the individual measures shows that the inquired EMA data provides the strongest link to the current mood level. The results suggest that the unobtrusive measures only minorly contribute to the mood inference. However, the analyzed data did not contain GPS, accelerometer measures for activity detection or stress sensing. Such higher-level information can considerably improve the inference of mood (Mohr, Zhang, and Schueller, 2017). Although these measures could be battery consuming and raise privacy concerns, they can considerably improve the prediction. Furthermore, the analyzed data only consist of university students, thus the results might not be generalizable to the whole demographics. In general, research in the field of unobtrusive EMA assessment using passive mobile phone measures has shown to provide little accuracy for many clients (Pratap et al., 2019). Such study outcomes could put an unexpected stop to the research of unobtrusive measures for EMA assessment and deem them as too unreliable for practical use.

Similarly, Chapter 5 discusses the prediction of future mood utilizing EMA data. For this analysis, a variety of machine learning methods are utilized to predict clients' mood levels. Additionally, the influence of the individual EMA measures on the mood prediction and their relation to depression scores are analyzed. Using a variety of methods is beneficial to provide a comparison of the predictions under equal conditions. Results among different studies are difficult to compare because each study makes assumptions regarding the preprocessing of the data, uses a different data set, and chooses different model parameters. By selecting different preprocessing steps, for example, imputation of missing values, results vary even on the same data set. Standardization of the test conditions could be researched which would provide the basis for comparable results. Otherwise, the problem of comparing results among different studies remains. Further, a relationship between EMA measures and depression scores was found but the relationship was not explored in great detail. Despite the correlation among EMA measures (Bremer, Funk, and Riper, 2019; Parsey and Schmitter-Edgecombe, 2019) and regularly assessed health questionnaires (Aguilera, Schueller, and Leykin, 2015; Nahum et al., 2017), the predictive accuracy might be too low for clinical use. On one hand, these correlations provide opportunities to explore predictive models and medical support tools, on the other hand, a low predictive accuracy might render them unreliable. Therefore, prediction of future EMA measures could share a similar fate as unobtrusive EMA assessment.

Chapter 6 proposes a two-step approach to infer a client's daily mood from a free-text diary. The first step, extracts daily activities from free-text diaries, followed by the second step, which infers the mood level based on the extracted activities. The results indicate a negative influence of sickness and rumination and a positive influence of social activities on an individual's mood level. Furthermore, the model can provide a more accurate prediction of the clients' mood level than the utilized reference measure. Although the analysis suggests the validity of this method, a more generic approach for feature extraction can provide a more accurate inference of the mood. However, the drawback would be lower model interpretability, which might reduce clinicians' trust in the model. A lack of understanding regarding the models' reasoning can hinder their use for clinical decision making, whereas interpretable models provide reasons to accept or reject a prediction or recommendation (Ahmad, Teredesai, and Eckert, 2018).

Chapter 7 analyzes the relation between costs and symptom improvement on a client individual level for two types of treatment. It has been shown that the prediction of the expected treatment outcome is possible and can be used to assign clients to their most beneficial treatment. Further, the prediction of the expected outcome and costs enables to optimize the treatment with respect to both objectives. The results allow the design of a valuable tool to plan and design future studies in terms of costs and treatment outcomes. Such a tool can be beneficial for treatment providers and clients (Isinkaye, Folajimi, and Ojokoh, 2015) because it enables to improve treatment outcomes for clients and reduce costs for treatment providers. However, this study also has some shortcomings. First, the model requires validation in clinical practice because it is not guaranteed that the actual use of a predictive model will enhance medical decision making in clinical practice (Kappen et al., 2018). Although the results suggest that it works in principle, it is difficult to verify because only the outcome of the assigned treatment type can be observed. Second, the utilized data consists of studies that have been conducted in 9 different European countries. This results in a heterogeneous data set which suggests a hierarchical analysis. A hierarchical analysis could improve the predictions of the expected outcome and cost. Finally, many values in the analyzed data set are missing, as a result potentially important features were imputed or had to be discarded. This illustrates a common problem with data originating from social science in general (Gyimah, 2001) but it also implies that the predictive performance might even be improved with the consideration of additional items and questionnaires.

Model evaluation and parameter estimation

Chapter 8 considers the statistical evaluation of predictive models. The evaluated model is a temporal causal model that allows to predict the course of multiple EMA factors. The research on this subject has been conducted because evaluation methods for multi-objective models are not sufficiently accounted for in the literature. The article demonstrates a thorough evaluation and combines study data and simulation analysis. The simulation analysis can help to verify if the model provides the expected prediction accuracy on the study data and allows to further investigate reasons for low model performance. The article further shows the evaluation of a predictive model that has been suggested for use in online treatments for clients' future mood prediction. However, the here presented model evaluation suggests that the model is not superior to a mean value prediction. The article outlines how a rigorous model evaluation can be conducted and demonstrates the importance of proper model evaluation. Overall, predictive models can improve the diagnosis and treatment of mental health conditions and the field is expected to continue to grow with novel applications (Shatte, Hutchinson, and Teague, 2019; Durstewitz, Koppe, and Meyer-Lindenberg, 2019). This emphasizes the need for rigorous model development and evaluation guidelines that account for these types of models to prevent overconfidence and provide reliable models.

Utilizing client individual models that can predict client improvement or detect situations that need momentary intervention, require client-specific parameters. Mobile devices can be used for client tracking and provide momentary intervention when required. However, model parameter estimation can be a time-consuming task. For individual model use, parameter estimation with Bayesian optimization with a linear surrogate function is researched in Chapter 9. Although the algorithm can provide more accurate parameters than a random search on the test problems, the assumption of linearity is a severe drawback despite the faster evaluation time. Therefore, it has to be evaluated if the algorithm can provide an efficient way of parameter estimation in real-life applications. Heuristic methods such as genetic algorithms might provide more cost-effective results with respect to the computational demand than the presented algorithm.

1.5 Conclusion

The amount of online treatment is rapidly growing and provides new technological innovations and treatment tools for mental disorders, which become increasingly socially accepted. These innovations allow to collect a vast amount of data that can immediately be used to derive insights into clients' mental health states. This dissertation, however, also illustrated that the large amount of data needs careful consideration to identify predictors for health-related conditions. Even with the estimation of significant relationships, the predictive power can be low, as shown in the analyses for mood prediction. These challenges fuel the research for new methods in terms of predictive models, evaluation methods, and parameter estimation. The presented articles were grouped into three parts and contribute to the research of predictive modeling and decision support in online treatment.

In the first part, the public perception of online treatment is analyzed followed by the possibilities for using predictive models in the domain of online treatment. Loss of personal information is a major concern of the German population besides the absence of knowledge about the clinical effectiveness of online-based interventions. The literature review about the utilized models in online treatment revealed that despite the use of many data analysis techniques they are not applied during the active treatment. The inclusion of predictive models into the active treatment process only appears gradually. Although, the methods are readily available and provide promising results, a slow adaptation process aids in preventing the use of immature models that could lead to harmful outcomes. Despite these

technology advancements, one should not depict online treatment as a universal cure for mental illness. There will always be a need for medical professionals because it requires humans to understand human emotions.

The second part considers the analysis of client data. Although it is possible to infer and predict future mood based on dairy data, mobile phone measures, and EMA, it still might not be accurate enough for practical use. The conducted research demonstrates that prediction of mood is a challenging task and that today's models cannot consider all the influences and potential factors to predict it reliable. Human emotions are very complex and influenced by many daily factors. EMA and mobile phone measures represent only a fraction of the influencing factors. Despite this finding, the relation between individual EMA measures, daily activities on the mood, and the relationship between EMA and depression scores enables to infer trends and estimate if clients are responding to the treatment. Mental illness is very complex, therefore, these results can already provide valuable insides into client behavior and treatment progress. These models cannot only support therapists but also clients using self-help platforms. They can help clients to increase their self-awareness and assist them on their way to recovery.

In the third part, parameter estimation and model evaluation, it was demonstrated that a thorough model evaluation is crucial to avoid overconfidence in the models developed. The developed models become increasingly complex which makes it more difficult to understand how predictions where derived and it makes them more challenging to evaluate properly. An understanding of the models' reasoning is, however, required to recognize their potential and limitations. The models that have been carefully statistically evaluated on different data sets then require to be verified under real-world conditions to test their performance in practice. Client individual models that are used on mobile devices appear beneficial in tracking client behavior and providing help everywhere. However, with more complex models the number of model parameters increases too. Efficient individual parameter estimation might be a future challenge to overcome. For today's models, this is already a challenging task and estimation can require days. This, however, would be unfeasible for real-world applications, which require fast model parameter estimation. Thus, a real-world application might favor simpler models.

In summary, the presented articles contribute to the field of e-mental health. E-mental health is an opportunity to bring help to many people, where computer science can provide a valuable contribution to the health sector. The research presented here provides an overview of the possibilities of predictive modeling in online treatment and proposes models that can be used to implement treatment tools for decision support.

References

Aan het Rot, Marije, Koen Hogenelst, and Robert A. Schoevers (2012). "Mood disorders in everyday life: A systematic review of experience sampling and ecological momentary assessment studies". In: *Clinical Psychology Review* 32.6, pp. 510–523. ISSN: 02727358. DOI: 10.1016/j.cpr.2012.05.007 (cit. on pp. 3, 140).

Adelman, Caroline B. et al. (2014). "A meta-analysis of computerized cognitive-behavioral therapy for the treatment of DSM-5 anxiety disorders". In: *Journal of Clinical Psychiatry* 75.7. ISSN: 01606689. DOI: 10.4088/JCP.13r08894 (cit. on p. 2).

Adibi, Sasan (2015). *Mobile Health: A Technology Road Map*. Springer Publishing Company, Incorporated. ISBN: 3319128167, 9783319128160 (cit. on p. 3).

Aguilera, Adrian, Stephen M. Schueller, and Yan Leykin (2015). "Daily mood ratings via text message as a proxy for clinic based depression assessment". In: *Journal of Affective Disorders* 175, pp. 471–474. ISSN: 15732517. DOI: 10.1016/j.jad.2015.01.033 (cit. on pp. 6, 26, 88, 90, 98).

Ahmad, Muhammad Aurangzeb, Ankur Teredesai, and Carly Eckert (2018). "Interpretable machine learning in healthcare". In: *Proceedings - 2018 IEEE International Conference on Healthcare Informatics, ICHI 2018*, p. 447. ISBN: 9781538653777. DOI: 10.1109/ICHI.2018.00095. arXiv: arXiv:1705.10301 (cit. on p. 26).

Aichele, Stephen, Patrick Rabbitt, and Paolo Ghisletta (2016). "Think Fast, Feel Fine, Live Long". In: *Psychological Science* 27.4, pp. 518–529. ISSN: 0956-7976 (cit. on p. 6).

Alahmed, Salman, Irfan Anjum, and Emad Masuadi (2018). "Perceptions of mental illness etiology and treatment in Saudi Arabian healthcare students: A cross-sectional study". In: *SAGE Open Medicine* 6, p. 205031211878809. ISSN: 2050-3121. DOI: 10.1177/2050312118788095 (cit. on p. 25).

Altaf Hussain Abro, Michel Klein (2016). "Validation of a Computational Model for Mood and Social Integration". In: *Lecture Notes in Computer Science*. Lecture Notes in Computer Science 10046.November 2016. Ed. by Emma Spiro and Yong-Yeol Ahn, pp. 361–375. DOI: 10.1007/978-3-319-47880-7 (cit. on pp. 7, 89, 93, 141, 142, 149).

Andrews, Gavin et al. (2010). *Computer therapy for the anxiety and depressive disorders is effective, acceptable and practical health care: A meta-analysis*. DOI: 10.1371/journal.pone.0013196 (cit. on pp. 1, 5).

Apolinário-Hagen, Jennifer, Jessica Kemper, and Carolina Stürmer (2017). "Public Acceptability of E-Mental Health Treatment Services for Psychological Problems: A Scoping Review". In: *JMIR Mental Health* 4.2, e10. ISSN: 2368-7959. DOI: 10.2196/mental.6186 (cit. on p. 5).

Apolinario-Hagen, Jennifer Anette and Siegfried Tasseit (2015). "Access to Psychotherapy in the Era of Web 2.0 – New Media, Old Inequalities? / Zugang zur Psychotherapie in der Ära des Web 2.0 – Neue Medien, Alte Ungleichheiten?" In: *International Journal of Health Professions* 2.2, pp. 119–129. ISSN: 2296-990X. DOI: 10.1515/ijhp-2015-0010 (cit. on p. 5).

Arean, Patricia A. (2012). "Personalizing behavioral interventions: The case of late-life depression". In: *Neuropsychiatry* 2.2, pp. 135–145. ISSN: 17582008. DOI: 10.2217/npy.12.15 (cit. on p. 4).

Arnberg, Filip K. et al. (2014). "Internet-delivered psychological treatments for mood and anxiety disorders: A systematic review of their efficacy, safety, and cost-effectiveness". In: *PLoS ONE* 9.5. ISSN: 19326203. DOI: 10.1371/journal.pone.0098118 (cit. on pp. 2, 5).

Aung, Min Hane, Mark Matthews, and Tanzeem Choudhury (2017). "Sensing behavioral symptoms of mental health and delivering personalized interventions using mobile technologies". In: *Depression and Anxiety* 34.7, pp. 603–609. ISSN: 10914269. DOI: 10.1002/da.22646. arXiv: 15334406 (cit. on pp. 3, 4, 140).

Bak, Maarten et al. (2016). "An n=1 Clinical network analysis of symptoms and treatment in psychosis". In: *PLoS ONE* 11.9, pp. 1–15. ISSN: 19326203. DOI: 10.1371/journal.pone.0162811 (cit. on pp. 1, 4, 140).

Barak, Azy and John M. Grohol (2011). "Current and Future Trends in Internet-Supported Mental Health Interventions". In: *Journal of Technology in Human Services* 29.3, pp. 155–196. ISSN: 15228991. DOI: 10.1080/15228835.2011.616939 (cit. on p. 4).

Barry, Kristen Lawton (1999). "Brief Interventions and Brief Therapies for Substance Abuse". In: *Substance Abuse and Mental Health Services Administration (US)* (cit. on p. 5).

Beal, MJ Matthew J (2003). "Variational algorithms for approximate Bayesian inference". In: *PhD Thesis* May, pp. 1–281. ISSN: 16000870. DOI: https://www.cse.buffalo.edu/faculty/mbeal/thesis/ (cit. on p. 10).

Beintner, Ina, Corinna Jacobi, and Craig Barr Taylor (2012). "Effects of an internet-based prevention programme for eating disorders in the USA and Germany -A Meta-analytic Review". In: *European Eating Disorders Review* 20.1, pp. 1–8. ISSN: 10724133. DOI: 10.1002/erv.1130 (cit. on p. 5).

Ben-Zeev, Dror et al. (2015). "Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health." In: *Psychiatric Rehabilitation Journal* 38.3, pp. 218–226. ISSN: 1559-3126. DOI: 10.1037/prj0000130 (cit. on p. 4).

Berry, Natalie, Sandra Bucci, and Fiona Lobban (2017). "Use of the Internet and Mobile Phones for Self-Management of Severe Mental Health Problems: Qualitative Study of Staff Views". In: *JMIR Mental Health* 4.4, e52. ISSN: 2368-7959. DOI: 10.2196/mental.8311 (cit. on p. 3).

Bishop, Christopher M (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc. ISBN: 0387310738 (cit. on pp. 12, 13, 19, 80).

Board, The Executive (2012). "Global burden of mental disorders and the need for a comprehensive , coordinated response from health and social sectors at the country level". In: *World Health* January, pp. 6–9 (cit. on pp. 5, 64).

Bouwmeester, Walter et al. (2012). "Reporting and methods in clinical prediction research: A systematic review". In: *PLoS Medicine* 9.5. ISSN: 15491277. DOI: 10.1371/journal.pmed.1001221 (cit. on pp. 7, 141).

Bremer, Vincent, Burkhardt Funk, and Heleen Riper (2019). "Heterogeneity Matters: Predicting Self-Esteem in Online Interventions Based on Ecological Momentary Assessment Data". In: *Depression Research and Treatment* 2019. ISSN: 2090133X. DOI: 10.1155/2019/3481624 (cit. on pp. 26, 90, 97, 142).

Bzdok, Danilo and Andreas Meyer-Lindenberg (2018). *Machine Learning for Precision Psychiatry: Opportunities and Challenges*. DOI: 10.1016/j.bpsc.2017.11.007. arXiv: 1705.10553 (cit. on p. 6).

Caliskan, Aylin, Joanna J. Bryson, and Arvind Narayanan (2017). "Semantics derived automatically from language corpora contain human-like biases". In: *Science* 356.6334, pp. 183–186. ISSN: 10959203. DOI: 10.1126/science.aal4230. arXiv: 1608.07187 (cit. on p. 7).

Cameron, Peter A and David R Thompson (2005). "Changing the health-care workforce". In: *International Journal of Nursing Practice* 11.1, pp. 1–4. ISSN: 1322-7114. DOI: 10.1111/j.1440-172X.2005.00499.x (cit. on p. 5).

Cannon, D. S. and S. N. Allen (2000). "A Comparison of the Effects of Computer and Manual Reminders on Compliance with a Mental Health Clinical Practice Guideline". In: *Journal of the American Medical Informatics Association* 7.2, pp. 196–203. ISSN: 1067-5027. DOI: 10.1136/jamia.2000.0070196 (cit. on p. 3).

Chang, Keng-hao, Drew Fisher, and John Canny (2011). "AMMON: A Speech Analysis Library for Analyzing Affect, Stress, and Mental Health on Mobile Phones". In: *Proceedings of the 2011 PhoneSense conference*. DOI: 10.1.1.232.365 (cit. on pp. 4, 49, 52).

Cheng, Helen and Adrian Furnham (2003). "Personality, self-esteem, and demographic predictions of happiness and depression". In: *Personality and Individual Differences* 34.6, pp. 921–942. ISSN: 01918869. DOI: 10.1016/S0191-8869(02)00078-8 (cit. on pp. 6, 88, 97).

Christensen, Helen and Kathleen M Griffiths (2002). "The prevention of depression using the Internet". In: *The Medical journal of Australia* 177 Suppl, S122—5. ISSN: 0025-729X (cit. on p. 3).

Christensen, Helen and Ian B. Hickie (2010). *E-mental health: A new era in delivery of mental health services*. DOI: 10.5694/j.1326-5377.2010.tb03684.x (cit. on p. 1).

Cohen, Sheldon, David A.J. Tyrrell, and Andrew P. Smith (1991). "Psychological Stress and Susceptibility to the Common Cold". In: *New England Journal of Medicine* 325.9, pp. 606–612. ISSN: 0028-4793. DOI: 10.1056/NEJM199108293250903 (cit. on p. 6).

Committee on Quality Health Care in America, Institute of Medicine (2011). "Crossing the Quality Chasm: A New Health System for the 21st Century". In: *Journal For Healthcare Quality* 24.5, p. 52. ISSN: 1062-2551. DOI: 10.1111/j.1945-1474.2002.tb00463.x (cit. on p. 3).

Constantinides, Marios et al. (2018). "Personalized versus Generic Mood Prediction Models in Bipolar Disorder". In: *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers - UbiComp '18*. New York, New York, USA: ACM Press, pp. 1700–1707. ISBN: 9781450359665. DOI: 10.1145/3267305.3267536 (cit. on pp. 8, 141).

Cooper, Gregory F. (1990). "The computational complexity of probabilistic inference using bayesian belief networks". In: *Artificial Intelligence* 42.2-3, pp. 393–405. ISSN: 00043702. DOI: 10.1016/0004-3702(90)90060-D (cit. on p. 10).

Cuijpers, Pim et al. (2008). "Internet-administered cognitive behavior therapy for health problems: a systematic review". In: *J Behav.Med* 31.0160-7715 (Print), pp. 169–177. ISSN: 0160-7715. DOI: 10.1007/s10865-007-9144-1 (cit. on pp. 1, 3).

Cuijpers, Pim et al. (2014). "Guided Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: a systematic review and meta-analysis". In: *World Psychiatry* 13.3, pp. 288–295. DOI: 10.1002/wps.20151 (cit. on p. 5).

D'Alfonso, Simon et al. (2017). "Artificial intelligence-assisted online social therapy for youth mental health". In: *Frontiers in Psychology* 8.JUN, pp. 1–13. ISSN: 16641078. DOI: 10.3389/fpsyg.2017.00796 (cit. on pp. 2, 4).

Davis, F D (1985). "A technology acceptance model for empirically testing new end-user information systems: Theory and results". In: *Management* Ph.D. P. 291. ISSN: 0025-1909. DOI: oclc/56932490 (cit. on pp. 5, 65).

De Graaf, L. E. et al. (2009). "Clinical effectiveness of online computerised cognitive-behavioural therapy without support for depression in primary care: Randomised trial". In: *British Journal of Psychiatry* 195.1, pp. 73–80. ISSN: 00071250. DOI: 10.1192/bjp.bp.108.054429 (cit. on pp. 1, 3, 53, 140).

Deen, Tisha L., John C. Fortney, and Gary Schroeder (2013). "Patient Acceptance of and Initiation and Engagement in Telepsychotherapy in Primary Care". In: *Psychiatric Services* 64.4, pp. 380–384. ISSN: 1075-2730. DOI: 10.1176/appi.ps.201200198. arXiv: NIHMS150003 (cit. on p. 5).

Dey, Anind K. (2001). "Understanding and Using Context". In: *Personal Ubiquitous Comput.* 5.1, pp. 4–7. ISSN: 1617-4909. DOI: 10.1007/s007790170019 (cit. on p. 1).

Dölemeyer, Ruth et al. (2013). "Internet-based interventions for eating disorders in adults: a systematic review." In: *BMC psychiatry* 13.1, p. 207. ISSN: 1471-244X. DOI: 10.1186/1471-244X-13-207 (cit. on p. 5).

Donker, Tara et al. (2013). "Smartphones for smarter delivery of mental health programs: A systematic review". In: *Journal of Medical Internet Research* 15.11, pp. 1–13. ISSN: 14388871. DOI: 10.2196/jmir.2791 (cit. on p. 5).

Drugowitsch, Jan (2013). "Variational Bayesian inference for linear and logistic regression". In: ISSN: 13652338. DOI: 10.1111/j.1365-2338.1972.tb02128.x. arXiv: 1310.5438 (cit. on p. 12).

Durstewitz, Daniel, Georgia Koppe, and Andreas Meyer-Lindenberg (2019). *Deep neural networks in psychiatry*. DOI: 10.1038/s41380-019-0365-9 (cit. on p. 27).

Ebert, D. D. et al. (2015). "Increasing the acceptance of internet-based mental health interventions in primary care patients with depressive symptoms. A randomized controlled trial". In: *Journal of Affective Disorders* 176, pp. 9–17. ISSN: 15732517. DOI: 10.1016/j.jad.2015.01.056 (cit. on p. 3).

Eekhout, Iris et al. (2012). "Missing data: A systematic review of how they are reported and handled". In: *Epidemiology* 23.5, pp. 729–732. ISSN: 10443983. DOI: 10.1097/EDE.0b013e3182576cdb (cit. on p. 9).

Fisher, Aaron J. and James F. Boswell (2016). "Enhancing the Personalization of Psychotherapy With Dynamic Assessment and Modeling". In: *Assessment* 23.4, pp. 496–506. ISSN: 15523489. DOI: 10.1177/1073191116638735 (cit. on pp. 1, 4, 140).

Furmark, Tomas et al. (2009). "Guided and unguided self-help for social anxiety disorder: Randomised controlled trial". In: *British Journal of Psychiatry* 195.5, pp. 440–447. ISSN: 00071250. DOI: 10.1192/bjp.bp.108.060996 (cit. on p. 5).

Gainsbury, Sally and Alex Blaszczynski (2011). "A systematic review of Internet-based therapy for the treatment of addictions". In: *Clinical Psychology Review* 31.3, pp. 490–498. ISSN: 02727358. DOI: 10.1016/j.cpr.2010.11.007 (cit. on p. 5).

Geman, S and D Geman. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, pp. 721–741. ISSN: 0162-8828. DOI: 10.1109/TPAMI.1984.4767596 (cit. on pp. 10, 11).

Gibbons, Chris J. (2017). "Turning the page on pen-and-paper questionnaires: Combining ecological momentary assessment and computer adaptive testing to transform psychological assessment in the 21st century". In: *Frontiers in Psychology* 7.JAN, pp. 1–4. ISSN: 16641078. DOI: 10.3389/fpsyg.2016.01933 (cit. on pp. 3, 4, 88, 140).

Giosan, Cezar et al. (2017). "Reducing depressive symptomatology with a smartphone app: Study protocol for a randomized, placebo-controlled trial". In: *Trials* 18.1, pp. 1–12. ISSN: 17456215. DOI: 10.1186/s13063-017-1960-1. arXiv: NIHMS150003 (cit. on p. 1).

Gyimah, S (2001). "Missing Data in Quantitative Social Research". In: *PSC Discussion Papers Series* 15.14, pp. 1–28 (cit. on p. 26).

Hastings, W. K. (1970). "Monte Carlo sampling methods using Markov chains and their applications". In: *Biometrika* 57.1, pp. 97–109. DOI: 10.1093/biomet/57.1.97. eprint: http://biomet.oxfordjournals.org/cgi/reprint/57/1/97.pdf (cit. on pp. 10, 11).

Hemingway, Harry et al. (2013). "Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes". In: *BMJ* 346. DOI: 10.1136/bmj.e5595. eprint: https://www.bmj.com/content/346/bmj.e5595.full.pdf (cit. on p. 5).

Heron, Kristin E and Joshua M Smyth (2010). "Ecological Momentary Interventions: Incorporating Mobile Technology Into Psychosocial and Health Behavior Treatments". In: *British Journal of Health Psychology* 15.Pt 1, pp. 1–39. DOI: 10.1348/135910709X466063.Ecological (cit. on pp. 6, 46).

Hilvert-Bruce, Zita et al. (2012). "Adherence as a determinant of effectiveness of internet cognitive behavioural therapy for anxiety and depressive disorders". In: *Behaviour Research and Therapy* 50.7–8, pp. 463–468. ISSN: 0005-7967. DOI: http://dx.doi.org/10.1016/j.brat.2012.04.001 (cit. on p. 4).

Hutter, Frank, Holger H. Hoos, and Kevin Leyton-Brown (2011). "Sequential model-based optimization for general algorithm configuration". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 6683 LNCS, pp. 507–523. ISBN: 9783642255656. DOI: 10.1007/978-3-642-25566-3_40. arXiv: 9610124v1 [arXiv:hep-th] (cit. on p. 8).

Huys, Quentin J M, Tiago V Maia, and Michael J Frank (2016). "Computational psychiatry as a bridge from neuroscience to clinical applications". In: *Nature Neuroscience* 19, p. 404 (cit. on p. 6).

Idoga, Patience E. et al. (2019). "Assessing factors militating against the acceptance and successful implementation of a cloud based health center from the healthcare professionals' perspective: a survey of hospitals in Benue state, northcentral Nigeria". In: *BMC Medical Informatics and Decision Making* 19.1, p. 34. ISSN: 1472-6947. DOI: 10.1186/s12911-019-0751-x (cit. on p. 25).

Iniesta, R., D. Stahl, and P. McGuffin (2016). *Machine learning, statistical learning and the future of biological research in psychiatry*. DOI: 10.1017/S0033291716001367 (cit. on p. 7).

Inza, Iñaki, Pedro Larrañaga, and Basilio Sierra (2001). "Feature subset selection by Bayesian networks: A comparison with genetic and sequential algorithms". In: *International Journal of Approximate Reasoning* 27.2, pp. 143–164. ISSN: 0888613X. DOI: 10.1016/S0888-613X(01)00038-X (cit. on p. 10).

Isinkaye, F. O., Y. O. Folajimi, and B. A. Ojokoh (2015). *Recommendation systems: Principles, methods and evaluation*. DOI: 10.1016/j.eij.2015.06.005 (cit. on p. 26).

Ivanescu, A E et al. (2016). "The importance of prediction model validation and assessment in obesity and nutrition research". In: *International Journal of Obesity* 40.6, pp. 887–894. ISSN: 0307-0565. DOI: 10.1038/ijo.2015.214 (cit. on pp. 7, 140).

Janz, Nancy K. and Marshall H. Becker (1984). "The Health Belief Model: A Decade Later". In: *Health Education Quarterly* 11.1, pp. 1–47. ISSN: 0195-8402 (cit. on p. 25).

Jaques, Natasha et al. (2017). *Predicting Tomorrow's Mood, Health, and Stress Level using Personalized Multitask Learning and Domain Adaptation*. Tech. rep., pp. 17–33 (cit. on pp. 8, 93, 141).

Jordan, Michael I. et al. (1999). "An Introduction to Variational Methods for Graphical Models". In: *Mach. Learn.* 37.2, pp. 183–233. ISSN: 0885-6125. DOI: 10.1023/A:1007665907178 (cit. on p. 10).

Kappen, Teus H. et al. (2018). "Evaluating the impact of prediction models: lessons learned, challenges, and recommendations". In: *Diagnostic and Prognostic Research* 2.1, pp. 1–11. DOI: 10.1186/s41512-018-0033-6 (cit. on pp. 26, 140).

Kawamoto, K. et al. (2005). "Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success". In: *BMJ (Clinical research ed.)* 330.7494, p. 765. ISSN: 1756-1833; 0959-535X. DOI: 10.1136/bmj.38398.500764.8F (cit. on p. 3).

Kleiboer, Annet et al. (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". In: *Trials* 17.1,

p. 387. ISSN: 1745-6215. DOI: 10.1186/s13063-016-1511-1 (cit. on pp. 6, 89, 91, 123, 126, 131, 141, 144).

Knapp, M (1999). "Economic Evaluation and Mental Health : Sparse Past . . . Fertile Future ?" In: *The Journal of Mental Health Policy and Economics* 2.4, pp. 163–167 (cit. on pp. 7, 122).

Kretzschmar, Kira et al. (2019). "Can Your Phone Be Your Therapist? Young People's Ethical Perspectives on the Use of Fully Automated Conversational Agents (Chatbots) in Mental Health Support". In: *Biomedical Informatics Insights* 11, p. 117822261982908. ISSN: 1178-2226. DOI: 10.1177/1178222619829083 (cit. on p. 2).

Kullback, S. and R. A. Leibler (1951). "On Information and Sufficiency". In: *Ann. Math. Statist.* 22.1, pp. 79–86 (cit. on p. 11).

Kumar, Vikram et al. (2017). "The Effectiveness of Internet-Based Cognitive Behavioral Therapy in Treatment of Psychiatric Disorders". In: *Cureus* 9.8. ISSN: 2168-8184. DOI: 10.7759/cureus.1626 (cit. on p. 1).

Kurian, Benji T et al. (2009). "A computerized decision support system for depression in primary care." In: *Primary care companion to the Journal of clinical psychiatry* 11.4, pp. 140–146. ISSN: 1523-5998. DOI: 10.4088/PCC.08m00687 (cit. on p. 3).

Kyrios, Michael et al. (2014). "Study protocol for a randomised controlled trial of internet-based cognitive-behavioural therapy for obsessive-compulsive disorder". In: *BMC Psychiatry* 14.1, pp. 96–110. ISSN: 1471244X. DOI: 10.1186/1471-244X-14-209 (cit. on p. 1).

Lal, Shalini and Carol E. Adair (2013). "E-Mental Health: A Rapid Review of the Literature". In: *Psychiatric Services* 65.1, pp. 24–32. ISSN: 1075-2730. DOI: 10.1176/appi.ps.201300009 (cit. on p. 5).

Lambert, Michael J (2012). "The Outcome Questionnaire-45". In: *Integrating Science and Practice* 2.1, pp. 24–27. ISSN: 1438-8871. DOI: 10.2196/jmir.954 (cit. on pp. 1, 140).

Lee, Yong-Ho, Heejung Bang, and Dae Jung Kim (2016). "How to Establish Clinical Prediction Models". In: *Endocrinology and Metabolism* 31.1, p. 38. ISSN: 2093-596X. DOI: 10.3803/EnM.2016.31.1.38 (cit. on p. 5).

Lenhard, Fabian et al. (2017). "Cost-effectiveness of therapist-guided internet-delivered cognitive behaviour therapy for paediatric obsessive-compulsive disorder: Results from a randomised controlled trial". In: *BMJ Open* 7.5, pp. 1–10. ISSN: 20446055. DOI: 10.1136/bmjopen-2016-015246 (cit. on p. 1).

Lester, Jonathan, Tanzeem Choudhury, and Gaetano Borriello (2006). "A Practical Approach to Recognizing Physical Activities". In: pp. 1–16. ISSN: 03029743. DOI: 10.1007/11748625_1 (cit. on p. 4).

Leykin, Yan et al. (2014). "Results from a trial of an unsupported internet intervention for depressive symptoms". In: *Internet Interventions* 1.4, pp. 175–181. ISSN: 22147829. DOI: 10.1016/j.invent.2014.09.002 (cit. on p. 2).

Little, Roderick J.A. and Donald B. Rubin (1989). "The Analysis of Social Science Data with Missing Values". In: *Sociological Methods & Research* 18.2-3, pp. 292–326. ISSN: 15528294. DOI: 10.1177/0049124189018002004 (cit. on p. 4).

Löwe, Bernd et al. (2004). "Monitoring Depression Treatment Outcomes With the Patient Health Questionnaire-9". In: *Medical Care* 42.12, pp. 1194–1201. ISSN: 0025-7079. DOI: 10.1097/00005650-200412000-00006 (cit. on pp. 7, 92).

Lu, Hong et al. (2012). "StressSense: Detecting Stress in Unconstrained Acoustic Environments using Smartphones". In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing - UbiComp '12*, p. 351. DOI: 10.1145/2370216.2370270 (cit. on pp. 4, 49, 52).

MacKay (1992). "Baysian Interpolation". In: *MIT Press Journal* 447, pp. 415–447 (cit. on p. 13).

Maji, Subhransu, Alexander C Berg, and Jitendra Malik (2008). "Classification using Intersection Kernel Support Vector Machines is Efficient Classification using Intersection Kernel Support Vector Machines is Efficient". In: *Slides*, pp. 1–8. ISBN: 9781424422432. DOI: 10.1109/CVPR.2008.4587630 (cit. on pp. 8, 155, 157).

McCrone, Paul et al. (2004). "Cost-effectiveness of computerised cognitive-behavioural therapy for anxiety and depression in primary care: Randomised controlled trial". In: *British Journal of Psychiatry* 185.JULY, pp. 55–62. ISSN: 00071250. DOI: 10.1192/bjp.185.1.55 (cit. on p. 5).

Mcgrath, Patrick et al. (2017). "RE-AIMing e-Mental Health : A Rapid Review of Current". In: *Mental Health Commission of Canada* July, p. 18 (cit. on p. 1).

Melville, Katherine M, Leanne M Casey, and David J Kavanagh (2010). "Dropout from Internet-based treatment for psychological disorders." In: *The British journal of clinical psychology / the British Psychological Society* 49.Pt 4, pp. 455–71. ISSN: 0144-6657. DOI: 10.1348/014466509X472138 (cit. on p. 7).

Metz, C E (1978). "Basic principles of ROC analysis." In: *Seminars in nuclear medicine* 8.4, pp. 283–298. ISSN: 0001-2998. DOI: http://dx.doi.org/10.1016/S0001-2998(78)80014-2 (cit. on p. 9).

Mohr, David C, Mi Zhang, and Stephen M Schueller (2017). "Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning". In: *Annual Review of Clinical Psychology* 13.1, pp. 23–47. ISSN: 1548-5943. DOI: 10.1146/annurev-clinpsy-032816-044949 (cit. on pp. 3, 25, 140).

Mohr, David C. et al. (2013). "Behavioral Intervention Technologies: Evidence review and recommendations for future research in mental health". In: *General Hospital Psychiatry* 35.4, pp. 332–338. ISSN: 01638343. DOI: 10.1016/j.genhosppsych.2013.03.008 (cit. on p. 6).

Moons, Karel G M et al. (2009b). "Prognosis and prognostic research: what, why, and how?" In: *BMJ* 338. ISSN: 0959-8138. DOI: 10.1136/bmj.b375. eprint: https://www.bmj.com/content (cit. on pp. 5, 7, 140).

Morris, Tim P., Ian R. White, and Michael J. Crowther (2019). "Using simulation studies to evaluate statistical methods". In: *Statistics in Medicine* 38.11, pp. 2074–2102. ISSN: 10970258. DOI: 10.1002/sim.8086. arXiv: 1712.03198 (cit. on p. 7).

Murphy, Kevin P. (2013). *Machine learning : a probabilistic perspective*. 1st ed. MIT Press. ISBN: 0262018020 (cit. on pp. 6, 19).

Musiat, P. and N. Tarrier (2014). "Collateral outcomes in e-mental health: a systematic review of the evidence for added benefits of computerized cognitive behavior therapy interventions for mental health". In: *Psychological Medicine* 44.15, pp. 3137–3150. ISSN: 0033-2917. DOI: 10.1017/S0033291714000245 (cit. on p. 5).

Musiat, Peter, Philip Goldstone, and Nicholas Tarrier (2014). "Understanding the acceptability of e-mental health - attitudes and expectations towards computerised self-help treatments for mental health problems". In: *BMC Psychiatry* 14.1, p. 109. ISSN: 1471244X. DOI: 10.1186/1471-244X-14-109 (cit. on pp. 1, 5, 65).

Myhr, Gail and Krista Payne (2006). *Cost-effectiveness of cognitive-behavioural therapy for mental disorders: Implications for public health care funding policy in Canada*. DOI: 10.1177/070674370605101006 (cit. on p. 1).

Nahum, Mor et al. (2017). "Immediate Mood Scaler: Tracking Symptoms of Depression and Anxiety Using a Novel Mobile Mood Scale". In: *JMIR mHealth and uHealth* 5.4, e44. DOI: 10.2196/mhealth.6544 (cit. on pp. 26, 90).

Olthuis, Janine V et al. (2016). *Therapist-supported Internet cognitive behavioural therapy for anxiety disorders in adults*. DOI: 10.1002/14651858.CD011565.pub2 (cit. on p. 5).

Oromendia, Pablo et al. (2016). "Internet-based self-help treatment for panic disorder: a randomized controlled trial comparing mandatory versus optional complementary psychological support". In: *Cognitive Behaviour Therapy* 45.4, pp. 270–286. ISSN: 16512316. DOI: 10.1080/16506073.2016.1163615 (cit. on p. 1).

Parsey, Carolyn M. and Maureen Schmitter-Edgecombe (2019). "Using actigraphy to predict the ecological momentary assessment of mood, fatigue, and cognition in older adulthood: Mixed-methods study". In: *Journal of Medical Internet Research* 21.1, pp. 1–12. ISSN: 14388871. DOI: 10.2196/11331 (cit. on pp. 26, 90).

Patel, Meenal J., Alexander Khalaf, and Howard J. Aizenstein (2016). "Studying depression using imaging and machine learning methods". In: *NeuroImage: Clinical* 10, pp. 115–123. ISSN: 22131582. DOI: 10.1016/j.nicl.2015.11.003 (cit. on p. 4).

Pearl, Judea (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN: 0-934613-73-7 (cit. on p. 10).

Peñate, Wenceslao and Ascensión Fumero (2016). "A meta-review of Internet computer-based psychological treatments for anxiety disorders". In: *Journal of Telemedicine and Telecare* 22.1, pp. 3–11. ISSN: 17581109. DOI: 10.1177/1357633X15586491 (cit. on p. 5).

Perna, G. et al. (2018). "The revolution of personalized psychiatry: Will technology make it happen sooner?" In: *Psychological Medicine* 48.5, pp. 705–713. ISSN: 14698978. DOI: 10.1017/S0033291717002859 (cit. on pp. 1, 4).

Petty, Richard E. and John T. Cacioppo (1986). *Communication and Persuasion*. Vol. 51. 4. New York, NY: Springer New York, pp. 438–439. ISBN: 978-1-4612-9378-1. DOI: 10.1007/978-1-4612-4964-1 (cit. on p. 25).

Peyrou, Bruno, Jean-Jacques Vignaux, and Arthur André (2018). "Artificial Intelligence and Health Care". In: vol. 7508. December, pp. 29–40. DOI: 10.1007/978-3-319-98216-8_3 (cit. on p. 4).

Postel, Marloes G. et al. (2010). "Effectiveness of a web-based intervention for problem drinkers and reasons for dropout: Randomized controlled trial". In: *Journal of Medical Internet Research* 12.4. ISSN: 14388871. DOI: 10.2196/jmir.1642 (cit. on p. 5).

Pratap, Abhishek et al. (2019). "The accuracy of passive phone sensors in predicting daily mood". In: *Depression and Anxiety* 36.1, pp. 72–81. ISSN: 15206394. DOI: 10.1002/da.22822 (cit. on p. 25).

Pritchard, Daryl E et al. (2017). *Strategies for integrating personalized medicine into healthcare practice*. DOI: 10.2217/pme-2016-0064 (cit. on p. 4).

Proudfoot, J. et al. (2003). "Computerized, interactive, multimedia cognitive-behavioural program for anxiety and depression in general practice". In: *Psychological Medicine* 33.2, S0033291702007225. ISSN: 00332917. DOI: 10.1017/S0033291702007225 (cit. on p. 3).

Rao, Sally and Indrit Troshani (2007). "A conceptual framework and propositions for the acceptance of mobile services". In: *Journal of Theoretical and Applied Electronic Commerce Research* 2.2, pp. 61–73. ISSN: 07181876 (cit. on p. 3).

Riper, Heleen et al. (2010). *Theme issue on E-mental health: A growing field in internet research*. DOI: 10.2196/jmir.1713 (cit. on pp. 1, 5).

Ritterband, Lee M. et al. (2009). "A behavior change model for internet interventions". In: *Annals of Behavioral Medicine* 38.1, pp. 18–27. ISSN: 08836612. DOI: 10.1007/s12160-009-9133-4 (cit. on p. 1).

Robert, Christian P. and George Casella (2005). *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Berlin, Heidelberg: Springer-Verlag. ISBN: 0387212396 (cit. on p. 10).

Robinson, Michael D and Gerald L Clore (2002). "Belief and feeling: evidence for an accessibility model of emotional self-report." In: *Psychological bulletin* 128.6, pp. 934–960. ISSN: 0033-2909. DOI: 10.1037/0033-2909.128.6.934 (cit. on pp. 3, 140).

Rollman, Bruce L. et al. (2001). "The Electronic Medical Record". In: *Archives of Internal Medicine* 161.2, p. 189. ISSN: 0003-9926. DOI: 10.1001/archinte.161.2.189 (cit. on p. 3).

Romero-Sanchiz, Pablo et al. (2017). "Economic evaluation of a guided and unguided internet-based CBT intervention for major depression: Results from a multicenter, three-armed randomized controlled trial conducted in primary care". In: *PLoS ONE* 12.2, pp. 1–15. ISSN: 19326203. DOI: 10.1371/journal.pone.0172741 (cit. on p. 1).

Runyan, Jason D. et al. (2013). "A Smartphone Ecological Momentary Assessment / Intervention "App" for Collecting Real-Time Data and Promoting Self-Awareness". In: *PLoS ONE* 8.8. ISSN: 19326203. DOI: 10.1371/journal.pone.0071325 (cit. on pp. 6, 46, 49, 52).

Rutledge, Robb B., Adam M. Chekroud, and Quentin JM Huys (2019). "Machine learning and big data in psychiatry: toward clinical applications". In: *Current Opinion in Neurobiology* 55, pp. 152–159. ISSN: 18736882. DOI: 10.1016/j.conb.2019.02.006 (cit. on pp. 2, 7).

Ryder, HF et al. (2009). "Decision Analysis and Cost-effectiveness Analysis". In: *Semin Spine Surg* 21.4, pp. 216–222. DOI: 10.1053/j.semss.2009.08.003.Decision (cit. on pp. 7, 122, 124).

Saeb, Sohrab et al. (2015a). "Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study". In: *Journal of Medical Internet Research* 17.7, e175. ISSN: 1438-8871. DOI: 10.2196/jmir.4273 (cit. on pp. 4, 77, 88).

Saeb, Sohrab et al. (2015b). "The Relationship between Clinical, Momentary, and Sensor-based Assessment of Depression." In: *International Conference on Pervasive Computing Technologies for Healthcare : [proceedings]. International Conference on Pervasive Computing Technologies for Healthcare* 2015, pp. 7–10. ISSN: 2153-1633. DOI: 10.4108/icst.pervasivehealth.2015.259034 (cit. on pp. 4, 49, 51).

Sandstrom, Gillian M. et al. (2016). "Opportunities for smartphones in clinical care: The future of mobile mood monitoring". In: *Journal of Clinical Psychiatry* 77.2, e135–e137. ISSN: 01606689. DOI: 10.4088/JCP.15com10054 (cit. on p. 4).

Schnyer, David M. et al. (2017). "Evaluating the diagnostic utility of applying a machine learning algorithm to diffusion tensor MRI measures in individuals with major depressive disorder". In: *Psychiatry Research - Neuroimaging* 264, pp. 1–9. ISSN: 18727506. DOI: 10.1016/j.pscychresns.2017.03.003 (cit. on p. 4).

Shatte, Adrian B.R., Delyse M. Hutchinson, and Samantha J. Teague (2019). "Machine learning in mental health: A scoping review of methods and applications". In: *Psychological Medicine*. ISSN: 14698978. DOI: 10.1017/S0033291719000151 (cit. on pp. 2, 27).

Silva, Bruno M. C. et al. (2014). "Towards a cooperative security system for mobile-health applications". In: *Electronic Commerce Research* October, pp. 1–26. ISSN: 1389-5753. DOI: 10.1007/s10660-014-9154-3 (cit. on p. 3).

Sinclair, Craig et al. (2013). "Online Mental Health Resources in Rural Australia: Clinician Perceptions of Acceptability". In: *Journal of Medical Internet Research* 15.9, e193. ISSN: 14388871. DOI: 10.2196/jmir.2772 (cit. on p. 1).

Snoek, Jasper, Hugo Larochelle, and Ryan P Adams (2012). "Practical Bayesian Optimization of Machine Learning Algorithms". In: *Adv. Neural Inf. Process. Syst. 25*, pp. 1–9. ISSN: 10495258. DOI: 2012arXiv1206.2944S. arXiv: arXiv:1206.2944v2 (cit. on pp. 8, 155).

Soares Teles, Ariel et al. (2017). "Enriching Mental Health Mobile Assessment and Intervention with Situation Awareness". In: *Sensors* 17.12, p. 127. ISSN: 1424-8220. DOI: 10.3390/s17010127 (cit. on p. 1).

Steyerberg, E.W. (2009). "Applications of prediction models". In: *Springer*. Vol. 36, pp. 11–31. ISBN: 9783642296505. DOI: 10.1007/978-0-387-77244-8_2. arXiv: arXiv:1011.1669v3 (cit. on p. 3).

Steyerberg, Ewout W. and Yvonne Vergouwe (2014). *Towards better clinical prediction models: Seven steps for development and an ABCD for validation*. DOI: 10.1093/eurheartj/ehu207 (cit. on pp. 7, 141).

Stodden, Victoria (2015). "Reproducing Statistical Results". In: DOI: 10.1146/annurev-statistics-010814-020127 (cit. on p. 4).

Stone, Arthur A. and Saul Shiffman (1994). "Ecological Momentary Assessment (Ema) in Behavioral Medicine". In: *Annals of Behavioral Medicine* 16.3, pp. 199–202. ISSN: 0883-6612. DOI: 10.1093/abm/16.3.199 (cit. on pp. 3, 76, 140).

Stone, Arthur A, Saul S Shiffman, and Marten W DeVries (1999). "Ecological momentary assessment." In: *Well-being: The foundations of hedonic psychology*. New York, NY, US: Russell Sage Foundation, pp. 26–39. ISBN: 0-87154-424-5 (Hardcover) (cit. on p. 3).

Taherdoost, Hamed (2018). "A review of technology acceptance and adoption models and theories". In: *Procedia Manufacturing*. Vol. 22. Elsevier B.V., pp. 960–967. DOI: 10.1016/j.promfg.2018.03.137 (cit. on p. 2).

Tate, Deborah F. et al. (2009). "Cost effectiveness of internet interventions: Review and recommendations". In: *Annals of Behavioral Medicine* 38.1, pp. 40–45. ISSN: 08836612. DOI: 10.1007/s12160-009-9131-6 (cit. on p. 3).

Thomas, H V et al. (2004). "Computerised patient-specific guidelines for management of common mental disorders in primary care: A randomised controlled trial". In: *British Journal of General Practice* 54.508, pp. 832–837. ISSN: 0960-1643 (cit. on p. 3).

Tibshirani, Robert (1996). *Regression Selection and Shrinkage via the Lasso*. DOI: 10.2307/2346178. arXiv: 11/73273 [1369-7412] (cit. on pp. 9, 79, 125).

Tipping, Michael E. (2000). "The Relevance Vector Machine". In: *Advances in Neural Information Processing Systems (NIPS' 2000)*. 1, pp. 652–658. ISBN: 0-262-19450-3. DOI: 10.1.1.34.4986. arXiv: 1502.02761 (cit. on pp. 13, 80).

Triñanes, Yolanda et al. (2015). "Development and impact of computerised decision support systems for clinical management of depression: A systematic review". In: *Revista de Psiquiatría y Salud Mental (English Edition)* 8.3, pp. 157–166. ISSN: 21735050. DOI: 10.1016/j.rpsmen.2015.05.004 (cit. on pp. 3, 122).

Trull, Timothy J and Ulrich Ebner-Priemer (2013). "Ambulatory Assessments". In: *Annu Rev Clin Psychol* 9, pp. 151–176. ISSN: 1548-5943. DOI: 10.1146/annurev-clinpsy-050212-185510.Ambulatory (cit. on pp. 1, 42).

Tzhak, O H N et al. (2004). "Thee Treatment Gap in Mental Health Care Health". In: *Bulletin of the World Health Organization* 82.11, pp. 858–866 (cit. on pp. 1, 5).

Varshney, Upkar (2014). "Mobile Health". In: *Decis. Support Syst.* 66.C, pp. 20–35. ISSN: 0167-9236. DOI: 10.1016/j.dss.2014.06.001 (cit. on p. 3).

Veenhoven, R. (2008). "Healthy happiness: Effects of happiness on physical health and the consequences for preventive health care". In: *Journal of Happiness Studies* 9.3, pp. 449–469. ISSN: 13894978. DOI: 10.1007/s10902-006-9042-1 (cit. on p. 6).

Venkatesh, Viswanath and Fred D. Davis (2000a). "A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies". In: *Management Science* 46.2, pp. 186–204. ISSN: 0025-1909. DOI: 10.1287/mnsc.46.2.186.11926 (cit. on p. 6).

Vogenberg, F Randy (2009). "Predictive and Prognostic Models: Implications for Healthcare Decision-Making in a Modern Recession". In: *Am Health Drug Benefits* 2.6 (cit. on pp. 3, 4).

Wang, Chong and David M. Blei (2013). "Variational inference in nonconjugate models". In: *Journal of Machine Learning Research* 14.1, pp. 1005–1031. ISSN: 15324435. arXiv: 1209.4360 (cit. on p. 12).

Wang, ziyu, Shakir Mohamed, and Nando de Freitas (2013). "Adaptive Hamiltonian and Riemann Manifold Monte Carlo Samplers". In: arXiv: 1302.6182 (cit. on p. 8).

Warmerdam, Lisanne et al. (2010). "Online cognitive behavioral therapy and problem-solving therapy for depressive symptoms: Exploring mechanisms of change". In: *Journal of Behavior Therapy and Experimental Psychiatry* 41.1, pp. 64–70. ISSN: 00057916. DOI: 10.1016/j.jbtep.2009.10.003 (cit. on p. 7).

Weinstein, Sally M and Robin Mermelstein (2008). "Role of Autonomy". In: *Journal of Clinical Child and Adolescent Psychology* 36.2, pp. 182–194 (cit. on pp. 7, 103, 104, 108, 110, 114, 115).

Wichers, M. et al. (2011). "Momentary assessment technology as a tool to help patients with depression help themselves". In: *Acta Psychiatrica Scandinavica* 124.4, pp. 262–272. ISSN: 0001690X. DOI: 10.1111/j.1600-0447.2011.01749.x (cit. on pp. 1, 42, 46, 88).

Williams, Christopher et al. (2016). "Online CBT life skills programme for low mood and anxiety: study protocol for a pilot randomized controlled trial". In: *Trials* 17.1, p. 220. ISSN: 1745-6215. DOI: 10.1186/s13063-016-1336-y (cit. on p. 1).

Wipf, David and Srikantan Nagarajan. "New View of Automatic Relevance Determination". In: *Advances in Neural Information Processing Systems 20*, pp. 1625–1632. ISBN: 160560352X (cit. on p. 13).

Wittink, Marsha N. et al. (2013). "Towards Personalizing Treatment for Depression". In: *The Patient - Patient-Centered Outcomes Research* 6.1, pp. 35–43. ISSN: 1178-1653. DOI: 10.1007/s40271-013-0003-6. arXiv: 15334406 (cit. on p. 4).

Wu, Jianxin, Wc Tan, and Jm Rehg (2011). "Efficient and effective visual codebook generation using additive kernels". In: *The Journal of Machine Learning Research* 12, pp. 3097–3118. ISSN: 15324435 (cit. on pp. 8, 155, 157).

Zarmpou, Theodora et al. (2012). "Modeling users' acceptance of mobile services". In: *Electronic Commerce Research* 12.2, pp. 225–248. ISSN: 13895753. DOI: 10.1007/s10660-012-9092-x (cit. on p. 3).

Zebley, Benjamin et al. (2016). "Resting-state connectivity biomarkers define neurophysiological subtypes of depression". In: *Nature Medicine* 23.1, pp. 28–38. ISSN: 1078-8956. DOI: 10.1038/nm.4246. arXiv: 15334406 (cit. on p. 6).

Part I

Acceptance and predictive modeling in online treatment

This part discusses the literature concerning decision support tools in online mental health treatment and provides background information regarding the online treatment process. In the first article, the possibilities of computer-aided support in online mental health are explored. It further, provides an overview of the online treatment process, collected data, and predictive model types. In the second article, the acceptance of mobile treatment applications and inhibiting factors for their use in the young German population are investigated.

Chapter 2

Predictive modeling in e-mental health: A common language framework

Becker, D., van Breda, W., Funk, B., Hoogendoorn, M., Ruwaard, J., and Riper, H. (2018). Internet Interventions, Volume 20, Issue 7.

Abstract: *Recent developments in mobile technology, sensor devices, and artificial intelligence have created new opportunities for mental health care research. Enabled by large datasets collected in e-mental health research and practice, clinical researchers and members of the data mining community increasingly join forces to build predictive models for health monitoring, treatment selection, and treatment personalization. This paper aims to bridge the historical and conceptual gaps between the distant research domains involved in this new collaborative research by providing a conceptual model of common research goals. We first provide a brief overview of the data mining field and methods used for predictive modeling. Next, we propose to characterize predictive modeling research in mental health care on three dimensions: 1) time, relative to treatment (i.e., from screening to post-treatment relapse monitoring), 2) types of available data (e.g., questionnaire data, ecological momentary assessments, smartphone sensor data), and 3) type of clinical decision (i.e., whether data are used for screening purposes, treatment selection or treatment personalization). Building on these three dimensions, we introduce a framework that identifies four model types that can be used to classify existing and future research and applications. To illustrate this, we use the framework to classify and discuss published predictive modeling mental health research. Finally, in the discussion, we reflect on the next steps that are required to drive forward this promising new interdisciplinary field.*

2.1 Introduction

Mental health problems have huge impacts on those affected, their social network, and society at large (Health, 2010). By 2030, global mental health costs are expected to rise to about 6 trillion dollars per year, which by then will be more than the predicted health costs related to cancer, diabetes and respiratory diseases combined (Bloom et al., 2011). Ample research has been devoted to understanding the underlying causes of (specific) mental health problems, the factors that play a role in recovery, as well as the effectiveness of various therapies. This evidence-based movement has led to substantial improvements in mental health care. At the same time, however, there is a consensus that improvements should be made to more effectively address the global burden of mental health conditions.

E-mental health, the application of information technology in mental health care for the prevention and treatment of psychological disorders, may provide an answer to the

global burden of mental health problems. The number of online treatment applications is increasing and can provide help to people who otherwise would remain untreated (Robinson et al., 2010; Titov et al., 2015). Recent advancements in computer and communication science have resulted in a rapid development of the field, which has led to new research opportunities and treatment models. Based on data that clients generate when using an online treatment, predictive models can be developed which support clients as well as therapists.

An example of a development fueled by the technological advancements includes so-called Ecological Momentary Assessments (EMA). Using EMA, a wealth of coarse- and fine-grained data is collected, from heart-rate sensors, physical activity sensors, and other mobile applications, to assess the dynamics of symptoms, affect, behavior and cognition over time, in the natural habitat of the patient (Trull and Ebner-Priemer, 2013). Initially designed as a pen and paper measurement, currently, EMA measures are dominantly collected using electronic devices such as the users' smartphone. EMA enables researchers to measure the users' current state and behavior while engaged in their daily routine (Shiffman, Stone, and Hufford, 2008; Wichers et al., 2011). Such type of data, however, is also partially collected when users engage into an online based treatment. This information allows to provide users with situational and personalized interventions (Burns et al., 2011) that are tailored to specific user needs. To analyze EMA data and other fine-grained types of data (e.g., log-level data), it is not straight forward to employ traditional statistical approaches such as t-tests, Analysis of Variance, or Ordinary Least Squares (OLS) regression. Although these methods allow to understand the relationship among measures and their influence on the treatment outcome, they neither account for the specific temporal (or sequential) nature of the data nor do they consider complex interaction effects among various measures. This is where new analytical methods come into play.

Promising methods include predictive modeling techniques such as Decision Trees, Bayesian approaches (Langley, Iba, and Thompson, 1992), Support Vector Machines (Cortes and Vapnik, 1995), and Artificial Neural Networks (Haykin, 2008). In order to apply these methods to the type of sequential data collected in e-mental health applications, the data has to be pre-processed (Hoogendoorn and Funk, 2018). The main objective of this pre-processing step is to derive meaningful variables that can then be used in predictive modeling. To succeed in leveraging the potential of predictive models for the purpose of developing effective interventions, each of the steps requires goal setting, understanding the data, construction of such models, as well as interpretation of the models, understanding user perception of these models, and finally deployment of the models developed. To fulfill all these requirements, an intense interdisciplinary collaboration between clinical researchers and computer scientists is mandatory.

The data science community has to understand therapists' needs, their work process with clients, and decision points. On the other hand, therapists have to understand the technical capabilities, what data might be beneficial and what type of predictions can be derived from it. Only when both sides can communicate effectively, the potential of predictive modeling in improving treatment outcomes can be realized. The success of this collaboration will stand or fall with the development of a common language that we think is necessary to successfully plan new research studies and implement more sophisticated online treatments, which incorporates predictive modeling that utilizes the assessed data during the treatment process, in routine practice. Note that predictive modeling based on observational data should only be used to generate hypothesis on causal effects. Whether these causal effects really exist and could inform clinical decision making has to be studied in subsequent, carefully designed experiments.

In this paper, we introduce a conceptual framework that aims to categorize predictive modeling e-mental health research. In Section 2.2, we introduce predictive modeling methods and techniques. Section 2.3 presents the framework, which categorizes the various uses

of predictive modeling in mental health research into four classes. In Section 2.4, we discuss illustrative examples of each model type from the existing literature. Finally, in the discussion, we reflect on identified research gaps, the way in which the framework may help to address these gaps, and expected future challenges.

2.2 Methods in Predictive Modeling

In this section, we provide a short introduction to the predictive modeling domain. We define basic terminology, briefly describe supervised learning and evaluation methods to access the performance of such models.

2.2.1 Terminology

Predictive modeling can be positioned under the broader umbrella term "statistical modeling" (Shmueli, 2010). While traditional statistical approaches focus on explaining data in terms of causality or identify relations, predictive modeling strives to find the model that provides the most accurate predictions. Predictions are derived from so-called attributes or features that are derived from observations. These features are not unlike the variables utilized in traditional statistical approaches. However, traditional statistical approaches, focusing on high explanatory power, do not necessarily lead to a high predictive power (Dawes, 1979; Forster and Sober, 1994).

2.2.2 Supervised learning approaches

Building predictive models is often done through so-called supervised learning. Figure 2.1 provides a general overview of the procedure of this approach. Supervised learning requires "historical" data providing predictive attributes and a target value (e.g., the occurrence of a depressive episode), that we want to predict from new data, where the attributes, but not the target values, are known. Targets can either be continuous values (referred to as regression) or categorical (classification). Data is commonly pre-processed to derive appropriate features from the raw attributes and to compose a training and test set. In addition, an assumption is made on the type of function that may describe the relationship between the features and the target. The set of all possible functions that can explain the observed data is called the hypothesis space. Next, a learning algorithm is applied that uses the training dataset in combination with an error metric to select a final hypothesis from the set of all hypotheses. To estimate the generalizability of this final hypothesis, it is evaluated against a test dataset that was not used in the learning process.

Various learning techniques and types of target functions exist. We follow the categorization of Witten, Frank, and Hall (2011). They distinguish between (1) probabilistic modeling methods such as *Naïve Bayes classifiers* (see, e.g., Langley, Iba, and Thompson (1992)), (2) divide-and-conquer modeling methods such as *decision trees* (S. Rasoul Safavian and David Landgrebe, 1991), (3) (extended) linear modeling methods such as generalized linear models (e.g., logistic regression), *support vector machines* (Cortes and Vapnik, 1995), or *artificial neural networks* (see Haykin (2008)), and (4) instance-based learning methods such as *k-nearest neighbors* (Altman, 1992). A discussion of these methods is beyond the scope of this paper, but the references provide a good starting point for detailed information.

Supervised learning is a so-called *bottom-up approach* as it is driven by the data. It is complemented by the *top-down* approach, which is knowledge-driven. In the latter approach, existing theories are formalized into computational models that connect theoretical concepts from a domain to executable code. Therefore, the top-down approach starts from a theory and typically utilizes model structures that are based on theoretical and empirical insights

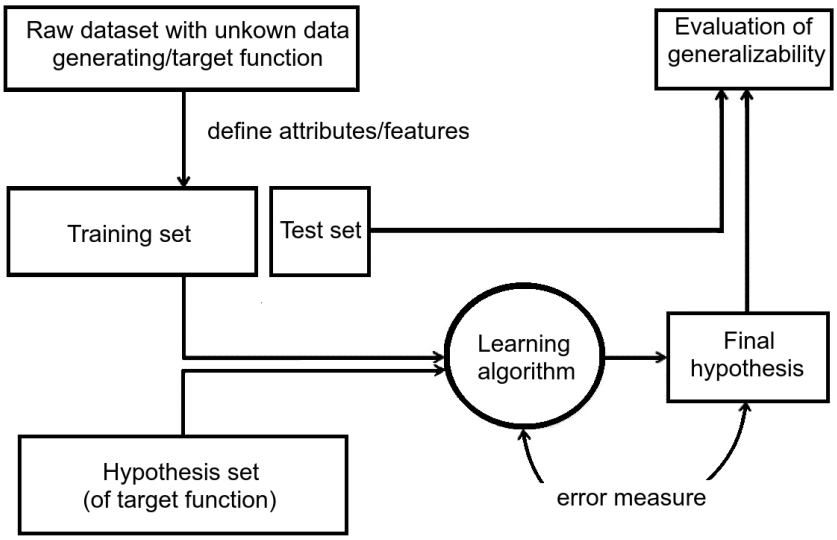


FIGURE 2.1: Generic model of a supervised learning approach based on Abu-Mostafa (2012)

into the problem domain. There are a multitude of techniques ranging from more mathematical types of modeling such as systems of differential equations (Zeigler, Praehofer, and Kim, 2000) to rule-based or agent-based systems (Hoek and Wooldridge, 2008).

2.2.3 Evaluating predictive models

The purpose of predictive modeling is to find a model that generalizes well beyond the training data. Evaluation methods are different for regression and classification models. When evaluating regression models, two objective functions that are often used are the *mean absolute error (MAE)* and the *mean squared error (MSE)*. The right type of error measure is dependent on what type of result is most desirable. The MAE is more suitable if you do not want your performance measure to be highly sensitive to outliers, and the MSE is more suitable if large errors are particularly undesirable.

For evaluating classification models, a great variety of measures such as accuracy (percentage of correctly classified cases), precision (fraction of correct classifications out of all cases attributed to a certain class by the predictive model), or recall (fraction of cases correctly found to be in that class out of all elements in the data that were in the class) are used. When the classification performance of a model depends on a threshold value, as is often the case, the *receiver operating characteristic (ROC) curve* and the *Area Under the Curve (AUC)* can be used as performance measures (see e.g., Hanley and McNeil (1982)).

As mentioned, the generalizability of models is often tested on a test dataset, which is kept apart from the training set used to fit the model. When data is limited, a validation method called *cross-validation* can be used. The method divides the original sample set into k subsets, where k-1 subsets are used as training samples and one subset is used as a validation and/or test sample. The process is repeated k times, where each of the subsets acts as the test set once, and the performance results are averaged (Olson and Delen, 2008).

2.3 Framework

To describe and categorize existing applications of predictive modeling within the mental health domain, we propose a framework. The framework (Figure 2.2) provides a common

language for researchers and forms the basis for informing treatment decisions and designing new interventions and experiments. The framework has four major elements, which are described below: time (phases), data, decisions, and model types.

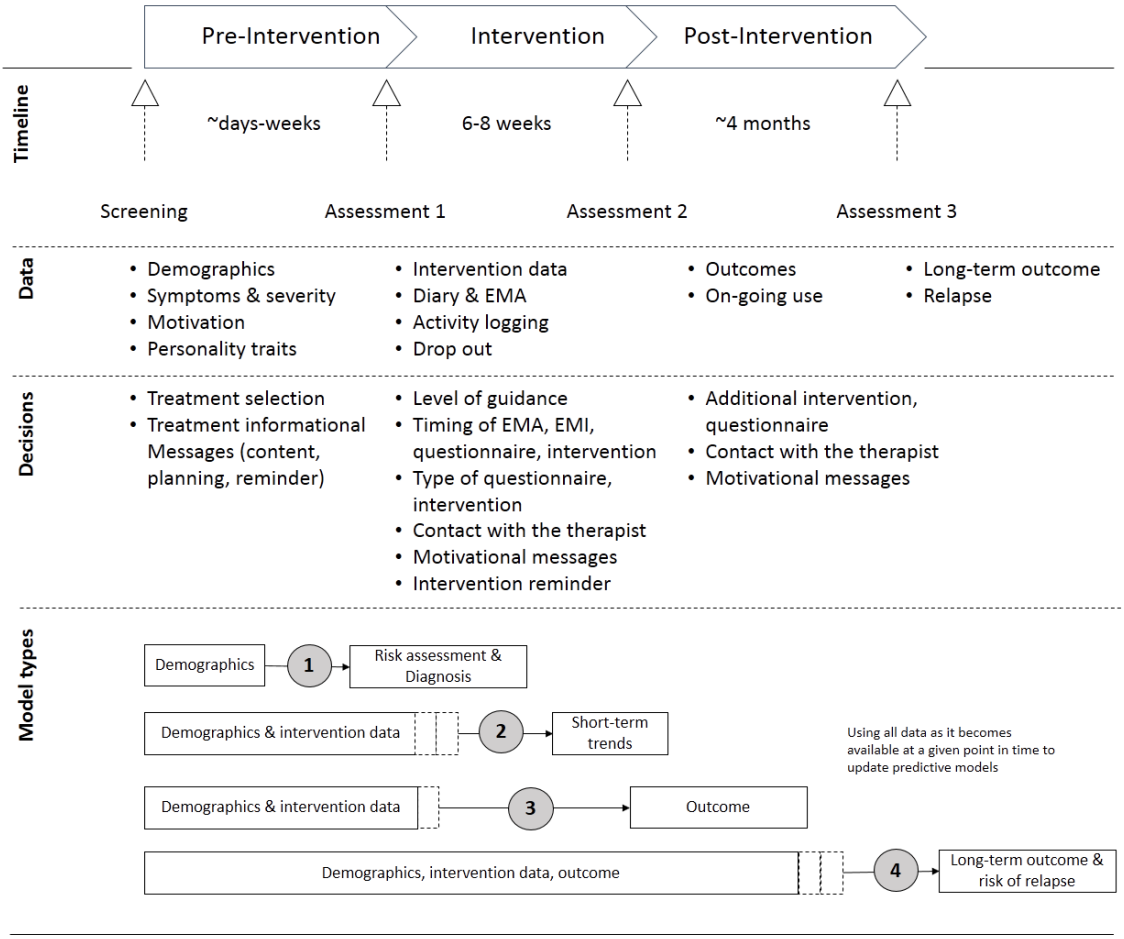


FIGURE 2.2: Proposed Framework to categorize predictive modeling in e-Mental-Health

2.3.1 Time

Predictive modeling in mental health naturally tries to uncover meaningful predictive sequential relationships. Hence, the framework is arranged along core phases of the timeline of the psychological intervention. We refer to these phases using the generic, yet recognizable, terms *pre-intervention*, *intervention*, and *post-intervention* to make them applicable across a wide range of mental health settings.

2.3.2 Data

Along the different phases, the framework identifies different types of *data* that are collected from participating clients, therapists and the technical systems involved. To predict states related to an individual, not only the data related to that individual could be relevant, but possibly the data from other individuals who have gone through this (or a similar) intervention can also hold predictive power.

Different data are collected in the different phases. In the *pre-intervention phase* questionnaires or interviews are used to determine the severity of past and current problems.

Additionally, socio-demographic data, personality traits, motivation and attitude towards the treatment can be assessed as part of this assessment process.

The largest amount of data is usually collected in the *intervention phase*. This encompasses, for example, the client's text responses to exercises in the interventions, their logins to the health systems, and other system interactions including technical loggings. Daily diaries are an option for recording the client's development during the intervention (e.g., Bolger et al. (1989), Burton, Weller, and Sharpe (2009), and Jacelon and Imperio (2005). Smartphone-based EMA enables clients to regularly report relevant variables a few times a day (Wichers et al., 2011). Smartphones can also serve as a rich source of data when usage data, such as call logs and app usage, and sensor data, such as GPS and accelerometer data, are collected (e.g., Lane et al. (2010) and Eagle, Pentland, and Lazer (2009)) . The high sampling frequency can quickly lead to a large amount of data per client. Whether and when more data leads to better predictive models is the subject of debate (Stange and Funk, 2016). The intervention often ends with a final screening, which can consist of a variety of questionnaires to estimate the remaining symptom severity and to evaluate the improvement of the client.

The *post-intervention phase* can consist of follow-up screenings and additional or repeated interventions to reduce the risk of relapse (Kessler and Chiu, 2005; Vittengl et al., 2009).

2.3.3 Decisions

The data described in the previous subsection fuel recommendations and therapeutic decisions. These decisions may influence the course of the treatment and the client's adherence and should aim to improve the therapeutic outcome.

In the *pre-intervention phase,* one has to decide what intervention and what level of guidance would best suit the client's needs. With moderate symptoms, for example, the client might be best of with a preventive self-help intervention without any personal guidance. When symptoms are severe or the when problems are complex, a guided or blended intervention would perhaps be more appropriate.

During the *intervention phase*, earlier decisions may need to be corrected on the basis of new information. For instance, the level of guidance might be increased if a client is at risk of dropping out. Keeping the intervention relevant for the client by personalizing it based on his or her behavior could prevent this. The right timing of EMA (Armey et al., 2015; Moskowitz and Young, 2006; Smyth and Stone, 2003) and Ecological Momentary Intervention (EMI) (Heron and Smyth, 2010; Runyan et al., 2013) also represent important decisions during the intervention phase. An optimal timing can minimize the intrusiveness and enhance the client's perception of the intervention.

In the *post-treatment phase*, additional interventions can be scheduled to maintain the outcome and prevent relapse (Kessler and Chiu, 2005; Vittengl et al., 2009). Usually, Internet-based treatment during this phase also consists of symptom screening and interventions to guide clients after their active treatment (Kok et al., 2014; Lord et al., 2016) . Decisions that are seen in all phases are adherence measures that include motivational messages and feedback as well as reminders or contact to the therapist.

2.3.4 Model types

Predictive models can serve a variety of purposes. Traditionally, predictive models have been limited to prediction based on the data of a certain stage (e.g., at the pre-intervention) to predict what is going to happen in the next stage (e.g., success of the intervention). However, given the higher granularity and quality of the current assessments, more detailed models have become available that may enable predictions over shorter time periods. These models can drive personalized interventions that are fully tailored to observations collected from

a specific client. The framework recognizes this by distinguishing four model types. It arranges the model types on the basis of the phase of treatment, the type of data that is available to the models, and the clinical decisions that are made on the basis of the model output. Table 2.1 presents an overview of the four model types.

Model type	**Predictors**	**Purpose**	**Usage**
1	Pre-intervention	Risk assessment and diagnosis	Identify clients at risk for mental health problems Support diagnostic process Support selection ofintervention Estimate level of guidance, EMA, EMI
2	Pre-intervention and intervention	Short-term trends	Support selection of therapy by therapist Facilitate personalized therapy Support selection of individual decisions, such as next intervention, screening, motivation and level of guidance Identify risk of drop-out Adapt intervention to maximize short-term outcomes Track therapeutic progress
3	Pre-intervention and intervention	Predict therapy outcome	Support selection of therapy by therapist Facilitate personalized therapy Support selection of individual decisions, such as next intervention, screening, motivation and level of guidance Adapt intervention to maximize treatment outcomes
4	Pre-intervention, intervention, Post-intervention	Stabilize results, prevent relapse	Identifying clients with high relapse risk Facilitate personalized after-care Adapt after-care to maximize long-term outcomes

Table 2.1: Proposed model types of the framework

Type 1 models predict *risk of mental illness* and can be used to identify *best treatment options*. There is extensive research on the predictive power of variables with respect to risk assessment and diagnosis, focusing on a wide variety of variables related to socio-demographic characteristics (see e.g., Tovar et al. (2014) and Mittendorfer-Rutz et al. (2014), personality traits (see e.g., Magidson et al. (2014)), and illness characteristics (see e.g., Otto et al. (2001) and Wade, Treat, and Stuart (1998)). In the past, this type of research has often employed standard statistical methods like OLS regression and logistic regression. Data available to Type 1 models can be collected through questionnaires, EMA assessments, or shared electronic health records. The time-span of the data collection varies from one hour (one-time screening instrument administrations), to days (e.g., a one-week diary), or weeks (e.g., EMA assessments).

Type 2 models predict *short-term changes* of the health status and treatment adherence to *optimize treatment and treatment progress*. As clients progress through the intervention, more data becomes available, as illustrated in Figure 2 by the little-dotted boxes. Whenever new data arrive, the client-specific model can be updated, which, on average, increases the power to predict the short-term evolution of key concepts such as the valence of mood or rumination. Based on predicted short-term changes, the next steps of the intervention

can be planned. Furthermore, such models might contribute to the client's adherence to the intervention. When the current engagement level is understood and the predicted probability for drop-out is high, motivational interventions can be applied, such as direct contact with the therapist, or a reminding message.

Type 3 models predict the *outcome* of the intervention phase, based on available data such as socio-demographic data and observations made during the intervention phase. Here, models that estimate and minimize dropout risk are also relevant, but in this case, the focus should be on drop-out over a longer term (which may require qualitatively different interventions than those deployed to reduce short-term drop-out).

Type 4 models aim to predict *relapse*. This class of models uses data from the pre-intervention phase and the intervention phase to provide a prediction of relapse risk (Kessing, 2004). Type 4 models can help to determine whether patients need after-care, or to decide on an optimal assessment schedule to regularly assess the stability of short-term treatment results.

2.4 Applying the Framework to Published Research

In this section, we use the introduced framework to categorize published e-mental health research in which predictive models were applied, to illustrate that 1) the framework works in an insightful way and is useful, and 2) the framework covers work done within the field. Discussed papers, their key characteristics, and the relationship to the framework are summarized in Table 2.

Model Type	Study	Data	Prediction	Method	Comment
1	Ankarali, Canan, and Akkus (2007)	Social-demographic data	Postpartum depression	Classification tree, logistic regression	Estimation of risk for postpartum depression in women.
1	Werf et al. (2006)	Time-to-event data	Estimation of recovery probability	Sequential-phase model	Models transitions from non-depressed to depressed states in the general population. Identifies factors that lead to depression.
1	Gittelman et al. (2015)	Facebook likes	Life expectancy is estimated from Facebook likes	Principal component analysis, Linear regression	Facebook likes indicate habits and activities, which are used to predict life expectancy.
1	Pestian and Nasrallah (2010)	Suicide notes	Possible suicide	Support vector machine	The genuinely of suicide notes is estimated.
1	Burns et al. (2011)	EMA	EMA ratings are inferred from the collected smartphone data	Various types of regression trees	Provides EMI to the user in stressful situations.
1	Saeb et al. (2015b)	GPS movement data	Correlation between GPS and phone usage	K-means for location clustering, and elastic net regularization for prediction of depressive symptoms	The relationship among mobile phone GPS sensor features, EMA ratings, and the PHQ-9 scores is analyzed.
1	Kim et al. (2015)	Activity measures	Estimation of mood	Hierarchical model	Mood is inferred from physical activity.
1	Doryab et al. (2014b)	Noise level, Movement and Location data, Light intensity, Phone usage	Correlation between depression and phone usage and sleep behavior	Tertius algorithm (see (Flach and Lachiche, 2001))	Preliminary study on detection of behavior change in people with depression.
1	Mestry et al. (2015)	Smartphone measures	Estimation of mental state	Correlation	Smartphone data is analyzed for correlation with the mental state such as depression, anxiety or stress.
1	Demirci, Akgönül, and Akpinar (2015)	Pittsburgh Sleep Quality Index, Beck Depression Inventory, Beck Anxiety Inventory	Sleep quality, depression and anxiety score	Correlation	Smartphone use correlates with sleep quality and symptoms of anxiety and depression.
1	Ma et al. (2012a)	Smartphone measures	Estimation of mood	Factor graph	Mood is inferred from data recorded by the personal smartphone.
1	LiKamWa et al. (2013)	Smartphone measures	Estimation of mood	Regression model	Mood is inferred from data recorded by the personal smartphone.
1	Panagiotakopoulos et al. (2010)	EMA rating and contextual Information	Mental state	Bayesian network	Mental states such as depression, anxiety and stress are estimated from contextual data.
1	Lu et al. (2012)	Voice samples	Stress estimation from voice samples	Mixture of hidden Markov models	Stress level can be estimated from voice-based features.
1	Chang, Fisher, and Canny (2011)	Speech samples	Emotion recognition	Support vector machine	Library that runs on a smartphone and estimates emotion from speech samples.
1	Sluis, Dijkstra, and Broek (2011)	Speech samples	Stress estimation	Regression	Stress level of Post-Traumatic Stress Disorder patients is estimated from voice samples.
1	Asselbergs et al. (2016)	Smartphone measures	Estimation of mood	Regression model	Failed replication of Likamwa et al, 2013
2	Runyan et al. (2013)	Mood measures	Estimation of mood swing cycles and treatment influence	Dynamic Model	Simulation of treatment and coupling behavior for bipolar individuals.
2	Both et al. (2008) and Both and Hoogendoorn (2011)	EMA	Simulates client's symptom trajectory	Dynamic Model	Allows predictions about the client's recovery curve and simulates the influence of various therapy forms.
2	Touboul, Romagnoni, and Schwartz (2010)	EMA	Estimation of recovery curve and relapse risk	Dynamic Model	The identification of underlying model parameters allow client specific predictions.
2	Demic and Cheng (2014)	EMA, clinical data	Prediction of depressive episodes and recovery chance	Finite-state machine , Dynamical system	Modeling of occurrence of depressive episodes, and influence of treatment.
2	Noble (2014)	Questionnaires	Scheduling of face-to-face interventions	Control-theoretic model	Control-theoretic scheduling of psychotherapy based on client individual data.
2 (4)	Patten 2005	Depression score	Estimated recovery curves	Markov model	Prevalence and recovery from major depressive episodes are estimated with a Markov model
2	Becker et al. (2016a)	Phone usage data, EMA data	Mood of the next day	Lasso regression, Support vector machines, linear regression, Bayesian hierarchical regression	Try to predict the mood level of the next day based on reported EMA data and phone usage data
2	van Breda et al. (2016)	EMA data	Mood of the next day	Linear regression with a bagging approach	Predict the mood of the next day by means of EMA data collected during previous days. Optimize the historical period used for predictions.
2	Bremer et al. (2017a)	Diary data	Current mood	Text mining and Bayesian regression	Clients' activity diary data is used to infer the current mood.
2	Osmani et al. (2013)	Smartphone measures	Depressive state	Correlation	Smartphone measured of the activity are correlated to the depressive state of bipolar individuals.
3	Karyotaki et al. (2015)	Demographics	Estimation of drop-out risk factors	Hierarchical Poisson regression modeling	Individual Patient Data Meta-Analysis: raw data from various trials were analyzed to identify drop-out risk factors for web-based interventions.

Model Type	Study	Data	Prediction	Method	Comment
3	Kegel and Flückiger (2015)	Self-esteem, Mastery, Clarification, Global Alliance	Treatment dropout	Hierarchical regression	Clients with lower levels of self-esteem, fewer clarifying experiences, and absence of therapeutic alliance are more likely to dropout.
3	Meulenbeek, Seeger, and Klooster (2015)	Socio-demographic, personal, and illness-related variables	Treatment dropout	Logistic regression	Dropout estimation for clients with mild panic disorder.
3	Proudfoot et al. (2013)	Perceived self-efficacy questionnaire	Symptom improvement	Correlation	Perceived self-efficacy reported at the beginning of web-based treatment indicates outcome.
3	Van Gemert-Pijnen, Kelders, and Bohlmeijer (2014)	Hamilton Rating Scale for Depression	Treatment failure	Logistic regression	Early improvement can be used to predict therapy outcome.
3	Priebe et al. (2011)	Therapeutic relationship	Treatment outcome	χ^2 Analysis	Outcome of 9 studies was compared to estimate the predictive capability of therapeutic relationship.
3	Donkin and Glozier (2012)	Application usage data	Completion of interventions	Logistic regression	Usage data and its influence on the outcome.
3	Van Gemert-Pijnen, Kelders, and Bohlmeijer (2014)	Login frequency	Prediction of outcome after therapy	Linear regression model	Client login frequency was correlated with improvement after the therapy.
3	Whitton et al. (2015)	Collected data about used program features	Outcome prediction	Correlation	Usage of (some) program features was correlated with treatment outcome.
3	Bennett and Doub (2010)	Session based outcome questionnaire	Treatment outcome	Various methods	Treatment outcome estimation based on session based questionnaires.
3	Perlis (2013)	Socio-Demographics, Self- reported clinical data	Treatment resistance	Naïve Bayes, Logistic regression, Support vector machine, Random forest	Treatment resistance is predicted based on self-reported data.
3	Hoogendoorn et al. (2017)	Socio-Demographics, Emails sent by patient	Treatment success	Logistic Regression, Decision tree, Random forest	Treatment success is predicted based on the text contained in the emails sent by the patient to the therapist.
4	Kessing (2004)	ICD-10 Depression rating	Relapse Risk, Suicide Risk	Cox-Regression	The risk of relapse is significantly related to the severity of baseline and post-treatment depression.
4	Busch et al. (2012)	Demographics, Medication, clinical data	Predict one year follow up outcome	Hierarchical logistic regression	The outcome of bipolar clients at one year follow up is predicted using clinical data.
4	Farren and McElroy (2010)	Demographics, Previous drinking characteristics, Comorbidity	Alcoholic relapse Risk	Logistic regression	Relapse after 3 or 6 month of clients with alcohol-dependence and depression or bipolar disorder.
4	Farren et al. (2013)	Demographics, Previous drinking characteristics, Comorbidity	Alcoholic relapse Risk	Logistic regression	Longitudinal outcome after 2 years of clients with alcohol-dependence and depression or bipolar disorder.
4	Pedersen and Hesse (2009)	Demographics, Previous drinking characteristics	Alcoholic relapse Risk	Logistic regression	Based on demographics and previous drinking behavior the alcoholic relapse risk is predicted.
4	Van Voorhees et al. (2008)	Mood, social and cognitive vulnerability	Relapse Risk	Regression trees	Estimation of depression relapse risk.
4	Gustafson et al. (2011)	GPS Position	Trigger EMI	Previously entered locations	EMI is triggered in locations where alcohol was obtained in the past.
4	Chih et al. (2014)	Weekly assessed EMA ratings	Predict relapse risk in coming week	Bayesian network model	Bases on a weekly surveys, the relapse risk in the coming week of previously alcohol-dependent clients is predicted.
4	Aziz, Klein, and Treur (2009)	Ambient measures	Triggers EMI, notifies family members or supervisors	Temporal trace langue rules	A support agent that triggers EMI or notifies medical staff based on monitoring techniques designed to identify risk of relapse.

Table 2.2: Predictive models used in e-mental health

2.4.1 Model Type 1: Risk assessment

Type 1 models aim for the prediction of mental health problems. One way of achieving this goal is to apply data mining techniques to (re-)analyze existing epidemiological data, to identify illness risk factors that might not be easily detected with more traditional statistical techniques. For example, Ankarali, Canan, and Akkus (2007) compared the performance of classification trees, a supervised modeling technique, to standard logistic regression in determining social-demographic risk factors for postpartum depression in women. The classification tree identified six risk factors, while the logistic regression model found only three. Another example is the study by Werf et al. (2006) who used a large Dutch epidemiological database to create a mathematical model of the recovery from depression in the general population.

Relevant data for identifying people with health risks can also be collected through popular social media, such as Facebook and Twitter. In a study conducted by Gittelman et al. (2015), for example, Facebook 'likes' were collected to predict mortality, using Principal Component Analysis (to reduce the high-dimensional predictor space), followed by Bootstrap Regression (a regression technique that is less vulnerable to violations of statistical assumptions). The model with Facebook likes and basic demographic features was better than either one alone. In this context, advances in automated text analysis are also promising. For example, Pestian and Nasrallah (2010) explored whether text mining techniques could be used to identify genuine suicide notes. In their study, the predictive model proved to be better in this task than mental health professionals (78% vs. 63%, respectively).

Technological progress allows to continuously acquire data on an individual level. For example, several studies suggest that depressive symptoms could possibly be monitored unobtrusively, without explicit user input, using predictive models built from smartphone log-file data. Building on a pilot study of Burns et al. (2011), Saeb et al. (2015b) analyzed the connection between users' daily movement patterns, estimated from their phones' GPS records, and the presence of depressive symptoms. Various features of the users' movements were derived from these data, such as the variance in location visits. To determine client-specific locations of importance, Saeb et al. (2015b) used a variant of the K-means clustering algorithm (David Arthur, 2007), an unsupervised learning technique that finds optimal partitions of multivariate data. Patterns in location cluster visits were found to be correlated with depressive symptoms as assessed by a self-report questionnaire. As a result, the presence of clinical levels of depression could be detected with a high degree of accuracy. The relevance of mobility and activity monitoring was further demonstrated in studies of Kim et al. (2015) and Osmani et al. (2013).

Several other studies explored the usefulness of smartphone logfiles to detect mental health problems in healthy participants. Doryab et al. (2014a) used phone call logs and phone light sensor data as proxies of users' social and sleeping behavior. Findings suggested that clear changes in outgoing calls were associated with changes in depressive symptoms. Similar findings were reported by Mestry et al. (2015), who concluded that predicting a possible range of depression, stress and anxiety is more feasible than predicting absolute values. Demirci, Akgönül, and Akpinar (2015) investigated the relationship between smartphone usage and sleep quality, depression, and anxiety in a student sample. Increased smartphone usage was found to be related to more symptoms of depression, anxiety, and lower sleep quality. Similar results were reported by Ma et al. (2012a) and LiKamWa et al. (2013), who were able to predict mood scores with an accuracy of 50% and 93%, respectively, based on variables that were collected on the smartphone.

The predictive accuracy of the models may be further increased by capturing contextual information, either through unobtrusive assessment or through prompted self-report questionnaires. Panagiotakopoulos et al. (2010) collected such data from 27 anxiety clients

over 30 days, five times a day (allowing participants to make additional ratings freely at any time). Using Bayesian networks, they were able to infer stress levels from EMA data with an average accuracy of 84%.

Voice recordings, which can be unobtrusively sampled through the microphones of smartphones, might provide another source for health screening applications. Lu et al. (2012) analyzed voice recordings to estimate stress levels and found changes in pitch to be correlated with stress levels. Chang, Fisher, and Canny (2011) showed that voice analysis programs can estimate current affection or stress levels, using smartphone-based voice analysis software. Sluis, Dijkstra, and Broek (2011) found that, in addition to pitch, other speech features such as amplitude, zero crossings, power, and high-frequency power are also useful predictors of stress levels of patients with post-traumatic stress disorder (PTSD).

As suggested by the studies discussed, Type 1 models most probably will find their way into future e-mental health in the form of smartphone applications. For this, it is important that the promising preliminary findings, which are often based on small-sample pilot studies, are corroborated by adequately powered independent replication studies. The findings of LiKamWa et al. (2013), for example, could not be replicated in a follow-up study (Asselbergs et al., 2016).

2.4.2 Model Type 2: Short-term predictions during treatment

Type 2 models target the prediction of patients' states as they evolve *during* the intervention. Studies have shown that tracking patients' states can improve treatment adherence and treatment outcomes by providing a feedback cycle (Lambert et al., 2003; Miller et al., 2003). For this, traditional self-report questionnaires can be administered regularly, for instance, every two weeks. If necessary, these coarse-grained assessments can be complemented by more fine-grained daily smartphone-based assessments (Torous et al., 2015). The distinction between Type 1 and Type 2 models is not always clear-cut. When Type 1 models and applications are used *during treatment* to make short-term predictions of health states to inform *decisions on the course of treatment*, the models should be classified as Type 2. Type 2 models can also be more advanced, since more detailed data from the ongoing treatment is available to inform the modeling process (i.e., log-data of web-based treatment platforms, therapist input, extensive diagnostic data, etc.)

Type 2 models explicitly consider the patient in the context of treatment. Runyan et al. (2013) used oscillating differential equations to predict hypomanic, stable, and depressive episodes in patients diagnosed with bipolar disorder. Their model incorporates the effects of medication treatment and behavior coupling between patients. Medication treatment slows the rapid mood changes and dampens the amplitude of mood oscillations. Patients with a similar cycle synchronize over time, whereas patients with an opposite cycle remain in an opposite cycle.

Becker et al. (2016a) exploit phone usage data and previous EMA measurements to predict the reported value for mood for an experiment with a group of students. For phone usage data they include app usage, activity levels, number of phone calls, and number of text messages sent. They apply a variety of Data Mining techniques including Support Vector Machines, Lasso and Linear regression, and Bayesian Hierarchical Regression. They create a general model as well as user level models and are able to achieve a root mean squared error of 0.83 using the Lasso Regression approach with the user level approach.

Similarly, van Breda et al. (2016) try to predict mood for the next day using self-reported EMA data of depressed patients during previous days. They optimize the number of days taken into account in their features to predict the mood value for the next day. They build individual patient models and use a combination of linear regression models (using a so-called bagging approach).

Another possibility to understand clients' mood trajectories are online diaries that are part of online interventions. Bremer et al. (2017a) demonstrate that the clients' mood on a specific day can be inferred based on their diary data. In a 2-phase modeling approach, they employ techniques from text mining to first extract activities that were likely described in the diary entries by the user. In the subsequent step, the set of relevant activities is used to successfully predict the clients' mood using an ordered logit model.

Another clear Type 2 model is the 'virtual patient' model of Both et al. (2008). This model describes the relationships between different client states such as mood, thoughts, and coping skills as a system of differential equations, which allows simulation of the development of patient states over time. As discussed in our overview of predictive modeling approaches, such a model is developed using a top-down approach, where domain knowledge is formalized in a computational model. With the model, the development of internal states can be simulated to predict mood during treatment. The model incorporates hypothesized working mechanisms of different therapeutic interventions, such as activity scheduling or cognitive behavior therapy (CBT), to enable simulations of the potential effects of these interventions, which in turn could be used to identify the intervention that could benefit the patient the most Both and Hoogendoorn (2011). To do this effectively, additional model states are necessary to account for external variables that are not controllable by the client (Hoogendoorn et al., 2017). Similar approaches, using different computational techniques, have been proposed by Touboul, Romagnoni, and Schwartz (2010), Demic and Cheng (2014), Noble (2014), and Patten (2005).

2.4.3 Model Type 3: Predicting treatment outcome

Type 3 models predict treatment outcome (including drop-out) during active treatment. Rather than focusing on the relationship between short-term symptom dynamics and treatment decisions, these models aim to predict the outcome of the full therapeutic process (i.e., the post-test results) from data that are collected before and during treatment. This outcome has two aspects: 1) the change in health symptoms (which is typically known as the vector of continuous mental health questionnaire scores), and 2) whether or not treatment was provided as intended (which is typically a binary variable, flagging the treatment as drop-out for patients who did not complete a predetermined minimum percentage of treatment).

Using regression techniques, clinical researchers have been exploring type 3 models, using predictors that were collected at baseline (pre-test) assessments (e.g., DeRubeis et al. (2014), De Graaf et al. (2009), Karyotaki et al. (2015), Kegel and Flückiger (2015), and Meulenbeek, Seeger, and Klooster (2015), mental health questionnaire responses that were collected during treatment (e.g., Proudfoot et al. (2013) and Van Gemert-Pijnen, Kelders, and Bohlmeijer (2014), or repeated measures of the therapeutic relationship between the patient and the therapists (e.g., Priebe et al. (2011)). In these studies, researchers tend to focus more on the importance of individual predictors on a population level (i.e., risk factors), rather than on the predictive power of the model as a whole.

Recent studies suggest that logfiles of electronic treatment delivery systems might also provide useful detailed predictors of treatment outcome. The number of logins, actions per login, completed modules, and time spent are metrics that can identify differences between adherers and non-adherers (Donkin and Glozier, 2012). Van Gemert-Pijnen, Kelders, and Bohlmeijer (2014), for example, found user login frequency to be correlated with depressive symptoms after treatment. Whitton et al. (2015) examined the correlation between the usage of program features and outcomes in a study of a web-based self-help intervention for symptoms of depression and anxiety. The usage of diary functions and SMS reminding was not correlated with outcome, nor did the number of interventions (both started and completed) have an impact on symptom reduction. The symptom tracking functionality,

which enabled users to track their improvement, also had no influence on the final treatment result. However, the use of the reminder function of the symptom tracking tool had a positive influence on the outcome.

Advanced data mining techniques have been used to build Type 3 models as well. For example, Bennett and Doub (2010) used (naïve) Bayesian classifiers and Random Forest decision trees (see Breiman et al. (1984)) to predict treatment outcomes from health questionnaires that were collected at each treatment session. Early reductions in symptom levels were found to be a significant predictor of treatment outcome. Using 10-fold cross-validation, the predictive accuracy of the various classification models varied between 60% and 76%.

Early detection of clients at high risk for treatment resistance could also be helpful, for example, to decide on the optimal level of therapist guidance in a web-based treatment. Perlis (2013) used self-reported socio-demographic and clinical variables to predict treatment resistance. With multivariate models such as logistic regression, naïve Bayes and Support Vector Machines, the authors achieved an AUC of 0.71, indicating fair predictive performance.

An approach that utilizes free text written by patients is proposed in Hoogendoorn and Funk (2018). They extracted features from the text messages sent by patients suffering from an anxiety disorder to their therapist as part of an anxiety treatment. Features included word usage, sentiment, response rate, length of email, phrasing of the emails, and topics that patients wrote about in their emails. They were aimed at predicting a reliable improvement in the so-called Social Phobia Measure and were able to achieve AUCs of around 0.83 halfway through the therapy and similar scores at the end, which is significantly better than using baseline data only.

2.4.4 Model Type 4: Models for relapse prediction

Type 4 models focus on predicting the risk of relapse, that is, the re-occurrence of symptoms in the long term. Typically, the relapse risk determines the amount of aftercare required. The design of interventions and treatment platforms for relapse prevention training is similar to that of treatment platforms and interventions used in regular treatment (Barnes et al., 2007; Holländare et al., 2013; Lobban et al., 2015). To stabilize the condition of patients, aftercare can be provided in the form of screenings and interventions.

As with type 3 models, clinical researchers have used traditional regression techniques such as logistic regression to predict relapse in patients diagnosed with a variety of disorders, including depression (e.g., Kessing (2004)), bipolar depression (Busch et al., 2012), and alcohol misuse (Farren and McElroy, 2010; Pedersen and Hesse, 2009). More advanced statistical techniques have also been applied. For instance, Patten (2005) analyzed data from several clinical depression trials to estimate the parameters of a Markov model, and Van Voorhees et al. (2008) identified risk factors through regression tree modeling.

Type 4 models can support mental health specialists to assess the risk of relapse in their patients, to scale-up after-care when needed. More likely, however, is that these models will find their way into mobile self-help applications. If so, these models will probably perform better because the models will then be able to take current contextual information of the patient into account, resulting in more options to personalize interventions. According to Juarascio et al. (2015), context-aware interventions are expected to appear for treatment and relapse prevention for a variety of health problems, due to the rapid developments in sensor technology and mobile applications. Gustafson et al. (2011) and Chih et al. (2014) implemented Bayesian network modeling in an EMA/EMI smartphone application aimed at reducing relapse for recovering alcohol-dependent individuals. In this app, relapse-prevention interventions are triggered when the risk of relapse occurrence is estimated to be high based on the EMA ratings (including unobtrusive contextual variables). In a validation study, the model achieved an AUC of 0.91 in predicting relapse in the following week. A

similar EMA/EMI approach was proposed by (Aziz, Klein, and Treur, 2009), who developed a rule-based support agent that monitors client conditions to trigger EMI when the risk of relapse is predicted to be high, based on the client's mental health state, social interactions, estimated substance use, and physiological conditions.

2.5 Discussion

In this paper, we provided a brief introduction to predictive modeling methods, introduced a common framework for understanding applications of predictive modeling in mental health, and applied the framework to categorize published mental health research in which predictive modeling techniques were applied. Doing so, we aimed to contribute to the development of a common language between clinical researchers and members of the data mining community.

With this framework, opportunities can be identified to extend and improve an emerging field. For instance, our preliminary literature review suggests that e-mental health researchers should perhaps focus more on the validity of *model predictions* rather than the more traditional goal of identifying specific *predictors*. Demonstrating the relevance of a specific predictor has theoretical relevance. However, single predictors rarely provide a basis for predictions of variables that are relevant to clinical practice. In practice, often only a fraction of the variance of the target variable of interest is explained. In comparison with clinical researchers, data miners are more used (and more tolerant) to focusing on the accurate prediction of high-value target variables while focusing less on the specific predictors that play a role in this prediction. It would be interesting to learn more about the relative merits of this approach in e-mental health applications.

We also identified room for improvement in the proper application of predictive modeling methodology. We found that several studies do not use independent test sets to evaluate experimental modeling techniques. Proper testing (validation) of models is of utmost importance to investigate the generalizability of proposed models. Patterns in available data can often be modeled well, but more importantly testing a model on to unseen data is critical to test generalizability, as explained in the first section of this paper. The current wealth of data collected in e-mental health applications provides many opportunities to use test sets for model validation, and we urge researchers to apply this technique more. Furthermore, utilization of predictive modeling can have pitfalls such as estimation of a spurious correlation due to an involuntarily introduced bias or a systematic error during study design (Perlis, 2013; Sjölander, 2009). Especially, with the larger amount of data collected during online intentions this effect can magnify (Khoury and Ioannidis, 2014).This emphasis the need for a common language and understanding of the subject.

Some identified studies, relevant to mental health care, could not be readily classified with the proposed framework. These studies focused on the modeling of mental processes, without explicit links to psychotherapeutic intervention. By focusing on the therapeutic purpose of the modeling, the framework stresses the need for empirical validation of model performance in clinical settings, which is more stringent (i.e., prediction errors are more serious when they provide the basis for clinical treatment decisions). Nonetheless, these studies apply predictive modeling to processes that are relevant to mental health care. The framework may need to be extended to incorporate this type of research more easily.

We would like to stress that our review of available predictive modeling mental health research should not be considered exhaustive. Although we feel that the papers identified are representative for the field, we cannot rule out that important publications were missed, especially since the number of modeling studies seems to be rising each month. In our view, this illustrates the need for conceptual frameworks such as the one proposed in this paper,

so that the growing body of research initiatives can be more easily understood, categorized, and evaluated.

In this paper, we could only provide a brief overview of data mining methods. While our framework should promote a shared understanding of high-level research goals, we also acknowledge that productive collaboration in this field requires researchers to gain a deeper and better understanding of data mining methods. To promote this, we suggest that authors make a deliberate effort to thoroughly explain core aspects of data mining methods applied when they publish predictive modeling mental health care research (i.e., to contribute to the development of a common language by providing basic descriptions of core methods to which clinical researchers may be less familiar).

E-mental health is a rich interdisciplinary research field that enables many new research approaches, of which predictive modeling appears to be most promising. The framework proposed in this paper might serve to bridge the conceptual gap between psychologists and predictive modelers. The framework provides a common language for classifying predictive modeling mental health research, which may help to promote constructive and productive interdisciplinary research and to identify new research opportunities.

References

Altman, N. S. (1992). "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression". In: *The American Statistician* 46.3, pp. 175–185. DOI: 10.1080/00031305.1992.10475879. eprint: https://www.tandfonline.com/doi/pdf/10.1080/00031305.1992.10475879 (cit. on p. 43).

Ankarali, Handan, Ayse Canan, and Zeki Akkus (2007). "Comparison of logistic regression model and classification tree : An application to postpartum depression data". In: *Academic Emergency Medicine*. ISSN: 09574174. DOI: 10.1016/j.eswa.2006.02.022 (cit. on pp. 49, 51).

Armey, Michael F et al. (2015). "Ecological momentary assessment (EMA) of depression-related phenomena". In: *Current Opinion in Psychology* 4.401, pp. 21–25. ISSN: 2352250X. DOI: 10.1016/j.copsyc.2015.01.002 (cit. on p. 46).

Asselbergs, Joost et al. (2016). "Mobile Phone-Based Unobtrusive Ecological Momentary Assessment of Day-to-Day Mood: An Explorative Study". In: *Journal of medical Internet research* 18.3, e72. ISSN: 14388871. DOI: 10.2196/jmir.5505 (cit. on pp. 49, 52, 77, 78, 84, 89).

Aziz, Azizi A, Michel C.A. Klein, and Jan Treur (2009). "Modeling an ambient agent to support depression relapse prevention". In: *Proceedings - 2009 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Workshops, WI-IAT Workshops 2009*. Vol. 3, pp. 335–340. ISBN: 9780769538013. DOI: 10.1109/WI-IAT.2009.296 (cit. on pp. 50, 55).

Barnes, Caryl et al. (2007). *Evaluation of an online relapse prevention program for bipolar disorder: An overview of the aims and methodology of a randomized controlled trial*. English. DOI: 10.2165/00115677-200715040-00003 (cit. on p. 54).

Becker, Dennis et al. (2016a). "How to Predict Mood? Delving into Features of Smartphone-Based Data". In: *Proceedings of the 22nd Americas Conference on Information Systems (AMCIS)* August, pp. 1–10 (cit. on pp. 49, 52).

Bennett, Casey and T Doub (2010). "Data mining and electronic health records: Selecting optimal clinical treatments in practice". In: *Proceedings of the 6th International Conference on Data Mining*, pp. 313–318. arXiv: 1112.1668 (cit. on pp. 50, 54).

Bloom, David E. et al. (2011). "The Global Economic Burden of Noncommunicable Diseases". In: *World Economic Forum* September, pp. 1–46. ISSN: <null>. DOI: 10.1192/bjp.184.5.393 (cit. on p. 41).

Bolger, Niall et al. (1989). "Effects of daily stress on negative mood." In: *Journal of personality and social psychology* 57.5, pp. 808–818. ISSN: 0022-3514. DOI: 10.1037/0022-3514.57.5.808 (cit. on pp. 46, 107).

Both, Fiemke and Mark Hoogendoorn (2011). "Utilization of a virtual patient model to enable tailored therapy for depressed patients". In: *Neural Information Processing*, pp. 700–710 (cit. on pp. 49, 53).

Both, Fiemke et al. (2008). "Modeling the Dynamics of Mood and Depression". In: *Proceedings of the 18th European Conference on Artificial Intelligence, ECAI'08*. Ed. by M. Ghallab et al., pp. 266–270 (cit. on pp. 49, 53, 105).

Breiman, L et al. (1984). *Classification and Regression Trees*. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis. ISBN: 9780412048418 (cit. on p. 54).

Bremer, Vincent et al. (2017a). "Predicting the individual mood level based on diary data". In: *In: Proceedings of the Twenty-Fifth Conference on Information Systems (ECIS 2017)* (cit. on pp. 49, 53).

Burns, Michelle Nicole et al. (2011). "Harnessing context sensing to develop a mobile intervention for depression". In: *Journal of Medical Internet Research* 13.3. ISSN: 14388871. DOI: 10.2196/jmir.1838. arXiv: arXiv:1011.1669v3 (cit. on pp. 42, 49, 51, 77).

Burton, Christopher, David Weller, and Michael Sharpe (2009). "Functional somatic symptoms and psychological states: an electronic diary study." In: *Psychosomatic medicine* 71.1, pp. 77–83. ISSN: 0033-3174. DOI: 10.1097/PSY.0b013e31818f2acb (cit. on p. 46).

Busch, Alisa B et al. (2012). "Accurately Predicting Bipolar Disorder Mood Outcomes". In: *Medical Care* 50.4, pp. 311–319. ISSN: 0025-7079. DOI: 10.1097/MLR.0b013e3182422aec (cit. on pp. 50, 54).

Chang, Keng-hao, Drew Fisher, and John Canny (2011). "AMMON: A Speech Analysis Library for Analyzing Affect, Stress, and Mental Health on Mobile Phones". In: *Proceedings of the 2011 PhoneSense conference*. DOI: 10.1.1.232.365 (cit. on pp. 4, 49, 52).

Chih, Ming Yuan et al. (2014). "Predictive modeling of addiction lapses in a mobile health application". In: *Journal of Substance Abuse Treatment* 46.1, pp. 29–35. ISSN: 07405472. DOI: 10.1016/j.jsat.2013.08.004 (cit. on pp. 50, 54, 65, 105).

Cortes, Corinna and Vladimir Vapnik (1995). "Supprot-Vector Networks". In: *Machine Learning* 297.20, pp. 273–297. ISSN: 1747-0285. DOI: 10.1111/j.1747-0285.2009.00840.x. arXiv: arXiv:1011.1669v3 (cit. on pp. 42, 43).

David Arthur, Sergei Vassilvitskii (2007). "k-means ++ : The Advantages of Careful Seeding". In: *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1027–1035. ISSN: 0898716241. DOI: 10.1145/1283383.1283494 (cit. on p. 51).

Dawes, Robyn M. (1979). "The robust beauty of improper linear models in decision making." In: *American Psychologist* 34.7, pp. 571–582. ISSN: 0003-066X. DOI: 10.1037/0003-066X.34.7.571 (cit. on p. 43).

De Graaf, L. E. et al. (2009). "Clinical effectiveness of online computerised cognitive-behavioural therapy without support for depression in primary care: Randomised trial". In: *British Journal of Psychiatry* 195.1, pp. 73–80. ISSN: 00071250. DOI: 10.1192/bjp.bp.108.054429 (cit. on pp. 1, 3, 53, 140).

Demic, Selver and Sen Cheng (2014). "Modeling the Dynamics of Disease States in Depression". In: *PLoS ONE* 9.10, e110358. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0110358 (cit. on pp. 49, 53).

Demirci, Kadir, Mehmet Akgönül, and Abdullah Akpinar (2015). "Relationship of smartphone use severity with sleep quality, depression, and anxiety in university students". In: *Journal of Behavioral Addictions* 4.2, pp. 85–92. ISSN: 2062-5871. DOI: 10.1556/2006.4.2015.010 (cit. on pp. 49, 51).

DeRubeis, Robert J. et al. (2014). "The personalized advantage index: Translating research on prediction into individualized treatment recommendations. A demonstration". In: *PLoS ONE* 9.1, pp. 1–8. ISSN: 19326203. DOI: 10.1371/journal.pone.0083875 (cit. on p. 53).

Donkin, Liesje and Nick Glozier (2012). "Motivators and motivations to persist with online psychological interventions: A qualitative study of treatment completers". In: *Journal of Medical Internet Research* 14. ISSN: 14388871. DOI: 10.2196/jmir.2100 (cit. on pp. 50, 53).

Doryab, Afsaneh et al. (2014a). "Detection of behavior change in people with depression". In: *AAAI Workshops Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pp. 12–16 (cit. on p. 51).

Doryab, Afsaneh et al. (2014b). "Detection of Behavior Change in People with Depression Overview of the Data Collection app". In: *AAAI Workshop on Modern Artificial Intelligence for Health Analytics*, pp. 12–16 (cit. on p. 49).

Eagle, Nathan, Alex (Sandy) Pentland, and David Lazer (2009). "Inferring friendship network structure by using mobile phone data". In: *Proceedings of the National Academy of Sciences* 106.36, pp. 15274–15278. ISSN: 0027-8424. DOI: 10.1073/pnas.0900282106. eprint: https://www.pnas.org/content/106/36/15274.full.pdf (cit. on p. 46).

Farren, Conor K. and Sharon McElroy (2010). "Predictive factors for relapse after an integrated inpatient treatment programme for unipolar depressed and bipolar alcoholics". In: *Alcohol and Alcoholism* 45.6, pp. 527–533. ISSN: 07350414. DOI: 10.1093/alcalc/agq060 (cit. on pp. 50, 54).

Farren, Conor K. et al. (2013). "Prognostic factors of 2-year outcomes of patients with comorbid bipolar disorder or depression with alcohol dependence: Importance of early abstinence". In: *Alcohol and Alcoholism* 48.1, pp. 93–98. ISSN: 07350414. DOI: 10.1093/alcalc/ags112 (cit. on p. 50).

Flach, Peter A. and Nicolas Lachiche (2001). "Confirmation-guided discovery of first-order rules with Tertius". In: *Machine Learning* 42.1-2, pp. 61–95. ISSN: 08856125. DOI: 10.1023/A:1007656703224 (cit. on p. 49).

Forster, Malcolm and Elliott Sober (1994). "How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions". In: *British Journal for the Philosophy of Science* 45.1, pp. 1–35. ISSN: 00070882. DOI: 10.1093/bjps/45.1.1 (cit. on p. 43).

Gittelman, Steven et al. (2015). "A new source of data for public health surveillance: Facebook likes." In: *Journal of medical Internet research* 17.4, e98. ISSN: 1438-8871 (cit. on pp. 49, 51).

Gustafson, David H et al. (2011). "Explicating an evidence-based, theoretically informed, mobile technology-based system to improve outcomes for people in recovery for alcohol dependence." In: *Substance use & misuse* 46.1, pp. 96–111. ISSN: 1532-2491. DOI: 10.3109/10826084.2011.521413 (cit. on pp. 50, 54).

Hanley, J A and B J McNeil (1982). "The meaning and use of the area under a receiver operating characteristic (ROC) curve." In: *Radiology* 143.1, pp. 29–36. ISSN: 0033-8419. DOI: 10.1148/radiology.143.1.7063747. arXiv: NIHMS150003 (cit. on p. 44).

Haykin, Simon (2008). *Neural Networks and Learning Machines: A Comprehensive Foundation*, p. 906. ISBN: 9780131471399. DOI: 978-0131471399. arXiv: 1312.6199v4 (cit. on pp. 42, 43).

Health, Centre for mental (2010). "The economic and social costs of mental health problems in 2009 / 10". In: *Health (San Francisco)* (cit. on p. 41).

Heron, Kristin E and Joshua M Smyth (2010). "Ecological Momentary Interventions: Incorporating Mobile Technology Into Psychosocial and Health Behavior Treatments". In: *British Journal of Health Psychology* 15.Pt 1, pp. 1–39. DOI: 10.1348/135910709X466063.Ecological (cit. on pp. 6, 46).

Hoek, Wiebe van der and Michael Wooldridge (2008). "Multi-Agent Systems". In: *Canadian Journal of Nursing Research*. Vol. 33. 2, pp. 887–928. ISBN: 0844-5621 (Print) 0844-5621 (Linking). DOI: 10.1016/S1574-6526(07)03024-6 (cit. on p. 44).

Holländare, Fredrik et al. (2013). "Two-year outcome of internet-based relapse prevention for partially remitted depression". In: *Behaviour Research and Therapy* 51.11, pp. 719–722. ISSN: 00057967. DOI: 10.1016/j.brat.2013.08.002 (cit. on p. 54).

Hoogendoorn, Mark and Burkhardt Funk (2018). *Machine Learning for the Quantified Self*. Vol. 35. ISBN: 978-3-319-66307-4. DOI: 10.1007/978-3-319-66308-1 (cit. on pp. 42, 54).

Hoogendoorn, Mark et al. (2017). "Predicting Social Anxiety Treatment Outcome Based on Therapeutic Email Conversations". In: *IEEE Journal of Biomedical and Health Informatics* 21.5, pp. 1449–1459. ISSN: 2168-2194. DOI: 10.1109/JBHI.2016.2601123. arXiv: 15334406 (cit. on pp. 50, 53).

Jacelon, Cynthia S and Kristal Imperio (2005). "Participant diaries as a source of data in research with older adults." In: *Qualitative health research* 15.7, pp. 991–7. ISSN: 1049-7323. DOI: 10.1177/1049732305278603 (cit. on p. 46).

Juarascio, Adrienne S. et al. (2015). "Review of smartphone applications for the treatment of eating disorders". In: *European Eating Disorders Review* 23.1, pp. 1–11. ISSN: 10990968. DOI: 10.1002/erv.2327 (cit. on pp. 54, 88).

Karyotaki, E et al. (2015). "Predictors of treatment dropout in self-guided web-based interventions for depression: an 'individual patient data' meta-analysis". In: *Psychological Medicine* 45.13, pp. 2717–2726. ISSN: 0033-2917. DOI: 10.1017/S0033291715000665 (cit. on pp. 49, 53).

Kegel, Alexander F and Christoph Flückiger (2015). "Predicting Psychotherapy Dropouts: A Multilevel Approach". In: *Clinical Psychology and Psychotherapy* 22.5, pp. 377–386. ISSN: 10990879. DOI: 10.1002/cpp.1899 (cit. on pp. 50, 53).

Kessing, Lars Vedel (2004). "Severity of depressive episodes according to ICD-I0: Prediction of risk of relapse and suicide." In: *The British Journal of Psychiatry* 184.2, pp. 153–156 (cit. on pp. 48, 50, 54).

Kessler, Rc and Wt Chiu (2005). "Prevalence, Severity, and Comorbidity of Twelve-month DSM-IV Disorders in the National Comorbidity Survey Replication (NCS- R)". In: *Archives of general psychiatry* 62.6, pp. 617–627. ISSN: 0003990X. DOI: 10.1001/archpsyc.62.6.617.Prevalence (cit. on p. 46).

Khoury, Muin J and John P A Ioannidis (2014). *Big data meets public health*. DOI: 10.1126/science.aaa2709. arXiv: 15334406 (cit. on p. 55).

Kim, J et al. (2015). "Covariation of Depressive Mood and Spontaneous Physical Activity in Major Depressive Disorder: Toward Continuous Monitoring of Depressive Mood". In: *Biomedical and Health Informatics, IEEE Journal of* 19.4, pp. 1347–1355. ISSN: 2168-2194. DOI: 10.1109/JBHI.2015.2440764 (cit. on pp. 49, 51).

Kok, Gemma et al. (2014). "Mobile cognitive therapy: Adherence and acceptability of an online intervention in remitted recurrently depressed patients". In: *Internet Interventions* 1.2, pp. 65–73. ISSN: 22147829. DOI: 10.1016/j.invent.2014.05.002 (cit. on p. 46).

Lambert, Michael J. et al. (2003). *Is it time for clinicians to routinely track patient outcome? A meta-analysis*. DOI: 10.1093/clipsy/bpg025 (cit. on p. 52).

Lane, Nicholas D. et al. (2010). "A survey of mobile phone sensing". In: *IEEE Communications Magazine* 48.9, pp. 140–150. ISSN: 01636804. DOI: 10.1109/MCOM.2010.5560598. arXiv: 10[0163-6804] (cit. on p. 46).

Langley, Pat, Wayne Iba, and Kevin Thompson (1992). *An Analysis of Bayesian Classifiers* (cit. on pp. 42, 43).

LiKamWa, Robert et al. (2013). "MoodScope". In: *Proceeding of the 11th annual international conference on Mobile systems, applications, and services - MobiSys '13*, p. 389. DOI: 10.1145/2462456.2464449 (cit. on pp. 49, 51, 52, 77, 80, 81, 88, 89).

Lobban, F. et al. (2015). "Feasibility and acceptability of web-based enhanced relapse prevention for bipolar disorder (ERPonline): Trial protocol". In: *Contemporary Clinical Trials* 41, pp. 100–109. ISSN: 15592030. DOI: 10.1016/j.cct.2015.01.004 (cit. on p. 54).

Lord, Sarah et al. (2016). "Implementation of a Substance Use Recovery Support Mobile Phone App in Community Settings: Qualitative Study of Clinician and Staff Perspectives of Facilitators and Barriers." In: *JMIR mental health* 3.2, e24. ISSN: 2368-7959. DOI: 10.2196/mental.4927 (cit. on p. 46).

Lu, Hong et al. (2012). "StressSense: Detecting Stress in Unconstrained Acoustic Environments using Smartphones". In: *Proceedings of the 2012 ACM Conference on Ubiquitous Computing - UbiComp '12*, p. 351. DOI: 10.1145/2370216.2370270 (cit. on pp. 4, 49, 52).

Ma, Yuanchao et al. (2012a). "Daily mood assessment based on mobile phone sensing". In: *Proceedings - BSN 2012: 9th International Workshop on Wearable and Implantable Body Sensor Networks*, pp. 142–147. ISBN: 9780769546988. DOI: 10.1109/BSN.2012.3 (cit. on pp. 49, 51).

Magidson, Jessica F. et al. (2014). "Theory-driven intervention for changing personality: Expectancy value theory, behavioral activation, and conscientiousness." In: *Developmental Psychology* 50.5, pp. 1442–1450. ISSN: 1939-0599. DOI: 10.1037/a0030583. arXiv: NIHMS150003 (cit. on p. 47).

Mestry, Madhura et al. (2015). "Identifying associations between smartphone usage and mental health during depression, anxiety and stress". In: *Proceedings - 2015 International Conference on Communication, Information and Computing Technology, ICCICT 2015*, pp. 1–5. ISBN: 9781479955220. DOI: 10.1109/ICCICT.2015.7045656 (cit. on pp. 49, 51).

Meulenbeek, Peter, Kristin Seeger, and Peter M. ten Klooster (2015). "Dropout prediction in a public mental health intervention for sub-threshold and mild panic disorder". In: *The Cognitive Behaviour Therapist* 8, e5. ISSN: 1754-470X. DOI: 10.1017/S1754470X15000057 (cit. on pp. 50, 53).

Miller, Scott D et al. (2003). "The Outcome Rating Scale : A Preliminary Study of the Reliability , Validity , and Feasibility of a Brief Visual Analog Measure". In: *Journal of Brief Therapy* 2.2, pp. 91–100 (cit. on p. 52).

Mittendorfer-Rutz, Ellenor et al. (2014). "Association of socio-demographic factors, sickleave and health care patterns with the risk of being granted a disability pension among psychiatric outpatients with depression". In: *PLoS ONE* 9.6. ISSN: 19326203. DOI: 10.1371/journal.pone.0099869 (cit. on p. 47).

Moskowitz, Debbie S and Simon N Young (2006). "Ecological momentary assessment: what it is and why it is a method of the future in clinical psychopharmacology." In: *Journal of psychiatry & neuroscience : JPN* 31.1, pp. 13–20. ISSN: 1180-4882 (cit. on p. 46).

Noble, S L (2014). "Control-theoretic scheduling of psychotherapy and pharmacotherapy for the treatment of post-traumatic stress disorder". In: *Control Theory Applications, IET* 8.13, pp. 1196–1206. ISSN: 1751-8644. DOI: 10.1049/iet-cta.2013.0615 (cit. on pp. 49, 53).

Olson, David L. and Dursun Delen (2008). *Advanced data mining techniques*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 1–180. ISBN: 9783540769163. DOI: 10.1007/978-3-540-76917-0. arXiv: arXiv:1011.1669v3 (cit. on p. 44).

Osmani, Venet et al. (2013). "Monitoring activity of patients with bipolar disorder using smart phones". In: *Proceedings of International Conference on Advances in Mobile Computing & Multimedia - MoMM '13*, pp. 85–92. DOI: 10.1145/2536853.2536882 (cit. on pp. 49, 51).

Otto, Michael W. et al. (2001). "An effect-size analysis of the relative efficacy and tolerability of serotonin selective reuptake inhibitors for panic disorder". In: *American Journal of Psychiatry* 158.12, pp. 1989–1992. ISSN: 0002953X. DOI: 10.1176/appi.ajp.158.12.1989 (cit. on p. 47).

Panagiotakopoulos, T C et al. (2010). "A Contextual Data Mining Approach Toward Assisting the Treatment of Anxiety Disorders". In: *Information Technology in Biomedicine, IEEE Transactions on* 14.3, pp. 567–581. ISSN: 1089-7771. DOI: 10.1109/TITB.2009.2038905 (cit. on pp. 49, 51).

Patten, Scott B (2005). "Markov models of major depression for linking psychiatric epidemiology to clinical practice." In: *Clinical practice and epidemiology in mental health : CP & EMH* 1.1, p. 2. ISSN: 1745-0179. DOI: 10.1186/1745-0179-1-2 (cit. on pp. 53, 54).

Pedersen, Mads Uffe and Morten Hesse (2009). "A simple risk scoring system for prediction of relapse after inpatient alcohol treatment." In: *The American journal on addictions / American Academy of Psychiatrists in Alcoholism and Addictions* 18.6, pp. 488–493. ISSN: 1521-0391. DOI: 10.3109/10550490903205983 (cit. on pp. 50, 54).

Perlis, Roy H. (2013). "A clinical risk stratification tool for predicting treatment resistance in major depressive disorder". In: *Biological Psychiatry*. ISSN: 00063223. DOI: 10.1016/j.biopsych.2012.12.007 (cit. on pp. 50, 54, 55).

Pestian, J and H Nasrallah (2010). "Suicide note classification using natural language processing: A content analysis". In: *Biomedical . . .* 2010.3, pp. 19–28 (cit. on pp. 49, 51).

Priebe, Stefan et al. (2011). "Does the therapeutic relationship predict outcomes of psychiatric treatment in patients with psychosis? A systematic review." In: *Psychotherapy and psychosomatics* 80.2, pp. 70–7. ISSN: 1423-0348. DOI: 10.1159/000320976 (cit. on pp. 50, 53).

Proudfoot, Judith et al. (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". In: *BMC Psychiatry* 13.1, p. 312. ISSN: 1471-244X. DOI: 10.1186/1471-244X-13-312 (cit. on pp. 50, 53).

Robinson, Emma et al. (2010). "Internet treatment for generlized anxiety disorder: A randomized controlled trial comparing clinician vs. technician assistance". In: *PLoS ONE* 5.6. ISSN: 19326203. DOI: 10.1371/journal.pone.0010942 (cit. on pp. 42, 64).

Runyan, Jason D. et al. (2013). "A Smartphone Ecological Momentary Assessment / Intervention "App" for Collecting Real-Time Data and Promoting Self-Awareness". In: *PLoS ONE* 8.8. ISSN: 19326203. DOI: 10.1371/journal.pone.0071325 (cit. on pp. 6, 46, 49, 52).

S. Rasoul Safavian and David Landgrebe (1991). "A survey of decision tree classifier methodology". In: 21.3, pp. 660–674. DOI: 10.9790/5933-0612104113 (cit. on p. 43).

Saeb, Sohrab et al. (2015b). "The Relationship between Clinical, Momentary, and Sensor-based Assessment of Depression." In: *International Conference on Pervasive Computing Technologies for Healthcare : [proceedings]. International Conference on Pervasive Computing Technologies for Healthcare* 2015, pp. 7–10. ISSN: 2153-1633. DOI: 10.4108/icst.pervasivehealth.2015.259034 (cit. on pp. 4, 49, 51).

Shiffman, Saul, Arthur A. Stone, and Michael R. Hufford (2008). "Ecological momentary assessment". In: *Annual review of clinical psychology* 4.5, pp. 1–32. ISSN: 1548-5943; 1548-5943. DOI: 10.1146/annurev.clinpsy.3.022806.091415 (cit. on pp. 42, 88).

Shmueli, Galit (2010). "To Explain or to Predict?" In: *Statistical Science* 25.3, pp. 289–310. ISSN: 0883-4237. DOI: 10.1214/10-STS330. arXiv: 1101.0891 (cit. on p. 43).

Sjölander, Arvid (2009). "Propensity scores and M-structures". In: *Statistics in Medicine* 28.9, pp. 1416–1420. ISSN: 02776715. DOI: 10.1002/sim.3532 (cit. on p. 55).

Sluis, Frans Van Der, Ton Dijkstra, and Egon L Van Den Broek (2011). "COMPUTER AIDED DIAGNOSIS FOR MENTAL HEALTH CARE On the Clinical Validation of Sensitive Machines". In: pp. 493–498 (cit. on pp. 49, 52).

Smyth, Joshua M and Arthur a Stone (2003). "Ecological momentary assessment research in behavioral medicine". In: *Journal of Happiness Studies* 4.1, pp. 35–52. ISSN: 1389-4978; 1573-7780. DOI: 10.1023/A:1023657221954 (cit. on pp. 46, 104).

Stange, Martin and Burkhardt Funk (2016). "How Big Does Big Data Need To Be?" In: *Enterprise Big Data Engineering, Analytics, and Management*, pp. 1–15. ISBN: 9781522502937. DOI: 10.4018/978-1-5225-0293-7.ch001 (cit. on p. 46).

Titov, Nickolai et al. (2015). "MindSpot Clinic: An Accessible, Efficient, and Effective Online Treatment Service for Anxiety and Depression". In: *Psychiatric Services* 66.10, pp. 1043–1050. ISSN: 1075-2730. DOI: 10.1176/appi.ps.201400477 (cit. on p. 42).

Torous, John et al. (2015). "Utilizing a Personal Smartphone Custom App to Assess the Patient Health Questionnaire-9 (PHQ-9) Depressive Symptoms in Patients With Major Depressive Disorder". In: *JMIR Mental Health* 2.1, e8. DOI: 10.2196/mental.3889 (cit. on pp. 52, 65, 90).

Touboul, Jonathan, Alberto Romagnoni, and Robert Schwartz (2010). "On the Dynamic Interplay between Positive and Negative Affects". In: pp. 1–31. arXiv: 1004.4856 (cit. on pp. 49, 53).

Tovar, Alison et al. (2014). "Baseline Socio-demographic Characteristics and Self-Reported Diet and Physical Activity Shifts Among Recent Immigrants Participating in the Randomized Controlled Lifestyle Intervention: "Live Well"". In: *Journal of Immigrant and Minority Health* 16.3, pp. 457–465. ISSN: 1557-1912. DOI: 10.1007/s10903-013-9778-8 (cit. on p. 47).

Trull, Timothy J and Ulrich Ebner-Priemer (2013). "Ambulatory Assessments". In: *Annu Rev Clin Psychol* 9, pp. 151–176. ISSN: 1548-5943. DOI: 10.1146/annurev-clinpsy-050212-185510.Ambulatory (cit. on pp. 1, 42).

van Breda, Ward et al. (2016). "Exploring and comparing machine learning approaches for predicting mood over time". In: *Innovation in Medicine and Healthcare 2016*. Vol. 60. Smart Innovation, Systems and Technologies. Springer Science and Business Media Deutschland GmbH, pp. 37–47. ISBN: 9783319396866. DOI: 10.1007/978-3-319-39687-3_4 (cit. on pp. 49, 52).

Van Gemert-Pijnen, Julia E W C, Saskia M. Kelders, and Ernst T. Bohlmeijer (2014). "Understanding the usage of content in a mental health intervention for depression: An analysis of log data". In: *Journal of Medical Internet Research* 16.1, e27. ISSN: 14388871. DOI: 10.2196/jmir.2991 (cit. on pp. 50, 53).

Van Voorhees, Benjamin W. et al. (2008). "Predicting Future Risk of Depressive Episode in Adolescents: The Chicago Adolescent Depression Risk Assessment (CADRA)". In: *Annals of Family Medicine* 6.6, pp. 503–512. DOI: 10.1370/afm.887.INTRODUCTION (cit. on pp. 50, 54, 88).

Vittengl, Jeffrey R et al. (2009). "Reducing Relapse and Recurrence in Unipolar Depression: A Comparative Meta-Analysis of Cognitive–Behavioral Therapy's Effects". In: 75.June 2006, pp. 475–488. DOI: 10.1037/0022-006X.75.3.475.Reducing (cit. on p. 46).

Wade, Wendy A., Teresa A. Treat, and Gregory L. Stuart (1998). "Transporting an empirically supported treatment for panic disorder to a service clinic setting: A benchmarking strategy." In: *Journal of Consulting and Clinical Psychology* 66.2, pp. 231–239. ISSN: 1939-2117. DOI: 10.1037/0022-006X.66.2.231 (cit. on p. 47).

Werf, S Y van der et al. (2006). "Major depressive episodes and random mood". In: *Archives of General Psychiatry* 63(5).May 2006, pp. 509–518. ISSN: 0003-990X. DOI: 10.1001/archpsyc.63.5.509 (cit. on pp. 49, 51).

Whitton, Alexis E et al. (2015). "Breaking Open the Black Box : Isolating the Most Potent Features of a Web and Mobile Phone-Based Intervention for Depression , Anxiety , and Stress". In: 2, pp. 1–13. DOI: 10.2196/mental.3573 (cit. on pp. 50, 53).

Wichers, M. et al. (2011). "Momentary assessment technology as a tool to help patients with depression help themselves". In: *Acta Psychiatrica Scandinavica* 124.4, pp. 262–272. ISSN: 0001690X. DOI: 10.1111/j.1600-0447.2011.01749.x (cit. on pp. 1, 42, 46, 88).

Witten, Ian H., Eibe Frank, and Mark A. Hall (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd ed. Morgan Kaufmann Series in Data Management Systems. Amsterdam: Morgan Kaufmann. ISBN: 978-0-12-374856-0 (cit. on p. 43).

Zeigler, Bernard P., Herbert Praehofer, and Tag Gon Kim (2000). *Theory of Modeling [Modelling] and Simulation Integrating Discrete Event*. January. ISBN: 0127784551 (cit. on p. 44).

Chapter 3

Acceptance of mobile mental health treatment applications

Becker, D. (2016). In Procedia Computer Science, Volume 58, pp. 220–227.

Abstract: *Mobile mental health applications are regarded as a promising solution to meet increasing demands in mental health treatment. They are used to treat mental disorders and can only be successful if the treatment population accepts and appreciates them. This research analyses the acceptance of mobile mental health applications by young adults in Germany in order to identify inhibiting factors regarding their use. To describe people's intentions to use mobile treatment applications, an extended version of the technology acceptance model (TAM) is applied. In the past, TAM has already been used to access the acceptance and adaption of new medical applications. The findings suggest that knowledge about the existence and clinical effectiveness of mobile mental health applications are considerably low. Even though, mobile applications are considered easy to use, their effectiveness in treating mental disorders is questioned by the young adults. Furthermore, concerns that personal information can potentially be revealed arise. This can additionally inhibit the acceptance of these applications. To improve the acceptance and increase future usage, mobile mental health applications should be promoted as a supporting tool that is always available for anyone and can facilitate mental treatment.*

3.1 Introduction

Mental health disorders such as anxiety, depression, social anxiety, or substance abuse are an increasing problem in our society. According to the World Health Organization, the gap between the need for treatment of mental disorders and the accessibility of treatment is rising, and already between 35% and 50% of mentally ill clients receive no treatment because appropriate treatment places are rare (Board, 2012).

One possible solution to meet the demand for mental health treatment can be online treatment (White, Krousel-Wood, and Mather, 2001). Internet or computer-based cognitive behaviour therapy programs have proven clinically effective for the treatment of a variety of mental disorders (Cuijpers et al., 2009; Andersson and Cuijpers, 2009; Olatunji, Cisler, and Deacon, 2010). An advantage of online treatment is its time and cost effectiveness. The amount of time that clinicians require for each client is considerably less than in regular face-to-face treatment (Vernmark et al., 2010; Robinson et al., 2010), which means that more clients can be supervised than in a conventional therapy setting. Despite the compelling evidence regarding the effectiveness, of computer or Internet-based treatment, the acceptance of Internet treatment outside the health sector is considerably lower (Wootton et al., 2011). Studies regarding the acceptability of web-based treatment programs among the population

report mixed results. It appears that the wide-spread opinion among the population is that online treatment is only effective in cases of mild and moderate symptoms(Gun, Titov, and Andrews, 2011) and moreover restricted to certain diseases (Musiat, Goldstone, and Tarrier, 2014). However, there are also positive beliefs about online treatment of mental diseases. Former participants in web-based treatment report higher acceptability after using the application compared to before. They are also more inclined to use such services in the future again (Gun, Titov, and Andrews, 2011). In the case of anxiety and depression, it appears that online treatment would even be a preferable treatment option because of anonymity concerns (Wootton et al., 2011). In addition, the convenience of accessing online treatment from home, and the fact that it does not require waiting time to start with the therapy are reasons in favour of online treatment (Musiat, Goldstone, and Tarrier, 2014).

However, the majority of the research that has been conducted in this area is concerned about the usage of web-based online intervention programs that require a stand-alone PC. But today's mental health applications are mostly mobile phone applications that are carried around in one's pocket and are accessible any time. Another advantage of these smart-phone applications is that they do not only provide useful interventions and screenings to track a user's improvement, they can also make use of the smart-phone's sensors to measure current location, activity and recent calls. With these measures, the client's current condition can be assessed and momentary interventions can be triggered to assist the client in difficult and stressful situations (Torous et al., 2015; Chih et al., 2014).

This study analyses the acceptance and intention to use mobile mental health treatment applications by young adults in the Germany population. Adults between the ages of 18 and 35 are focused in this research because it might take some years until mobile mental health treatment applications are widely available. They also represent the future target population that might require mental health treatment. Additionally, young adults are open to new technology, already familiar with the use of mobile phones, and adapting to the use of mobile mental health treatment applications might require less effort for them than for older people. To infer the current acceptance and future intentional use of such applications, the technology acceptance model (TAM) (Davis, 1985) is used. The results lead to implications for promoting and developing greater acceptance of mobile mental health applications because the success of these applications depends on understanding peoples concerns and identifying the factors that promote or inhibit their use.

3.2 Method

3.2.1 Structural equation model

To describe people's intentions to use mobile mental health applications, a structural equation model was developed. This model is based on the technology acceptance model (TAM) (Davis, 1985) and on previous research about acceptance of mobile services. In previous research, TAM was introduced to estimate acceptance of technological innovations and predict their future use in companies. The main components in TAM that describe the intention to use a new technology are perceived usefulness and perceived ease of use. Perceived usefulness is the impact a user expects on their performance due to their system use; perceived ease of use describes the users anticipated effort in using the new system.

The number of concepts that explain the acceptance of new technologies were further refined and extended in TAM2 (Venkatesh and Davis, 2000b; Venkatesh, 2000), Unified Theory of Acceptance and Use of Technology (UTAUT) (Venkatesh et al., 2003), and UTAUT2 (Venkatesh, Thong, and Xu, 2012). Adaptations of TAM have already been used in the context of medical applications for evaluating a variety of technologies such as a fictional online diagnosis program (Lanseng and Andreassen, 2007), use of virtual reality

as a therapeutic tool (Bertrand and Bouchard, 2008), and intention to use telepsychotherapy (Monthuy-Blanc et al., 2013).

The centre of the model developed here is represented by the perceived ease of use and perceived usefulness of mobile mental health applications. These concepts mainly influence a client's intention to use such an application when facing mental health problems (H1, H2). Furthermore, in TAM the perceived ease of use also influences the perceived usefulness (H3) of the application. To the concepts of perceived usefulness and ease of use, the concept of social influence is added.

The concept of social influence is part of UTAT, TAM3, and various research studies that evaluate the future of mobile services (Kaasinen, 2005; Martignoni et al., 2008). Social influence describes to which extent users perceive that their social environment, such as family members, friends, and colleagues, believe the application should be used. Therefore, social influence is modelled to mediate the general technology perception (H4) as well as directly influence the behavioural intention (H5).

Another concept that is part of TAM3 and included into the model is self-efficacy. Since electronic treatment requires a high amount of self-dedication compared to regular face-to-face therapy, self-efficacy is an important factor for clients considering online treatment. Clients who are well aware of the fact that they cannot work through the exercises on their own should have a reduced intention to use an e-mental health solution even if these clients are convinced of the benefits of online treatment. This phenomenon is reported for online learning applications (Tung and Chang, 2008) and acceptance of mobile health services (Sun et al., 2013). Self-efficacy directly influences the behavioural intention (H6) to use a new application.

A significant influencing factor for the success of online applications is trust. Lack of trust in the application and the security of personal data might adversely affect people who consider online solutions (Pedersen and Nysveen, 2003; Malhotra, Kim, and Agarwal, 2004). During online treatment, sensitive data about the client is collected, which can lead to privacy and similar concerns about online applications. Thus, this might be an discouraging factor for clients comparing to using an online application or face-to-face treatment. As the latter might provide more security concerning personal details. Therefore, the concept of trust is incorporated into the model, even though it is not default in UTAUT. The concept of trust is expected to directly influence the intention to use a mobile mental health application (H7).

Although the task-technology fit aspect is not part of the latest TAM development, it is added to the present research model. The task-technology fit model (TTF) (Goodhue, Thompson, and Goodhue, 2014) assumes that people will use technology that fits a task well. Initially, TTF was developed to evaluate workspace technologies, but was adapted to fit other purposes as well. It has already been used in combination with TAM (Klopping and Mckinney, 2004) and the combined model has proven to be superior than either one alone (Dishaw and Strong, 1999). TTF does not influence behavioural intention directly but rather effects perceived ease of use (H8) and perceived usefulness (H9). The research hypotheses are summarised in Table 3.1.

Research hypotheses	Source
H1: Perceived ease of use influences behavioural intention	(Davis, 1993)
H2: Perceived usefulness influences behavioural intention	(Davis, 1993)
H3: Perceived ease of use influences perceived usefulness	(Davis, 1993)
H4: Social influence influences perceived usefulness	(Kaasinen, 2005; Martignoni et al., 2008)
H5: Social influence directly influences behavioural intention	(Venkatesh et al., 2003)
H6: Perceived self-efficacy directly influences behavioural intention	(Tung and Chang, 2008; Sun et al., 2013)
H7: Trust in the application's security influences behavioural intention	(Pedersen and Nysveen, 2003; Malhotra, Kim, and Agarwal, 2004)
H8: Task-technology fit influences perceived ease of use	(Klopping and Mckinney, 2004; Dishaw and Strong, 1999)
H9: Task-technology fit influences perceived usefulness of the application	(Goodhue, Thompson, and Goodhue, 2014; Klopping and Mckinney, 2004; Dishaw and Strong, 1999)

TABLE 3.1: Hypotheses tested in this research.

3.2.2 Measurement tool

To evaluate the extended TAM, a structured questionnaire is created that is based on questionnaires that have been used in previous studies. The final questionnaire includes 33 items that measure the 7 different concepts. In the following, a short overview of the origin of the questions for the individual concepts is given. A detailed listing of the final questionnaire can be found in 3.5. The concept of perceived usefulness is measured with 6 items and the questions are adapted from studies about mobile commerce (Pedersen, 2001; Nysveen, 2005) and user satisfaction (Wixom and Todd, 2005). The concept of ease of use is measured with 5 items and the questions are adapted from the original TAM (Davis, Bagozzi, and Warshaw, 1989) and further refinements were taken from a study by Bagozzi and Richard (2002) (Bagozzi, 2002). The seven questions used to measure perceived task-technology fit were previously used by Jarupathirun and Zahedi (2007). Social influence was measured with 5 items and was utilised by Nysveen (2005) in a study on intention to use mobile services. Trust and self-efficacy consist of 4 items each. The tool to measure self-efficacy was previously applied by Park (2009) and the measurement items for trust are adapted from UTAUT (Venkatesh et al., 2003). The outcome variable behavioural intention is measured with two questions and these are adopted from Parks (Park, 2009) analysis regarding the acceptance of e-learning. For all questions, a 7-point Likert-type scale ranging from 1 for strongly disagree to 7 for strongly agree was applied.

3.2.3 Data collection and analysis

For participant recruitment of German participants of both genders within an age of 18 to 35 years, requests in various facebook survey groups were posted. Typically, students and companies use these groups for participant recruitment. In total, 125 people were recruited from December 16th, 2015, to February 16th, 2016.

The statistical analysis was done using R (R Core Team, 2015a), and the structural equation model is fitted with maximum likelihood estimation routines provided by the R package lavaan (Rosseel, 2012). The lavaan package is also used to calculate a variety of model fit measurements, to access the fit of the model.

3.3 Results

The general characteristics of the participants are shown in Table 3.2. The majority of participants are of female gender, between 18 and 30 years old, and students. Five of the participants reported to already have experience with online treatment.

Variable	Frequency	Percent (%)	Cumulative (%)	Variable	Frequency	Percent (%)	Cumulative (%)
Gender				Education			
Female	80	64.0	64.0	Primary	3	2.4	2.4
Other	45	36.0	100.0	Secondary	36	28.8	31.2
Age				College/University	86	68.8	100.0
18–25	57	45.6	45.6	Persons in household			
26–30	55	44.0	89.6	One	29	23.2	23.2
31–35	13	10.4	100.0	Two	52	41.6	64.8
Occupation				More than two	44	35.2	100.0
Student	98	78.4	78.4	Online treatment			
Working	23	18.4	96.8	Already used	5	4.0	4.0
Other	4	3.2	100.0	Never used	120	96.0	100.0

Table 3.2: Demographic information for the participants.

For the assessment of the model fit to the empirical data, Table 3.3 shows a summary of various model fit measures of the estimated structural model. Based on the calculated measures, the model appears to not fit the empirical data well. Only the RMSEA measurement can be satisfied although 0.1 is considered the highest possible cutoff. A value above this threshold indicates a poor fitting model (MacCallum, Browne, and Sugawara, 1996).

Fit measure	Value	Recommended value	Fit measure	Value	Recommended value
χ^2	1032.850 (P < 0.00)	P > 0.05	NFI	0.646	>0.90
RMR	0.238	<0.05	CFI	0.768	>0.93
RMSEA	0.096	<0.10	TLI	0.745	>0.90

Table 3.3: Goodness-of-fit measures for the model.

The estimations of the connections among the concepts are summarised in Table 3.4. The significance level of each research hypothesis and their influence on behavioural intention or preceding concepts are listed as well. Significant estimates support the initially assumed hypotheses proposed by the literature.

Hypotheses	Endogenous variable	Exogenous variable	Standardised estimate	SE	P value
H8	Perceived ease of use	Task-technology fit	0.310	0.095	0.008**
H9	Perceived usefulness	Task-technology fit	0.600	0.072	<0.001**
H3		Perceived ease of use	0.232	0.060	0.002**
H4		Social influence	0.277	0.076	0.001**
H6	Behavioural intention	Self-efficacy	0.322	0.133	0.020*
H7		Trust	-0.329	0.105	0.004**
H2		Perceived ease of use	-0.207	0.105	0.069
H1		Perceived usefulness	0.002	0.165	0.989
H5		Social influence	0.212	0.144	0.147

Table 3.4: Standardised estimates of the structural model (*P < .05, **P < .01).

The standardised estimates of the latent variables are shown in Table 3.5 as well as the estimated Cronbach alpha values for the questionnaires to measure the concepts. The estimated Cronbach alpha values indicate that the internal consistency of responses to the questionnaires ranges from good to acceptable. The only concept indicated to have questionable consistency is social influence with a Cronbach alpha value of 0.647.

Variable	Latent variable	Standardised estimate	Cronbach alpha	Mean (STD)	Variable	Latent variable	Standardised estimate	Cronbach alpha	Mean (STD)
Task-technology fit	TTF1	0.831	0.884	3.42 (1.39)	Behavioural intention	BI1	0.859	0.835	2.34 (1.51)
	TTF2	0.865		3.45 (1.49)		BI2	0.850		2.66 (1.73)
	TTF3	0.858		4.21 (1.55)	Trust	T1	0.691	0.722	4.78 (1.87)
	TTF4	0.305		2.36 (1.46)		T2	0.185		5.27 (1.58)
	TTF5	0.851		3.21 (1.34)		T3	-0.736		3.28 (1.67)
	TTF6	0.787		4.01 (1.45)		T4	-0.860		3.22 (1.39)
	TTF7	0.613		5.34 (1.28)	Perceived ease of use	EOU1	0.457	0.700	4.16 (1.38)
Perceived usefulness	PU1	0.856	0.884	4.17 (1.53)		EOU2	-0.533		2.65 (1.42)
	PU2	0.635		2.91 (1.35)		EOU3	-0.322		3.64 (1.40)
	PU3	0.911		3.98 (1.55)		EOU4	0.847		4.80 (1.28)
	PU4	0.504		4.17 (1.48)		EOU5	0.697		5.42 (1.20)
	PU5	0.776		4.80 (1.47)	Self-efficacy	SE1	0.814	0.758	4.46 (1.44)
	PU6	0.763		4.03 (1.52)		SE2	0.396		4.62 (1.27)
Social influence	SI1	0.869	0.647	3.21 (1.25)		SE3	0.792		3.63 (1.62)
	SI2	0.751		3.42 (1.42)		SE4	0.545		4.92 (1.41)
	SI3	0.741		3.23 (1.21)					
	SI4	0.267		1.94 (1.29)					
	SI5	-0.137		4.46 (1.74)					

Table 3.5: Standardised estimates of the latent variables, mean and standard deviation from participant's responses.

3.4 Discussion

In this study, theories of technology acceptance are used to build a structural equation model which is then evaluated with empirical data in order to analyse the acceptance of mobile applications for the treatment of mental disorders in the German population. The results suggest that the concept of trust and self-efficacy show a possible direct impact on the acceptance and future use of mobile mental health applications.

Thus, the concept of trust could be of great importance for mobile mental health applications as it may contribute to a person's consideration when facing mental health problems. Leakage or loss of personal data is still a major concern as such online applications will use sensitive data. Many people greatly fear divulging personal information online. The influence of perceived self-efficacy when considering mobile mental health applications is supported by this analysis. A lack of obligation and an absence of expected commitment, might parallel to online education programs lead to low success for mobile treatment applications.

This analysis shows no significant direct influence of perceived usefulness on behavioural intention. Nonetheless, it still arguably has an indirect influence. This is because perceived usefulness reflects the population's knowledge about the new application. Knowledge about these applications, their clinical effectiveness, possibility for treating a wide range of mental disorders such as depression, stress or substance dependence, and availability is still quite low in the targeted population. When young adults are more aware and informed about mobile mental health applications, willingness to use such applications in the future will possibly increase. Furthermore, the results indicate that mobile mental health applications should not be promoted as a replacement for personal therapy and qualified treatment but rather as a supporting and quickly available tool. The participants indicate that mobile applications are perceived as useful and can provide helpful information regarding mental health problems, but disagree that mobile applications are sufficient for treatment of mental disorders. Therefore, task-technology fit can influence the perceived usefulness of these applications and the perceived ease of use.

The hypothesis that social influence has a direct effect on behavioural intention to use mobile treatment applications cannot be supported, although their direct influence was suggested in the original UTAUT model (Venkatesh et al., 2003). This may because mental health treatment is considered more personal than other technology adaption. On the other hand, certain social influences could discourage mobile mental health application uptake.

Yet mobile phone use is ubiquitous, and use of these applications can be kept private even from family or friends, so neither of these social factors should affect use. Unlike behavioural intention, the empirical analysis suggests that social influence does affect the perceived usefulness of these applications. Opinion and experience of friends and family contribute to the perception of technologies. Therefore, openness in discussing mental disorders and their treatment with mobile application or the support of treatment with mobile applications further contributes to the acceptance of mobile treatment and can increase their future use.

Apparently, the usability of mobile applications is not a concern for the targeted users. As initially assumed in this study, the younger German population is experienced with the use of smart-phones and does not consider the use of mobile treatment application a challenge. But the influence of the perceived usefulness on perceived ease of use, as suggested by the TAM model (Davis, 1993), is supported by this analysis. Therefore, the expectation that mobile applications are easy and intuitive to use might indirectly influence the perception of mobile mental health treatment.

3.4.1 Limitations

A participant selection bias is possible since all participants are Internet users and, thus more likely to be experienced with technology and think favourably about it. Also the research demographic age between 18 and 35 years introduces an additional bias. Therefore, a generalisation of the findings to the entire German population is not feasible. Second, cultural differences are not considered in this study, but this still permits an analysis of acceptance in the German population overall. Still, cultural differences might influence the choice of mobile mental health applications so these results may not translate to other nations or cultures. Finally, the sample size is low for the estimation of this model. A higher sample size should permit more accurate estimation of the influence of the individual concepts that have been indicated to be relevant by this study.

3.5 Conclusion

The main contribution of this paper is the proposed model for user acceptance of mobile mental health applications and its analysis. The model aims to describe the acceptance and adoption of mobile mental health treatment by the population. For the evaluation of the model, data was collected in an online survey targeting young adults in Germany. The model estimated from the data suggests that trust, social influence, and task-technology fit may influence people's adoption of mobile mental health applications. However, young adults in Germany are little aware that mobile mental health treatment already exists and that these applications are effective for a wide range of diseases. People are possibly worried that mobile treatment is a cheap and inadequate replacement for conventional psychotherapy and human interaction. But this is not the case. Mobile applications are a support tool that can provide a little bit of help at any time. They could even become a valuable tool in regular face-to-face therapy. Mobile mental health applications are likely to become widely accepted and used in the future when the population is better informed about their possibilities and security of personal data can be ensured.

This paper also makes a contribution in integrating additional concepts such as self-efficacy and task-technology fit into TAM. The data supports the hypothesis that task-technology fit can influence perceived usefulness as well as perceived ease of use. Self-efficacy might also influence behavioural intention in the case of mobile mental health treatment. In future, the proposed structural model for acceptance of mobile mental health treatment scenarios can be further refined to determine influential factors in subsequent studies or to re-evaluate acceptance after a promotion campaign.

References

Andersson, Gerhard and Pim Cuijpers (2009). "Internet-based and other computerized psychological treatments for adult depression: a meta-analysis." In: *Cognitive behaviour therapy* 38.4, pp. 196–205. ISSN: 1650-6073. DOI: 10.1080/16506070903318960 (cit. on p. 64).

Bagozzi, Richard P (2002). "An Attitudinal Model of Technology - Based Self-Service: Moderating Effects of Consumer Traits and Situational Factors". In: *Journal of the Academy of Marketing Science* 30.3, pp. 184–201. ISSN: 0092-0703. DOI: 10.1177/0092070302303001 (cit. on pp. 67, 74).

Bertrand, Manon and Stéphane Bouchard (2008). "Applying the technology acceptance model to vr with people who are favorable to its use". In: *Journal of Cyber Therapy and Rehabilitation* 1, pp. 200–207. ISSN: 17849934 (cit. on p. 66).

Board, The Executive (2012). "Global burden of mental disorders and the need for a comprehensive , coordinated response from health and social sectors at the country level". In: *World Health* January, pp. 6–9 (cit. on pp. 5, 64).

Chau, Patrick Y. K. and Paul J. Hu (2002). "Examining a Model of Information Technology Acceptance by Individual Professionals: An Exploratory Study". In: *J. Manage. Inf. Syst.* 18.4, pp. 191–229. ISSN: 0742-1222 (cit. on p. 74).

Cheong, Je Ho and Myeong-Cheol Park (2005). "Mobile internet acceptance in Korea". In: *Internet Research* 15.2, pp. 125–140. ISSN: 1066-2243. DOI: 10.1108/10662240510590324 (cit. on p. 74).

Chih, Ming Yuan et al. (2014). "Predictive modeling of addiction lapses in a mobile health application". In: *Journal of Substance Abuse Treatment* 46.1, pp. 29–35. ISSN: 07405472. DOI: 10.1016/j.jsat.2013.08.004 (cit. on pp. 50, 54, 65, 105).

Cuijpers, Pim et al. (2009). "Computer-aided psychotherapy for anxiety disorders: a meta-analytic review." In: *Cognitive behaviour therapy* 38.2, pp. 66–82. ISSN: 1650-6073. DOI: 10.1080/16506070802694776 (cit. on p. 64).

Davis, F D (1985). "A technology acceptance model for empirically testing new end-user information systems: Theory and results". In: *Management* Ph.D. P. 291. ISSN: 0025-1909. DOI: oclc/56932490 (cit. on pp. 5, 65).

Davis, Fd, Rp Bagozzi, and Pr Warshaw (1989). *User acceptance of computer technology: a comparison of two theoretical models*. DOI: 10.1287/mnsc.35.8.982. arXiv: /www.jstor.org/stable/2632151 [http:] (cit. on pp. 67, 74).

Davis, Fred D Fd (1993). *User acceptance of information technology: system characteristics, user perceptions and behavioral impacts*. DOI: 10.1006/imms.1993.1022 (cit. on pp. 67, 70).

Dishaw, Mt and Dm Strong (1999). "Extending the Technology Acceptance Model with Task-Technology Fit Constructs". In: *Information & Management* 36, pp. 9–21. ISSN: 03787206. DOI: 10.1016/S0378-7206(98)00101-3 (cit. on pp. 66, 67).

Goodhue, Dale L, Ronald L Thompson, and By Dale L Goodhue (2014). "Task-Technology Fit and Individual Performance". In: *MIS Quarterly* 19, pp. 213–236 (cit. on pp. 66, 67).

Gu, Linwu and Jianfeng Wang (2009). "A study of exploring the "Big Five" and task technology fit in web-based decision support systems". In: *Issues in Information System* 10.2, pp. 210–217 (cit. on p. 74).

Gun, Shih Ying, Nickolai Titov, and Gavin Andrews (2011). "Acceptability of Internet treatment of anxiety and depression." In: *Australasian psychiatry : bulletin of Royal Australian*

and New Zealand College of Psychiatrists 19, pp. 259–264. ISSN: 1440-1665. DOI: 10.3109/10398562.2011.562295 (cit. on p. 65).

Jarupathirun, Suprasith and Fatemeh Mariam Zahedi (2007). "Exploring the influence of perceptual factors in the success of web-based spatial DSS". In: *Decision Support Systems* 43.3, pp. 933–951. ISSN: 01679236. DOI: 10.1016/j.dss.2005.05.024 (cit. on pp. 67, 74).

Kaasinen, Eija (2005). "User acceptance of mobile services - Value, ease of use, trust and ease of adoption". In: *VTT Publications*. ISSN: 12350621 (cit. on pp. 66, 67).

Klopping, Inge M and Earl Mckinney (2004). "Extending the Technology Acceptance Model and the Task-Technology Fit Model T". In: *Information Technology, Learning, and Performance Journal* 22, pp. 35–48 (cit. on pp. 66, 67).

Lanseng, Even J. and Tor W. Andreassen (2007). "Electronic healthcare: a study of people's readiness and attitude toward performing self-diagnosis". In: *International Journal of Service Industry Management* 18, pp. 394–417. ISSN: 0956-4233. DOI: 10.1108/09564230710778155 (cit. on p. 65).

MacCallum, Robert C., Michael W. Browne, and Hazuki M. Sugawara (1996). "Power analysis and determination of sample size for covariance structure modeling." In: *Psychological Methods* 1.2, pp. 130–149. ISSN: 1082-989X. DOI: 10.1037//1082-989X.1.2.130 (cit. on p. 68).

Malhotra, Naresh K, Sung S Kim, and James Agarwal (2004). *Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model*. DOI: 10.1287/isre.1040.0032 (cit. on pp. 66, 67).

Martignoni, Robert et al. (2008). "Evaluation of Future Mobile Services based on the Technology Acceptance Model". In: *16th European Conference on Information System* 3747.3747 (cit. on pp. 66, 67).

Monthuy-Blanc, Johana et al. (2013). "Factors influencing mental health providers' intention to use telepsychotherapy in First Nations communities." In: *Transcultural psychiatry* 50.2, pp. 323–43. ISSN: 1461-7471. DOI: 10.1177/1363461513487665 (cit. on p. 66).

Musiat, Peter, Philip Goldstone, and Nicholas Tarrier (2014). "Understanding the acceptability of e-mental health - attitudes and expectations towards computerised self-help treatments for mental health problems". In: *BMC Psychiatry* 14.1, p. 109. ISSN: 1471244X. DOI: 10.1186/1471-244X-14-109 (cit. on pp. 1, 5, 65).

Nysveen, H. (2005). "Intentions to Use Mobile Services: Antecedents and Cross-Service Comparisons". In: *Journal of the Academy of Marketing Science* 33.3, pp. 330–346. ISSN: 0092-0703. DOI: 10.1177/0092070305276149 (cit. on pp. 67, 74).

Olatunji, Bunmi O., Josh M. Cisler, and Brett J. Deacon (2010). "Efficacy of cognitive behavioral therapy for anxiety disorders: A review of meta-analytic findings". In: *Psychiatric Clinics of North America* 33, pp. 557–577. ISSN: 0193953X. DOI: 10.1016/j.psc.2010.04.002 (cit. on p. 64).

Park, Sung Youl (2009). "An Analysis of the Technology Accept ance Model in Understanding University Students' Behavioral Intention to Use e-Learning". In: *Educational Technology & Society* 12, pp. 150–162. ISSN: 14364522. DOI: 10.1007/s00340-009-3513-0 (cit. on pp. 67, 74).

Pedersen, Per E . (2001). *Adoption of mobile commerce : An exploratory analysis*. 51, p. 90. ISBN: 8249101758 (cit. on pp. 67, 74).

Pedersen, Per E. and Herbjørn Nysveen (2003). "Usefulness and self-expressiveness: extending TAM to explain the adoption of a mobile parking service". In: *16th Electronic Commerce*, pp. 705–717 (cit. on pp. 66, 67).

R Core Team (2015a). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria (cit. on p. 67).

Robinson, Emma et al. (2010). "Internet treatment for generlized anxiety disorder: A randomized controlled trial comparing clinician vs. technician assistance". In: *PLoS ONE* 5.6. ISSN: 19326203. DOI: 10.1371/journal.pone.0010942 (cit. on pp. 42, 64).

Rosseel, Yves (2012). "{lavaan}: An {R} Package for Structural Equation Modeling". In: *Journal of Statistical Software* 48.2, pp. 1–36 (cit. on p. 67).

Sun, Yongqiang et al. (2013). "Understanding the Acceptance of Mobile Health Services: a Comparison and Integration of Alternative Models". In: *Journal of Electronic Commerce Research* 14.2, pp. 183–200. ISSN: 19389027 (ISSN) (cit. on pp. 66, 67).

Torous, John et al. (2015). "Utilizing a Personal Smartphone Custom App to Assess the Patient Health Questionnaire-9 (PHQ-9) Depressive Symptoms in Patients With Major Depressive Disorder". In: *JMIR Mental Health* 2.1, e8. DOI: 10.2196/mental.3889 (cit. on pp. 52, 65, 90).

Tung, Feng-Cheng and Su-Chao Chang (2008). "Nursing students' behavioral intention to use online courses: a questionnaire survey." In: *International journal of nursing studies* 45.9, pp. 1299–309. ISSN: 0020-7489. DOI: 10.1016/j.ijnurstu.2007.09.011 (cit. on pp. 66, 67).

Venkatesh et al. (2003). "User Acceptance of Information Technology: Toward a Unified View". In: *MIS Quarterly* 27.3, p. 425. ISSN: 02767783. DOI: 10.2307/30036540. arXiv: arXiv:1011.1669v3 (cit. on pp. 65, 67, 69, 74).

Venkatesh, Viswanath (2000). "Determinants of Perceived Ease of Use: Integrating Control, Intrinsic Motivation, and Emotion into the Technology Acceptance Model". In: *Information Systems Research* 11.4, pp. 342–365. DOI: 10.1287/isre.11.4.342.11872 (cit. on p. 65).

Venkatesh, Viswanath and Fred D. Davis (2000b). "A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies". In: *Manage. Sci.* 46.2, pp. 186–204. ISSN: 0025-1909. DOI: 10.1287/mnsc.46.2.186.11926 (cit. on p. 65).

Venkatesh, Viswanath, James Y. L. Thong, and Xin Xu (2012). "Consumer Acceptance and Use of Information Technology : Extending the Unified Theory". In: *MIS Quarterly* 36.1, pp. 157–178 (cit. on p. 65).

Vernmark, Kristofer et al. (2010). "Internet administered guided self-help versus individualized e-mail therapy: A randomized trial of two versions of {CBT} for major depression". In: *Behaviour Research and Therapy* 48.5, pp. 368–376. ISSN: 0005-7967. DOI: http://dx.doi.org/10.1016/j.brat.2010.01.005 (cit. on p. 64).

White, L a, M a Krousel-Wood, and F Mather (2001). "Technology meets healthcare: distance learning and telehealth." In: *The Ochsner journal* 3.January 1997, pp. 22–29. ISSN: 1524-5012 (cit. on p. 64).

Wixom, Barbara H. and Peter a. Todd (2005). "A theoretical integration of user satisfaction and technology acceptance". In: *Information Systems Research* 16.1, pp. 85–102. ISSN: 10477047. DOI: 10.1287/isre.1050.0042 (cit. on pp. 67, 74).

Wootton, Bethany M et al. (2011). "The acceptability of internet-based treatment and characteristics of an adult sample with obsessive compulsive disorder: An internet survey". In: *PLoS ONE* 6.6, pp. 1–6. ISSN: 19326203. DOI: 10.1371/journal.pone.0020548 (cit. on pp. 64, 65).

Appendix A. Survey questionnaire

Construct	Abbreviation	Measurement items	Adopted from
Task-	TTF1	Mobile mental health applications are adequate for the described scenario.	(Jarupathirun and Zahedi, 2007)
technology	TTF2	Mobile mental health applications are compatible with the task of treating mentally ill clients.	
fit	TTF3	Mobile mental health applications are helpful.	
	TTF4	Mobile mental health applications are sufficient.	
	TTF5	Mobile mental health applications fit the task well.	
	TTF6	Mobile mental health applications are useful for treating people.	
	TTF7	Mobile mental health applications are useful to provide information to people.	
Social	SI1	Your friends and family think that mobile mental health applications are a useful thing.	(Nysveen, 2005)
influence	SI2	Your friends and family think that mobile mental health applications would be useful for you.	
	SI3	Your friends and family would also use mobile mental health applications.	
	SI4	Do you often discuss the advantages of mobile treatment with your friends/family.	
	SI5	Would your friends and family be surprised if you use a mobile mental health application.	
Ease of	EOU1	I find it easy to get the benefits from a mobile mental health application.	(Bagozzi, 2002; Davis, Bagozzi, and Warshaw, 1989)
use	EOU2	Using an mobile mental health application will be complicated.	
	EOU3	Using an mobile mental health application will take a lot of effort	
	EOU4	I find mobile mental health applications are easy to use	
	EOU5	Learning to operate a mobile mental health application would be / is ease for me.	
Perceived	PU1	I find mobile mental health to be useful to improve my life in general.	(Davis, Bagozzi, and Warshaw, 1989)
usefulness	PU2	Using a mobile mental health application would improve my life quickly.	
	PU3	I would find mobile mental health applications useful.	
	PU4	Using a mobile mental health application would make me save time.	(Pedersen, 2001; Nysveen, 2005)
	PU5	I think that mobile mental health applications provide very useful services.	
	PU6	Mobile mental health applications are an improvement to the services it supersedes.	(Wixom and Todd, 2005)
Trust	T1	I feel apprehensive about using a mobile mental health application.	(Venkatesh et al., 2003)
	T2	Using mobile mental health applications would not divulge my personal information.	(Gu and Wang, 2009)
	T3	Using mobile mental health applications is entirely within my control.	(Chau and Hu, 2002)
	T4	I think that mobile mental health applications are secure to use.	(Cheong and Park, 2005)
Self-	SE1	I feel confident finding information and advice in a mobile mental health application.	(Park, 2009)
efficacy	SE2	I have the necessary skills for using an mobile mental health application successfully.	
	SE3	I feel confident using the mobile mental health application regularly.	
	SE4	I feel confident to work through all interventions that the application provides me.	
Intention	BI1	I intend to use a mobile mental health application.	(Park, 2009)
to use	BI2	I intend to check the availability of a suited mobile mental health application.	

TABLE 3.6: Survey questionnaire.

Part II

Predictive models in online treatment

The next part encompasses analyses of various types of data originating from different online treatments. For the here presented analyses, a variety of predictive models are utilized to infer and predict client individual mood levels. Specifically, the inference of mood levels using mobile-phone measures and online diaries are presented, as well as future mood prediction using EMA. Furthermore, the relationship between EMA measures and clinical depression scores is examined. The last article in this part discusses the prediction of the expected treatment outcome and cost on a client individual level.

Chapter 4

How to predict mood? Delving into features of smartphone-based data

Becker, D., Bremer, V., Funk, B., Asselbergs, J., Riper, H., and Ruwaard, J. (2016). In Proceedings of the 22nd Americas Conference on Information Systems, AMCIS 2016.

Abstract: *Smartphones are increasingly utilized in society and enable scientists to record a wide range of behavioral and environmental information. These information, referred to as Unobtrusive Ecological Momentary Assessment Data, might support prediction procedures regarding the mood level of users and simultaneously contribute to an enhancement of therapy strategies. In this paper, we analyze how the mood level of healthy clients is affected by unobtrusive measures and how this kind of data contributes to the prediction performance of various statistical models (Bayesian methods, Lasso procedures, etc.). We conduct analyses on a non-user and a user level. We then compare the models by utilizing introduced performance measures. Our findings indicate that the prediction performance increases when considering individual users. However, the implemented models only perform slightly better than the introduced mean model. Indicated by feature selection methods, we assume that more meaningful variables regarding the outcome can potentially increase prediction performance.*

4.1 Introduction

A good state of mental health is an important factor for every individual as it promotes a healthy social environment, provides general motivation in achieving life goals, and can even benefit the economy. Using today's modern resources, individuals are being more proactive about their mental health. The Internet is increasingly being used to search for medical information and thus individuals are becoming better informed about their health (Spil and Schuring, 2006) as the demand for medical knowledge in the population is steadily growing (Andreassen et al., 2007). Seeking "new" ways to treat depression and other mental diseases, healthcare treatments supported by electronic processes such as e-mental-health have been established. Internet-based treatments have the potential to provide high quality treatment by sustaining deeper insight into the daily lives of the participants and thus enhancing the overall therapy success (Eysenbach, 2001b). These treatments additionally allow for the creation of relatively new kinds of data.

Ecological Momentary Assessments (EMA) are data collection methods that provide researchers with data in regard to symptoms, behavior, and cognition close in time to the participants' experience in their natural environment (Iida et al., 2012; Stone and Shiffman, 1994). Due to technological development, instead of the traditional pen and paper diary or submission through a bulky at-home desktop, EMA data can now take on the form of accessible and convenient electronic devices that can lead to a greater ecological validity

(Iida et al., 2012). Smartphones, in particular, provide advanced computing and storage capabilities that in turn lead to the ability to record a wide range of behavioral and environmental information (Gaggioli et al., 2013; Gimpel, Regal, and Schmidt, 2015), can possibly be a beneficial factor for gaining deeper insight into clients' behavior and simultaneously provide guidance for an improvement of future therapy strategies. Thus, smartphones are increasingly utilized as sensors for physical, social, and other activities humans are occupied with (Asselbergs et al., 2016).

The Smartphone sensing process, also referred to as unobtrusive EMA, silently accumulates the clients' data without prompting the user for additional information. Various unobtrusive measures can therefore be collected and utilized when predicting or inferring certain psychological concepts such as academic performance (Wang et al., 2014), sleep duration (Chen et al., 2013), and depression (Saeb et al., 2015a). Therefore, this data can be utilized as additional information besides subjectively reported measures by the clients (Gaggioli et al., 2013) or even as predictor to forecast traditional measurements. This collection method has certain advantages. It obviously reduces usability issues tremendously since no interaction between the client and software is required. Furthermore, almost everyone can potentially be monitored since nowadays a smartphone is a fundamental device which is wide spread in the society and part of every day life (Chen et al., 2013; Abdelzaher et al., 2007).

Previous studies have already investigated the usage of smartphone-based data. Burns et al. (2011), for example, develop a mobile application that assists depressive clients in difficult situations with Ecological Momentary Interventions. This application infers the users' mental state using measures of location, activity, social environment, and smartphone usage. Based on the measurements, this model then determines if a supportive intervention is triggered. Furthermore, Saeb et al. (2015a) demonstrate that daily movement patterns, estimated from a phones' GPS record, help in determining depressive symptoms. Features such as the variance in visited locations, frequency of visited locations, and time spend there are being generated from the GPS data. These movement patterns have then been shown to correlate with depressive symptoms. Moreover, Ma et al. (2012b) develop a program called MoodMiner. This is a smartphone application that infers the owner's mood. LiKamWa et al. (2013) also create a program for achieving a similar goal (MoodScope). However, even though many studies exist in this field and LiKamWa et al. (2013) find that smartphone usage data similar to ours correlates well with the users' mood level, it is not assured that unobtrusive measures can generally contribute significantly towards predictions of any kind. Asselbergs et al. (2016), for example, demonstrate that unobtrusive measures may not contribute as much as the study of LiKamWa et al. (2013) suggests. Thus, in this paper we seek to analyze this aspect even further.

We make an attempt to predict the mood level of healthy Dutch participants by using the aforementioned mobile phone usage data and simultaneously compare different statistical methods. Furthermore, we seek to reveal the importance of various unobtrusive measures and specify their contribution and influence on the prediction performance. The analysis of unobtrusive mobile phone measurements and their contribution to prediction performance is yet to be analyzed more intensively in research. Additionally, the fact that we utilize data of healthy participants might be an interesting factor because the gained insight into healthy individuals' behavior might provide guidance for improving actions of unhealthy clients. In the following chapters, we introduce the data used for our approach, illustrate the utilized models, conclude our results, and briefly outline future procedures.

4.2 Data & Methods

4.2.1 The Data

We utilize smartphone-based obtrusive and unobtrusive data from an explorative uncontrolled pilot study for our analyses (Asselbergs et al., 2016). The dataset consists of 27 healthy Dutch students who reported their mood level on their smartphones for six weeks at a frequency of 5 times per day and simultaneously provided unobtrusive measures by the usage of a smartphone application called iYouVU that silently collects data in the background (Asselbergs et al., 2016). Usually, the clients provided traditional mood measures (one-dimensional mood measure– 10-point scale; two-dimensional measures valence and arousal– -2 to 2) at specific times: 9am, 12pm, 3pm, 6pm, and 9pm; however, the clients were allowed to push back the measurement requests which results in differing measurement times amongst the participants. Asselbergs et al. (2016) already utilized and aggregated the data on a daily basis in an explorative study. However, we only aggregate the unobtrusive data up to the point of each mood request. This procedure results in 1335 observations of the 27 patients. Because no treatment was provided for the participants, the data reflect the natural course of mood over time. Appendix A illustrates the attributes of the dataset in more detail.

4.2.2 The Approach

For analyzing the data, we conduct several analyses including different statistical methods. We first analyze the data on a non-user level. Afterwards, we conduct analyses on a user level in order to possibly reveal advantages of the hierarchical structure. In the following accompanying information, we illustrate the different approaches:

First, we perform analyses in the instances in which every data point is being considered as independent from the other touchpoints - individuals are not taken into account. We call this analysis the non-user level analysis. For this method, we split the dataset into a training (75%) and a test dataset (25%). As mentioned before, this splitting process does not consider any user specific touchpoints. Then, based on the specific model utilized and the used independent variables, we make an attempt to predict the mood level of the test dataset. Additionally, a feature ranking is used to reveal the contribution of each phone measure to the mood prediction. All phone measures are then sorted and listed according to their importance.

Second, we analyze the data on a personalized level. Particularly, we consider individuals in this case. We also introduce the weekday and time of the request as additional variables, which explains why the first week of the data points is required for model training. We then start the estimation of the daily mood based on the data points of the previous week. After the first mood prediction, the next data point is added to the training set, the model is retrained, and the next day is predicted based on all previous data. Specifically, when estimating the mood level on the 8^{th} day, data from day one to seven are used for training. The records of the variables on the 8^{th} day are then used to predict the mood level on the 8^{th} day based on the trained model. We call this method the user level analysis because again, we do consider individual clients. In the following sub-chapters, we introduce the utilized models for the prediction of the participant's mood level, for the creation of feature importance rankings, and subsequently the used performance measures for model comparison.

4.2.3 The Mean Model

As method for comparison, we use a mean model in our analyses. Specifically, we utilize the outcome mean of the observations for predictions. For the non-user level analysis,

we use the mean value over all mood values of the training set and utilize this mean for predictions. However, since we consider individuals in the user level analysis, we calculate the mean based on every participants' mood observations upon the current measurement point. Particularly, when estimating the mood level on the 8^{th} day, we utilize the mean of day one to seven for this individual as the mean model prediction.

4.2.4 Linear Regression

We use linear models and generalized linear models to predict the mood level of the participants. We also utilize the glmnet-package (Friedman, Hastie, and Tibshirani, 2009) in R (R Core Development Team, 2014) for feature selection purpose. The relationships between the outcome variable (mood level in this case) and the independent variables (all unobtrusive measures) are represented by the coefficients. The intercept illustrates the prediction the model creates if all independent variables were zero. The other coefficients (one for each attribute) represent the change in the predicted value of the outcome per unit of change in the corresponding independent variable (i.e. app usage) while all other x variables are fixed (Rencher and Christensen, 2012, chap. 10).

4.2.5 Support Vector Machine

We also utilize support vector machines (SVM). Support vector machines exhibit great classification performance and have been used in various fields (Burges, 1998). They are often applied to supervised learning problems (Vapnik, 1998). For predicting the participants' mood, we use the so called ε-Support Vector regression (Vapnik, 2000). Support vector regression seeks to find a linear function that does not penalize mood values that are smaller than a specific ε value. By not penalizing the training samples within this ε bound, it is ensured that most samples lie within this bound. Simultaneously, the estimated weights are supposed to be as small as possible to produce a flat solution (Smola and Schölkopf, 2004). During training, a balance between the flatness and a small ε parameter is found in order to produce a solution that sufficiently fits both requirements.

4.2.6 Lasso Regression

Furthermore, we use LASSO regression in our project (Least Absolute Shrinkage and Selection Operator). LASSO regression is a linear regression algorithm including an additional linear penalty term. The cost function of regression, which is to be optimized, consists of the mean square error of the misclassified samples. By minimizing the classification error, over-fitting can possibly occur. In this case, the training error steadily decreases, essentially improving predictions. However, the error on a new test set increases because the algorithm generalizes poorly. An additional penalty term is introduced to the cost function to prevent over-fitting. Specifically, the Lasso regression penalizes the absolute value of the regression coefficients (Tibshirani, 1996). This linear penalty is the sum of the absolute values over all weights and enforces useless coefficients to shrink towards zero in order to produce a sparse solution. The optimization problem that arises is shown in Equation 4.1:

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \left(||\boldsymbol{Y} - \boldsymbol{x}\boldsymbol{\beta}||^2 + \lambda ||\boldsymbol{\beta}||_1 \right) \tag{4.1}$$

The λ term 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 equal zero).

4.2.7 Bayesian Hierarchical Linear Regression

To consider the hierarchical structure of the data and include the phone measures of all clients in every estimation, we develop a Bayesian hierarchical linear regression model. In a clinical setting, it is expected that people would not be willing to constantly contribute mood measures for training purposes. In this case and similar to the study of LiKamWa et al. (2013), our model can consider data of additional clients in order to improve predictions for another client.

The implemented model is similar to the model used by MacKay (1996), Neal (1996), and Tipping (2000). However, we additionally implement a hierarchical prior on the individual client weights. Since the model is trained for each client and each day in the user level analysis, we implement a variational approximation for speed benefits (Bishop, 2006). This approximation is possible because only conjugate priors are used. Figure 4.1 illustrates the plate notation of the model:

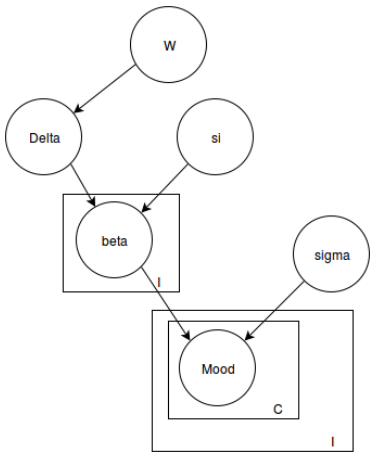


FIGURE 4.1: Plate Notation

Every user has its own set of beta coefficients. These coefficients are dependent on the hierarchical uninformative prior Delta that is sampled with variance w. These individual gamma priors w are used to estimate the influence of each weight and determine their importance. Specifically, w stands for the variance of each individual weight and simultaneously allows for the creation of a feature importance ranking. Additionally, one gamma prior si is used for the individual weights for all users. This prior defines how strongly the coefficients are allowed to differ from the hierarchical prior Delta. Moreover, sigma (the variance for mood) represents the additional noise in the mood level.

4.2.8 Performance Measures

We use the Root Mean Square Error (RMSE) as comparison and performance measure. The RMSE can be defined as the average distance between predicted and observed values. Based on the RMSE, a confidence interval can be derived. The interval limits for the 95% confidence interval lie approximately on +/-2·RMSE. Thus, the actual value (the value that is attempted of being predicted) lies within this confidence interval with a probability of 95%. In addition to the RMSE, we utilize a specific performance measure that is illustrated by Equation 4.2:

$$\text{Performance} = \frac{1}{N}\sum_{i=1}^{N} 1(\text{residual}_i^2 < 0.25) \tag{4.2}$$

This performance measure illustrates the percentage of correctly classified predictions. It classifies forecasts as correct when the prediction lies within a certain boundary around the true value, in this case .5 (LiKamWa et al., 2013). In the next chapter, we present the results from our analyses. We subdivide the chapters into our non-user and user level analyses.

4.3 Results

4.3.1 Analysis – Non-User Level

Table 1 illustrates the results of the non-user level analysis. We execute the methods already introduced in the previous chapter. Furthermore, we illustrate the feature importance ranking of the lasso procedures in Appendix B.

Methods	RMSE	Performance
Mean Model	1.11	0.38
SVM	0.87	0.41
Regression	0.83	0.40
Lasso	0.86	0.39
Lasso (IE)	0.84	0.38

Table 4.1: Results Non-User Level

In our non-user level analyses, we fit linear models by using the glm function and apply support vector machine procedures by utilizing the e1071-package (Meyer et al., 2015) in R (R Core Development Team, 2014). In a first attempt, we use all variables as predictors (SVM & Regression). Afterwards, we perform an analysis that executes lasso procedures in a linear regression and simultaneously selects features by using the glmnet-package (Friedman, Hastie, and Tibshirani, 2009). Additionally, we multiply all columns with one another and center the results to consider potential interaction effects between the different concepts. For feature selection in this case, we repeatedly take advantage of the glmnet-package (Friedman, Hastie, and Tibshirani, 2009).

The RMSE values of our analysis are between .83 and 1.11. The mean value prediction model achieves a RMSE of 1.11 and a performance measure of .38, as shown in Table 4.1. The predictions of this model are solely based on the mean mood of the training data. The RMSE and performance measures of the other models, which use more features besides the mood measure, perform slightly better. Their RMSE error is smaller than the RSME of the mean model. Consequently, the additional measurements might contain information about the current mood level. Therefore, we do correctly classify more values and perform slightly better by utilizing various models compared to the mean model. However, the results of the interaction effect analysis indicate that only the psychological concepts valence and arousal influence the mood significantly. Almost all other attributes either only slightly contribute or completely fail to contribute to the prediction at all. The importance ranking in Appendix B supports this result even further. To conclude, although the linear regression model that is inclusive of all variables demonstrates the best performance, we do not correctly predict significantly more values than the mean model- irrespective of which model is used. As these results are insufficient, we also examine the provided dataset on a user level in an attempt to potentially further reduce the RMSE by considering individuals and the hierarchical structure.

4.3.2 Analysis – User Level

For this analysis, we add further features to the dataset. Specifically, we add the weekday (extracted from the time stamp) and a feature that indicates the n-th measure of this particular day (measureOfDay). The results are illustrated in Table 4.2:

Methods	RMSE	Performance
Mean Model	0.90	0.49
SVM	0.85	0.51
Regression	0.61	0.57
Lasso	0.53	0.56
Bayesian Hierarchical	0.58	0.57

TABLE 4.2: Results User Level

In the user level analysis, the RMSE differs between .53 and .90. Therefore, Table 4.2 indicates that considering individual participants can in fact improve the results of analyses. The applied models perform slightly better than the mean model. As indicated in Table 4.2, the improvement is also higher compared to the non-user level analysis. The hierarchical model results in a smaller RSME than the regression and SVM model. Nevertheless, the lasso regression results in an even lower RSME. This is unexpected because the hierarchical model uses the data of all users and lasso regression utilizes only the current users' data. However, when inspecting both algorithms, it is noticeable that the lasso procedures set the weights to 0 based on the lambda value. This procedure is not to be found in the hierarchical model. Even tough the weights are all individually penalized to enforce sparseness and provide a feature ranking, they never truly become zero. Therefore, the utilization of all the non-contributing variables might be responsible for the slightly higher RSME. Even though the models perform better than the models in the non-user level analysis and even slightly better than the mean model, we do not deem these results sufficient. Therefore, we seek to gain more insight as to which features can potentially be useful for mood prediction. As a result, we turn to a feature ranking using all data estimated from the lasso regression and the hierarchical model.

	Lasso Analysis		Bayesian Analysis	
Rank	Feature Name	Lambda	Feature Name	Prior
1	valence	.657	valence	1.06
2	arousal	.085	activity	98.14
3	activity	.058	arousal	116.54
4	measureOfDay	.049	weekday7	132.52
5	appCat.entertainment	.040	weekday2	141.29
6	appCat.weather	.037	weekday3	144.18
7	weekday7	.033	weekday5	146.86
8	call	.019	weekday4	152.81
9	appCat.builtin	.016	measureOfDay	155.61
10	weekday3	.013	weekday6	159.13
11	sms	.011	sms	160.27
12	appCat.social	.010	appCat.weather	160.52
13	appCat.office	.009	appCat.finance	161.16
14	appCat.finance	.008	appCat.travel	161.17
15	appCat.other	.008	call	161.35
16	appCat.unknown	.007	appCat.office	161.35
17	weekday5	.006	appCat.game	161.36
18	weekday2	.004	appCat.utilities	161.36
19	appCat.utilities	.003	appCat.other	161.36
20	screen	.003	appCat.social	161.36
21	appCat.game	.002	appCat.enterntainment	161.37
22	weekday4	.002	appCat.unknown	161.37
23	weekday6	.002	screen	161.37
24	appCat.travel	.002	appCat.builtin	161.37
25	appCat.communication	.002	appCat.communication	161.37

TABLE 4.3: Feature ranking

As indicated by both feature rankings in Table 4.3, the coefficient of valence, arousal, and activity appear useful when inferring the mood from mobile phone measurements. The influence of valence and arousal are not surprising because these measurements correlate with mood and are taken at the same time the mood level is reported. Both rankings also indicate that weekday and measureOfDay might influence the current mood. To analyze this phenomenon further, we implement a hierarchical Bayesian regression model in JAGS (Appendix C). Hence, we are enabled to reveal the significant and influencing factors regarding the mood level. Table 4.4 illustrates the results of this analysis:

Attributes	Mean	2.5% & 97.5%
arousal	0.09	(0.04; 0.14)
valence	1.00	(0.91;1.10)
activity	0.08	(0.04; 0.13)
Monday	-0.06	(-0.13;-0.001)
Saturday	0.07	(0.002; 0.13)
measureOfDay4	0.09	(0.02; 0.15)
measureOfDay5	0.14	(0.08; 0.20)

TABLE 4.4: Significant features for predicting mood - Bayesian hierarchical regression

According to Table 4.4, the activity level and Saturdays have a significant positive effect on the mood level. Furthermore, Mondays seems to affect the mood in a slightly negative manner. This can possibly be due to the fact that individuals often have a hard time returning to work after a weekend (Helliwell and Wang, 2015). Our results also show that the mood level of the clients steadily increases in time. Specifically, the fourth measure of mood (measureOfDay4) is taken after work and the fifth measure (measureOfDay5) is taken around 9pm when participants are presumably at home or engaged in leisure time activities. Since an individuals' mood is increasing after work or on days off, the influences of time and weekday on mood can potentially be referred to work activities.

4.4 Conclusion

We perform a multitude of different statistical methods on a dataset that includes UEMA and EMA observations of healthy Dutch participants. We try to reveal important variables and seek to predict the future mood level of the clients. In conclusion, we find that individuals generally have an increased mood level on weekends and during leisure time. Additionally, the weekday Monday influences healthy individuals negatively. However, since the utilized dataset consists of only students, we might not be able to generalize the findings and assume that an older population would be influenced similarly. We further illustrate that analyzing this dataset on a user level, which means the consideration of the hierarchical structure and therefore individuals, can improve analyses. However, the implemented models only perform slightly better than the introduced mean model. We are still not satisfied with these results. Although the implementation of the more complex Bayesian hierarchical approximation model and the LASSO procedures enhance the prediction performance in the user level analysis, only a slight improvement is achieved. The feature ranking analyses further raises the assumption that the variables do not tremendously contribute to the predictions. This circumstance would support the inability of creating better prediction performance. Since the UEMA method is considered a young field (Asselbergs et al., 2016), this kind of data might still be a powerful tool for future findings. In our case, with the provided dataset and the implemented models, we are not able to reach a significantly better prediction performance. In the future, we aim to build models that are able to predict the mood level of participants more accurately and find ways to select important features more reliably. Implementing more individualized hierarchical models that account for heterogeneity amongst the participants might be one option to increase the prediction performance of the mood level. Moreover, we assume that obtaining more meaningful features that might contribute to the forecast more intensively can boost prediction performance as well.

References

Abdelzaher, Tarek et al. (2007). "Mobiscopes for human spaces". In: *IEEE Pervasive Computing* 6.2, pp. 20–29. ISSN: 15361268. DOI: 10.1109/MPRV.2007.38 (cit. on p. 77).

Andreassen, Hege K et al. (2007). "European citizens' use of E-health services: a study of seven countries." In: *BMC public health* 7.53. ISSN: 14712458. DOI: 10.1186/1471-2458-7-53 (cit. on p. 76).

Asselbergs, Joost et al. (2016). "Mobile Phone-Based Unobtrusive Ecological Momentary Assessment of Day-to-Day Mood: An Explorative Study". In: *Journal of medical Internet research* 18.3, e72. ISSN: 14388871. DOI: 10.2196/jmir.5505 (cit. on pp. 49, 52, 77, 78, 84, 89).

Bishop, Christopher M (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc. ISBN: 0387310738 (cit. on pp. 12, 13, 19, 80).

Burges, Chris (1998). "A Tutorial on Support Vector Machines for Pattern Recognition". In: *Data Mining and Knowledge Discovery* 2.2, pp. 121–167. ISSN: 13845810. DOI: 10.1023/A:1009715923555. arXiv: 1111.6189v1 (cit. on pp. 79, 125).

Burns, Michelle Nicole et al. (2011). "Harnessing context sensing to develop a mobile intervention for depression". In: *Journal of Medical Internet Research* 13.3. ISSN: 14388871. DOI: 10.2196/jmir.1838. arXiv: arXiv:1011.1669v3 (cit. on pp. 42, 49, 51, 77).

Chen, Zhenyu et al. (2013). "Unobtrusive Sleep Monitoring using Smartphones". In: *Proceedings of the ICTs for improving Patients Rehabilitation Research Techniques*. January. ISBN: 978-1-936968-80-0. DOI: 10.4108/icst.pervasivehealth.2013.252148. arXiv: arXiv:1407.5910v1 (cit. on p. 77).

Eysenbach, Gunther (2001b). "What is e-health?" In: *Journal of Medical Internet Research* 3.2, pp. 1–5. ISSN: 14388871. DOI: 10.2196/jmir.3.2.e20. arXiv: Eriksen2004a.pdf (cit. on p. 76).

Friedman, Jerome, Trevor Hastie, and Rob Tibshirani (2009). "Regularization paths for generalized linear models via coordinate descent". In: *Journal of Statistical Software* (cit. on pp. 79, 81).

Gaggioli, Andrea et al. (2013). "A mobile data collection platform for mental health research". In: *Personal and Ubiquitous Computing* 17.2, pp. 241–251. ISSN: 1617-4909. DOI: 10.1007/s00779-011-0465-2 (cit. on p. 77).

Gimpel, Henner, Christian Regal, and Marco Schmidt (2015). "myStress: Unobtrusive Smartphone Based Stress Detection". In: *Ecis* 4801.2015, pp. 0–12 (cit. on p. 77).

Helliwell, John F. and Shun Wang (2015). "How Was the Weekend? How the Social Context Underlies Weekend Effects in Happiness and Other Emotions for US Workers". In: *PLOS ONE* 10.12. Ed. by Christopher M. Danforth, e0145123. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0145123 (cit. on p. 84).

Iida, M et al. (2012). "Using Diary Methods in Psychological Research". In: *APA Handbook of Research Methods in Psychology: Vol. 1. Foundations, Planning, Measures and Psychometrics*. Ed. by H. Cooper et al. Washington, DC: US: American Psychological Association, pp. 277–305. ISBN: 1-4338-1004-2. DOI: 10.1037/13619-016 (cit. on pp. 76, 77, 104).

LiKamWa, Robert et al. (2013). "MoodScope". In: *Proceeding of the 11th annual international conference on Mobile systems, applications, and services - MobiSys '13*, p. 389. DOI: 10.1145/2462456.2464449 (cit. on pp. 49, 51, 52, 77, 80, 81, 88, 89).

Ma, Yuanchao et al. (2012b). "Daily Mood Assessment Based on Mobile Phone Sensing". In: *2012 Ninth International Conference on Wearable and Implantable Body Sensor Networks*. IEEE, pp. 142–147. ISBN: 978-0-7695-4698-8. DOI: 10.1109/BSN.2012.3 (cit. on pp. 77, 88).

MacKay, David J. C. (1996). "Bayesian Methods for Backpropagation Networks". In: *Models of neural networks III*. Physics of. Springer, pp. 211–254. DOI: 10.1007/978-1-4612-0723-8_6 (cit. on p. 80).

Meyer, David et al. (2015). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien* (cit. on p. 81).

Neal, Radford M. (1996). *Bayesian Learning for Neural Networks*. Vol. 118. Lecture Notes in Statistics. New York, NY: Springer New York. ISBN: 978-0-387-94724-2. DOI: 10.1007/978-1-4612-0745-0 (cit. on p. 80).

R Core Development Team (2014). *R: a language and environment for statistical computing, 3.1.2 ed.* R Foundation for Statistical Computing. Vienna, Austria. ISBN: 3-900051-07-0 (cit. on pp. 79, 81, 159).

Rencher, Alvin C. and William F. Christensen (2012). *Methods of Multivariate Analysis*. Wiley Series in Probability and Statistics. Hoboken, NJ, USA: John Wiley & Sons, Inc. ISBN: 9781118391686. DOI: 10.1002/9781118391686 (cit. on p. 79).

Saeb, Sohrab et al. (2015a). "Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study". In: *Journal of Medical Internet Research* 17.7, e175. ISSN: 1438-8871. DOI: 10.2196/jmir.4273 (cit. on pp. 4, 77, 88).

Smola, Alexander J and Bernhard Schölkopf (2004). "A Tutorial on Support Vector Regression". In: *Statistics and Computing* 14.3, pp. 199–222. ISSN: 09603174. DOI: Doi10.1023/B:Stco.0000035301.49549.88 (cit. on p. 79).

Spil, Ton and R.W. Schuring, eds. (2006). *E-Health Systems Diffusion and Use*. IGI Global. ISBN: 9781591404231. DOI: 10.4018/978-1-59140-423-1 (cit. on p. 76).

Stone, Arthur A. and Saul Shiffman (1994). "Ecological Momentary Assessment (Ema) in Behavioral Medicine". In: *Annals of Behavioral Medicine* 16.3, pp. 199–202. ISSN: 0883-6612. DOI: 10.1093/abm/16.3.199 (cit. on pp. 3, 76, 140).

Tibshirani, Robert (1996). *Regression Selection and Shrinkage via the Lasso*. DOI: 10.2307/2346178. arXiv: 11/73273 [1369-7412] (cit. on pp. 9, 79, 125).

Tipping, Michael E. (2000). "The Relevance Vector Machine". In: *Advances in Neural Information Processing Systems (NIPS' 2000)*. 1, pp. 652–658. ISBN: 0-262-19450-3. DOI: 10.1.1.34.4986. arXiv: 1502.02761 (cit. on pp. 13, 80).

Vapnik, Vladimir (1998). "Statistical Learning Theory". In: *Adaptive and learning Systems for Signal Processing, Communications and Control*, pp. 1–740 (cit. on p. 79).

Vapnik, Vladimir N. (2000). *The Nature of Statistical Learning Theory*. New York, NY: Springer New York. ISBN: 978-1-4419-3160-3. DOI: 10.1007/978-1-4757-3264-1 (cit. on p. 79).

Wang, Rui et al. (2014). "StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones". In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '14 Adjunct*. New York, New York, USA: ACM Press, pp. 3–14. ISBN: 9781450329682. DOI: 10.1145/2632048.2632054 (cit. on p. 77).

Appendix A. Description of analysed features

Attribute	Description
valence	Two dimensional construct (-2 - 2)
arousal	Two dimensional construct (-2 - 2)
activity	average percentage of high accelerometer data points
measureOfDay	which number of data record a day
appCat.entertainment	app use frequency, entertainment app category
appCat.office	app use frequency, office app category
appCat.finance	app use frequency, business tools category
appCat.other	app use frequency, other app category
appCat.unknown	app use frequency, unknown app category
appCat.weather	app use frequency, weather app category
appCat.builtin	app use frequency, builtin app category
appCat.social	app use frequency, social app category
appCat.utilities	app use frequency, utilities app category
appCat.game	app use frequency, games app category
appCat.communication	app use frequency, communication app category
appCat.travel	app use frequency, travel app category
call	number of calls made (5 most frequent contacts)
sms	number of SMS's sent (5 most frequent contacts)
screen	screen-on frequency
weekday2	weekday – Monday
weekday3	weekday – Tuesday
weekday4	weekday – Wednesday
weekday5	weekday – Thursday
weekday6	weekday – Friday
weekday7	weekday – Saturday

TABLE 4.5: Description of analysed features

Appendix B. Estimated Lasso feature importance ranking

Rank	Feature	Lambda
1	valence	.63
2	arousal	.12
3	activity	.003
4	appCat.builtin	.003
5	appCat.weather	.003
6	call	.003
7	appCat.game	.002
8	appCat.social	.002
9	appCat.communication	.002
10	appCat.utilities	.001
11	appCat.unknown	.001
12	appCat.finance	.001
13	appCat.office	.001
14	appCat.travel	.001

TABLE 4.6: Lasso feature ranking

Appendix C. Hierarchical Bayesian regression model in JAGS

```
model{
        for( j in 1:N ){
                mu[j] <- inprod(X[j,], beta[id[j],] )
                mood[j] ~ dnorm( mu[j], tau )
        }

        for(u in 1:USERS){
                for(i in 1:p){
                        beta[u,i] ~ dnorm(delta[i],indiv_gamma[i])
                }
        }
        for(j in 1:p){
        indiv_gamma[j] ~ dgamma(.001, .001)
        }

        delta ~ dmnorm(mubar, sigma)

        tau ~ dgamma(.001, .001)
}
```

Chapter 5

The predictive power of EMA data for mood and depression score prediction

Becker, D., Bremer, V., Funk, B. (2019). Working paper.

Abstract: *Ecological Momentary Assessment (EMA) data are self-reported measures that allow inference and prediction of clients' future mental health states. Despite promising results, the predictive performance reported in the literature varies considerably. Reasons for the variance among studies could be explained by use of different datasets, data preprocessing, and utilized methods. To eliminate these causes, we analyze an EMA dataset from a European depression study and apply the same preprocessing for the estimation of the predictive power of EMA data regarding short-term mood and medium-term depressive symptoms (PHQ9) prediction. The results of the short-term mood prediction suggest that measures of mood are considerably stable and that EMA measures alone are not able to predict daily mood fluctuations. Contrary, the results on the PHQ9 prediction suggest that medium-term prediction is feasible. We find that features created from EMA measures of mood, self-esteem, and conducted activities are predictive for the PHQ9 score.*

5.1 Introduction

In computer-based therapies mobile phones allow clients to regularly report their symptoms through Ecological Momentary Assessments (EMA) (Shiffman, Stone, and Hufford, 2008). EMA measures are inquired by regularly posing questions regarding mood, worries or symptoms, which are rated by clients on a numeric scale (Wichers et al., 2011; Gibbons, 2017). Self-reported measures are an important link to clinical health consequences. For example, self-reported personality and self-esteem measures are predictive for depression (Cheng and Furnham, 2003). These measures provide therapists and researchers with the opportunity to understand how clients are progressing during treatment. It has been shown that EMA measures exhibit a strong correlation with weekly assessed depression questionnaires (Saeb et al., 2015a) and mood ratings have been significantly related to clinical depression scores (Aguilera, Schueller, and Leykin, 2015). Accordingly, EMA data have been utilized for the prediction of therapy success (Bremer et al., 2018; Breda et al., 2018), risk of future depressive episodes (Van Voorhees et al., 2008; Ebert et al., 2019), or relapse prediction (Beckjord and Shiffman, 2014; Juarascio et al., 2015; Jones et al., 2018).

During treatment, EMA can be utilized to predict clients' future health states (Van Breda et al., 2016a; Mikus et al., 2017). However, the predictive accuracy reported for mood level prediction varies tremendously, it ranges from barely existing (Ma et al., 2012b) to very high(LiKamWa et al., 2013). These differences can originate from using different datasets, experimental settings, and different methods for modeling. Specifically, differences among

datasets can result from the way the study was conducted and how the treatment effects were measured (Veroniki et al., 2016). When utilizing the same dataset, differences in the reported results can arise from using different model parameters or different data preprocessing (Crone, Lessmann, and Stahlbock, 2006). Therefore, for the prediction of short-term and medium-term measures using EMA data, we utilize data from a large scale European study delivering cognitive behavioral therapy to clients with depressive symptoms (Kleiboer et al., 2016) and apply the same preprocessing to each analysis. This allows an objective comparison of the results for each method. In the analysis of short-term prediction, we predict clients' mood levels on the following day and utilize algorithms that have been reported in the literature to perform well for this task. Specifically, we compare the predictive performance of a support vector regression (Breda et al., 2016), recurrent neural network (Mikus et al., 2017), social integration model (Altaf Hussain Abro, 2016), and hierarchical Bayesian regression (Becker et al., 2016b). For the analysis of medium-term prediction, we use the EMA dataset to predict PHQ9 depression scores (Kroenke, Spitzer, and Williams, 2001) that were assessed over the course of the treatment. Particularly, EMA data is a time series that represents clients' progress over weeks and months. To utilize the information in the EMA data, we extract features from the EMA data preceding a PHQ9 assessment to predict the depression score utilizing a mixed effect model.

5.2 Related work

Across the literature, EMA data has been utilized to answer various research questions. It has been utilized to predict other EMA concepts, depression scores, and future mood. Similarly, unobtrusive mobile phone measures have been utilized to predict users' current mood levels. Research on binary mood level prediction for utilizing mobile phone data reports results ranging from 70% to 76% accuracy (Grünerbl et al., 2015; Jaques et al., 2015). Although these results suggest that mood prediction is feasible, this only allows distinguishing between low and high mood levels, which does not provide any granularity. Research regarding a more fine-grained mood prediction based on mobile phone measures such as communication history and usage patterns has been conducted by LiKamWa et al. (2013). Here, a client-unspecific classifier could achieve an overall accuracy of 66% and a client-specific classifier led to an accuracy of 93%. However, these results could not be replicated in a study by Asselbergs et al. (2016), who report an accuracy ranging from 55% to 76% and root mean square error of 0.76. Client-specific clusters (Breda et al., 2016) and client individual models (Becker et al., 2016b) have also been studied based on the same data utilized by Asselbergs et al. (2016) and achieved a prediction root mean square error of 0.64 and 0.53, respectively. The results suggest that mood prediction utilizing unobtrusive measures is a challenging task and that mobile measures are affected by high noise levels and user individual variances in behavior.

Therefore, EMA measures might be more reliable for the prediction of future mood levels. However, also for this task, the reported prediction accuracies vary considerably. Van Breda et al. (2016a) analyzed the prediction of clients' mood levels on the following day utilizing linear regression and obtained a root mean square errors range from 0.26 to 0.33. Similarly, (Altaf Hussain Abro, 2016) used a temporal causal model for mood prediction and achieved a root mean square error of 0.21. A comprehensive analysis for the task of tomorrow's mood prediction was conducted by Mikus et al. (2017) who utilized hierarchical clustering, linear data imputation, and recurrent neural networks for client individual mood prediction and report a root mean square error ranging from 0.065 to 0.11.

EMA measures have further been utilized to estimate different health-related measures such as mood, self-esteem, sleep, and suicidal ideation. Diary data, which can also be reported as free-text, has been utilized for mood inference. Bremer et al. (2017b) applied

text-mining methods to such free-text diary data to extract daily activities for the inference of daily mood levels. They estimated that reports of sickness and rumination have a significant negative influence on the mood level while reports of social activities are related to a higher mood rating. The relationship between depressive symptoms and mood has been analyzed by Wenze et al. (2013). They estimated that higher depression symptoms were predictive for lower mood levels but symptoms of anxiety were not influential for mood prediction. The relationship between EMA measures and self-esteem has been analyzed by Bremer, Funk, and Riper (2019). They utilized hierarchical models that account for client infinitival effect differences and estimated that measures of mood and enjoyed activities are significantly related to ratings of self-esteem. Similarly, the relationship between EMA measures and ratings of sleep quality have been analyzed by Parsey and Schmitter-Edgecombe (2019). They aimed to answer the question of how ratings of sleep affect the perceived work performance and estimated a significant relation between the previous day's sleep rating and fatigue. However, they did not find a significant influence on mood or perceived thinking abilities during the day. In contrast to this finding, Triantafillou et al. (2019) estimated a strong effect between sleep quality and the next days' mood and vice versa by analyzing EMA data collected using a mobile phone. Similarly, Sano et al. (2015) reported that ratings of sleep duration and sleep regularity are predictive of mood ratings. The relationship between sleep disturbance and suicidal ideation has been researched by Littlewood et al. (2019). They estimated that self-reported measures of poor sleep quality are predictive for higher severity of next-day suicidal ideation.

Besides the prediction of various self-reported measures, the relationship between EMA measures and the PHQ9 depression score has been analyzed in research. Torous et al. (2015) examined the influence of a daily inquired subset of the PHQ9 questions on traditionally administered PHQ-9 scores. The daily measures were collected using a mobile application and exhibited a strong correlation with the PHQ9 scores, where the daily collected scores were on average higher than in the traditionally administered PHQ-9 questionnaire. Further, the mobile collected scores were more likely to contain higher ratings of suicidal intention than obtained from the PHQ-9 questionnaire. Drake, Csipke, and Wykes (2013) analyzed the use of an online service called Moodscope, which allows tracking of daily mood. In this application, mood states are estimated by ratings of 20 positive and negative emotions such as proud, nervous, and determined. In their evaluation, they found a significant correlation between weekly scores derived with Moodscope and PHQ-9 scores. A different online mood tracker that captures mood states using 22 self-reported items has been evaluated by Nahum et al. (2017). They also found strong correlations among the assessed mood items and PHQ-9 depression scores. Likewise, ratings of mood are significantly related to PHQ9 depression scores in a study conducted by Aguilera, Schueller, and Leykin (2015). Aguilera, Schueller, and Leykin (2015) assessed clients' mood and PHQ9 ratings over weeks and analyzed the relation between the PHQ9 depression scores and preceding mood measures. The average mood score of the previous week is a significant predictor for the consecutive PHQ9 depression score. A limited predictive EMA performance was suggested by Breda et al. (2018), who analyzed the use of EMA measures and a variety of questionnaires to predict treatment success.

5.3 Data and Method

For short-term mood prediction, we utilize the EMA measures to predict the clients' mood on the following day and evaluate client individual models and a model for all clients. In the second analysis, the relationship between EMA measures and PHQ9 depression scores is examined. Here, the EMA measures are used to create time and frequency domain features

for use in a random intercept model (Hoogendoorn and Funk, 2017). An overview of the specific steps for both analyses is illustrated in Figure 5.1.

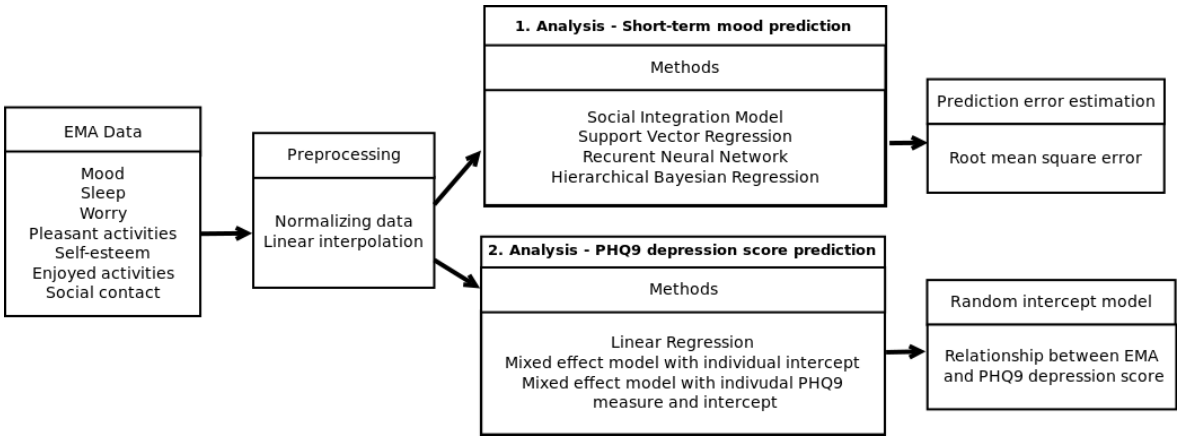


FIGURE 5.1: Overview of the conducted analyses.

5.3.1 Description of the analyzed data

The utilized data for our analysis originates from the EU funded project E-COMPARED (Kleiboer et al., 2016). The study compared bCBT (blended cognitive behavior therapy, experiment group) and face-to-face treatment (control group) for the treatment of depression. For the participants in the experimental group, EMA data was gathered and CBT was provided using a mobile phone between February 2015 and February 2018.

The EMA measures were inquired once a day, but clients were allowed to add additional mood measures at any time. The study protocol defines that a single measure of mood was inquired once a day at a random time between 10am and 10pm. A more complete assessment was done once a week, where all the psychological factors were assessed (shown in Table 8.1). During the initial week of the treatment and at the end of the suggested mobile application usage period all measures were inquired daily. However, the clients were also free to neglect any of these inquirements. The dataset contains measures of 324 clients.

EMA measure	**Assessment question**	**Assessment frequency**
Mood	How is your mood right now?	Daily
Worry	How much do you worry about things at the moment?	Daily during first and last week; random day in other weeks
Self-Esteem	How good do you feel about yourself right now?	Daily during first and last week; random day in other weeks
Sleep	How did you sleep tonight?	Daily during first and last week; random day in other weeks
Pleasant activities	To what extent have you carried out enjoyable activities today?	Daily during first and last week; random day in other weeks
Enjoyed activities	How much did you enjoy activities today?	First and last week; random day in other weeks
Social contact	How much were you involved in social interactions today?	Daily during first and last week; random day in other weeks

TABLE 5.1: Utilized EMA data.

Besides the EMA measures that have been inquired using the mobile application, a variety of questionnaires were used to assess the clients' condition. These questionnaires provide a more detailed assessment of the clients' condition and current depressive symptoms,

which were assessed at the beginning of the treatment and after 3, 6, and 12 months. In particular, for the analysis of the relation between these EMA measures and depressive symptoms, we utilize the PHQ9 questionnaire score that measures the severity of depressive symptoms (Kroenke, Spitzer, and Williams, 2001). The PHQ9 questionnaire comprises of 9 questions that are ranked on a Likert scale ranging from 0 to 3, where 0 represents no symptoms and 3 represents symptoms almost every day. Besides the interpretation of the individual questions, the overall score can be used to estimate the severity of depressive symptoms. This questionnaire can be particularly useful to measure the course of the symptoms and the treatment efficacy (Löwe et al., 2004).

5.3.2 Preprocessing and short-term mood prediction

For preprocessing the EMA measures, we create a time series for each client containing one measure per day. In the case of multiple mood measures in one day, we utilize the mean value. Since some clients only have a few observations, we only consider clients that have at least measures over a period of 27 days. This ensures that sufficient data for each client is available to enable a split into train and test data. After this procedure, the EMA data of 292 clients are retained. After creating the time series for each client, all EMA measures are normalized between $[0, 1]$ and missing values are linearly interpolated as used in the study conducted by Mikus et al. (2017). In general, the imputation of missing data has been shown to provide a valuable alternative to removing missing observations and can produce low errors (Penone et al., 2014).

For the prediction of the following day's mood level, we use the first half of a clients' time series for training and the remaining half as the test set. To utilize the applied algorithms for mood prediction, we use a sliding time window that captures the last 7 days to predict the 8th. This is equivalent to an autoregressive model, where the next day is predicted as a weighted linear combination of the past 7 days. The choice of a window size of 7 is based on the findings of Mikus et al. (2017), who estimated that the use of the past 7 days resulted in the best prediction performance. Similarly, Suhara, Xu, and Pentland (2017) revealed that the last mood measure has the highest importance while then decreasing over time. However, the periodic measures on days 7 and 14 also exhibit increased importance. This might suggest a periodic behavior in the mood, where the measure 7 days ago would allow to capture a weekly cycle.

For the analysis of short-term mood prediction, we estimate the prediction error for a one-fits-all model and a client individual model. In addition, we retrain the client individual model after each prediction to achieve a higher predictive performance.

Since we use the first half of a client's time series for the training of the models and the second half for testing, we interpolate these two parts differently. The first half of a client's time series is assumed as observed and can thus be linearly interpolated. The second part of the time series (the test set) is only interpolated up to the point of prediction. Specifically, the complete time series cannot be interpolated at once, because this would incorporate knowledge of future measures for interpolation. Instead, the last observed measure is repeated in the case of missing values. For the estimation of the prediction error, the interpolated values are not considered because they would artificially reduce the average prediction error.

Utilized algorithms for short-term mood prediction

Recurrent neural networks For the prediction of future clients' EMA ratings, recurrent neural networks have been shown to provide low prediction errors (Suhara, Xu, and Pentland, 2017; Mikus et al., 2017; Mikelsons et al., 2017). Recurrent neural networks (RNN) are an

extension of a feed-forward neural network, which can handle a sequence of inputs. Typically, the RNN processes a sequence of variable-length by having a recurrent hidden state. The output of such a network is dependent on the current input and the hidden state which is dependent on the previous input. Internally, recurrent networks can consist of many layers of recurrent units. One type of recurrent network unit is the Gated Recurrent Unit (GRU) that has been introduced in 2014 by Cho et al. (2014). The output of a GRU is a linear combination of the current input and the hidden state. GRUs can provide better performance than similar recurrent units such as Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) on smaller datasets (Chung et al., 2014). A GRU is less complex than LSTM, which might explain the better performance on small datasets and might, therefore, provide a better prediction performance for a client individual mood prediction. Mikus et al. (2017) tested the performance of GRU and LSTM for future mood prediction and could achieve the best prediction performance with a single layer GRU followed by a recurrent projection layer. Therefore, we will utilize the same model for the estimation of the prediction error in our analysis.

Support vector regression Support Vector Regression (Vapnik, 1995) (SVR) aims to find a function that represents the training data within a small error margin. For training, the support vector regression neglects all errors as long as they are within this margin. This algorithm is often referred to as ϵ Support Vector Regression. Besides, the SVR utilizes a Kernel for non-linear feature transformation. This provides a measure of similarity in a higher-dimensional space compared to the actual input data, where the data might be linearly separable. The use of the Support Vector Regression with a radial basis kernel has been applied for the prediction of future mood by Breda et al. (2016). In their study, this method provided the best prediction performance in comparison to the other applied algorithms. SVR has also been applied for the clients' prediction depression score based on GPS movement patterns (Canzian and Musolesi, 2015), for forecasting of mood changes in bipolar clients (Valenza et al., 2015), and as a reference measure for mood prediction in depression (Mikus et al., 2017).

Hierarchical Bayesian regression Since all clients in the analyzed dataset possibly share similarities regarding their mood trajectory, the use of methods that model this hierarchical structure can provide a better prediction performance (Jaques et al., 2017; Jaques et al., 2016). Therefore, we use a hierarchical Bayesian regression model that has been utilized for the prediction of future mood based on mobile phone usage data (Becker et al., 2016b). The model defines a hierarchical prior that captures the similarities among clients.

Social integration model The social integration model (Altaf Hussain Abro, 2016) (SIM) employs a temporal-causal modeling approach (Treur, 2016), which uses differential equations to describe the relationships among various EMA measures. This type of modeling used prior domain knowledge by describing the relationship between a client's mood and its social interactions. In contrast to the previous methods, the SIM is a top-down approach because it defines the relationships among the EMA measures based on knowledge derived in the field of psychology. Additionally, this model only utilizes a subset of the overall available EMA measures in the dataset. Specifically, only the measures of social contact, enjoyed activities, and pleasant activities are represented in the social integration model.

Utilized reference measures For comparison of the obtained prediction error, we also report two different naive prediction methods, which we refer to as reference measures. These methods only use the mood measure and allow to estimate a baseline prediction error.

The first method uses the mean mood value of the training data as the prediction for each mood value in the test set. Further, we report the prediction error for re-training this model, which results in a moving average prediction. The second reference measure predicts the last observed value as the next mood level. Specifically, it predicts tomorrow's mood equal to today's mood.

5.3.3 PHQ9 prediction

For the prediction of the PHQ9 depression score using the EMA measures, we create a variety of time and frequency domain features using the linear interpolated time series. We split the time series into parts according to the scheduled assessment of the PHQ9 questionnaire after 3, 6, and 12 months in order to link a client's time series to the PHQ9 score. The first split of the time series contains the first EMA measure up to 90 days. The second part is from day 90 up to day 180, and the third part is from 180 to 360 days. If there are EMA measures for this period and a PHQ9 score, we consider the PHQ9 scores as the outcome for the time series. This procedure results in multiple time series including a PHQ9 score for each client. 311 samples from 246 different clients are considered; 239 samples for 3-months, 70 for 6-months, and 7 for the PHQ9 score assessed after 12-months.

We then proceed to extract time series features for the application of a mixed-effect model with a client individual intercept. As time domain-based features, we estimated the mean, variance, skewness, and kurtosis for the individual measures. The mean can be considered as the average EMA value in the considered time frame and the variance reflects the fluctuation of the measure. The skewness of the measure captures the asymmetry of the distribution. A positive value of the skewness reflects that the distribution is leaned towards the lower end of the scale and a negative value reflects a distribution that has more measures towards the higher values. The kurtosis describes how peaked the distribution is, where a negative value indicates a more peaked distribution in comparison to a standard normal distribution and a positive value a flatter peak.

Furthermore, the normalized sum of absolute values of consecutive changes in the series was estimated as a feature. This feature is derived by summing all absolute gradients and dividing it by the number of measures. It indicates how many changes occur within the time series. Similarly, we calculate the normalized absolute energy of the time series, which is the sum over the squared values divided by the number of samples.

These time series features have been normalized to ensure that a longer time series does not result in larger values, which might naturally be associated with a lower PHQ9 score because more time has passed. Additionally, we added the minimum and maximum value of the time series as a feature. These features represent, for example, the lowest and highest mood within the considered time frame. For the mood time series, we compute a regression model and include the estimated intercept and slope as features. For the other measures, we did not use these features because they contain many missing values and a regression model based on many imputations might not be representative.

For the frequency-based features, we calculated the Fourier transformation of the mood time series and estimate the various moments of the real part of the frequency spectrum. Similar to the time domain features, we estimated the mean, variance, skewness, and kurtosis.

For the analysis of the relationship between the extracted features and the PHQ9 score, we considered a regression model and two linear mixed effect models. Linear mixed models are an extension of the fixed-effect model that allow incorporating random effects. This type of model is particularly useful because it allows to model the hierarchical structure of the data. Mixed-effects regression models are a widely applied method for analysis of longitudinal data (Molenberghs and Verbeke, 2001), which includes the analysis of EMA data (Hedeker, Mermelstein, and Demirtas, 2008). For the prediction of the PHQ9 score, we

compared a linear regression model, a mixed-effect model with client individual intercept, and a mixed-effect model with a client individual intercept and an individual intercept for the assessment time of the PHQ9 (3, 6, or 12 months). An ANOVA comparison of these three models showed that the model consisting only of a client individual intercept has the lowest AIC and BIC. Therefore, we utilize the linear mixed effect model with the client individual intercept for this analysis.

5.4 Results

5.4.1 Short-term mood prediction analysis

The results of the individual methods for short-term mood prediction are listed in Table 5.2.

Method	Individual model		One-fits all model
	RMSE	RMSE (re-train)	RMSE
SIM (only mood)	0.148	0.144	0.145
SIM (additional EMA)	0.148	0.144	0.142
RNN (only mood)	0.128	0.126	0.141
RNN (additional EMA)	0.157	0.135	0.163
SVR (only mood)	0.144	0.131	0.133
SVR (additional EMA)	0.158	0.140	0.129
Bayesian Reg. (only mood)	0.135	0.136	0.127
Bayesian Reg. (additional EMA)	0.190	0.190	0.134
Mean of observed values	0.148	0.125	—
Previous mood value	—	0.148	—

TABLE 5.2: Average root mean square error for the individual methods.

The lowest prediction error could be obtained using the RNN for each client and re-training after each new observation. However, the moving average prediction achieved a comparable prediction error and a paired t-test suggests that there is no difference between the prediction performance of both methods (p-value = 0.8896). The second best prediction performance has been achieved with the hierarchical model and utilizing the training data of all clients. However, the retrained RNN still provides a lower prediction error as confirmed by a paired t-test (p-value = 0.03728) than the Bayesian hierarchical regression. It further can be noticed that the SVR benefits from using the training data of all clients, in contrast to the RNN, which benefits from using a client individual model.

Significant EMA measure for mood prediction

For an estimation of the contribution of each utilized EMA measure on the mood prediction, we estimate the Bayesian regression with Gibbs sampling and illustrate the hierarchical weights with density intervals in Figure 5.2.

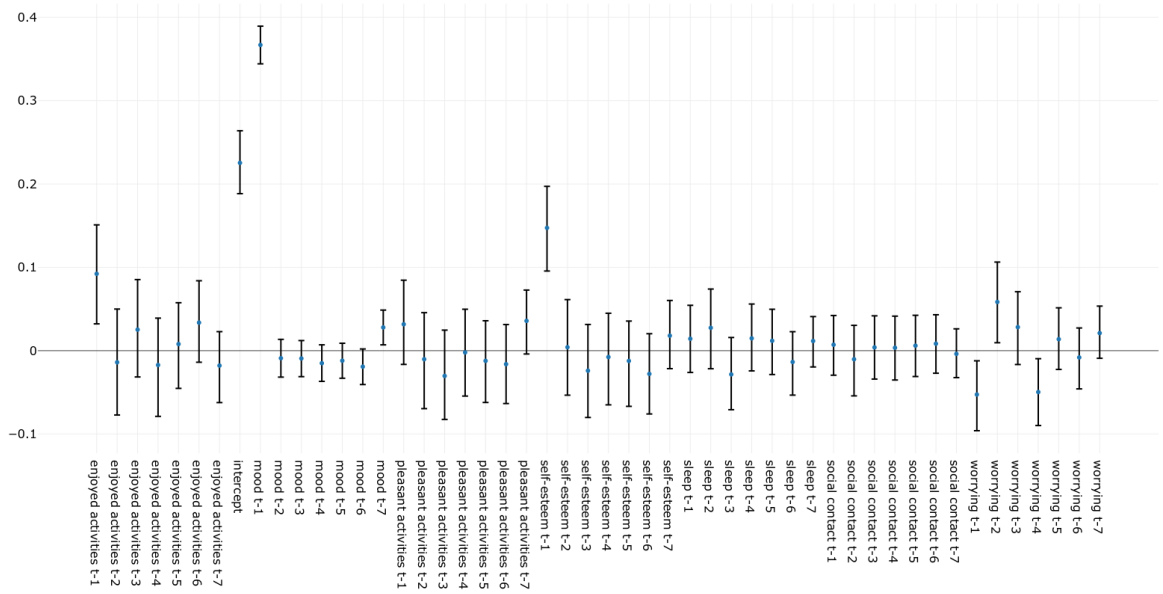


FIGURE 5.2: The estimates of the hierarchical prior for mood prediction with 95% density interval.

For the weight estimates of the hierarchical prior, we notice that the past mood measure, as well as the mood measure one week ago, are significant for the prediction of tomorrow's mood. Further, the past measure of self-esteem, enjoyed activities, and past two measures of worrying are significant.

5.4.2 Depression score prediction using EMA

The results of the mixed-effect model with a client individual intercept are shown in Table 5.3.

Feature	Estimate	Std. Error	Pr(>|t|)	Feature	Estimate	Std. Error	Pr(>|t|)
(Intercept)	8.86	0.29	0.00	pleasant activities (variance)	-0.73	0.69	0.29
spectral (mean)	-0.75	0.42	0.07	mood (mean)	1.73	2.28	0.45
spectral (var)	3.12	2.20	0.16	worrying (mean)	-0.75	2.19	0.73
spectral (skewness)	-2.63	2.93	0.37	self-esteem (mean)	-1.56	3.03	0.61
spectral (kurtosis)	1.41	2.97	0.63	enjoyed activities (mean)	-1.82	3.03	0.55
mood (min)	-2.22	0.62	0.00	sleep (mean)	1.68	2.62	0.52
worrying (min)	-0.17	0.81	0.83	social contact (mean)	1.93	2.58	0.45
self-esteem (min)	1.36	0.99	0.17	pleasant activities (mean)	1.65	2.66	0.54
enjoyed activities (min)	0.12	0.97	0.90	mood (absChanges)	-0.16	0.54	0.76
sleep (min)	0.24	0.82	0.77	worrying (absChanges)	-0.42	0.91	0.65
social contact (min)	-0.27	0.77	0.72	self-esteem (absChanges)	2.21	1.02	0.03
pleasant activities (min)	-0.29	0.88	0.74	enjoyed activities (absChanges)	0.12	1.11	0.92
mood (max)	-2.33	0.87	0.01	sleep (absChanges)	-0.61	0.77	0.43
worrying (max)	-0.74	1.18	0.53	social contact (absChanges)	-0.71	0.96	0.46
self-esteem (max)	-0.14	1.65	0.93	pleasant activities (absChanges)	-0.15	0.93	0.87
enjoyed activities (max)	-0.93	1.71	0.59	mood (absEnergy)	-1.78	2.93	0.54
sleep (max)	-1.09	1.67	0.52	worrying (absEnergy)	1.31	1.58	0.41
social contact (max)	2.48	1.27	0.05	self-esteem (absEnergy)	-0.96	2.02	0.63
pleasant activities (max)	-0.94	1.55	0.54	enjoyed activities (absEnergy)	1.60	2.08	0.44
mood (variance)	-1.31	0.69	0.06	sleep (absEnergy)	-1.19	1.72	0.49
worrying (variance)	0.79	0.54	0.14	social contact (absEnergy)	-2.36	1.81	0.19
self-esteem (variance)	-0.16	0.52	0.75	pleasant activities (absEnergy)	-0.41	1.84	0.82
enjoyed activities (variance)	1.25	0.63	0.05	mood (regression Slope)	-1.29	0.41	0.00
sleep (variance)	0.57	0.55	0.30	skewness (mood)	0.96	0.39	0.02
social contact (variance)	0.39	0.61	0.52	kurtosis (mood)	0.12	0.32	0.72

TABLE 5.3: Regression results of extracted EMA time series features to predict the PHQ9 score.

Some of the extracted time series features appear to be significant for the prediction of the depression score. The minimal and maximal mood values exhibit some relation to the PHQ9 measure. A high minimal and maximal mood value (mood (min), mood (max)) indicate

high mood ratings. Thus, these high mood ratings affect the PHQ9 score negatively, which suggests a potential reduction in depressive symptoms. Similarly, the skewness of the mood shows a connection with the PHQ9 questionnaire. If the mood is skewed towards higher mood values, the skewness is negative, which leads to a reduction of the expected PHQ9 score. Whereas a tendency to lower mood measures results in a positive skewness of the mood distribution and suggests a higher PHQ9 score. The regression slope calculated on the mood time series appears to influence the PHQ9 score. This is not surprising because a positive slope might indicate that the client is improving due to the received treatment. The absolute change in self-esteem also appears to be associated with the measured PHQ9 score. Large and rapid changes in self-esteem might, therefore, indicate a higher PHQ9 score and are associated with symptoms of depression. Similarly, higher variance in enjoyed activities indicates a higher depression score.

This suggests that the extracted information might be helpful to estimate the PHQ9 score based on the extracted time series features. For estimation of the additional benefit of the features in contrast to a depression score mean value prediction, we train a SVR on the extracted features and estimate the leave one out cross-validation error. This results in a root mean square error of 4.93 for the SVR to predict the PHQ9 score. To compare this estimate, we repeat this procedure and use the mean of PHQ9 scores in the training set as the prediction. This simple reference model achieves a root mean square error of 5.60. A paired t-test shows that there is a significant difference between the predictions of both models ($p=2.504 \cdot 10^{-5}$). This further indicates that the extracted features can be utilized for the inference of the PHQ9 score based on the EMA data. However, this analysis assumes independence among the samples and, therefore, does not account for the hierarchical structure of the data.

5.5 Discussion

The short-term mood prediction analysis showed that the utilized algorithms could not provide a significant lower prediction error than the moving average mood prediction. Although the RNN could provide a comparable prediction performance by utilizing retained information within the GRU layer for its prediction, it does not allow to predict daily mood fluctuations. Similar results are obtained for the Bayesian hierarchical regression, which has the advantage of utilizing data across the clients. Also, the addition of further EMA measures besides mood could not improve the prediction. This might, however, be related to the sparse assessment of measures besides mood.

The estimation of the significant EMA measures for mood prediction of the hierarchical model showed that the client individual intercept and previous mood level provide the most information for the prediction of tomorrow's mood level. This, and the low prediction error of the moving average prediction, suggest that clients' mood is considerable stable over short periods and that mood fluctuations cannot be predicted based on the utilized EMA measures. Further, the mood rating one week ago exhibited significance, which suggests periodic mood behavior. The findings agree with the results reported by Suhara, Xu, and Pentland (2017) and suggest why Mikus et al. (2017) found that utilization of the past 7 days led to the best prediction of future mood. The measure of self-esteem and enjoyed activities on the previous day appears to contribute to the prediction of the next day's mood level. A predictive relationship between measures of self-esteem and mood have already been reported in the literature (Cheng and Furnham, 2003; Bremer, Funk, and Riper, 2019).

A possible reason for the varying short-term mood prediction errors among the conduced studies could be due to the differences in the applied preprocessing. For example, in the case of many missing values in the test data, one could interpolate the entire time series before the analysis. This, however, would use information from the test data and would reflect the

assumption that the imputed measures were indeed observed. Another problem in the case of missing values is the question of which values are to be used for prediction. One can use the model prediction or an imputed value. If the model prediction is used, the prediction trajectory is continued but the model prediction has to be limited to the valid range. By using the interpolated data, this does not occur but an observed mood value after many missing values might not be predicted accurately.

In the PHQ9 depression score prediction analysis, we estimated significant time series features to predict the depression score. In contrast to the study conducted by Aguilera, Schueller, and Leykin (2015), the average mood value was not found predictive but other features created from mood measures were. Specifically, the minimal, maximal, skewness, and regression slope of the mood time series are indicators for the overall PHQ9 score. Furthermore, the number of absolute changes in self-esteem and variance in enjoyed activities appear to have a significant relation to the PHQ9 score. By utilizing the created time series features, we were able to predict PHQ9 scores more reliably than a mean value prediction. The created features appear to provide additional information for the prediction.

5.6 Conclusion

In this study, we analyzed EMA data originating from a depression treatment for short-term mood prediction and regularly assessed PHQ9 depression scores. We compared algorithms that have been reported in the literature about future mood prediction to provide low prediction errors. Furthermore, we used extracted features from the EMA measures and analyzed their relationship to the PHQ9 score.

The analyzed algorithms could not predict the daily fluctuation in mood more reliable than a moving average prediction. Despite the estimation of a significant relation of self-esteem, worries, and enjoyed activities, these were not sufficient to predict daily mood fluctuations accurately. However, the results suggest that one could design different EMA assessment protocols that reduce the number of inquiries without significant information loss. Daily mood assessment might not be necessary since it appears considerably stable and a client's mood level might be best summarized by a weekly average. Furthermore, the correlation among the EMA measures could be utilized for EMA assessment. Since mood measures are related to measures of self-esteem, enjoyed activities, and worries, these measures might not require an assessment on the same day. Highly correlated measures could be assessed alternating since they are predictive for each other to reduce a client's workload. The analysis of the extracted time series features shows that measures of mood and self-esteem are predictive for the PHQ9 score. Therefore, features derived from these EMA measures can indicate clients' depression scores in clinical practice.

References

Aguilera, Adrian, Stephen M. Schueller, and Yan Leykin (2015). "Daily mood ratings via text message as a proxy for clinic based depression assessment". In: *Journal of Affective Disorders* 175, pp. 471–474. ISSN: 15732517. DOI: 10.1016/j.jad.2015.01.033 (cit. on pp. 6, 26, 88, 90, 98).

Altaf Hussain Abro, Michel Klein (2016). "Validation of a Computational Model for Mood and Social Integration". In: *Lecture Notes in Computer Science*. Lecture Notes in Computer Science 10046.November 2016. Ed. by Emma Spiro and Yong-Yeol Ahn, pp. 361–375. DOI: 10.1007/978-3-319-47880-7 (cit. on pp. 7, 89, 93, 141, 142, 149).

Asselbergs, Joost et al. (2016). "Mobile Phone-Based Unobtrusive Ecological Momentary Assessment of Day-to-Day Mood: An Explorative Study". In: *Journal of medical Internet research* 18.3, e72. ISSN: 14388871. DOI: 10.2196/jmir.5505 (cit. on pp. 49, 52, 77, 78, 84, 89).

Becker, Dennis et al. (2016b). "How to Predict Mood? Delving into Features of Smartphone-Based Data". In: *Twenty-second Americas Conference on Information Systems*. San Diego (USA) (cit. on pp. 89, 93, 105).

Beckjord, Ellen and Saul Shiffman (2014). "Background for real-time monitoring and intervention related to alcohol use." In: *Alcohol Research: Current Reviews* 36.1, pp. 9–18. ISSN: 21694796 (cit. on p. 88).

Breda, Ward van et al. (2016). "Exploring and comparing machine learning approaches for predicting mood over time". In: *Smart Innovation, Systems and Technologies*. Vol. 60. November 2017, pp. 37–47. ISBN: 9783319396866. DOI: 10.1007/978-3-319-39687-3_4 (cit. on pp. 89, 93).

Breda, Ward van et al. (2018). "Predicting therapy success for treatment as usual and blended treatment in the domain of depression". In: *Internet Interventions* 12.August, pp. 100–104. ISSN: 22147829. DOI: 10.1016/j.invent.2017.08.003 (cit. on pp. 88, 90).

Bremer, Vincent, Burkhardt Funk, and Heleen Riper (2019). "Heterogeneity Matters: Predicting Self-Esteem in Online Interventions Based on Ecological Momentary Assessment Data". In: *Depression Research and Treatment* 2019. ISSN: 2090133X. DOI: 10.1155/2019/3481624 (cit. on pp. 26, 90, 97, 142).

Bremer, Vincent et al. (2017b). "Predicting the individual mood level based on diary data". In: *Proceedings of the 25th European Conference on Information Systems, ECIS 2017*. Vol. 2017. October, pp. 1161–1177. ISBN: 9780991556700 (cit. on p. 89).

Bremer, Vincent et al. (2018). "Predicting Therapy Success and Costs for Personalized Treatment Recommendations Using Baseline Characteristics : Data-Driven Analysis". In: *Journal of Medical Internet Research* 20.8, e10275. DOI: 10.2196/10275 (cit. on p. 88).

Canzian, Luca and Mirco Musolesi (2015). "Trajectories of depression". In: *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '15*, pp. 1293–1304. ISBN: 9781450335744. DOI: 10.1145/2750858.2805845 (cit. on p. 93).

Cheng, Helen and Adrian Furnham (2003). "Personality, self-esteem, and demographic predictions of happiness and depression". In: *Personality and Individual Differences* 34.6, pp. 921–942. ISSN: 01918869. DOI: 10.1016/S0191-8869(02)00078-8 (cit. on pp. 6, 88, 97).

Cho, Kyunghyun et al. (2014). "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation". In: ISSN: 09205691. DOI: 10.3115/v1/D14-1179. arXiv: 1406.1078 (cit. on p. 93).

Chung, Junyoung et al. (2014). "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling". In: pp. 1–9. ISSN: 2161-4393. DOI: 10.1109/IJCNN.2015.7280624. arXiv: 1412.3555 (cit. on p. 93).

Crone, Sven F., Stefan Lessmann, and Robert Stahlbock (2006). "The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing". In: *European Journal of Operational Research* 173.3, pp. 781–800. ISSN: 03772217. DOI: 10.1016/j.ejor.2005.07.023 (cit. on p. 89).

Drake, G., E. Csipke, and T. Wykes (2013). "Assessing your mood online: Acceptability and use of Moodscope". In: *Psychological Medicine* 43.7, pp. 1455–1464. ISSN: 00332917. DOI: 10.1017/S0033291712002280 (cit. on p. 90).

Ebert, David D. et al. (2019). "Prediction of major depressive disorder onset in college students". In: *Depression and Anxiety* 36.4, pp. 294–304. ISSN: 15206394. DOI: 10.1002/da.22867 (cit. on p. 88).

Gibbons, Chris J. (2017). "Turning the page on pen-and-paper questionnaires: Combining ecological momentary assessment and computer adaptive testing to transform psychological assessment in the 21st century". In: *Frontiers in Psychology* 7.JAN, pp. 1–4. ISSN: 16641078. DOI: 10.3389/fpsyg.2016.01933 (cit. on pp. 3, 4, 88, 140).

Grünerbl, Agnes et al. (2015). "Smartphone-based recognition of states and state changes in bipolar disorder patients". In: *IEEE Journal of Biomedical and Health Informatics* 19.1, pp. 140–148. ISSN: 21682194. DOI: 10.1109/JBHI.2014.2343154 (cit. on p. 89).

Hedeker, Donald, Robin J. Mermelstein, and Hakan Demirtas (2008). "An application of a mixed-effects location scale model for analysis of ecological momentary assessment (EMA) data". In: *Biometrics* 64.2, pp. 627–634. ISSN: 0006341X. DOI: 10.1111/j.1541-0420.2007.00924.x. arXiv: NIHMS150003 (cit. on p. 94).

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long Short-Term Memory". In: *Neural Comput.* 9.8, pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735 (cit. on p. 93).

Hoogendoorn, Mark and Burkhardt Funk (2017). *Machine Learning for the Quantified Self: On the Art of Learning from Sensory Data*. English. Springer. ISBN: 978-3-319-66307-4 (cit. on p. 91).

Jaques, Natasha et al. (2015). "Predicting students' happiness from physiology, phone, mobility, and behavioral data". In: *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015*, pp. 222–228. ISBN: 9781479999538. DOI: 10.1109/ACII.2015.7344575. arXiv: 15334406 (cit. on p. 89).

Jaques, Natasha et al. (2016). *Multi-task Learning for Predicting Health, Stress, and Happiness* (cit. on p. 93).

Jaques, Natasha et al. (2017). *Predicting Tomorrow's Mood, Health, and Stress Level using Personalized Multitask Learning and Domain Adaptation*. Tech. rep., pp. 17–33 (cit. on pp. 8, 93, 141).

Jones, Andrew et al. (2018). "Do daily fluctuations in inhibitory control predict alcohol consumption? An ecological momentary assessment study". In: *Psychopharmacology* 235.5, pp. 1487–1496. ISSN: 14322072. DOI: 10.1007/s00213-018-4860-5 (cit. on p. 88).

Juarascio, Adrienne S. et al. (2015). "Review of smartphone applications for the treatment of eating disorders". In: *European Eating Disorders Review* 23.1, pp. 1–11. ISSN: 10990968. DOI: 10.1002/erv.2327 (cit. on pp. 54, 88).

Kleiboer, Annet et al. (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". In: *Trials* 17.1,

p. 387. ISSN: 1745-6215. DOI: 10.1186/s13063-016-1511-1 (cit. on pp. 6, 89, 91, 123, 126, 131, 141, 144).

Kroenke, K, RL Spitzer, and JB Williams (2001). "The PHQ-9: validity of a brief depression severity measure". In: *J Gen Intern Med* 16.9, pp. 606–13 (cit. on pp. 89, 92, 123, 126).

LiKamWa, Robert et al. (2013). "MoodScope". In: *Proceeding of the 11th annual international conference on Mobile systems, applications, and services - MobiSys '13*, p. 389. DOI: 10.1145/2462456.2464449 (cit. on pp. 49, 51, 52, 77, 80, 81, 88, 89).

Littlewood, Donna L. et al. (2019). "Short sleep duration and poor sleep quality predict next-day suicidal ideation: An ecological momentary assessment study". In: *Psychological Medicine* 49.3, pp. 403–411. ISSN: 14698978. DOI: 10.1017/S0033291718001009 (cit. on p. 90).

Löwe, Bernd et al. (2004). "Monitoring Depression Treatment Outcomes With the Patient Health Questionnaire-9". In: *Medical Care* 42.12, pp. 1194–1201. ISSN: 0025-7079. DOI: 10.1097/00005650-200412000-00006 (cit. on pp. 7, 92).

Ma, Yuanchao et al. (2012b). "Daily Mood Assessment Based on Mobile Phone Sensing". In: *2012 Ninth International Conference on Wearable and Implantable Body Sensor Networks*. IEEE, pp. 142–147. ISBN: 978-0-7695-4698-8. DOI: 10.1109/BSN.2012.3 (cit. on pp. 77, 88).

Mikelsons, Gatis et al. (2017). "Towards Deep Learning Models for Psychological State Prediction using Smartphone Data: Challenges and Opportunities". In: *CoRR* abs/1711.0.Nips, pp. 1–6. arXiv: 1711.06350 (cit. on p. 92).

Mikus, Adam et al. (2017). *Predicting short term mood developments among depressed patients using adherence and ecological momentary assessment data*. DOI: 10.1016/j.invent.2017.10.001 (cit. on pp. 88, 89, 92, 93, 97, 141).

Molenberghs, Geert and Geert Verbeke (2001). "A review on linear mixed models for longitudinal data, possibly subject to dropout". In: *Statistical Modeling* 1.4, pp. 235–269. ISSN: 1471082X. DOI: 10.1177/1471082X0100100402 (cit. on p. 94).

Nahum, Mor et al. (2017). "Immediate Mood Scaler: Tracking Symptoms of Depression and Anxiety Using a Novel Mobile Mood Scale". In: *JMIR mHealth and uHealth* 5.4, e44. DOI: 10.2196/mhealth.6544 (cit. on pp. 26, 90).

Parsey, Carolyn M. and Maureen Schmitter-Edgecombe (2019). "Using actigraphy to predict the ecological momentary assessment of mood, fatigue, and cognition in older adulthood: Mixed-methods study". In: *Journal of Medical Internet Research* 21.1, pp. 1–12. ISSN: 14388871. DOI: 10.2196/11331 (cit. on pp. 26, 90).

Penone, Caterina et al. (2014). "Imputation of missing data in life-history trait datasets: Which approach performs the best?" In: *Methods in Ecology and Evolution* 5.9, pp. 1–10. ISSN: 2041210X. DOI: 10.1111/2041-210X.12232 (cit. on p. 92).

Saeb, Sohrab et al. (2015a). "Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study". In: *Journal of Medical Internet Research* 17.7, e175. ISSN: 1438-8871. DOI: 10.2196/jmir.4273 (cit. on pp. 4, 77, 88).

Sano, Akane et al. (2015). "Prediction of Happy-Sad mood from daily behaviors and previous sleep history". In: *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS* 2015-Novem, pp. 6796–6799. ISSN: 1557170X. DOI: 10.1109/EMBC.2015.7319954 (cit. on p. 90).

Shiffman, Saul, Arthur A. Stone, and Michael R. Hufford (2008). "Ecological momentary assessment". In: *Annual review of clinical psychology* 4.5, pp. 1–32. ISSN: 1548-5943; 1548-5943. DOI: 10.1146/annurev.clinpsy.3.022806.091415 (cit. on pp. 42, 88).

Suhara, Yoshihiko, Yinzhan Xu, and Alex Sandy Pentland (2017). "DeepMood: Forecasting Depressed Mood Based on Self-Reported Histories via Recurrent Neural Networks." In: *Www*, pp. 715–724. DOI: 10.1145/3038912.3052676 (cit. on pp. 92, 97).

Torous, John et al. (2015). "Utilizing a Personal Smartphone Custom App to Assess the Patient Health Questionnaire-9 (PHQ-9) Depressive Symptoms in Patients With Major

Depressive Disorder". In: *JMIR Mental Health* 2.1, e8. DOI: 10.2196/mental.3889 (cit. on pp. 52, 65, 90).

Treur, Jan (2016). "Dynamic modeling based on a temporal-causal network modeling approach". In: *Biologically Inspired Cognitive Architectures* 16.April, pp. 131–168. ISSN: 2212683X. DOI: 10.1016/j.bica.2016.02.002 (cit. on pp. 93, 141).

Triantafillou, Sofia et al. (2019). "Relationship between sleep quality and mood: Ecological momentary assessment study". In: *Journal of Medical Internet Research* 21.3, pp. 1–10. ISSN: 14388871. DOI: 10.2196/12613 (cit. on p. 90).

Valenza, Gaetano et al. (2015). "Predicting Mood Changes in Bipolar Disorder through Heartbeat Nonlinear Dynamics : a Preliminary Study natingfrom depression to (hypo-) manic , including mixed uations exclusively . To overcome this limitation , here we heartbeat nonlinear dynamics . Su". In: pp. 801–804 (cit. on p. 93).

Van Breda, Ward et al. (2016a). "A feature representation learning method for temporal datasets". In: *2016 IEEE Symposium Series on Computational Intelligence, SSCI 2016*. ISBN: 9781509042401. DOI: 10.1109/SSCI.2016.7849890 (cit. on pp. 88, 89).

Van Voorhees, Benjamin W. et al. (2008). "Predicting Future Risk of Depressive Episode in Adolescents: The Chicago Adolescent Depression Risk Assessment (CADRA)". In: *Annals of Family Medicine* 6.6, pp. 503–512. DOI: 10.1370/afm.887.INTRODUCTION (cit. on pp. 50, 54, 88).

Vapnik, Vladimir N (1995). *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag New York, Inc. ISBN: 0-387-94559-8 (cit. on p. 93).

Veroniki, Areti Angeliki et al. (2016). "Methods to estimate the between-study variance and its uncertainty in meta-analysis". In: *Research Synthesis Methods* 7.1, pp. 55–79. ISSN: 17592887. DOI: 10.1002/jrsm.1164 (cit. on p. 89).

Wenze, Susan J et al. (2013). "Biases in Short-Term Mood Prediction in Individuals with Depression and Anxiety Symptoms." In: *Individual differences research : IDR* 11.3, pp. 91–101. ISSN: 1541-745X (cit. on p. 90).

Wichers, M. et al. (2011). "Momentary assessment technology as a tool to help patients with depression help themselves". In: *Acta Psychiatrica Scandinavica* 124.4, pp. 262–272. ISSN: 0001690X. DOI: 10.1111/j.1600-0447.2011.01749.x (cit. on pp. 1, 42, 46, 88).

Chapter 6

Predicting the individual mood level based on diary data

Bremer, V., Becker, D., Funk, B., and Lehr, D. (2017). In Proceedings of the 25th European Conference on Information Systems, ECIS 2017.

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 neuronal 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.*

6.1 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, 2008).

These events and subsequently mood levels have a stake in determining well-being and cognitive functions such as problem solving (Isen, Daubman, and Nowicki, 1987), creativity, and the performance level (Nadler, Rabi, and Minda, 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, 2008), 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, 2008). 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 (Iida et al., 2012; Eysenbach, 2001a).

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.

6.2 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, 2008), 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 (Lewinsohn and Amenson, 1978; Grosscup and Lewinsohn, 1980). Researchers have also identified the existence of a relationship between specific activities such as exercise (Wang et al., 2012) or social activities (Clark and Watson, 1988; Byrne and Byrne, 1993) 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 (Nesse, 2000; Both et al., 2008; 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. (2016b) 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, Mishne, and Rijke (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, Guillory, and Hancock (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 (Hippisley-Cox et al., 2008; Chih et al., 2014; Wilson and Lankton, 2004; Hah and Bharadwaj, 2012). 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.

6.3 Setting, Predictors, & 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" (Ebert et al., 2014;

Buntrock 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 (Ebert et al., 2014; Buntrock et al., 2014). All clients gave their informed consent. Our dataset is based on an activity diary that has been kept by 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).

6.3.1 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, preventing low mood levels (Qian, Yarnal, and Almeida, 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

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

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, Powell, and Christenson, 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; Wang et al., 2012; Kanning and Schlicht, 2010). 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, 2008). 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, 2008). 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.

6.3.2 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 neuronal 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 6.1 illustrates our approach.

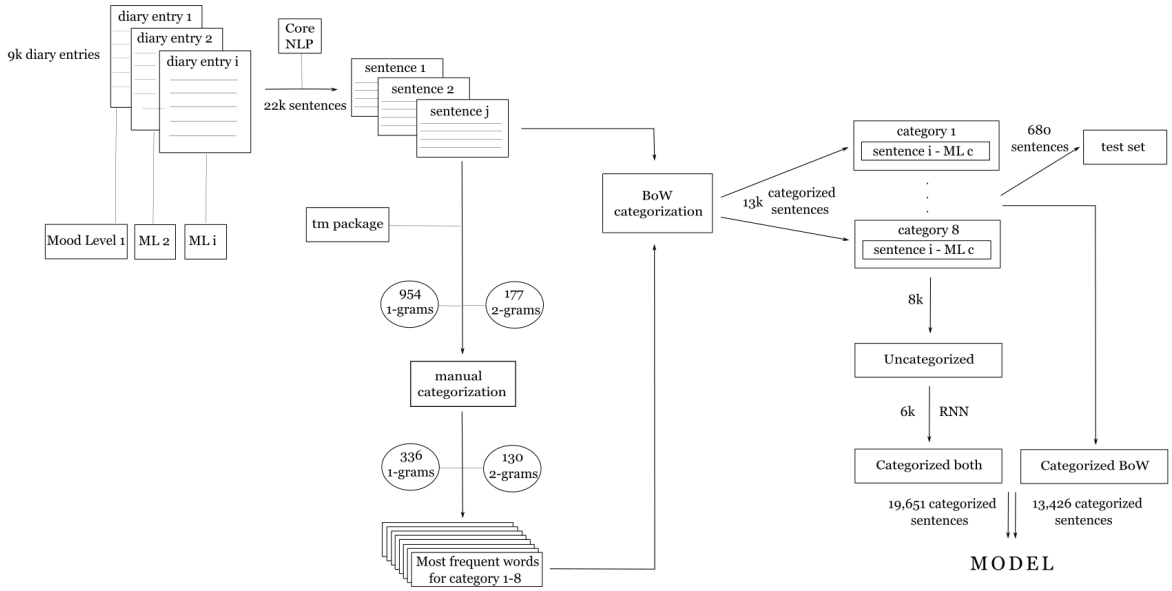


FIGURE 6.1: Process of the text-mining approaches for the categorization of diary entries.

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, Hornik, and Meyer, 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 neuronal network.
Why do we choose RNN? RNN architectures have been shown to produce strong results in language processing (Karafi and Kombrink, 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.

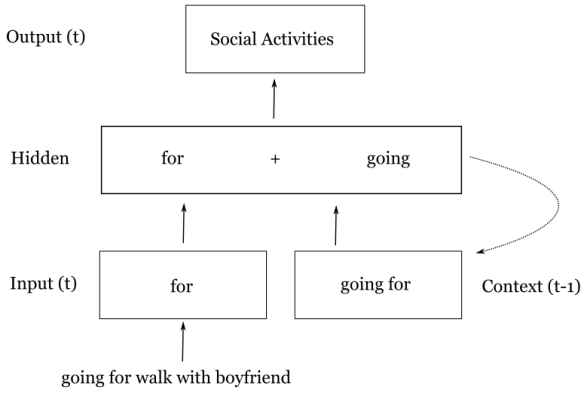


FIGURE 6.2: Visualization of the Elman Network.

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 neuronal network that subsequently tries to estimate the category of the sentence (output layer). Figure 6.2 illustrates an example of the classification process of the neuronal 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 6.2. 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.

6.4 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, Reis, and Elliot, 2000; Weinstein and Mermelstein, 2008). 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 (McCullagh, 1980; Peterson and Harrell, 1990; Liu and Koirala, 2012). 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.

$$\begin{aligned}\ln(\theta_{ijt}) = \alpha_i - (&\beta_{social_i}\, x_{social_{jt}} + \beta_{work_i}\, x_{work_{jt}} + \beta_{recreational}\, x_{recreational_{jt}} \\ &+ \beta_{necessary_i}\, x_{necessary_{jt}} + \beta_{exercise_i}\, x_{exercise_{jt}} + \beta_{sickness_i}\, x_{sickness_{jt}} \\ &+ \beta_{sleep}\, x_{sleep_{jt}} + \beta_{rumination_i}\, x_{rumination_{jt}}).\end{aligned} \tag{6.1}$$

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 6.1 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, Allenby, and McCulloch (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, Allenby, and McCulloch, 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.

6.4.1 Prior Settings & 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, Meyer, and Yu, 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, Meyer, and Yu, 2004). This method has been shown to perform adequately regarding a variety of examples (Spiegelhalter, Best, and Carlin, 1998; Berg, Meyer, and Yu, 2004). 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.

6.5 Results & 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 6.1 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 6.1: 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 6.2: Prediction performance for each model and text-mining approach.

As illustrated in Table 6.2, 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 6.2 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 6.3: Estimated model parameters (significant parameters in bold).

As illustrated in Table 6.3, 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; Weinstein and Mermelstein, 2008; Sonnentag, 2001). 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; Yeung, 1996; Netz and Lidor, 2003; Kanning and Schlicht, 2010). 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, 2008).

6.6 Limitations & 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 V Dhar (2014). "Big Data , Data Science , and Analytics : The Opportunity and Challenge for IS Research". In: *Information Systems Research* 25.3, pp. 443–448 (cit. on p. 104).

Ando, Tomohiro (2007). "Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models". In: *Biometrika* 94.2, pp. 443–458. ISSN: 00063444. DOI: 10.1093/biomet/asm017 (cit. on p. 112).

Balog, Krisztian, Gilad Mishne, and Maarten de Rijke (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*. Stroudsburg, PA (USA), pp. 207–2010 (cit. on p. 105).

Becker, Dennis et al. (2016b). "How to Predict Mood? Delving into Features of Smartphone-Based Data". In: *Twenty-second Americas Conference on Information Systems*. San Diego (USA) (cit. on pp. 89, 93, 105).

Berg, Andreas, Renate Meyer, and Jun Yu (2004). "Deviance Information Criterion for Comparing Stochastic Volatility Models". In: *Journal of Business & Economic Statistics* 22.1, pp. 107–120. ISSN: 0735-0015. DOI: 10.1198/07350010328861943O (cit. on p. 112).

Bolger, Niall et al. (1989). "Effects of daily stress on negative mood." In: *Journal of personality and social psychology* 57.5, pp. 808–818. ISSN: 0022-3514. DOI: 10.1037/0022-3514.57.5.808 (cit. on pp. 46, 107).

Both, Fiemke et al. (2008). "Modeling the Dynamics of Mood and Depression". In: *Proceedings of the 18th European Conference on Artificial Intelligence, ECAI'08*. Ed. by M. Ghallab et al., pp. 266–270 (cit. on pp. 49, 53, 105).

Bright, Tiffani J. et al. (2012). "Effect of Clinical Decision-Support Systems: A Systematic Review". In: *Ann Intern Med* April, pp. 0003–4819–157–1–201207030–00450–. ISSN: 0003-4819. DOI: 10.1059/0003-4819-157-1-201207030-00450 (cit. on p. 105).

Buntrock, Claudia et al. (2014). "Evaluating the efficacy and cost-effectiveness of web-based indicated prevention of major depression: design of a randomised controlled trial." In: *BMC psychiatry* 14, pp. 25–34. ISSN: 1471-244X. DOI: 10.1186/1471-244X-14-25 (cit. on pp. 103, 104, 106).

Byrne, A. and D.G. Byrne (1993). "The effect of exercise on depression, anxiety and other mood states: A review". In: *Journal of Psychosomatic Research* 37.6, pp. 565–574. ISSN: 00223999. DOI: 10.1016/0022-3999(93)90050-P (cit. on pp. 105, 107, 114).

Caspersen, C. J., K. E. Powell, and G. M. Christenson (1985). "Physical activity, exercise, and physical fitness: definitions and distinctions for health-related research". In: *Public Health Rep* 100.2, pp. 126–131 (cit. on p. 107).

Chih, Ming Yuan et al. (2014). "Predictive modeling of addiction lapses in a mobile health application". In: *Journal of Substance Abuse Treatment* 46.1, pp. 29–35. ISSN: 07405472. DOI: 10.1016/j.jsat.2013.08.004 (cit. on pp. 50, 54, 65, 105).

Christensen, R H B (2015). *Ordinal - Regression Models for Ordinal Data* (cit. on p. 110).

Clark, L and D Watson (1988). "Mood and the mundane: Relations between daily life events and self-reported mood". In: *Journal of personality and social psychology* 54, pp. 296–308 (cit. on pp. 105, 108, 114).

Cropley, Mark and Frh Zijlstra (2011). "Work and rumination". In: *Handbook of stress in the occupations*. Ed. by Janice Langan-Fox and Cary L. Cooper. Cheltenham, UK: Edward Elgar Publishing. Chap. 24, pp. 487–499. ISBN: 978-0-85793-114-6. DOI: 10.4337/9780857931153.00061 (cit. on pp. 106, 115).

David, J et al. (1997). "Differential roles of neuroticism, extraversion, and event desirability for mood in daily life: An integrative model oftop-down and bottom-up influences". In: *Journal of personality and social psychology* 73, pp. 149–159 (cit. on pp. 108, 114).

Demerouti, Evangelia et al. (2012). "Work-related flow and energy at work and at home: A study on the role of daily recovery". In: *Journal of Organizational Behavior* 33.2, pp. 276–295. DOI: 10.1002/job.760 (cit. on p. 106).

Dinges, D F et al. (1997). "Cumulative sleepiness, mood disturbance, and psychomotor vigilance performance decrements during a week of sleep restricted to 4-5 hours per night." In: *Sleep* 20.4, pp. 267–277. ISSN: 0161-8105 (cit. on p. 107).

Donaldson, C. and D. Lam (2004). "Rumination, mood and social problem-solving in major depression". In: *Psychol Med* 34.7, pp. 1309–1318 (cit. on p. 105).

Ebert, David Daniel et al. (2014). "GET.ON Mood Enhancer: efficacy of Internet-based guided self-help compared to psychoeducation for depression: an investigator-blinded randomised controlled trial." In: *Trials* 15.1, p. 39. ISSN: 1745-6215. DOI: 10.1186/1745-6215-15-39 (cit. on pp. 105, 106).

Elman, J L (1990). "Finding Structure in Time". In: *Cognitive Science* 14.2, pp. 179–211. ISSN: 03640213. DOI: 10.1207/s15516709cog1402_1 (cit. on pp. 108, 109).

Eysenbach, G (2001a). "What is e-health?" In: *Journal of Medical Internet Research* 3.2, e20. DOI: 10.2196/jmir.3.2.e20 (cit. on p. 104).

Farewell, V. T. (1982). "A note on regression analysis of ordinal data with variability of classification". In: *Biometrika* 69.3, pp. 533–538. ISSN: 00063444. DOI: 10.1093/biomet/69.3.533 (cit. on p. 111).

Feinerer, I, K Hornik, and D Meyer (2008). "Text Mining Infrastructure in R". In: *Journal of Statistical Software* 25.5, pp. 1–54 (cit. on p. 108).

Gable, S L, H T Reis, and a J Elliot (2000). "Behavioral activation and inhibition in everyday life." In: *Journal of personality and social psychology* 78.6, pp. 1135–1149. ISSN: 0022-3514. DOI: 10.1037//0022-3514.78.6.1135 (cit. on p. 110).

Garg, A.X. et al. (2005). "Effects of Computerized Clinical Decision Support Systems on Practitioner Performance and Patient Outcomes". In: *Journal of the American Medical Association* 293.10, pp. 1223–1238. ISSN: 1538-3598 (Electronic) 0098-7484 (Linking). DOI: 10.1001/jama.293.10.1223 (cit. on p. 105).

Grosscup, S J and P M Lewinsohn (1980). *Unpleasant and pleasant events, and mood.* (Cit. on p. 105).

Gustavsson, Anders et al. (2011). "Cost of disorders of the brain in Europe 2010". In: *European Neuropsychopharmacology* 21.10, pp. 718–779. ISSN: 0924977X. DOI: 10.1016/j.euroneuro.2011.08.008 (cit. on p. 103).

Hah, Hyeyoung and Anandhi Bharadwaj (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) (cit. on p. 105).

Hannan, Corinne et al. (2005). "A lab test and algorithms for identifying clients at risk for treatment failure". In: *Journal of Clinical Psychology* 61.2, pp. 155–163. ISSN: 00219762. DOI: 10.1002/jclp.20108 (cit. on p. 105).

Hippisley-Cox, J et al. (2008). "Predicting cardiovascular risk in England and Wales: prospective derivation and validation of QRISK2". In: *BMJ* 336.7659, pp. 1475–1482 (cit. on p. 105).

Hornik, Kurt (2016). "NLP: Natural Language Processing Infrastructure." In: (cit. on p. 108).

Iida, M et al. (2012). "Using Diary Methods in Psychological Research". In: *APA Handbook of Research Methods in Psychology: Vol. 1. Foundations, Planning, Measures and Psychometrics*. Ed. by H. Cooper et al. Washington, DC: US: American Psychological Association, pp. 277–305. ISBN: 1-4338-1004-2. DOI: 10.1037/13619-016 (cit. on pp. 76, 77, 104).

Ingram, Rick E. and Timothy W. Smith (1984). "Depression and internal versus external focus of attention". In: *Cognitive Therapy and Research* 8.2, pp. 139–151. ISSN: 01475916. DOI: 10.1007/BF01173040 (cit. on p. 107).

Isen, Alice M., Kimberly a. Daubman, and Gary P. Nowicki (1987). "Positive affect facilitates creative problem solving." In: *Journal of Personality and Social Psychology* 52.6, pp. 1122–1131. ISSN: 0022-3514. DOI: 10.1037/0022-3514.52.6.1122 (cit. on p. 104).

Jardim, Sandra (2013). "The Electronic Health Record and its Contribution to Healthcare Information Systems Interoperability". In: *Procedia Technology* 9, pp. 940–948 (cit. on p. 104).

Johnson, Timothy R. (2003). "On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style". In: *Psychometrika* 68.4, pp. 563–583. ISSN: 0033-3123. DOI: 10.1007/BF02295612 (cit. on p. 111).

Kanning, Martina and Wolfgang Schlicht (2010). "Be active and become happy: an ecological momentary assessment of physical activity and mood." In: *Journal of sport & exercise psychology* 32.2, pp. 253–261. ISSN: 0895-2779 (cit. on pp. 107, 114).

Karafi, Martin and Stefan Kombrink (2011). "Recurrent Neural Network based Language Modeling in Meeting Recognition". In: *Interspeech* 3.January, pp. 2877–2880 (cit. on p. 109).

Kramer, Adam D I, Jamie E Guillory, and Jeffrey T Hancock (2014). "Experimental evidence of massive-scale emotional contagion through social networks". In: *Proceedings of the National Academy of Sciences of the United States of America* 111.24, pp. 8788–8790 (cit. on p. 105).

Landis, J Richard and Gary G Koch (2008). "The Measurement of Observer Agreement for Categorical Data". In: *International Biometric Society* 33.1, pp. 159–174. ISSN: 0006341X. DOI: 10.2307/2529310 (cit. on p. 109).

Larsen, Randy J. and Gregory S. Cowan (1988). "Internal focus of attention and depression: A study of daily experience". In: *Motivation and Emotion* 12.3, pp. 237–249. ISSN: 01467239. DOI: 10.1007/BF00993113 (cit. on p. 107).

Leger, D (1994). "The cost of sleep-related accidents: a report for the National Commission on Sleep Disorders Research." In: *Sleep* 17.1, pp. 84–93. ISSN: 0161-8105 (cit. on p. 103).

Lewinsohn, Peter M. and Christopher S. Amenson (1978). "Some relations between pleasant and unpleasant mood-related events and depression." In: *Journal of abnormal psychology* 87.6, pp. 644–654. ISSN: 0021-843X. DOI: 10.1037/0021-843X.87.6.644 (cit. on p. 105).

Liu, Xing and Hari Koirala (2012). "Ordinal regression analysis: Using generalized ordinal logistic regression models to estimate educational data". In: *Journal of Modern Applied Statistical Methods* 11.1, pp. 242–254. ISSN: 15389472 (cit. on p. 110).

McCorkle, R and J Quint-Benoliel (1983). "Symptom distress, current concerns and mood disturbance after diagnosis of life-threatening disease". In: *Social science & medicine* 17.7, pp. 431–438. ISSN: 02779536. DOI: 10.1016/0277-9536(83)90348-9 (cit. on p. 107).

McCullagh, Peter (1980). "Regression Models for Ordinal Data". In: *Journal of the Royal Statistical Society* 42.2, pp. 109–142 (cit. on p. 110).

McMahon, Francis J. (2014). "Prediction of treatment outcomes in psychiatry-where do we stand?" In: *Dialogues in Clinical Neuroscience* 16.4, pp. 455–464. ISSN: 12948322. DOI: 10.1164/rccm.200408-1036SO. arXiv: arXiv:1011.1669v3 (cit. on p. 105).

Minden, S L (2000). "Mood disorders in multiple sclerosis: diagnosis and treatment." In: *Journal of neurovirology* 6.2, pp. 160–167. ISSN: 1355-0284 (cit. on p. 104).

Nadler, Ruby T, Rahel Rabi, and John Paul Minda (2010). "Better mood and better performance. Learning rule-described categories is enhanced by positive mood." In: *Psychological science : a journal of the American Psychological Society / APS* 21.12, pp. 1770–1776. ISSN: 0956-7976. DOI: 10.1177/0956797610387441 (cit. on p. 104).

Nesse, R M (2000). "Is depression an adaptation?" In: *Archives of general psychiatry* 57.1, pp. 14–20. ISSN: 0003990X. DOI: 10.1001/archpsyc.58.11.1086-a (cit. on p. 105).

Netz, Yael and Ronnie Lidor (2003). "Mood alterations in mindful versus aerobic exercise modes." In: *The Journal of psychology* 137.5, pp. 405–419. ISSN: 0022-3980. DOI: 10.1080/00223980309600624 (cit. on pp. 107, 114).

Nolen-Hoeksema, Susan and Jannay Morrow (1993). "Effects of rumination and distraction on naturally occurring depressed mood". In: *Cognition & Emotion* 7.6, pp. 561–570. ISSN: 0269-9931. DOI: 10.1080/02699939308409206 (cit. on p. 107).

Norusis, Marija J. (2010). "Ordinal Regression". In: *PASW Statistics 18.0 Advanced Statistical Procedures Companion*. Prentice Hall. Chap. 4, pp. 69–89 (cit. on p. 110).

Penninx, BW et al. (2001). "Depression and cardiac mortality: results from a community-based longitudinal study". In: *Arch Gen Psychiatry* 58.3, pp. 221–227 (cit. on p. 105).

Peterson, Bercedis and Frank E. Harrell (1990). "Partial Proportional Odds Models for Ordinal Response Variables". In: *Journal of the Royal Statistical Society* 39.2, pp. 205–217. DOI: 10.2307/2347760 (cit. on p. 110).

Plummer, Martyn (2003). "{JAGS}: A program for analysis of {Bayesian} graphical models using {Gibbs} sampling". In: *Proceedings of the 3rd International Workshop on Distributed Statistical Computing* (cit. on p. 111).

Qian, Xinyi, Careen Yarnal, and David M. Almeida (2014). "Does leisure time moderate or mediate the effect of daily stress on positive affect? An examination using eight-day diary data". In: *Journal of Leisure Research* 46.1, pp. 106–124 (cit. on p. 106).

Reis, Harry T et al. (2000). "Daily Well-Being: The Role of Autonomy, Competence, and Relatedness". In: *Personality and Social Psychology Bulletin* 26, pp. 419–435 (cit. on p. 105).

Richardson, S. (2002). "Discussion of a paper by D. J. Spiegelhalter et al." In: *J.R. Statist. Soc.* B.64, pp. 626–7 (cit. on p. 112).

Robinson, R G et al. (1984). "Mood disorders in stroke patients. Importance of location of lesion." In: *Brain : a journal of neurology* 107 (Pt 1, pp. 81–93. ISSN: 0006-8950. DOI: 10.1093/brain/107.1.81 (cit. on p. 107).

Rook, John W. and Fred R. H. Zijlstra (2006). "The contribution of various types of activities to recovery". In: *European Journal of Work and Organizational Psychology* 15.2, pp. 218–240. ISSN: 1359-432X. DOI: 10.1080/13594320500513962 (cit. on p. 106).

Rosen, Ilene M et al. (2006). "Evolution of sleep quantity, sleep deprivation, mood disturbances, empathy, and burnout among interns." In: *Academic medicine* 81.1, pp. 82–85. ISSN: 1040-2446. DOI: 10.1097/00001888-200601000-00020 (cit. on p. 107).

Rossi, P E, G M Allenby, and R McCulloch (2012). *Bayesian Statistics and Marketing*. Wiley Series in Probability and Statistics. Wiley. ISBN: 9780470863688 (cit. on p. 111).

Sluiter, J K et al. (2003). "Need for recovery from work related fatigue and its role in the development and prediction of subjective health complaints". In: *Occupational and environmental medicine* 60.Suppl 1, pp. i62–i70. ISSN: 1351-0711. DOI: 10.1136/oem.60.suppl_1.i62 (cit. on p. 106).

Smyth, Joshua M and Arthur a Stone (2003). "Ecological momentary assessment research in behavioral medicine". In: *Journal of Happiness Studies* 4.1, pp. 35–52. ISSN: 1389-4978; 1573-7780. DOI: 10.1023/A:1023657221954 (cit. on pp. 46, 104).

Sonnentag, Sabine (2001). "Work, recovery activities and well-being: a diary study". In: *J Occup Health Psychol* 6.3, pp. 196–210. DOI: 10.1037/1076-8998.6.3.196 (cit. on p. 114).

Spiegelhalter, D., Nicola G. Best, and Bradley P. Carlin (1998). *Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models*. Tech. rep. , MRC Biostastatistics Unit, Cambridge (cit. on p. 112).

Spiegelhalter, David J. et al. (2002). "Bayesian measures of model complexity and fit". In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 64.4, pp. 583–616. ISSN: 13697412. DOI: 10.1111/1467-9868.00353 (cit. on p. 112).

Stewart, Wendy and Julian Barling (1996). "Daily work stress, mood and interpersonal job performance: A mediational model". In: *Work & Stress* 10.4, pp. 336–351. ISSN: 0267-8373. DOI: 10.1080/02678379608256812 (cit. on p. 106).

Stone, A (1987). "Event content in a daily survey is differentially associated with concurrent mood". In: *Journal of personality and social psychology* 52, pp. 56–58 (cit. on p. 106).

Tadic, Maja et al. (2013). "Daily Activities and Happiness in Later Life: The Role of Work Status". In: *Journal of Happiness Studies* 14.5, pp. 1507–1527. ISSN: 13894978. DOI: 10.1007/s10902-012-9392-9 (cit. on pp. 104, 106).

Thomsen, Dorthe Kirkegaard et al. (2003). "Rumination—relationship with negative mood and sleep quality". In: *Personality and Individual Differences* 34.7, pp. 1293–1301. ISSN: 01918869. DOI: 10.1016/S0191-8869(02)00120-4 (cit. on p. 107).

Wang, Feng et al. (2012). "Long-term association between leisure-time physical activity and changes in happiness: Analysis of the prospective National Population Health Survey". In: *American Journal of Epidemiology* 176.12, pp. 1095–1100. ISSN: 00029262. DOI: 10.1093/aje/kws199 (cit. on pp. 105, 107).

Watkins, E. and S. Baracaia (2001). "Why do people ruminate in dysphoric moods?" In: *Personality and Individual Differences* 30, pp. 723–734 (cit. on p. 107).

Weinstein, Sally M and Robin Mermelstein (2008). "Role of Autonomy". In: *Journal of Clinical Child and Adolescent Psychology* 36.2, pp. 182–194 (cit. on pp. 7, 103, 104, 108, 110, 114, 115).

Wilson, Vance and Nancy Lankton (2004). "Modeling Patients' Acceptance of Provider-delivered E-health". In: *J Am Med Inform Assoc* 11.4, pp. 241–248 (cit. on p. 105).

Wittchen, H. U. et al. (2011). "The size and burden of mental disorders and other disorders of the brain in Europe 2010". In: *European Neuropsychopharmacology* 21.9, pp. 655–679. ISSN: 0924977X. DOI: 10.1016/j.euroneuro.2011.07.018 (cit. on p. 103).

Yeung, Robert R (1996). "Review the Acute Effects of Exercise on Mood State". In: *Journal of psychosomatic research* 40.2, pp. 123–141. ISSN: 00223999. DOI: 10.1016/0022-3999(95)00554-4 (cit. on p. 114).

Chapter 7

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). Journal of Medical Internet Research, Volume 20, Issue 8.

Abstract: *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. 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. 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. 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%). 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.*

7.1 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., 2016b). 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, Duan, and Braslow, 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 (Breda et al., 2017). 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.

7.2 Methods

7.2.1 Data & 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 7.1 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 7.1: 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 (Kleiboer et al., 2016; Rush et al., 2003; Kroenke, Spitzer, and Williams, 2001; EuroQolGroup, 1990; Hakkaart-van Roijen et al., 2002). 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 7.1. 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 7.6 (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 7.7 and Table 7.8 (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.

7.2.2 Approach & 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 7.1. 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 predicting costs due to the challenging nature of this activity (Jones et al., 2007; Diehr et al., 1999; 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.

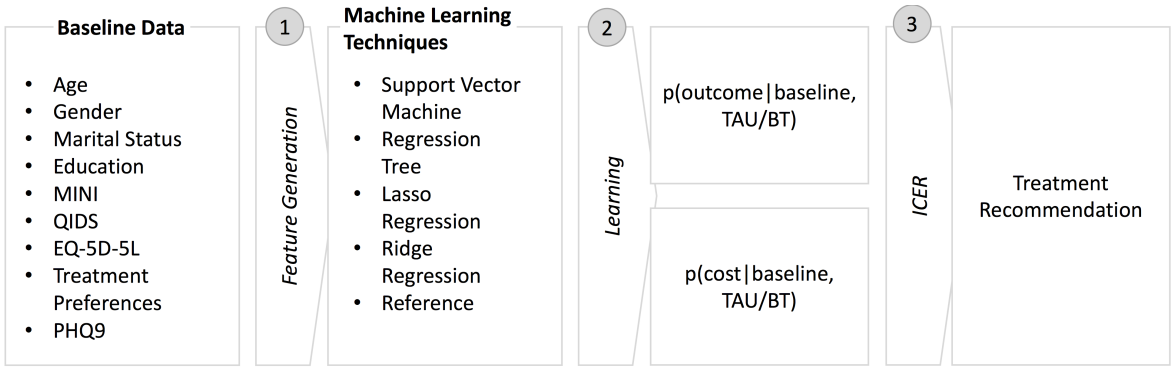


FIGURE 7.1: Process for deriving treatment recommendations for individuals. BT: blended treatment; ICER: incremental cost-effectiveness ratio; TAU: treatment as usual.

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 (Willmott, Matsuura, and Robeson, 2009; Chai and Draxler, 2014).

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:

$$ICER = \frac{(Cost_{BT} - Cost_{TAU})}{(Improvement_{BT} - Improvement_{TAU})}.$$

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, 2015b).

7.3 Results

7.3.1 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, Spitzer, and Williams, 2001). Because we only investigated the improvements for a 6-month period, these results are not final; however, they can indicate a trend. Table 7.2 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
Mean Start PHQ9	15.42	15.35
Mean End PHQ9	9.49	7.85
#Improvements	140	154
#No Improvements	38	18

Table 7.2: 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).

7.3.2 Outcome & Cost Prediction

Table 7.3 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	0.0697	0.0990	5933.847	9287.728
2	Regression Tree	0.0698	0.0992	6573.935	9406.110
3	Ridge Regression	0.0711	0.1000	6557.693	9187.775
4	Reference Measure	0.0770	0.1017	7024.11	9539.54

TABLE 7.3: 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 7.9, Table 7.10, Table 7.11, and Table 7.12 (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 7.4: 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 7.4). 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 7.2 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,

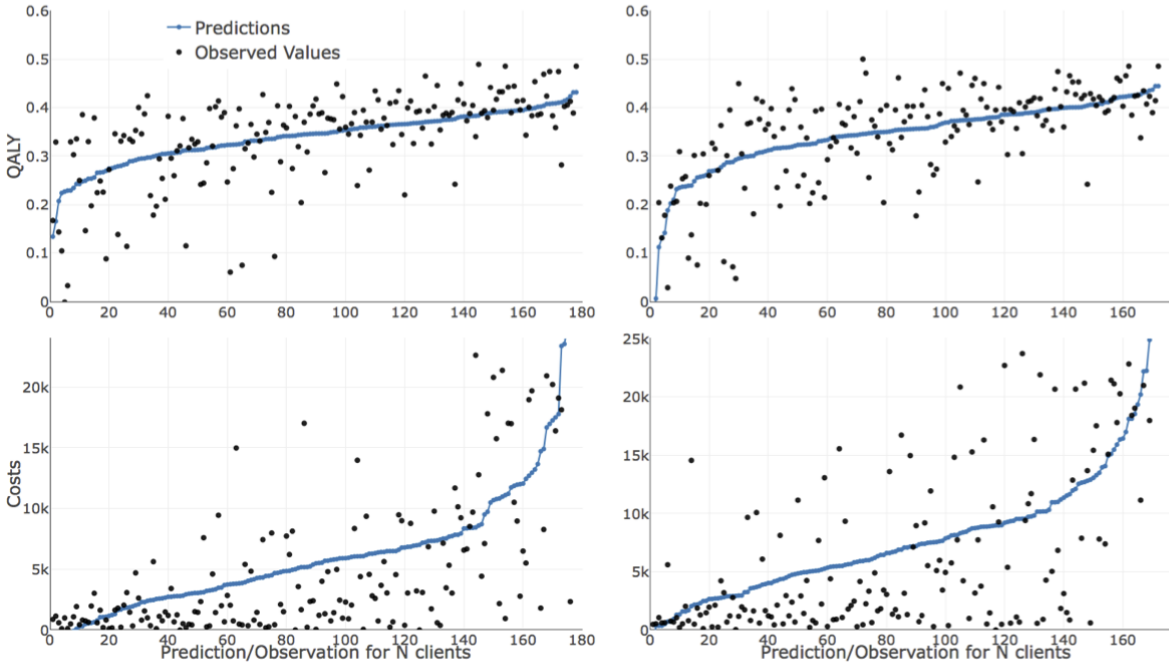


FIGURE 7.2: Predicted and observed values for QALY and costs and both treatment types (left panels for treatment as usual and right panels for blended treatment).

and as illustrated by the increased performance compared with the reference measure; this result indicates a step in the right direction.

7.3.3 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 7.3 (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 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

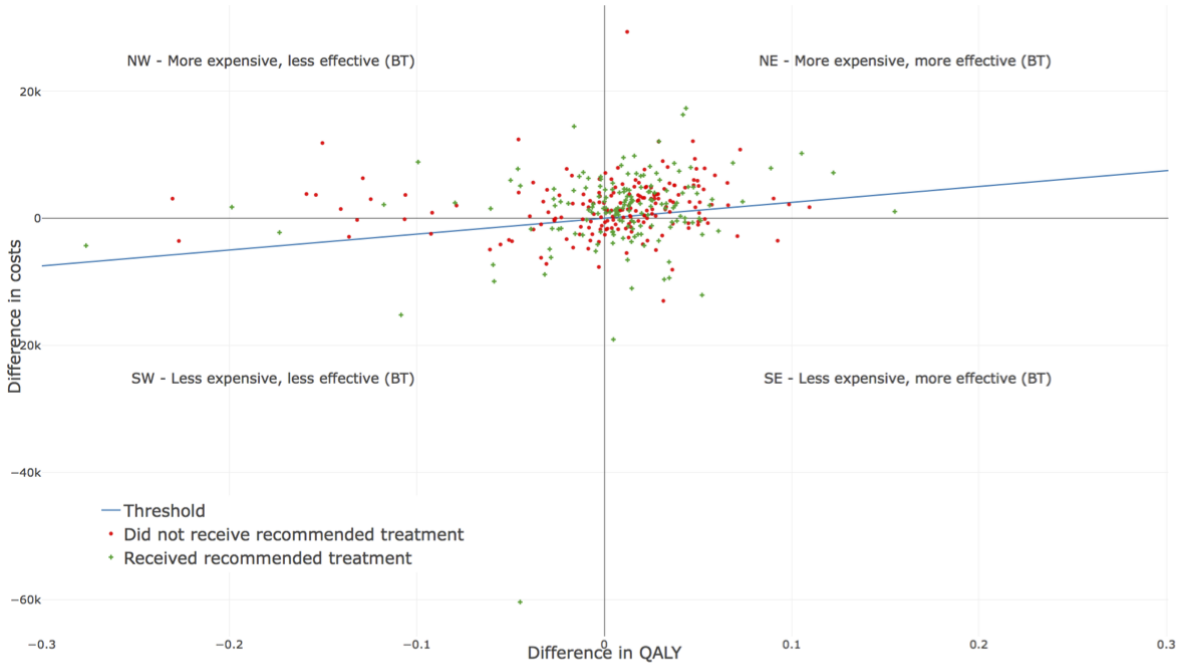


FIGURE 7.3: Expected improvement for all patients in relation to costs. The x-axis illustrates the difference in quality-adjusted life years (blended treatment - treatment as usual) and the y-axis the difference in costs (blended treatment - treatment as usual).

high differences when calculating the ICER. The point for the participant at the bottom of Figure 7.3 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 7.5 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 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 7.5: Treatment recommendation for all patients (N=350).

7.4 Discussion

7.4.1 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., 2016b). 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, Laking, and Fischer, 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.

7.4.2 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.

7.4.3 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, CJ and VJ Peel (1998). "Transforming a hospital through growing, not building, an electronic patient record system". In: *Methods Inf Med* 37.3, pp. 285–93 (cit. on p. 131).

Berg, M (2001). "Implementing information systems in health care organizations: myths and challenges". In: *International Journal of Medical Informatics* 64.2-3, pp. 143–56 (cit. on p. 131).

Black, W C (1990). "The CE plane: a graphic representation of cost-effectiveness". In: *Medical Decision Making* 10.3, pp. 212–4. ISSN: 0272-989X. DOI: 10.1177/0272989X9001000308 (cit. on pp. 126, 128).

Breda, Ward van et al. (2017). "Assessment of temporal predictive models for health care using a formal method". In: *Computers in Biology and Medicine* 87.November 2016, pp. 347–357. ISSN: 00104825. DOI: 10.1016/j.compbiomed.2017.06.014 (cit. on pp. 122, 141, 143).

Burges, Chris (1998). "A Tutorial on Support Vector Machines for Pattern Recognition". In: *Data Mining and Knowledge Discovery* 2.2, pp. 121–167. ISSN: 13845810. DOI: 10.1023/A:1009715923555. arXiv: 1111.6189v1 (cit. on pp. 79, 125).

Chai, T. and R. R. Draxler (2014). "Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature". In: *Geoscientific Model Development* 7.3, pp. 1247–1250. ISSN: 19919603. DOI: 10.5194/gmd-7-1247-2014 (cit. on p. 125).

Diehr, P et al. (1999). "Methods for analyzing health care utilization and costs". In: *Annu. Rev. Public Health* 20, pp. 125–144 (cit. on p. 124).

Dunlop, BW (2015). "Prediction of treatment outcomes in major depressive disorder." In: *Expert review of clinical pharmacology* 8.6, pp. 669–72. ISSN: 1751-2441. DOI: 10.1586/17512433.2015.1075390 (cit. on p. 123).

EuroQolGroup (1990). "EuroQol-a new facility for the measurement of health-related quality of life". In: *Health policy* 16.3, pp. 199–208 (cit. on p. 123).

Hakkaart-van Roijen, L et al. (2002). *Trimbos/iMTA Questionnaire for Costs Associated with Psychiatric Illness (TiC-P)*. Institute for Medical Technology Assessment, Erasmus University (cit. on pp. 123, 124).

Hakkaart-van Roijen, L et al. (2015). *Kostenhandleiding: Methodologie van kostenonderzoek en referentieprijzen voor economische evaluaties in de gezondheidszorg* (cit. on p. 124).

Hartswood, M et al. (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*, pp. 96–105 (cit. on p. 131).

Ioannidis, John P A and Alan M. Garber (2011). "Individualized cost-effectiveness analysis". In: *PLoS Medicine* 8.7, e1001058 (cit. on pp. 122, 123).

Jones, J et al. (2007). "Predicting costs of mental health care: a critical literature review". In: *Psychological Medicine* 37.4, pp. 467–477. DOI: 10.1017/S0033291706009676 (cit. on pp. 122, 124).

Kautzky, A. et al. (2018). "Refining Prediction in Treatment-Resistant Depression: Results of Machine Learning Analyses in the TRD III Sample". In: *J Clin Psychiatry* 79.1, p. 16m11385 (cit. on p. 122).

Kleiboer, Annet et al. (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". In: *Trials* 17.1,

p. 387. ISSN: 1745-6215. DOI: 10.1186/s13063-016-1511-1 (cit. on pp. 6, 89, 91, 123, 126, 131, 141, 144).

Klok, R.M. and M.J. Postma (2004). "Four quadrants of the cost-effectiveness plane: Some considerations on the south-west quadrant". In: *Expert Review of Pharmacoeconomics and Outcomes Research* 4.6, pp. 599–601. ISSN: 1473-7167. DOI: 10.1586/14737167.4.6.599 (cit. on p. 128).

Knapp, M (1999). "Economic Evaluation and Mental Health : Sparse Past . . . Fertile Future ?" In: *The Journal of Mental Health Policy and Economics* 2.4, pp. 163–167 (cit. on pp. 7, 122).

Kolovos, S. et al. (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". In: *Journal of affective disorders* 200, pp. 284–292 (cit. on p. 124).

Kravitz, R L, N Duan, and J Braslow (2004). "Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages". In: *Milbank Q* 82.4, pp. 661–87 (cit. on p. 122).

Kroenke, K, RL Spitzer, and JB Williams (2001). "The PHQ-9: validity of a brief depression severity measure". In: *J Gen Intern Med* 16.9, pp. 606–13 (cit. on pp. 89, 92, 123, 126).

Lord, J, G Laking, and A Fischer (2006). "Non-linearity in the cost-effectiveness frontier". In: *Health Econ* 15.6, pp. 565–77 (cit. on p. 130).

Manning, Willard G and John Mullahy (2001). "Estimating log models : to transform or not to transform?" In: *Journal of Health Economics* 20.4, pp. 461–94 (cit. on p. 124).

National Institute for Health and Care Excellence (NICE) (2013). "Guide to the methods of technology appraisal". In: *Process and Methods Guides* 9 (cit. on p. 128).

O'Hagan, Anthony and John W Stevens (2002). "The probability of cost-effectiveness". In: *BMC Medical Research Methodology* 2, p. 5 (cit. on p. 123).

R Core Team (2015b). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria (cit. on p. 126).

Rush, A John et al. (2003). "The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression". In: *Biological psychiatry* 54.5, pp. 573–83 (cit. on p. 123).

Russell, LB et al. (1996). "The role of cost-effectiveness analysis in health and medicine". In: *JAMA* 276.4, pp. 1172–77 (cit. on p. 122).

Ryder, HF et al. (2009). "Decision Analysis and Cost-effectiveness Analysis". In: *Semin Spine Surg* 21.4, pp. 216–222. DOI: 10.1053/j.semss.2009.08.003.Decision (cit. on pp. 7, 122, 124).

Sculpher, Mark (2015). "Clinical trials provide essential evidence, but rarely offer a vehicle for cost-effectiveness analysis". In: *Value in Health* 18.2, pp. 141–142. ISSN: 15244733. DOI: 10.1016/j.jval.2015.02.005 (cit. on p. 122).

Sim, I et al. (2001). "Clinical decision support systems for the practice of evidence-based medicine". In: *Journal of the American Medical Association* 8.6, pp. 527–534 (cit. on p. 131).

Tibshirani, Robert (1996). *Regression Selection and Shrinkage via the Lasso*. DOI: 10.2307/2346178. arXiv: 11/73273 [1369-7412] (cit. on pp. 9, 79, 125).

Triñanes, Yolanda et al. (2015). "Development and impact of computerised decision support systems for clinical management of depression: A systematic review". In: *Revista de Psiquiatría y Salud Mental (English Edition)* 8.3, pp. 157–166. ISSN: 21735050. DOI: 10.1016/j.rpsmen.2015.05.004 (cit. on pp. 3, 122).

Van Breda, Ward et al. (2016b). "A feature representation learning method for temporal datasets". In: *IEEE Symposium Series on Computational Intelligence*, pp. 1–8. DOI: 10.1109/SSCI.2016.7849890 (cit. on pp. 122, 130).

Versteegh, Matthijs et al. (2016). "Dutch Tariff for the Five-Level Version of EQ-5D". In: *Value in Health* 19.4, pp. 343–352. ISSN: 15244733. DOI: 10.1016/j.jval.2016.01.003 (cit. on p. 124).

Willmott, Cort J., Kenji Matsuura, and Scott M. Robeson (2009). "Ambiguities inherent in sums-of-squares-based error statistics". In: *Atmospheric Environment* 43.3, pp. 749–752. ISSN: 13522310. DOI: 10.1016/j.atmosenv.2008.10.005 (cit. on p. 125).

7.5 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)
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)
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)

Item No.	Item Description
97	Who is providing this other medication (atreat13)

TABLE 7.6: 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	0.104	6471.05	9346.4
3	Ridge Regression	0.0647	0.087	6044.44	8395.56

TABLE 7.7: 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 7.8: 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
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 7.9: 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

Feature	Parameter Coefficient

TABLE 7.10: 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
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
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 7.11: 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
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

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

Part III

Model evaluation and parameter estimation

The following part presents an evaluation of a predictive model and introduces a method for model parameter estimation. The evaluation of predictive models for decision making is mandatory to ensure model validity and prevent mistakes. Accordingly, a model evaluation that combines study data and simulation analysis is presented. Predictive models can provide client individual predictions that can path the way to personalized treatment. Client individual models imply the necessity to estimate model parameters for each client, which can be computationally demanding. Therefore, the second article in this part presents a method for model parameter estimation applicable to mobile devices.

Chapter 8

Evaluation of a temporal causal model for predicting the mood of clients in an online therapy

Becker, D., Bremer, V., Funk, B., Hoogendoorn, M., Rocha, A., and Riper, H. (2020). Evidence-based mental health, Volume 23, Issue 1.

Abstract: *Self-reported client assessments during online treatments enable the development of statistical models for the prediction of client improvement and symptom development. Evaluation of these models is mandatory to ensure their validity. For this purpose, we suggest besides a model evaluation based on study data the use of a simulation analysis. The simulation analysis provides insight into the model performance and enables to analyze reasons for a low predictive accuracy. In this study, we evaluate a temporal causal model (TCM) and show that it does not provide reliable predictions of clients' future mood levels. Based on the simulation analysis we investigate the potential reasons for the low predictive performance, e.g. noisy measurements and sampling frequency. We conclude that the analyzed TCM in its current form is not sufficient to describe the underlying psychological processes. The results demonstrate the importance of model evaluation and the benefit of a simulation analysis.*

8.1 Introduction

Mobile devices provide new possibilities to deliver Internet-based cognitive behavioral therapy (ICBT) (De Graaf et al., 2009; Lambert, 2012) and to measure clients' mental health, behavior, and activities (Aung, Matthews, and Choudhury, 2017; Mohr, Zhang, and Schueller, 2017). Ecological momentary assessment (EMA) is the term used to describe the assessment of clients' mood and behavior throughout the day in their natural environment (Robinson and Clore, 2002; Stone and Shiffman, 1994; Aan het Rot, Hogenelst, and Schoevers, 2012). EMA can encompass a diversity of data such as diaries, open-text, and questions regarding the clients' symptoms and experiences using Likert-scaled responses (Gibbons, 2017). These collected time series data are a gateway to model symptom interaction and understand the psychological dynamics that occur in individuals over time (Fisher and Boswell, 2016; Bak et al., 2016). The collected EMA data provides patterns, which can be used to model relationships between symptoms and predict clients' future wellbeing.

With the increasing development of predictive models for both diagnostic and prognostic predictions, there is an intensified interest in the methodology on model evaluation (Moons et al., 2009b; Ivanescu et al., 2016). Besides statistical model evaluation utilizing study data, clinical evaluation is required (Kappen et al., 2018; Altman et al., 2009). However, clinical model evaluation requires substantial effort and money, therefore only a small portion of

available models can be evaluated in practice (Moons et al., 2009a; Steyerberg et al., 2013). Guidelines for model development and statistical evaluation provide methods to improve their validity and identify invalid models early. These guidelines encompass the definitions of prediction targets, predictors, statistical model evaluation, and reporting (Steyerberg and Vergouwe, 2014). Suboptimal adherence to evaluation guidelines can limit the reliability and applicability of predictive models (Bouwmeester et al., 2012). For the statistical model evaluation, such guidelines suggest the use of cross-validation and bootstrapping (Altman et al., 2009; Steyerberg and Vergouwe, 2014). Cross-validation indicates the expected model performance of unobserved samples from the same studyin practice and bootstrapping allows to infer the significance of parameters and variance of predictions. We, however, argue that these methods are not sufficient for statistical model evaluation. Therefore, we suggest the inclusion of a simulation analysis. The simulation analysis is used to estimate the expected model performance on study data and the model's sensitivity regarding changes in the data. Accordingly, the simulation can be utilized to investigate specific reasons for poor model prediction and provide insights into the models' prediction performance under varying study conditions. Furthermore, treatment decisions can consider multiple objectives, where an improvement in one dimension might lead to a reduction in another. Likewise, predictive models become increasingly complex and allow to predict multiple objectives simultaneously. Therefore, methods for comparing these models are required as well.

In this paper, we demonstrate a thorough model evaluation combining performance estimation on study data and simulation analysis. The examined predictive model is the so-called social integration model (SIM) (Altaf Hussain Abro, 2016), which in a preliminary model evaluation was suggested to provide reliable predictions for clients' future mood levels. It describes the relationship between social interactions of study participants and their mental well-being. The SIM is a temporal-causal model (TCM) (Treur, 2016; Araújo and Treur, 2016) that allows to predict the course of multiple EMA factors. In general, temporal causal models are continuous dynamic network models that describe a graph of connected states using differential equations. They allow to universally model any dynamic system, simulate its change over time, and have been shown to be applicable for a wide range of domains (Treur and Ziabari, 2018; Naze and Treur, 2011; Abro and Treur, 2017; Franke and Hosain, 2017).

For the estimation of the model performance on study data, we use the complete EMA data collected in a Europe-wide depression study (Kleiboer et al., 2016). A framework for temporal causal model comparisons Breda et al. (2017) is used for performance estimation and comparison to a reference model. The framework further allows to compare the performance with respect to all EMA measures and utilizes client individual model parameters, which have been shown to provide more accurate results than using the same model parameters for each client (Mikus et al., 2017; Jaques et al., 2017; Constantinides et al., 2018). The simulation analysis allows to estimate the theoretical model performance on the study dataset and to investigate reasons for differences in the performance. We investigated the literature regarding potential downsides of EMA measures and identified the influence of measurement noise (Sudman, Bradburn, and Schwarz, 1996; Clark and Schober, 1992) and too few factor assessments (Courvoisier, Eid, and Lischetzke, 2012; Sokolovsky, Mermelstein, and Hedeker, 2014; Papageorgiou et al., 2018) as possible reasons that can lead to low model performances. Therefore, we utilize the simulation analysis to investigate these reasons and use the study data to inform the simulation analyses. By systematically altering the noise and missing values on the simulated data, the model's sensitivity to these influences can be assessed. Specifically, for the analysis of noise, data with the same assessment frequency as the study data is generated. To analyze the influence of fewer EMA assessments, the EMA measures' standard deviation from the study data is used for data generation.

This study demonstrates a thorough model evaluation. We show that besides results

obtained on study data with utilizing client individual model parameters and a comparison to a reference model, a simulation analysis to assess the model's robustness for varying conditions is required. The simulation analysis is designed to reflect the study conditions and allows to infer the theoretical model performance that would be expected on the study data. If the results obtained in both analyses contradict each other, then there is reason to believe that the model does not represent the underlying dynamics accurately. The simulation further allows to investigate and eliminate possible reasons for the obtained differences in both analyses. By employing these methods, models that do not provide reliable predictions can be identified early.

8.2 Method

8.2.1 Social integration modeling

The social integration model (Altaf Hussain Abro, 2016) is a TCM that describes the relationship between clients' social contact and mood and allows to simulate their future behavior. Social integration exhibits a relation to wellbeing and social isolation can foster mental health issues (Nicholson, 2009). People that are socially well-integrated and have more social contact are usually happier than individuals with limited social contacts (Gariépy, Honkaniemi, and Quesnel-Vallée, 2016). Contrary, people with depression tend to report feelings of loneliness, are less likely to engage in social activities, and have fewer social contacts (Cacioppo et al., 2002). These relationships provide the basis for the model and an overview of the interaction is shown in Figure 8.1.

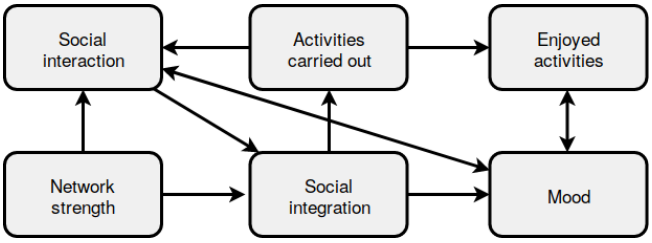


FIGURE 8.1: Visualization of the social integration model

In the social integration model, the mood level is influenced by the daily social interactions and how well the person is integrated into their social network. Specifically, the number of social activities and the perception of how enjoyable these activities have been. People with depression often perceive social interactions as less rewarding while interactions with close friends are more valued (Nezlek, Hampton, and Shean, 2000; Nezlek, Imbrie, and Shean, 1994). Likewise, the daily mood level can influence the motivation to conduct activities and the perception of social interactions. A low mood level sensitizes peoples' perception regarding social interactions and they are more likely to emphasize negative social interactions (Steger and Kashdan, 2010). This can create a vicious cycle, where a lower mood can lead to fewer conducted activities and less social contacts, which eventually can result in low model levels and depression (Cacioppo, Hawkley, and Thisted, 2010; Taylor et al., 2018).

The strength of the influence among the different factors, which are the model's parameters, can significantly vary between healthy and unhealthy individuals and among them in general (Spanakis et al., 2016; Bremer, Funk, and Riper, 2019). Therefore, the model parameters are estimated for each client to provide client individual predictions. The model predictions are derived for each EMA factor on a daily basis. This results in a multi-objective optimization problem for the parameter estimation and prediction error estimation for each measured EMA factor.

8.2.2 Evaluation of temporal predictive models

For model evaluation, we utilized a framework proposed by Breda et al. (2017). This framework assesses temporal models in the domain of mental health and allows to infer performance measures. These performance measures consider the models' fit and prediction of all EMA measures, and provide a comprehensive summary by considering each objective. It, therefore, provides methods to compare multi-objective models.

To infer these performance measures, client-specific model parameters are estimated by minimizing the root mean square error of each objective. This results in a multi-objective optimization problem. For solving this optimization problem, the framework utilizes the NSGA2 (non-dominated sorting genetic algorithm II) (Deb et al., 2002) optimization algorithm. NSGA2 is a genetic algorithm for multi-objective optimization that optimizes with respect to each model objective. This process results in a set of valid solutions. Each solution provides an error for each optimized objective which can be considered as a point in Euclidean space. These points represent the Pareto optimal front, where an improvement in one dimension leads to a reduction in another dimension. Therefore, optimizing one dimension leads to a trade-off among the conflicting optimization objectives. To summarize how well the optimization objective has been met, the dominated hyper-volume measures the multidimensional volume covered by the Pareto optimal front and a reference point. Specifically, the dominated hyper-volume provides a measure of how well the data can be approximated by the model.

This dominated hyper-volume is utilized by the framework for the estimation of a models' *descriptive performance*. The *descriptive performance* is calculated by normalizing the mean over all dominated hyper-volumes resulting from the estimated parameters for each client by its standard deviation. In other words, this performance measure indicates the models' fit to the data with respect to all objectives.

Similarly, a *predictive performance* is estimated by combining the prediction error of unobserved data and their correlation to the models' fit of the data. To represent both performances, the *predictive performance* consists of the mean of the absolute and relative predictive performance. The absolute predictive performance describes the fit of unobserved future data using the estimated parameters. For deriving this measure, the prediction error over all optimization objectives and estimated models is summarized by the mean and variance of the prediction error. It provides an estimate of how well the model represents future data. The relative predictive performance is calculated using the correlation between the fitting error and the prediction error. This measure ensures that models with a low fitting and prediction error receive a high relative predictive performance. In summary, the *descriptive performance* describes the fit of the model to the data and the *predictive performance* the models' capability of predicting future data. The performance measures are bound between 0 and 1, where a higher value indicates better performance. The performance measures provided by the framework are point estimates, therefore, we applied bootstrapping for confidence interval estimation.

Regarding the framework parameters, we chose a population size of 80 and 100 generations for NSAG2 for individual model parameter estimation. The remaining parameters for the NSGA2 algorithm were set in accordance with the suggestion of the framework (crossover probability of 0.7 and a mutation probability of 0.2). We inspected the improvement of the dominated hyper-volume over the consecutive generations to confirm that the algorithm converges to a stable solution, which ensures a robust performance measure estimation. Further, we chose to execute 5 independent runs for each client, which reduces variability in the estimated performances. For more information regarding this framework, see Breda et al. (2017).

8.2.3 Experimental setting and data

The utilized EMA data for model evaluation in this study originates from the EU funded project E-COMPARED (Kleiboer et al., 2016), which compared bCBT (blended cognitive behavior therapy, experiment group) and face-to-face treatment (control group) for depression. For the participants in the experimental group, EMA data was gathered and CBT was provided using a mobile phone between February 2015 and February 2018. An overview of the analyzed data is provided in Table 8.1.

Concept	Assessment question	Assessment frequency
Mood	How is your mood right now?	Daily
Activities carried out	To what extent have you carried out enjoyable activities today?	Daily during first and last week; random day in other weeks
Enjoyed activities	How much did you enjoy activities today?	First and last week; random day in other weeks
Social interaction	How much were you involved in social interactions today?	Daily during first and last week; random day in other weeks

Table 8.1: Utilized EMA data

The EMA measures have been assessed with varying granularity over the treatment period. The factor mood was inquired once a day at a random time between 10am and 10pm. The remaining factors were prompted once a week on a random day. However, during the first and last week of the treatment, the assessment was intensified and all factors were inquired daily. Since the study was conducted in eight different European countries, the suggested application use period differed among the participating countries and ranged between 6 and 20 weeks. Furthermore, clients were allowed to contribute mood measures at any time which allows for additional measures besides the one defined in the assessment protocol.

Study data analysis

To assess the model's performance on the study data, we chose a period of 6 weeks for model training and aimed to predict the data of the 7th week of treatment. Considering the prediction of the whole upcoming week might be a suitable time frame for therapists to inform short-term treatment decisions such as identifying drop-out risk or maximizing short-term outcomes. The dataset consists of 324 clients. However, by considering the first 7 weeks of treatment for all clients, 112 clients provide data for model training and prediction error estimation. In the case of multiple mood measures a day, the corresponding mean value for that day was used. Afterward, all EMA data was normalized between 0 and 1.

For deriving performance measures on the study data, we utilized the framework for model comparison to compute the descriptive and predictive performance of the social integration model for each client. These performances summarize the models' capabilities with respect to each EMA measure and allowed a comparison of the models' descriptive and predictive power to a reference model. This reference model, called *mean model* here, was defined as the mean value of the training data for each clients' psychological factor, which was then used as prediction. Comparison to a reference model allows an objective evaluation of the results. Furthermore, the root mean square error (RMSE) for each factor was additionally utilized to compare the performance on each EMA measure individually.

Simulation analysis

This simulation allows to estimate the expected performance on the study data and helps to analyze reasons for a low model performance. From the literature on EMA, we identified

that self-reported measures are affected by a **(1) high noise level**, which has an impact on the model performance. Potential causes for measurement errors include a lack of question comprehension by questionnaire participants, the influence of question order, or the number of response alternatives (Sudman, Bradburn, and Schwarz, 1996). To provide a meaningful answer to the EMA questions, clients' typically utilize contextual information to infer what the researcher might be interested in (Clark and Schober, 1992). Their interpretation of the question, therefore, goes further than their literal meaning. Additionally, clients could consider different periods to answer the question ranging from right now to the last few hours. This variability in the interpretation of the question suggests that clients could include relatively minor events into their answers which translates to a variance in ratings. A related issue is the clients' scale usage. They anchor the endpoints of the scale with a low and high event that they experienced. This leads to a relative ranking of the current event regarding the most extremes. Therefore, events can receive a lower rating the more intense the previous events have been that serve as a high anchor (Parducci, 1965; Daamen and Bie, 1992). Similarly, concurrent ratings will be ranked according to previously still highly memorable ratings and clients tend to adapt their rating according to their currently poor health (Riis et al., 2005).

Another reason for a low predictive performance, can result from **(2) too few data points** which can lead to biased parameter estimates and poorer model performance (Davey, 2005). EMA data are affected to a varying degree by missing data when clients struggle to respond to the inquiries (Courvoisier, Eid, and Lischetzke, 2012; Sokolovsky, Mermelstein, and Hedeker, 2014). The reported compliance rates among studies considerably vary with an expected compliance rate of 75% (Jones et al., 2019). Types of missing values are typically categorized as: missing completely at random, missing at random, and missing not at random (Sterne et al., 2009). For values missing completely at random, there are no systematic differences between observed and unobserved values. Measures that are missing at random imply a likelihood that can be derived from the data, such as people with higher age might be more forgetful about reporting their measures. Missing not at random describes data where the likelihood of not observing the value depends on itself. Consequently, a rating of depression could be missing because of severe depressive symptoms, which resulted in a missed inquirement (Papageorgiou et al., 2018). Further, this could also be the case if the assessment protocol did not define an assessment on that particular day.

The simulation analysis allowed us to control and inspect high noise levels and fewer data points separately. For both simulation analyses, we used the social integration model to simulate 100 time series with parameters randomly chosen from the parameters estimated in the study data analysis. This further links the simulation analysis to the study data analysis by utilizing parameters that are likely to be encountered from clients in a study. For the analysis of noise, we utilize the same number of measures per week as defined in the study protocol. Specifically, the simulated data includes five measures of mood and one measure of every other factor per week. To the simulated data, we add step-wise increasing Gaussian noise with a standard deviation ranging from 0 to 0.5 with a step size of 0.01. To analyze the influence of a varying number of weekly assessments, we simulated data ranging from 1 to 7 factor measurements per week. For this analysis, the amplitude of the additional noise was estimated for each EMA factor from the study data. Where we assume that the EMA factors in the study data are constant over the considered period of 7 weeks. Following, we estimate the standard deviation for each factor and utilize these to generate the additional noise in the simulated data. It is unlikely that these concepts are indeed constant over the considered period, thus, we consider the estimated noise levels as a worst-case estimate. The estimates are shown in Table 8.2. These simulations enabled to estimate the models' sensitivity to these influences and to compare the predictive capabilities on the study data and the results of the simulation. If, however, neither a **(1) a high noise level** or **(2) too few data points** can

explain a low model performance on the study data, it might be plausible that **(3) the social integration model does not represents the dynamics of mood development sufficiently**.

Concept name	Mood level	Enjoyed activities	Social interaction	Activities carried out
Average deviation (sd)	0.137 (0.047)	0.173 (0.060)	0.200 (0.066)	0.181 (0.066)

TABLE 8.2: Average standard deviation for each EMA factor

8.3 Results and discussion

8.3.1 Study data analysis

In the following, we estimated the predictive and descriptive performance of the social integration and *mean model* on the study data, which are shown in Table 8.3.

Model	Descriptive Performance (95%CI)	Predictive Performance (95%CI)
Social integration model	0.853 (0.837, 0.869)	0.492 (0.483, 0.502)
Mean model	0.839 (0.823, 0.856)	0.502 (0.447, 0.561)

TABLE 8.3: Performance measures on the study data

The performance scores indicate that both models provide similar predictive and descriptive performance. However, the descriptive performance of the social integration model is slightly higher whereas the mean model has a higher predictive performance. If we consider the 95% confidence intervals of both scores, we cannot suggest a significant difference among both models.

For comparing the average fitting and prediction error of each factor, we estimated the root mean square error on each EMA measure on the training and test data. The estimated errors are illustrated in Table 8.4.

Concept name	Mood level (sd)	Enjoyed activities (sd)	Social interaction (sd)	Activities carried out (sd)	Average (sd)
Training root mean square error					
Social integration model	0.129 (0.014)	0.152 (0.018)	0.175 (0.023)	0.167 (0.022)	0.157 (0.020)
Mean model	0.134 (0.014)	0.163 (0.020)	0.192 (0.024)	0.172 (0.025)	0.167 (0.024)
Prediction root mean square error					
Social integration model	0.142 (0.021)	0.189 (0.045)	0.208 (0.061)	0.203 (0.075)	0.187 (0.030)
Mean model	0.146 (0.022)	0.179 (0.040)	0.186 (0.043)	0.183 (0.051)	0.174 (0.019)

TABLE 8.4: Average prediction root mean square error for each factor

Inspection of the training errors suggests that the social integration model provides a slightly better fit to the training data compared to the *mean model* as suggested by the descriptive performance. A t-test on the average fit of both models does not provide evidence that there is a significant difference (p-value = 0.0698) between both models' fit to the data. There is, further, no evidence that the prediction error of the social integration model is lower than the prediction error of the mean model, which was similarly indicated by the predictive performance measure.

Therefore, the performance measures, as well as the RMSE, indicate the same trend: the social integration model might provide a closer fit to the data but does not provide a better prediction performance compared to the *mean model*. Although the social integration model has higher complexity in terms of free parameters than the mean model, it does not provide more accurate future predictions. The two possible reasons for this finding, which we previously defined ((1) high measurement noise and (2) too few data points) are examined more closely in the simulation analysis.

8.3.2 Simulation analysis of measurement noise

Figure 8.2 illustrates the influence of noise on the model performance and prediction error in the simulation analysis. With an increase in noise, the descriptive and predictive performance of the social integration model and *mean model* decreased. However, the predictive performance of the social integration model was higher than the predictive performance of the mean model even though they were approaching each other with an increase in noise. The same finding applied to the RMSE of both models, where up to a noise of 0.24 we estimated a significant difference using a one-tailed t-test (p-value=0.0002), with a higher noise level both models' prediction RMSE are indistinguishable.

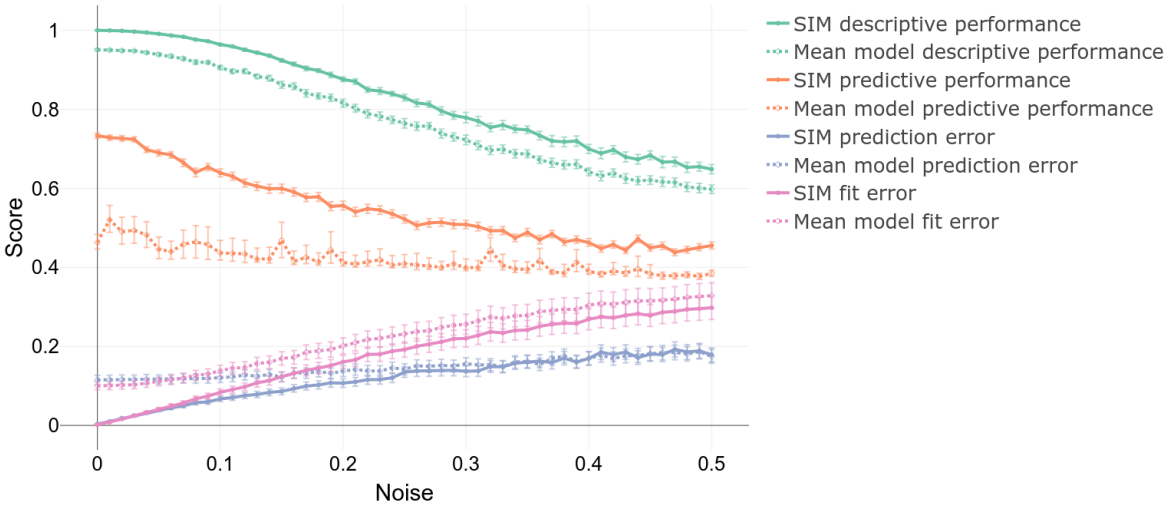


FIGURE 8.2: Influence of increasing noise on the performance measures

When comparing the average prediction RMSE on the simulated data with the averaged error on the study data (0.187), we noticed that at a noise level with a standard deviation of 0.4, the prediction RMSE of the social integration model approached the prediction RMSE from the study data. The simulation showed that at this noise level, the error of the social integration model should be below the error of the mean model. However, this is not the case in the study data analysis (0.187 vs 0.174). The same observation can be made for the predictive performance. At a noise with a standard deviation of 0.35 the predictive performance approaches the results on the study data. At this noise level, the predictive performance of the SIM is well above the predictive performance of the mean model in the simulation.

Additionally, a noise level with a standard deviation of 0.4 or 0.35 appears to be too high. For example, if we assume the rating to be normally distributed with a mean value of 0.5 and a standard deviation of 0.4, then there is approximately 21% of the probability mass outside the valid range of the ratings $[0, 1]$. This would result in many maximal and minimal ratings. However, in the study data, we observed that only 2.1% of the analyzed measures are such extreme points (ratings of 0 or 1). According to our assumption that the ratings are normally distributed with a mean of 0.5, this would rather suggest a standard deviation of 0.22. At this noise level, the simulation results in a lower prediction RMSE for the social integration model than the mean model.

The simulation showed that with data comparable to the study data, in terms of sparsity and additional noise, the social integration model should provide a lower prediction error than the mean model. Since this was not the case in our analysis we conclude that high noise level was not responsible for the low prediction performance of the social integration model.

8.3.3 Simulation analysis of weekly assessed measures

For analyzing the effect of missing values per week, we utilized the estimated standard deviations from the study data and simulated data with an increasing number of missing values per week. The estimated model performance on the simulated data is illustrated in Figure 8.3.

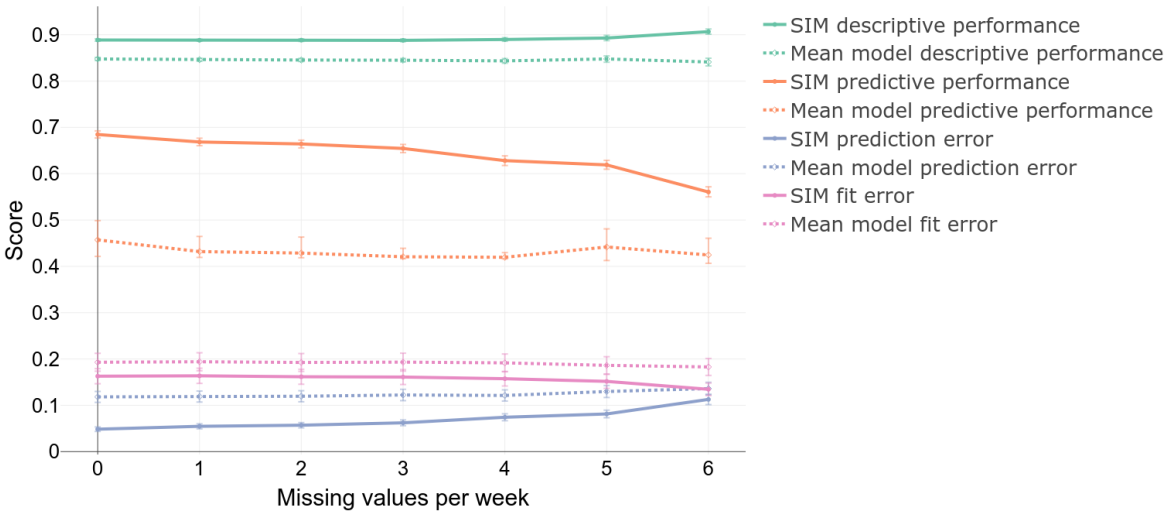


FIGURE 8.3: Influence of a reduction of weekly measures on the performance measures

A reduction in simulated measures led to a slight increase in the descriptive performance for the social integration model because the model can fit the reduced number of data points more accurately. The descriptive performance of the mean model appears mainly unaffected.

The predictive performance dropped with an increase in missing values for the social integration model because the model parameters that provide a close fit to the data do not allow accurate future predictions. The predictive performance of the mean model was slightly reduced with fewer measures. With fewer measures, the noise has a stronger impact on the estimated mean value. The change in prediction performance is also reflected in the prediction RMSE of both models. The RMSE of the fit to the data of the social integration model is smaller than for the mean model. Further, the predictive performance of the SIM in the simulation is higher than on the study data. In summary, by comparing the prediction RMSE and predictive performance in the simulation for one through six missing values per week to the results estimated on the study data, one notices that the prediction errors in the simulation are lower than on the study data.

This simulation shows that a reduction in measures cannot explain the lower performance of the social integration model on the study data. The simulation analyses suggested that for both analyzed cases the SIM should provide higher predictive performance than the mean model, which is not the case on the study data. We, therefore, conclude by the process of elimination that only the third reason for the low predictive performance of the social integration model remains as a valid option. That is, the social integration model in its current form does not represent the real-world relationships of psychological concepts to a degree that supports predictions.

8.4 Conclusion

In this analysis, we demonstrated a detailed evaluation of the predictive capabilities of a temporal-causal model. We employed an EMA dataset for predicting clients' upcoming EMA ratings for a week in advance using the social integration model (Altaf Hussain Abro, 2016). For the evaluation, we demonstrated the use of a simulation analysis and compared the social integration model to a reference model (i.e. the mean of EMA ratings observed so far). We evaluated the descriptive and predictive performance for both models on the study data and in a simulation analysis. Subsequently, the predictive capability of the social integration model was not superior to the mean model prediction on the study data. We argued that this finding could be explained as follows: (1) measurement errors (noise) in the EMA data, (2) sparsity of measurements (e.g. due to missing values), (3) or that the social integration model does not fully represent the psychological dynamics. To investigate these reasons, the performance of both models was analyzed for an increasing measurement error (1) and fewer weekly assessed measures (2) in the simulation analysis. For both simulations, the study data was used to inform the analysis. In the case of measurement noise, we simulated an increasing noise level until we matched the error on the study data and approximated the expected noise level based on clients' extreme ratings (rating of 0 or 1). In the case of fewer measures (2) we used the average variance of each factor among clients as an estimate for the measurement noise. For the simulated data we showed that regarding the measurement noise and reduced weekly measurements, the social integration model should provide a better descriptive as well as predictive performance than the reference model. Since this was not the case on the study data, we excluded (1) measurement errors and (2) the sparsity of observations as potential reasons for the experimental results that is the social integration model is not superior to the reference model. Thus, we conclude that the social integration model in its current form is limited when the goal is predicting future mental states of clients.

We proposed and applied a systematic model evaluation that combines study data and simulation analysis. For both evaluations, the use of client individual parameters and comparison to a reference model is required. This is accomplished by utilizing a model evaluation framework that provides methods for multi-objective model performance estimation and model comparison. The simulation analysis is used to analyze reasons for low model performances described in the literature. Further, the simulation is designed to reflect the study conditions to enable a comparison of the results. This provides insight into the theoretical model performance which should be in an agreement with the results obtained on study data to provide evidence for the model to be accurate. These analyses also provide insights into the model robustness under varying study conditions. The derived insights from a simulation analysis can be used to state requirements on the data assessment in studies to ensure the model performance. We demonstrated the benefits of a simulation analysis and suggest that a simulation can complement guidelines and requirements for statistical model evaluation. The presented setup enables researchers to examine the impact of measurement errors and missing values on the predictive capabilities of their model. We thus hope to provide a solid setup for model evaluation in subsequent studies.

References

Aan het Rot, Marije, Koen Hogenelst, and Robert A. Schoevers (2012). "Mood disorders in everyday life: A systematic review of experience sampling and ecological momentary assessment studies". In: *Clinical Psychology Review* 32.6, pp. 510–523. ISSN: 02727358. DOI: 10.1016/j.cpr.2012.05.007 (cit. on pp. 3, 140).

Abro, Altaf Hussain and Jan Treur (2017). "A cognitive agent model for desire regulation applied to food desires". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 10207 LNAI. November, pp. 251–260. ISBN: 9783319592930. DOI: 10.1007/978-3-319-59294-7_20 (cit. on p. 141).

Altaf Hussain Abro, Michel Klein (2016). "Validation of a Computational Model for Mood and Social Integration". In: *Lecture Notes in Computer Science*. Lecture Notes in Computer Science 10046.November 2016. Ed. by Emma Spiro and Yong-Yeol Ahn, pp. 361–375. DOI: 10.1007/978-3-319-47880-7 (cit. on pp. 7, 89, 93, 141, 142, 149).

Altman, Douglas G et al. (2009). "Prognosis and prognostic research: validating a prognostic model". In: *BMJ* 338. ISSN: 0959-8138. DOI: 10.1136/bmj.b605. eprint: https://www.bmj.com/content (cit. on pp. 140, 141).

Araújo, Eric Fernandes de Mello and Jan Treur (2016). "Analysis and refinement of a temporal-causal network model for absorption of emotions". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 9875 LNCS. August 2018, pp. 27–39. ISBN: 9783642166921. DOI: 10.1007/978-3-319-45243-2_3. arXiv: arXiv:1011.1669v3 (cit. on p. 141).

Aung, Min Hane, Mark Matthews, and Tanzeem Choudhury (2017). "Sensing behavioral symptoms of mental health and delivering personalized interventions using mobile technologies". In: *Depression and Anxiety* 34.7, pp. 603–609. ISSN: 10914269. DOI: 10.1002/da.22646. arXiv: 15334406 (cit. on pp. 3, 4, 140).

Bak, Maarten et al. (2016). "An n=1 Clinical network analysis of symptoms and treatment in psychosis". In: *PLoS ONE* 11.9, pp. 1–15. ISSN: 19326203. DOI: 10.1371/journal.pone.0162811 (cit. on pp. 1, 4, 140).

Bouwmeester, Walter et al. (2012). "Reporting and methods in clinical prediction research: A systematic review". In: *PLoS Medicine* 9.5. ISSN: 15491277. DOI: 10.1371/journal.pmed.1001221 (cit. on pp. 7, 141).

Breda, Ward van et al. (2017). "Assessment of temporal predictive models for health care using a formal method". In: *Computers in Biology and Medicine* 87.November 2016, pp. 347–357. ISSN: 00104825. DOI: 10.1016/j.compbiomed.2017.06.014 (cit. on pp. 122, 141, 143).

Bremer, Vincent, Burkhardt Funk, and Heleen Riper (2019). "Heterogeneity Matters: Predicting Self-Esteem in Online Interventions Based on Ecological Momentary Assessment Data". In: *Depression Research and Treatment* 2019. ISSN: 2090133X. DOI: 10.1155/2019/3481624 (cit. on pp. 26, 90, 97, 142).

Cacioppo, John T., Louise C. Hawkley, and Ronald A. Thisted (2010). "Perceived social isolation makes me sad: 5-year cross-lagged analyses of loneliness and depressive symptomatology in the chicago health, aging, and social relations study". In: *Psychology and Aging* 25.2, pp. 453–463. ISSN: 19391498. DOI: 10.1037/a0017216 (cit. on p. 142).

Cacioppo, John T. et al. (2002). "Loneliness and health: Potential mechanisms". In: *Psychosomatic Medicine* 64.3, pp. 407–417. ISSN: 00333174. DOI: 10.1097/00006842-200205000-00005 (cit. on p. 142).

Clark, Herbert H and Michael F Schober (1992). "Asking questions and influencing answers." In: *Questions about questions: Inquiries into the cognitive bases of surveys.* New York, NY, US: Russell Sage Foundation, pp. 15–48. ISBN: 0-87154-842-9 (Hardcover) (cit. on pp. 141, 145).

Constantinides, Marios et al. (2018). "Personalized versus Generic Mood Prediction Models in Bipolar Disorder". In: *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers - UbiComp '18*. New York, New York, USA: ACM Press, pp. 1700–1707. ISBN: 9781450359665. DOI: 10.1145/3267305.3267536 (cit. on pp. 8, 141).

Courvoisier, Delphine S., Michael Eid, and Tanja Lischetzke (2012). "Compliance to a cell phone-based ecological momentary assessment study: The effect of time and personality characteristics". In: *Psychological Assessment* 24.3, pp. 713–720. ISSN: 10403590. DOI: 10.1037/a0026733 (cit. on pp. 141, 145).

Daamen, Dancker D. L. and Steven E. de Bie (1992). "Serial Context Effects in Survey Interviews". In: *Context Effects in Social and Psychological Research*. New York, NY: Springer New York, pp. 97–113. DOI: 10.1007/978-1-4612-2848-6_8 (cit. on p. 145).

Davey, Adam (2005). "Issues in Evaluating Model Fit With Missing Data". In: *Structural Equation Modeling: A Multidisciplinary Journal* 12.4, pp. 578–597. ISSN: 1070-5511. DOI: 10.1207/s15328007sem1204_4 (cit. on p. 145).

De Graaf, L. E. et al. (2009). "Clinical effectiveness of online computerised cognitive-behavioural therapy without support for depression in primary care: Randomised trial". In: *British Journal of Psychiatry* 195.1, pp. 73–80. ISSN: 00071250. DOI: 10.1192/bjp.bp.108.054429 (cit. on pp. 1, 3, 53, 140).

Deb, Kalyanmoy et al. (2002). "A fast and elitist multiobjective genetic algorithm: NSGA-II". In: *IEEE Transactions on Evolutionary Computation* 6.2, pp. 182–197. ISSN: 1089778X. DOI: 10.1109/4235.996017 (cit. on p. 143).

Fisher, Aaron J. and James F. Boswell (2016). "Enhancing the Personalization of Psychotherapy With Dynamic Assessment and Modeling". In: *Assessment* 23.4, pp. 496–506. ISSN: 15523489. DOI: 10.1177/1073191116638735 (cit. on pp. 1, 4, 140).

Franke, Annelore and Rukshar Wagid Hosain (2017). *A Temporal-Causal Model for Spread of Messages in Disasters*. Vol. 10449, pp. 386–397. ISBN: 978-3-319-67076-8. DOI: 10.1007/978-3-319-67077-5. arXiv: arXiv:1011.1669v3 (cit. on p. 141).

Gariépy, Geneviève, Helena Honkaniemi, and Amélie Quesnel-Vallée (2016). *Social support and protection from depression: Systematic review of current findings in western countries*. DOI: 10.1192/bjp.bp.115.169094 (cit. on p. 142).

Gibbons, Chris J. (2017). "Turning the page on pen-and-paper questionnaires: Combining ecological momentary assessment and computer adaptive testing to transform psychological assessment in the 21st century". In: *Frontiers in Psychology* 7.JAN, pp. 1–4. ISSN: 16641078. DOI: 10.3389/fpsyg.2016.01933 (cit. on pp. 3, 4, 88, 140).

Ivanescu, A E et al. (2016). "The importance of prediction model validation and assessment in obesity and nutrition research". In: *International Journal of Obesity* 40.6, pp. 887–894. ISSN: 0307-0565. DOI: 10.1038/ijo.2015.214 (cit. on pp. 7, 140).

Jaques, Natasha et al. (2017). *Predicting Tomorrow's Mood, Health, and Stress Level using Personalized Multitask Learning and Domain Adaptation*. Tech. rep., pp. 17–33 (cit. on pp. 8, 93, 141).

Jones, Andrew et al. (2019). "Compliance with ecological momentary assessment protocols in substance users: a meta-analysis". In: *Addiction* 114.4, pp. 609–619. ISSN: 13600443. DOI: 10.1111/add.14503 (cit. on p. 145).

Kappen, Teus H. et al. (2018). "Evaluating the impact of prediction models: lessons learned, challenges, and recommendations". In: *Diagnostic and Prognostic Research* 2.1, pp. 1–11. DOI: 10.1186/s41512-018-0033-6 (cit. on pp. 26, 140).

Kleiboer, Annet et al. (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". In: *Trials* 17.1, p. 387. ISSN: 1745-6215. DOI: 10.1186/s13063-016-1511-1 (cit. on pp. 6, 89, 91, 123, 126, 131, 141, 144).

Lambert, Michael J (2012). "The Outcome Questionnaire-45". In: *Integrating Science and Practice* 2.1, pp. 24–27. ISSN: 1438-8871. DOI: 10.2196/jmir.954 (cit. on pp. 1, 140).

Mikus, Adam et al. (2017). *Predicting short term mood developments among depressed patients using adherence and ecological momentary assessment data*. DOI: 10.1016/j.invent.2017.10.001 (cit. on pp. 88, 89, 92, 93, 97, 141).

Mohr, David C, Mi Zhang, and Stephen M Schueller (2017). "Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning". In: *Annual Review of Clinical Psychology* 13.1, pp. 23–47. ISSN: 1548-5943. DOI: 10.1146/annurev-clinpsy-032816-044949 (cit. on pp. 3, 25, 140).

Moons, Karel G M et al. (2009a). "Prognosis and prognostic research: application and impact of prognostic models in clinical practice". In: *BMJ* 338. ISSN: 0959-8138. DOI: 10.1136/bmj.b606. eprint: https://www.bmj.com/content (cit. on p. 141).

Moons, Karel G M et al. (2009b). "Prognosis and prognostic research: what, why, and how?" In: *BMJ* 338. ISSN: 0959-8138. DOI: 10.1136/bmj.b375. eprint: https://www.bmj.com/content (cit. on pp. 5, 7, 140).

Naze, Sebastien and Jan Treur (2011). "A computational agent model for post-traumatic stress disorders". In: *Frontiers in Artificial Intelligence and Applications* 233.April, pp. 249–261. ISSN: 09226389. DOI: 10.3233/978-1-60750-959-2-249 (cit. on p. 141).

Nezlek, John B., Christianne P. Hampton, and Glenn D. Shean (2000). "Clinical depression and day-to-day social interaction in a community sample". In: *Journal of Abnormal Psychology* 109.1, pp. 11–19. ISSN: 0021843X. DOI: 10.1037/0021-843X.109.1.11 (cit. on p. 142).

Nezlek, John B., Mark Imbrie, and Glenn D. Shean (1994). "Depression and Everyday Social Interaction". In: *Journal of Personality and Social Psychology* 67.6, pp. 1101–1111. ISSN: 00223514. DOI: 10.1037/0022-3514.67.6.1101 (cit. on p. 142).

Nicholson, Nicholas R. (2009). "Social isolation in older adults: An evolutionary concept analysis". In: *Journal of Advanced Nursing* 65.6, pp. 1342–1352. ISSN: 03092402. DOI: 10.1111/j.1365-2648.2008.04959.x (cit. on p. 142).

Papageorgiou, Grigorios et al. (2018). "Statistical primer: How to deal with missing data in scientific research?" In: *Interactive Cardiovascular and Thoracic Surgery* 27.2, pp. 153–158. ISSN: 15699285. DOI: 10.1093/icvts/ivy102 (cit. on pp. 141, 145).

Parducci, Allen (1965). *Category judgment: A range-frequency model*. US. DOI: 10.1037/h0022602 (cit. on p. 145).

Riis, Jason et al. (2005). "Ignorance of hedonic adaptation to hemodialysis: A study using ecological momentary assessment". In: *Journal of Experimental Psychology: General* 134.1, pp. 3–9. ISSN: 00963445. DOI: 10.1037/0096-3445.134.1.3 (cit. on p. 145).

Robinson, Michael D and Gerald L Clore (2002). "Belief and feeling: evidence for an accessibility model of emotional self-report." In: *Psychological bulletin* 128.6, pp. 934–960. ISSN: 0033-2909. DOI: 10.1037/0033-2909.128.6.934 (cit. on pp. 3, 140).

Sokolovsky, Alexander W., Robin J. Mermelstein, and Donald Hedeker (2014). "Factors predicting compliance to ecological momentary assessment among adolescent smokers". In: *Nicotine and Tobacco Research* 16.3, pp. 351–358. ISSN: 14622203. DOI: 10.1093/ntr/ntt154 (cit. on pp. 141, 145).

Spanakis, Gerasimos et al. (2016). "Network Analysis of Ecological Momentary Assessment Data for Monitoring and Understanding Eating Behavior". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 9545. January, pp. 43–54. ISBN: 9783319291741. DOI: 10.1007/978-3-319-29175-8_5 (cit. on p. 142).

Steger, Michael F and Todd B Kashdan (2010). "Depression and Everyday Social Activity". In: *Journal of Counseling Psychology* 56.2, pp. 289–300. DOI: 10.1037/a0015416.Depression (cit. on p. 142).

Sterne, J. A C et al. (2009). "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls". In: *BMJ* 338.jun29 1, b2393–b2393. ISSN: 0959-8138. DOI: 10.1136/bmj.b2393 (cit. on p. 145).

Steyerberg, Ewout W. and Yvonne Vergouwe (2014). *Towards better clinical prediction models: Seven steps for development and an ABCD for validation*. DOI: 10.1093/eurheartj/ehu207 (cit. on pp. 7, 141).

Steyerberg, Ewout W. et al. (2013). "Prognosis Research Strategy (PROGRESS) 3: Prognostic Model Research". In: *PLoS Medicine* 10.2, e1001381. ISSN: 15491277. DOI: 10.1371/journal.pmed.1001381 (cit. on p. 141).

Stone, Arthur A. and Saul Shiffman (1994). "Ecological Momentary Assessment (Ema) in Behavioral Medicine". In: *Annals of Behavioral Medicine* 16.3, pp. 199–202. ISSN: 0883-6612. DOI: 10.1093/abm/16.3.199 (cit. on pp. 3, 76, 140).

Sudman, Seymour, Norman M Bradburn, and Norbert Schwarz (1996). *Thinking about answers: The application of cognitive processes to survey methodology.* San Francisco, CA, US: Jossey-Bass, pp. xiv, 304–xiv, 304. ISBN: 0-7879-0120-2 (Hardcover) (cit. on pp. 141, 145).

Taylor, Harry Owen et al. (2018). "Social Isolation, Depression, and Psychological Distress Among Older Adults". In: *Journal of Aging and Health* 30.2, pp. 229–246. ISSN: 15526887. DOI: 10.1177/0898264316673511 (cit. on p. 142).

Treur, Jan (2016). "Dynamic modeling based on a temporal-causal network modeling approach". In: *Biologically Inspired Cognitive Architectures* 16.April, pp. 131–168. ISSN: 2212683X. DOI: 10.1016/j.bica.2016.02.002 (cit. on pp. 93, 141).

Treur, Jan and S Sahand Mohammadi Ziabari (2018). *Computational Collective Intelligence*. Vol. 11056. Springer International Publishing, pp. 13–25. ISBN: 978-3-319-98445-2. DOI: 10.1007/978-3-319-98446-9. arXiv: arXiv:1011.1669v3 (cit. on p. 141).

Chapter 9

Analysis of the histogram intersection kernel for use in Bayesian Optimization

Becker, D. (2017). International Journal of Modeling and Optimization, Volume 7, Issue 6.

Abstract: *In optimization problems, a typical assumption is that the objective function is cheap to evaluate. However, many problems are not conform to this assumption. When the objective function is expensive to evaluate, Bayesian optimization is a powerful strategy to estimate the optimum of the objective function. Typically, Bayesian optimization utilizes a Gaussian process with an exponential kernel for the approximation of the objective function. However, the training time of the Gaussian process scales cubically and prediction and memory requirement quadratically. This poses limitations on the number of utilized data points for the training of the Gaussian process. To potentially overcome this drawback, this paper analyzes the use of the histogram intersection kernel with approximation methods, which has been shown to scale linearly, as covariance function for the Gaussian process. The resulting algorithm is compared to the EGO optimization algorithm, which utilizes an exponential kernel, and random sampling on common optimization problems. The results show that the linear approximation of the histogram intersection kernel does not accurately approximate the error surface and introduces false minima which can prevent the algorithm to identify the global minimum of a function. However, the implemented algorithm performs better than random sampling and its potential for fast evaluation might make it attractive for time-limited optimization tasks.*

9.1 Introduction

Non-convex optimization problems arise in many research fields and are often solved using black-box optimization algorithms. These algorithms typically do not rely on the computation of a gradient and are applicable to any optimization problem. The cost of optimizing such a problem is dominated by the number of objective function evaluations required to reach an acceptable solution. Difficulties in employing black-box optimization algorithms arise when the optimization function is time intensive to evaluate.

A possibility to improve the search is to utilize the already evaluated solutions to build a model that approximates the function that is to be optimized. This approximation is then optimized instead of the true optimization function. Afterward, the estimated solution of this model is estimated on the true cost function. This can often reduce the computation time to find an acceptable solution. Such models are also referred to as surrogates for the original

objective function. However, a prerequisite for using such models is that the expense of model construction and prediction is lower than evaluating the true optimization function.

An approach that aims at finding the global optimum with a small number of objective function evaluations is Bayesian optimization (Brochu, Cora, and De Freitas, 2010; Jones, Schonlau, and Welch, 1998; Kleijnen, 2014). Typically, Bayesian optimization approximates the true objective function with a Gaussian process (GP) and estimates the next location to evaluate the true objective function using an acquisition function. An acquisition function balances between exploration and exploiting by considering the expected value of the surrogate function and the variance. Considering the uncertainty in the expected costs allows to search in so far unexplored areas. Research on Bayesian optimization dates back to Kushner (1964) and Mockus, Tiesis, and Zilinskas (1978) but subsided shortly after. New interest arose when it was realized that Bayesian optimization provides a tool to estimate the hyperparameters of machine learning algorithms (Snoek, Larochelle, and Adams, 2012; Swersky, Snoek, and Adams, 2013; Gelbart, Snoek, and Adams, 2914). Hyperparameter estimation tends to be multi-modal and expensive to evaluate functions, where Bayesian optimization has been used to successfully reduce the computational demand.

Although the Gaussian process can model a variety of functions and produce reasonable predictions, training time of the model scales cubically and memory usage quadratically with increasing data samples. This severely limits the sample size that can be used for function approximation. Sparse approximations can often provide a solution for this problem where only a subset of the data is used to alleviate the computational burden (Williams and Seeger, 2001; Quiñonero-candela, Rasmussen, and Herbrich, 2005; Hensman, Fusi, and Lawrence, 2013). The utilized subset for the training of the model can consist of real training examples or pseudo inputs (Snelson and Ghahramani, 2006). Another possibility is the distributed Gaussian process (Deisenroth and Ng, 2015) where the computational and the memory load is distributed to many independent computational units. Each unit operates on a subset of the data and the results are recombined for an overall result.

An alternative to these methods to reduce the memory and computational burden, that will be analyzed in this paper, is the use of a Gaussian process with a histogram intersection kernel (HIK). This particular kernel has been utilized for computer vision due to its fast learning, classification and linear memory requirements (Wu, Tan, and Rehg, 2011; Wu, 2010; Maji, Berg, and Malik, 2008). Initially used as a kernel for support vector machines, it can also be used for the Gaussian process. Although the kernel only provides a piecewise linear approximation of the true function, its capability for large large-scale Gaussian process (Rodner et al., 2012; Rodner et al., 2016) inference appears attractive for the use as surrogate function.

The lower time and memory requirements could be beneficial for use on mobile devices or when the optimization task poses limitations on the available time for optimization. Since the histogram intersection kernel has not been used in the context of optimization so far, we explore and analyze the use of this kernel in the context of the surrogate function for Bayesian optimization. We compare the resulting Bayesian optimization algorithm with the EGO optimization algorithm (Jones, Schonlau, and Welch, 1998) that utilizes an exponential kernel and has been shown to estimate the global optima quickly. Furthermore, we outline the potential of parallelization of the estimation of the Gaussian process with HIK for additional speedup.

9.2 Method

9.2.1 Gaussian Process Regression

For the approximation of the true cost function, we consider the regression problem $y = f(\mathbf{x}) + \epsilon \in \mathbb{R}$. In this function, it is assumed that the observed data y_i is generated by an unknown function $f(\mathbf{x}_i)$, and potentially corrupted by additional independent noise $\epsilon \sim \mathcal{N}(0, \sigma_y^2)$. A Gaussian process (Rasmussen and Williams, 2006) defines a prior over functions that could have created the observed data with mean 0 and covariance given by $\Sigma_{ij} = \kappa(x_i, x_j)$. The covariance matrix is built by the kernel function κ which describes the similarity among samples. For the prediction of new data points, the posterior predictive distribution of the new data sample $\mathbf{X}_*$ is calculated by marginalizing over all possible functions. The posterior predictive distribution of the corresponding function value $y_* = f(\mathbf{x}_*)$ is a Gaussian with mean and variance given by,

$$p(\mathbf{f}_* \mid \mathbf{X}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}(\mathbf{f}_* \mid \boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*), \tag{9.1}$$

$$\boldsymbol{\mu}_* = \mathbf{K}_*^T \mathbf{K}_\mathrm{y}^{-1} \mathbf{y}, \tag{9.2}$$

$$\boldsymbol{\Sigma}_* = \mathbf{K}_{**} - \mathbf{K}_*^T \mathbf{K}_\mathrm{y}^{-1} \mathbf{K}_*, \tag{9.3}$$

where $\mathbf{K}_\mathrm{y} = \kappa(\mathbf{X}, \mathbf{X}) + \sigma_y^2 \mathbf{I} \in \mathbb{R}^{N \times N}$, $\mathbf{K}_* = \kappa(\mathbf{X}, \mathbf{X}_*) \in \mathbb{R}^{N \times N_*}$, and $\mathbf{K}_{**} = \kappa(\mathbf{X}_*, \mathbf{X}_*) \in \mathbb{R}^{N_* \times N_*}$. For the Gaussian process regression, all these computations can be executed in closed form. However, the creation of the $N \times N$ kernel matrix requires $O(N^2)$ space and inversion of this matrix $O(N^3)$ time. These requirements practically limit the use of the Gaussian process to data sets of size $O(10^4)$. The most widely used kernel in machine learning might be the squared exponential kernel (SEK) that is given by

$$\kappa_{\mathrm{SE}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\gamma ||\boldsymbol{x}_i - \boldsymbol{x}_j||^2\right). \tag{9.4}$$

Where the free parameter γ is called a hyperparameter that influences the fit of the Gaussian process to the data. Usually, the hyperparameters are inferred from the data by optimizing the likelihood of the Gaussian process given the data. An example of different settings of the γ value is shown in Figure 9.1. Usually, hyperparameters are inferred from the data by optimizing the likelihood of the Gaussian process.

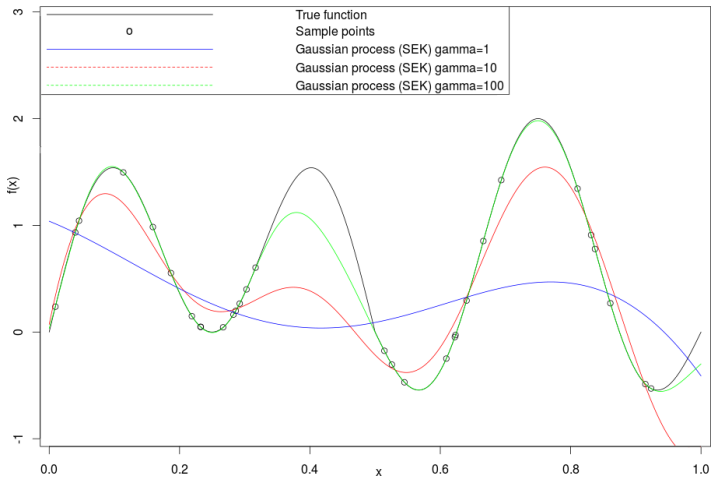


FIGURE 9.1: Example of influence of Gaussian process hyperparameter

9.2.2 Histogram Intersection Kernel

To reduce the computational effort of the Gaussian process, the histogram intersection kernel is utilized. The histogram intersection kernel is defined as,

$$\kappa_{\text{HIK}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{d=1}^{D} \min(\boldsymbol{x}_i(d), \boldsymbol{x}_j(d)). \tag{9.5}$$

It has been shown that properties of this kernel allow to speed up model learning and prediction of new data samples (Wu, Tan, and Rehg, 2011; Wu, 2010; Maji, Berg, and Malik, 2008). The multiplication of the kernel matrix $\mathbf{K}$ by an arbitrary vector v can be done without explicitly creating the kernel matrix, which enables sub quadratic calculation of the matrix-vector product. This alleviates computational and memory storage cost during the training of the model. Furthermore, the prediction of new data samples scales sub-linear, which improves prediction time respectively. However, these considerations only apply when each dimension of the kernel matrix is sorted. For illustration of these properties, we first consider the prediction of a new data sample given the matrix $\mathbf{K}_*$ and $\boldsymbol{\alpha} = \mathbf{K}_{\text{y}}^{-1}\mathbf{y}$,

$$\begin{aligned} \mathbf{K}_*^T \boldsymbol{\alpha} &= \sum_{i=1}^{N} \boldsymbol{\alpha}(i) \sum_{d=1}^{D} \min(\boldsymbol{x}_i(d), \boldsymbol{x}_*(d)), \\ &= \sum_{d=1}^{D} \left(\sum_{i:\boldsymbol{x}_i(d)<\boldsymbol{x}_*(d)} \boldsymbol{\alpha}(i)\boldsymbol{x}_i(d) + \boldsymbol{x}_*(d) \sum_{j:\boldsymbol{x}_j(d)\geq\boldsymbol{x}_*(d)} \boldsymbol{\alpha}(j) \right). \end{aligned} \tag{9.6}$$

For each dimension, the summation can be separated into two parts. The first part consists of the samples $\mathbf{x}$ whose values are smaller than the value of the new sample to predict $\mathbf{x}_*$. For the calculation of this term, the values of $\boldsymbol{\alpha}$ have to be multiplied by the values of $\mathbf{x}$ and are summed up. The second term consists of all samples where $\mathbf{x}_*$ is smaller than or equal to $\mathbf{x}$. For this part of the kernel matrix $\mathbf{K}_*$, all values will be equal to $\mathbf{x}_*$. This allows to sum the remaining values of alpha and multiply the result by the value of x_*. Similar observations apply to the multiplication of the kernel matrix $\mathbf{K}$ with an arbitrary vector $\boldsymbol{v}$. The following equation illustrates the calculation of one value of the kernel vector product for one dimension,

$$(\mathbf{K}\boldsymbol{v})_i = \sum_{j=1}^{N} \boldsymbol{v}(j) \cdot \kappa(\boldsymbol{x}_j, \boldsymbol{x}_i). \tag{9.7}$$

For the calculation of each value of the matrix-vector product, the kernel value will be $\mathbf{x}_j$ for all samples smaller than $\mathbf{x}_i$ and $\mathbf{x}_j$ for all remaining values. This allows the efficient multiplication of the kernel matrix with an arbitrary vector when the samples are sorted for each dimension. Therefore the kernel matrix has not to be stored, which reduces the memory requirements from $O(N^2)$ to $O(2ND)$ assuming the matrix with the sort indices is stored for later use.

The definition of the histogram intersection kernel can be further generalized to increase its flexibility. Boughorbel, Tarel, and Boujemaa (2005) have shown that any positive valued function $g(\cdot)$ can be applied to the data and the kernel remains positive definite. This allows the transformation of the data and is often referred to as the generalized histogram intersection kernel:

$$\kappa_{\text{GHIK}}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{d=1}^{D} \min(g(\boldsymbol{x}_i(d)), g(\boldsymbol{x}_j(d))). \tag{9.8}$$

As transformation function, we utilize the exponential transformation:

$$g_{\exp}(\boldsymbol{x}(d)) = \frac{\exp(\eta\,|\boldsymbol{x}(d)|) - 1}{\exp(\eta) - 1}. \tag{9.9}$$

This transformation retains the order of the samples and therefore the fast evaluation of the kernel vector product can still be applied. For each dimension, a different parameter $\boldsymbol{\eta}(d)$ is used in order to individually weight each dimension. Furthermore, we apply a positive shift to the data for each individual dimension. This allows to approximate functions that are also defined for negative values and circumvents that the prediction of the HIK at the origin will be zero.

9.2.3 Bayesian Optimization

For the optimization of a function $f : \mathcal{X} \to \mathbb{R}$ that is to be minimized on some domain $X \subseteq \mathcal{X}$, we wish to find

$$x^* = \arg\min_{x \in X} f(x). \tag{9.10}$$

Typically, we aim to find the global optimum of a potential multi-modal function. Commonly, for the optimization of some function, assumptions are made such as non-linearity or, in contrast, smoothness. Bayesian optimization, although not strictly required, assumes a Gaussian process prior on the function f: $p(f) = \mathcal{GP}(f; \mu; \mathbf{K})$. Given observations of the function, the Gaussian process is conditioned on the data. The Gaussian process then approximates the underlying function and can be used to estimate an x that might provide a better solution. A function that estimates the next best location to evaluate is called an acquisition function. The simplest acquisition function would be to use the minimum of the conditioned Gaussian process. But this procedure can easily lead to a local minimum because the acquisition function does not consider the uncertainty about the surface. Therefore, this acquisition function would focus on exploitation rather than balancing the trade-off between exploration and exploitation. Exploration represents a global search to identify further minima and exploitation emphasis on the local search for refining a solution.

There are a variety of acquisition functions (Brochu, Cora, and De Freitas, 2010) that consider the trade-off between exploration and exploitation such as the probability of improvement (Kushner, 1964), expected improvement, entropy search, and upper confidence bound (Cox and John, 1997). In the following, the expected improvement and upper confidence bound acquisition function are introduced because these will be utilized for the optimization.

For the calculation of the expected improvement, first the lowest observed function value so far $f_{min} = \min(y_1, \ldots, y_n)$ is needed. To estimate if the evaluation of the function at a new point y might be beneficial, the prediction of the mean and variance of the Gaussian process are required next. This prediction is considered as a random variable Y and the improvement is defined as $I(x) = \max(f_{min} - Y, 0)$. This again represents a random variable and the expected value of this variable is estimated to obtain the expected improvement:

$$\mathbf{E} \equiv \mathbf{E}[I(\mathbf{x}) = \max(f_{min} - Y, 0)]. \tag{9.11}$$

The expected improvement (EI) acquisition function can be evaluated analytically (Jones, Schonlau, and Welch, 1998):

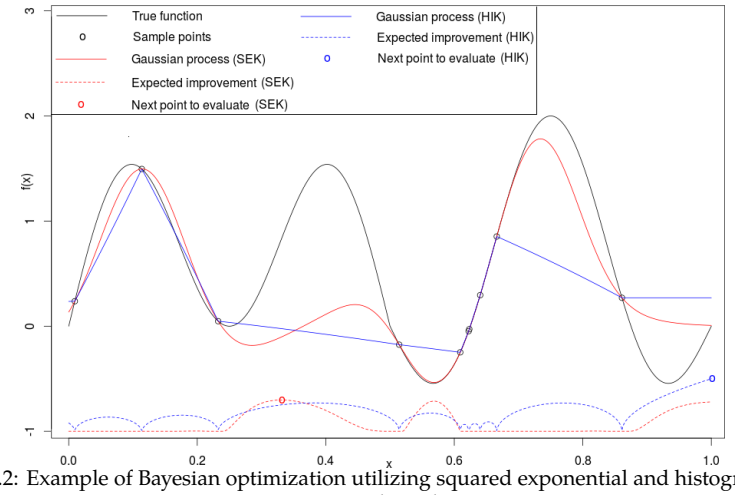


FIGURE 9.2: Example of Bayesian optimization utilizing squared exponential and histogram intersection kernel

$$\mathbf{E}[I(\mathbf{x})] = (f_{min} - \mu(\mathbf{x}))\Phi(Z) + \sigma(\mathbf{x})\phi(Z), \tag{9.12}$$

$$Z = \frac{f_{min} - \mu(\mathbf{x})}{\sigma(\mathbf{x})}, \tag{9.13}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the PDF and CDF of the standard normal distribution respectively.

Another acquisition function that considers the mean and the weighted variance is the upper confidence bound (UCB) (Cox and John, 1997). New sample points are selected based on $\mu(x) - \beta\sigma(x)$. Where the choice of the weight parameter $\beta \geq 0$ is left to the user.

An example of Bayesian optimization for the squared exponential and generalized histogram intersection kernel is shown in Figure 9.2. Different points from an unknown cost function are estimated and fitted using the Gaussian process. By optimizing the acquisition function, in this case, the expected improvement, the next point that is to be evaluated on the true cost function is identified.

9.2.4 Utilized Approximations

The implementation of the Gaussian process with HIK mostly follows the implementation described in Rodner et al. (2016). For the prediction of the mean value of new samples, the $\boldsymbol{\alpha}$ vector is required. The $\boldsymbol{\alpha}$ values are estimated using conjugate gradient descent, which allows to avoid estimation and inversion of kernel matrix $\mathbf{K_y}$ estimated from the data. In the absence of round-off errors, this would allow to obtain the exact solution after N iterations. In practice, however, the algorithm can be stopped significantly earlier when the norm of the residual drops below a predefined threshold.

For the estimation of the sample variance, which is required for the acquisition function of the Gaussian process, the inverse of the matrix $\mathbf{K_y}$ is needed. However, to reduce the computational demand, the variance can be approximated with a fewer number of eigenvectors. For the estimation of the eigenvalues, the Lanczos algorithm (Lanczos, 1950) with full reorthogonalization is used. To estimate the eigenvalues from the resulting tridiagonal matrix, the strum sequences are estimated and bounded by bisection. The corresponding eigenvectors are estimated using inverse iteration. Although this requires the inversion of the kernel matrix, this is achieved by applying the conjugate gradient descent again. The determinant of the kernel matrix, likelihood, and prediction variance are approximated using the approach presented in Rodner et al. (2016). The implementation of the Gaussian processes with the described approximations for the HIK is implemented as an *R* (R Core

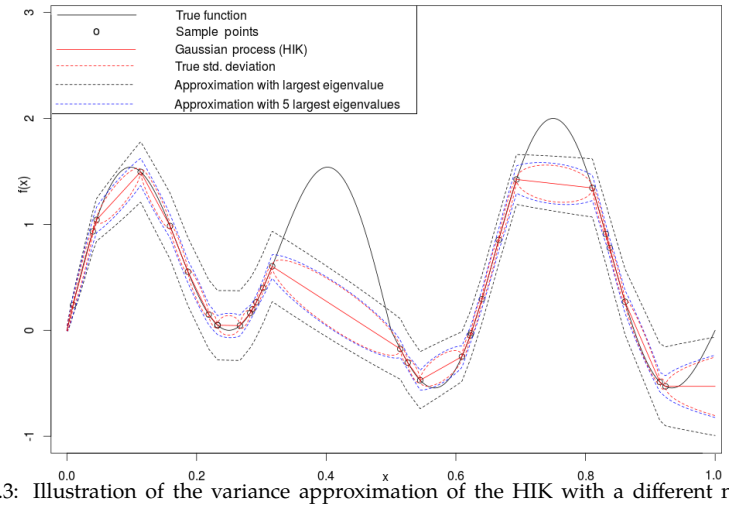


FIGURE 9.3: Illustration of the variance approximation of the HIK with a different number of eigenvectors

Development Team, 2014) package, to provide other researchers with easy access to the package for their research.

The Bayesian optimization algorithm is implemented in *R* using the implemented package of the Gaussian process. For the optimization of the acquisition function, the *genoud* (Goslee et al., 2007) package is used, which is a combination of a genetic algorithm for the global and quasi-Newton method for the local optimization. The hyperparameters that are required for the exponential transformation and shift of the data are estimated by maximizing the likelihood of the Gaussian process. The likelihood is optimized using the downhill simplex algorithm (Nelder and Mead, 1965), which is a gradient-free optimization algorithm. This circumvents estimation of the gradient, which would also require an inversion of the kernel matrix.

An example of the resulting approximation of the variance with a varying number of eigenvalues is displayed in Figure 9.3. From the unknown cost function, 30 points were randomly sampled. The Gaussian process with HIK is fitted to the data and the standard deviation, as well as their approximation with the largest eigenvector and largest five largest eigenvectors, are plotted for comparison.

9.3 Results

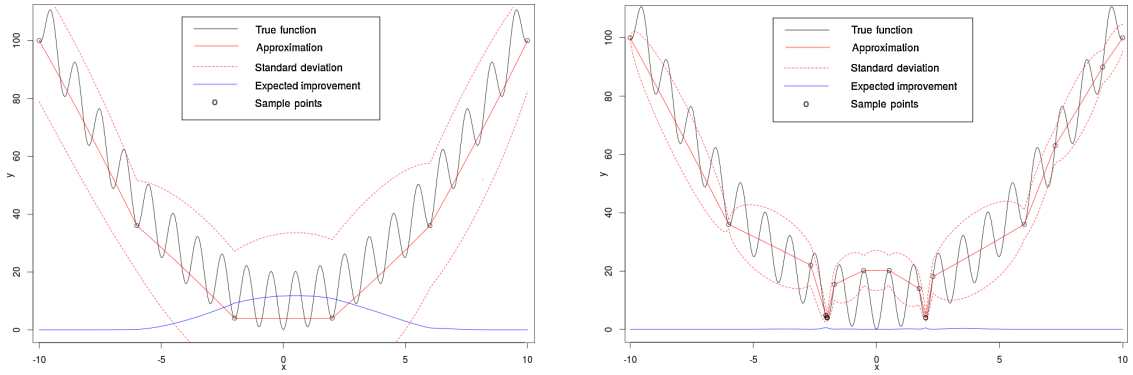
To analyze and visualize the fitting capabilities of the HIK, we consider a 1-D case of the Rastrigin function and 2-D case of the Branin function for optimization. Afterward, we will compare the results with the EGO algorithm because it uses a non-linear approximation of the error function and has been shown to find global optima surprisingly quick. Finally, we will conclude with the speedup that is provided by a parallel implementation of the fast kernel matrix-vector multiplication.

9.3.1 Visual analysis

For the illustration of the Bayesian optimization algorithm with HIK we visually analyze the optimization of the 1-D Rastrigin and 2-D Branin function. The EGO algorithm uses the expected improvement acquisition function to identify the next point to evaluate on the true objective function. Therefore, we will analyze the use of the expected improvement, but will also consider the use of the UCB acquisition function for reasons that will become apparent later. For the Rastrigin function to be optimized, we assume the function range [-10, 10]. Since the HIK can only be used for positive numbers, the data is shifted into the positive

range. As stated in the previous section, any transformation can be applied to the data while the kernel still remains positive definite. Additionally, the exponential transformation is used to increase the fitting capabilities and an additional positive shift of the data because otherwise, the HIK will always predict a value of 0 for an x-position of 0. The magnitude of the additional shift to the data is estimated from the data as additional hyperparameter.

For the optimization of the Rastrigin function, initially, six equally spaced points are chosen to start the Bayesian optimization. The number of utilized eigenvectors influences the quality of the predicted variance. Since the estimation of the eigenvectors is considered time intensive, we divide the number of samples by four and round up this value and estimate this amount of eigenvectors for the variance approximation. This appears to provide a good appropriation to the true variance especially in areas with only a few samples. However, in areas with many samples, the approximation loses accuracy, which can be hindering during the exploitation phase.



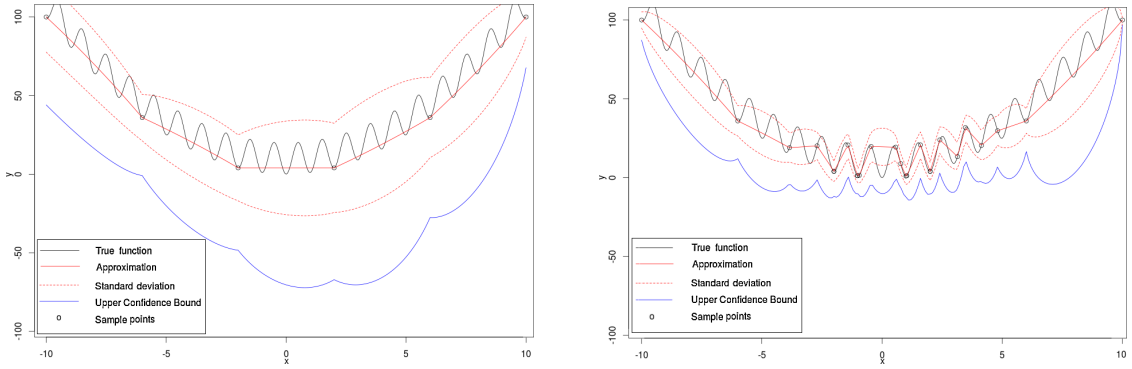
(a) Bayesian optimization of the Rastrigin function with an initial estimation of six points.

(b) Optimization results of the Rastrigin function after 20 function estimations utilizing the EI acquisition function.

FIGURE 9.4: Bayesian optimization of the Rastrigin function

The initial approximation with the six estimated function values is shown in Figure 9.4(a). The HIK can approximate the overall shape of the function and also the variance appears to capture the overall uncertainty of the Rastrigin function. Furthermore, the expected improvement signals where to estimate the next true function value and also indicates the correct location of the global minimum. The acquisition function is optimized to estimate the point with the highest expected improvement and estimated as the next point on the true cost function.

Figure 9.4(b) shows the status of the Bayesian optimization algorithm with 20 estimated true function values (including the six initial estimations). The algorithm was not able to locate the global minimum and got stuck in two local optima while underestimating the variance. This too small variance effectively prevents the algorithm to search near the area of the global optimum. This results in a flattened expected improvement function, which suggests that the optimal solution might already be found and prevents further search. The EI apparently underestimates the variance and stagnates at a local minimum. To counter this shortcoming, we apply the UCB acquisition function, which potentially emphasizes more on exploration. The UCB function requires a parameter β which is a factor that weights the standard deviation. It effectively balances the trade-off between exploration and exploitation. The choice of this parameter might require domain knowledge, and no clear recommendations can be given. However, for the following analysis of the Rastrigin function, a parameter of $\beta = 2.5$ is chosen to upscale the influence of the variance.



(a) Initial estimate of Rastrigin function utilizing the UCB acquisition function.

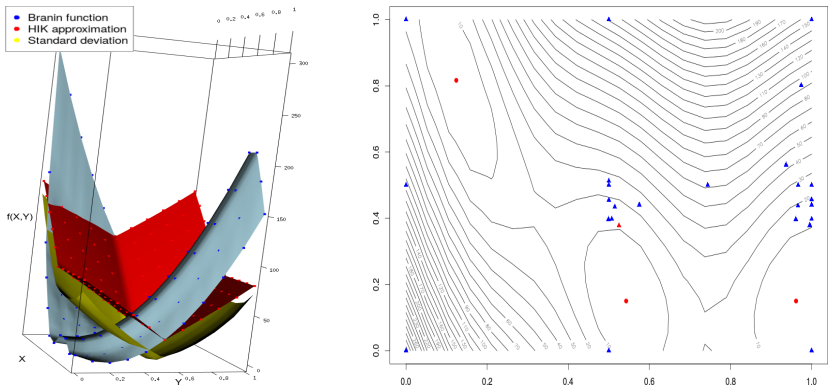
(b) Optimization results of the Rastrigin function after 20 function estimations utilizing the UCB acquisition function.

FIGURE 9.5: Bayesian optimization of the Rastrigin function utilizing the UCB acquisition function

Figure 9.5(a) shows the initial estimates of the UCB acquisition function. In this Figure, the acquisition function is represented by the mean value plus times 2.5 the predicted standard deviation. This emphasizes the uncertainty because it could be noticed from the previous analyzes, that the HIK kernel appears to underestimate the variance. In contrast to the expected improvement plot, the representation of the acquisition function is inverted to provide better visibility. Therefore, we wish to estimate the solution with the lowest UCB value as the next point on the true function.

Figure 9.5(b) shows the status of the Bayesian optimization after 20 function evaluations using the UCB acquisition function. Although the UCB acquisition function was not able to estimate the global optimum either, it can be noticed that in contrast to the expected improvement function, the local optimum can still be estimated due to the remaining uncertainty.

After this visual analysis of the one-dimensional case, we move on to the two-dimensional case. For the two-dimensional case, the 2-D Branin function in an interval $[0, 1]$ is considered. For the optimization of the Branin function, an initial grid of nine sample points is utilized and the resulting approximation of the HIK is displayed in Figure 9.6(a). It can be noticed that the approximation is quite poor and does not reflect the true function and that the sampled points are not represented accurately.



(a) Initial approximation of the Branin function with 9 sampled points.

(b) Bayesian optimization after 28 evaluations. Estimated points are shown in blue triangles, next point to evaluate is shown with red triangle. The location of the global optima are shown as red dots.

FIGURE 9.6: Bayesian optimization of the Branin function

Figure 9.6(b) illustrates the optimization results after a total of 28 function evaluations. It can be seen that the sampled points are all sampled in the area of the HIK approximation falsely introduced minima. Additionally, the variance of the HIK approximation is considerably low, which prevents the algorithm from searching in areas with a lower error. The considerably low variance could be explained with the assumptions of the linear kernel, which reflects the assumption made on the true cost function. It assumes that unknown values are within a linear range of two known values. This leads to an underestimation of the variance which effectively prevents the algorithm from exploring the optimal regions. Based on these observations, we assume that using the EI acquisition function does not allow to move away from the falsely introduced minima. Therefore the UCB acquisition function is applied in the following analysis.

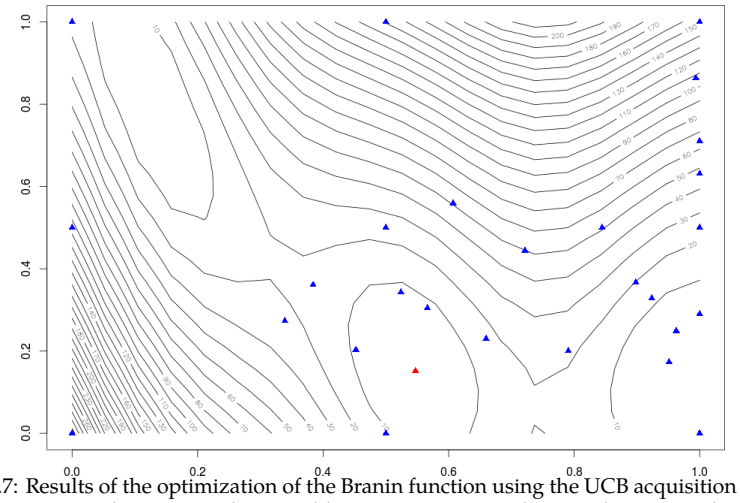
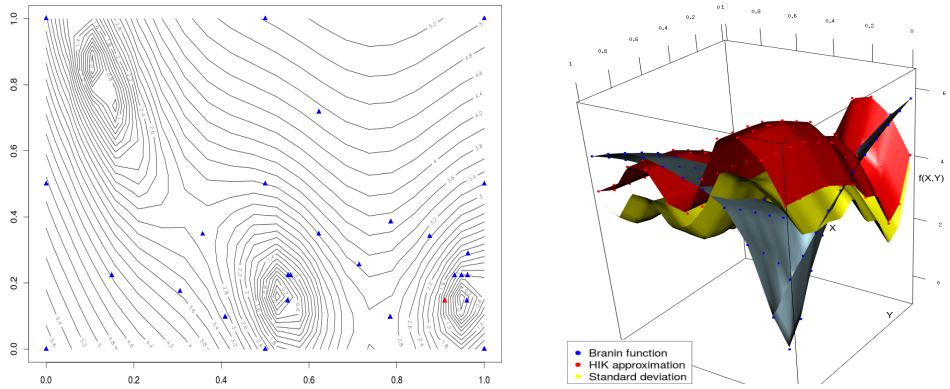


FIGURE 9.7: Results of the optimization of the Branin function using the UCB acquisition function. Estimated points are shown in blue, next point to evaluate is shown in red.

Figure 9.7 shows that the UCB acquisition function explores the error surface better than the EI acquisition function. Apparently, the search focuses on two of the three available minima. A further observation that can be made from the 2-D case is that the HIK kernel is not able to fit an abrupt increase in the cost function. It accordingly fails to approximate the true cost function and even the true measurement values. To alleviate the problem of rapidly increasing values, we apply the log-transformation to the problem. Results of this transformation for the UCB function are shown in Figure 9.8(a).

Applying the log-transformation allows the HIK to provide a more accurate fit of the surface and enables the Bayesian optimization algorithm to further reduce the error. Although the HIK appears to approximate the function to some degree, it still assumes very large values around the third minima as shown in Figure 9.8(b), which prevents the discovery of this minima.



(a) Contour plot of the log Branin function after 28 function evaluations. Estimated points are shown in blue, next point to evaluate is shown in red.

(b) 3-D plot of the approximation after 28 function evaluations.

FIGURE 9.8: Approximation of the log transformed Branin function after 28 function evaluations.

9.3.2 Comparison of optimization results

After the visual analysis of the fitting capabilities of the HIK for Bayesian optimization, we compare the algorithm to the EGO algorithm (Jones, Schonlau, and Welch, 1998). Therefore, we analyze the same test functions as presented in the paper of Jones, Schonlau, and Welch (1998) as well as their archived optimization results. For further comparison, the results of a Bayesian optimization with a squared experimental kernel applying the same settings as the HIK implementation are provided. To emphasize the exploration and avoid underestimation of the variance, a $\beta = 5.576$ is chosen for the UCB acquisition function. For all problems, we use one-fourth of the eigenvectors for variance approximation.

TABLE 9.1: Test function results for the EGO algorithm, the Bayesian optimization with HIK, and randomly taken samples.

Test problem	Samples evaluated	Error EGO	SEK UCB mean (std)	HIK UCB mean (std)	Random sample mean (std)
Branin	28	0.2%	8% (10%)	179% (270%)	407% (400%)
log Branin				94% (128%)	
Goldstein-Price	32	0.1%	25% (13%)	35% (10%)	34% (16%)
log Goldstein-Price				30% (6%)	
Hartman 3	34		4% (5%)	13% (7%)	16% (12%)
log Hartman 3		1.7%		5% (3%)	
Hartman 6	84		7% (6%)	49% (18%)	42% (14%)
log Hartman 6		1.9%		36% (9%)	

Table 9.1 shows the results on the test problems for the EGO, the Bayesian optimization with HIK and SEK utilizing the UCB acquisition function, and a reference measure where the same amount of samples is taken randomly from the search space. Since the implementation of the Bayesian optimizations does not provide deterministic results, due to the non-deterministic optimization of the acquisition function, the function was optimized 50 times and the mean and standard deviation are provided. The same applies to the reference measure of randomly taken samples.

9.3.3 Time requirements and parallelization

Finally, we analyze the improvement in each iteration and runtime of an implementation of the EGO algorithm and the implemented Bayesian optimization using the HIK. For the estimation of the runtime and error of the EGO algorithm, the *DiceOptim* (Roustant, Ginsbourger, and Deville, 2012) package for R is used. This package provides an implementation of the EGO optimization algorithm. However, in contrast to the original EGO algorithm presented in the paper by Jones, Schonlau, and Welch (1998), that utilizes a bound on the expected improvement function for their optimization, the *DiceOptim* package relies on the *genoud* package for the optimization of the acquisition function. To make both algorithms comparable, the same settings for the *genoud* package have been used in both algorithms. The population size was set to 20 and the number of generations was set to 12. The generation size of 12 is the default setting utilized in the *DiceOptim* package and a population size of 20 has been utilized in the example, provided in the package documentation, for the optimization of the Branin function. However, it might be argued that these settings are quite low for higher-dimensional problems. Since both algorithms will not provide deterministic outcomes, the optimization was repeated 50 times. The results for the log-transformed Hartman3 optimization function are shown Figure 9.9.

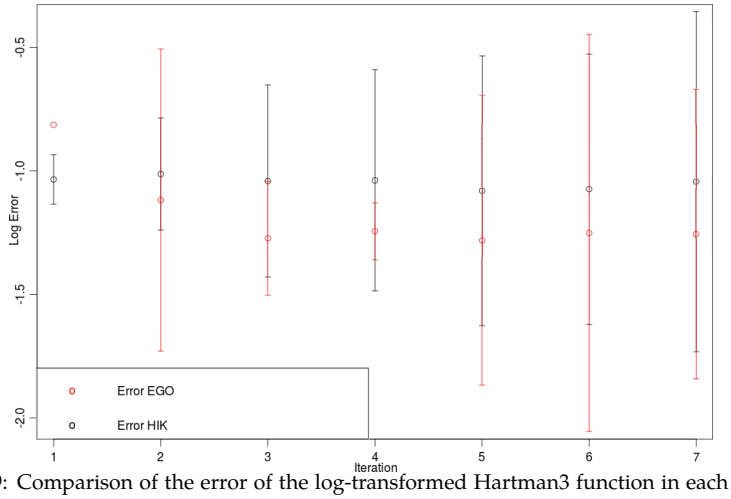


FIGURE 9.9: Comparison of the error of the log-transformed Hartman3 function in each iteration of the Bayesian optimization using EGO and HIK.

For the comparison on the log-transformed Hartman3 function, it appears that the Bayesian optimization with the HIK does not reduce the error as much as the EGO algorithm within the same amount of true objective function evaluations. The execution time of the EGO algorithm was on average 2.40 seconds with a standard deviation of 0.18 seconds. The average execution time of the Bayesian optimization implementation with HIK was 1.66 seconds with a standard deviation of 0.15 seconds.

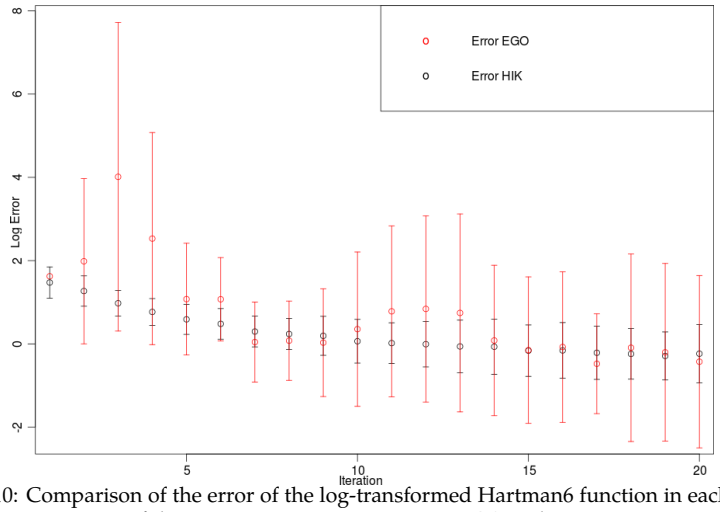


FIGURE 9.10: Comparison of the error of the log-transformed Hartman6 function in each iteration of the Bayesian optimization using EGO and HIK.

The comparison of both algorithms for the log-transformed Hartman6 problem is shown in Figure 9.10. In the early iterations, it appears that the EGO algorithm emphasis more on exploring than the HIK. With increasing iterations, the EGO algorithm can improve more than the HIK implementation and also yields an overall better optimization result, which is also indicated by the larger variance of the error. Surprisingly, the average execution time for the HIK was longer than the execution time of the EGO algorithm. The EGO implementation required on average 8.61 seconds with a standard deviation of 0.45 seconds. In comparison, the HIK implementation required on average 17.43 seconds with a standard deviation of 1.58 seconds. This might be due to the increasing number of eigenvectors that are to be estimated for the six-dimensional Hartman6 function. The influence of the number of estimated eigenvectors on the runtime is discussed next as well as a possibility to reduce this runtime.

The required execution time to estimate the Gaussian process utilizing the HIK can be reduced, by recognizing that the fast vector kernel multiplication is independent in each dimension. Therefore, the multiplication of the histogram intersection kernel matrix with an arbitrary vector can be parallelized. To analyze this potential speed improvement 1000 data points with 100 dimensions were randomly sampled from the Rastrigin function. Then the Gaussian process with HIK with an increasing number of eigenvectors was estimated. This has been repeated 25 times and the results for sequential matrix-vector and parallel matrix-vector multiplication are shown Figure 9.11. For the parallel estimation of the Gaussian process, 8 threads were used.

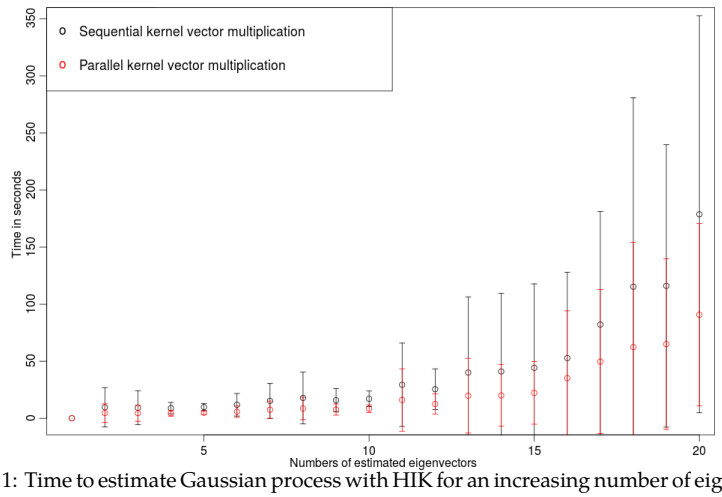


FIGURE 9.11: Time to estimate Gaussian process with HIK for an increasing number of eigenvectors.

It can be noticed that the runtime quadratically increases with the number of estimated eigenvectors. This complexity increase might mostly be due to the utilized Lanczos algorithm for the estimation of the eigenvalues. Furthermore, a large variance in the estimation time can be noticed. This is due to the varying convergence speed of the individual parts of the implementation. The convergence speed of the utilized conjugate gradient descent depends on the starting vector. Therefore, the algorithm requires more or fewer iterations till convergence. The same applies to the inverse iteration method for the estimation of the eigenvectors, which also utilizes the conjugate gradient descent function. Depending on the starting vector it requires more or fewer iterations until the vector has the desired accuracy. However, parallelization of the fast vector-matrix multiplication can speed up the estimation of the Gaussian process for an increasing number of estimated eigenvectors.

9.4 Discussion

The implemented Bayesian optimization algorithm with HIK for the approximation of the true cost function does not provide solutions that are comparable to the EGO algorithm, which utilizes a squared exponential kernel for function approximation. Furthermore, the analyses presented in this paper, show that the piecewise linear approximation of the HIK cannot accurately approximate the true cost function. The underestimation of the variance and inability to represent observed data points, as shown in the visual analysis, leads to the introduction of false minima. This suggests that poor approximation models can lead to large approximation errors and introduce false minima, which has already been shown in the case of genetic algorithms with surrogate function (Jin, Olhofer, and Sendhoff, 2000). The same problem appears to apply here for the optimization with Bayesian optimization and the HIK. In conclusion, this emphasizes the role of the covariance function as assumptions that are made on the cost function to be optimized. Although the use of a log-transformation to the problem, to smooth out steep increases in the cost function, improves the performance of the algorithm, the problems remain.

This analysis further suggests that the fitting capability of the HIK appears to decrease with an increase in dimensions. Where the illustration of the one-dimensional problem indicates a sufficient fit of the surface, in the two-dimensional case, the fitting capabilities greatly suffers. This is also suggested by the results of the comparison of different optimization problems. Where the difference between the remaining error of the HIK and the random sample decreases with an increase in dimension. However, it appears that the HIK can greatly benefit from a log-transformation of the objective function, and is then able to perform better than random sampling. The effect of introducing false minima and underestimation of the variance might be specifically apparent in the cause of the Goldstein-price optimization function. The Goldstein-price optimization function has 3 minima where one of them is the global minimum. Apparently, the optimization using the HIK focuses around one local minimum and the lack of approximation capabilities prevents the algorithm to explore the global optimum similar. This effect might be identical to the one illustrated for the log-transformed Branin function.

Although the implemented Bayesian optimization with the HIK has linear memory requirements, the estimation of the eigenvectors scales quadratically in terms of runtime for the utilized approximations. However, a higher number of eigenvalues is only required for a more accurate approximation of the sample variance. A lower number of eigenvalues might be sufficient for the envisioned optimization problem and the runtime for an increasing number of eigenvectors can be reduced utilizing parallelization.

9.5 Conclusion

In this paper, the use of Bayesian optimization with HIK was analyzed and compared to the results of the EGO optimization algorithm. The HIK can provide linear runtime and memory requirements, which makes it attractive for large problems. However, estimation of the eigenvectors that are required for prediction of the sample variance is computationally expensive. To improve the speed and memory requirements, approximations for the prediction variance and log-likelihood for hyperparameter estimation are utilized. Despite the benefits, the HIK possesses limitations and only provides a piecewise linear approximation of the true function. Although the definition of the kernel can be generalized to improve its fitting capabilities, it appears unsuited for the use as the surrogate function in optimization.

The visual inspection and comparison with the EGO algorithm show that a poor-fitting surrogate function can be hindering for optimization purposes. Although the HIK kernel can provide better results in image recognition tasks than other traditional kernels, potentially due to generalizing well among multiple dimensions, this generalizability is not suited for optimization tasks, where an accurate fit of the surrogate function is preferable.

When applying approximations for the predictive variance, the amount of the eigenvalues influences the performance in the context of Bayesian optimization. Although the estimation of fewer eigenvalues is beneficial in terms of runtime and memory requirement, the approximated variance can significantly differ from the variance estimated with all eigenvalues in areas with many samples. This mainly influences the exploitation phase, where the Bayesian optimization refines the solutions around a currently estimated minimum. Therefore, the choice of the numbers of eigenvalues influences the performance of the algorithm. Also, the choice of the β parameter when applying the UCB acquisition function influences the behavior and performance of the Bayesian optimization.

It appears that a well-fitting model that adequately represents the data and the variance is necessary to steer search into promising areas and to allow an efficient exploitation to enable good results.

References

Boughorbel, Sabri, Jean-Philippe Tarel, and Nozha Boujemaa (2005). "Generalized Histogram Intersection Kernel for Image Recognition". In: *{IEEE} International Conference on Image Processing* 3.October, pp. III–161. DOI: 10.1109/ICIP.2005.1530353 (cit. on p. 157).

Brochu, E, V M Cora, and N De Freitas (2010). "A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning". In: *ArXiv*, p. 49. DOI: 1012.2599. arXiv: 1012.2599 (cit. on pp. 155, 158).

Cox, Dennis D. and Susan John (1997). "SDO: A statistical method for global optimization". In: *Multidisciplinary Design Optimization*, pp. 315–329. DOI: 10.1109/ICSMC.1992.271617 (cit. on pp. 158, 159).

Deisenroth, Marc Peter and Jun Wei Ng (2015). "Distributed Gaussian Processes". In: *International Conference on Machine Learning* 37, p. 10. DOI: 10.1016/j.physa.2015.02.029. arXiv: 1502.02843 (cit. on p. 155).

Gelbart, Michael A., Jasper Snoek, and Ryan P. Adams (2914). "Bayesian Optimization with Unknown Constraints". In: *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*. UAI'14. Quebec City, Quebec, Canada: AUAI Press, pp. 250–259. ISBN: 978-0-9749039-1-0 (cit. on p. 155).

Goslee, Sarah C et al. (2007). "Journal of Statistical Software". In: *Journal Of Statistical Software* 22.11 (cit. on p. 160).

Hensman, James, N Fusi, and Neil D. Lawrence (2013). "Gaussian Processes for Big Data". In: *UAI*, pp. 282–290. ISBN: 978-1-4503-1285-1. DOI: 10.1162/089976699300016331. arXiv: 1309.6835 (cit. on p. 155).

Jin, Y, M Olhofer, and B Sendhoff (2000). "On Evolutionary Optimisation with Approximate Fitness Functions". In: *Proc. of the 2000 Genetic and Evolutionary Conference ({GECCO} 2000)* JANUARY, pp. 786–793 (cit. on p. 167).

Jones, Donald R, Matthias Schonlau, and William J Welch (1998). "Efficient Global Optimization of Expensive Black-Box Functions". In: *Journal of Global Optimization* 13, pp. 455–492. ISSN: 09255001. DOI: 10.1023/a:1008306431147. arXiv: 0005074v1 [arXiv:astro-ph] (cit. on pp. 155, 158, 164, 165).

Kleijnen, Jack P C (2014). "Simulation-optimization via Kriging and bootstrapping: a survey". In: *Journal of Simulation* 8.4, pp. 241–250. ISSN: 1747-7778. DOI: 10.1057/jos.2014.4 (cit. on p. 155).

Kushner, H J (1964). "A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise". In: *Journal of Basic Engineering* 86.1, pp. 97+. ISSN: 00219223. DOI: 10.1115/1.3653121 (cit. on pp. 155, 158).

Lanczos, C. (1950). "An iteration method for the solution of the eigenvalue problem of linear differential and integral operators". In: *Journal of Research of the National Bureau of Standards* 45.4, p. 255. ISSN: 0091-0635. DOI: 10.6028/jres.045.026 (cit. on p. 159).

Maji, Subhransu, Alexander C Berg, and Jitendra Malik (2008). "Classification using Intersection Kernel Support Vector Machines is Efficient Classification using Intersection Kernel Support Vector Machines is Efficient". In: *Slides*, pp. 1–8. ISBN: 9781424422432. DOI: 10.1109/CVPR.2008.4587630 (cit. on pp. 8, 155, 157).

Mockus, J, V Tiesis, and A Zilinskas (1978). "The application of Bayesian methods for seeking the extremum". In: *Towards global optimisation. II* December 2016, pp. 117–129 (cit. on p. 155).

Nelder, J. A. and R. Mead (1965). "A Simplex Method for Function Minimization". In: *The Computer Journal* 7.4, pp. 308–313. ISSN: 0010-4620. DOI: 10.1093/comjnl/7.4.308 (cit. on p. 160).

Quiñonero-candela, Joaquin, Carl Edward Rasmussen, and Ralf Herbrich (2005). "A unifying view of sparse approximate Gaussian process regression". In: *Journal of Machine Learning Research* 6, pp. 1935–1959. ISSN: 1533-7928 (cit. on p. 155).

R Core Development Team (2014). *R: a language and environment for statistical computing, 3.1.2 ed.* R Foundation for Statistical Computing. Vienna, Austria. ISBN: 3-900051-07-0 (cit. on pp. 79, 81, 159).

Rasmussen, CE. and CKI. Williams (2006). *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. Cambridge, MA, USA: MIT Press, p. 248 (cit. on p. 156).

Rodner, Erik et al. (2012). "Large-scale Gaussian process classification with flexible adaptive histogram kernels". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 7575 LNCS. PART 4, pp. 85–98. ISBN: 9783642337642. DOI: 10.1007/978-3-642-33765-9_7 (cit. on p. 155).

Rodner, Erik et al. (2016). *Large-Scale Gaussian Process Inference with Generalized Histogram Intersection Kernels for Visual Recognition Tasks*. DOI: 10.1007/s11263-016-0929-y (cit. on pp. 155, 159).

Roustant, Olivier, David Ginsbourger, and Yves Deville (2012). "DiceKriging, DiceOptim: Two R Packages for the Analysis of Computer Experiments by Kriging-Based Metamodeling and Optimization". In: *Journal of Statistical Software, Articles* 51.1, pp. 1–55. ISSN: 1548-7660. DOI: 10.18637/jss.v051.i01 (cit. on p. 165).

Snelson, Edward and Zoubin Ghahramani (2006). "Sparse Gaussian Processes using Pseudo-inputs". In: *Advances in Neural Information Processing Systems 18*. Ed. by Y Weiss, P B Schölkopf, and J C Platt. MIT Press, pp. 1257–1264 (cit. on p. 155).

Snoek, Jasper, Hugo Larochelle, and Ryan P Adams (2012). "Practical Bayesian Optimization of Machine Learning Algorithms". In: *Adv. Neural Inf. Process. Syst. 25*, pp. 1–9. ISSN: 10495258. DOI: 2012arXiv1206.2944S. arXiv: arXiv:1206.2944v2 (cit. on pp. 8, 155).

Swersky, Kevin, Jasper Snoek, and Ryan P. Adams (2013). "Multi-Task Bayesian Optimization". In: *Advances in Neural Information Processing Systems 26*, pp. 2004–2012. ISSN: 10495258. arXiv: arXiv:1406.3896v1 (cit. on p. 155).

Williams, Christopher and Matthias W. Seeger (2001). "Using the Nystrom Method to Speed Up Kernel Machines". In: *NIPS Proceedings* 13, pp. 682–688. ISSN: 1098-6596. DOI: 10.1017/CBO9781107415324.004. arXiv: arXiv:1011.1669v3 (cit. on p. 155).

Wu, Jianxin (2010). "A fast dual method for HIK SVM learning". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 6312 LNCS.PART 2, pp. 552–565. ISSN: 03029743. DOI: 10.1007/978-3-642-15552-9_40 (cit. on pp. 155, 157).

Wu, Jianxin, Wc Tan, and Jm Rehg (2011). "Efficient and effective visual codebook generation using additive kernels". In: *The Journal of Machine Learning Research* 12, pp. 3097–3118. ISSN: 15324435 (cit. on pp. 8, 155, 157).