Looking into the Future:

How to Use Advanced Statistical Methods for
Predicting Psychotherapy Outcomes in Routine Care

Carl-Johan Uckelstam

Looking into the Future:

How to Use Advanced Statistical Methods for Predicting Psychotherapy Outcomes in Routine Care

Carl-Johan Uckelstam

Linköping Studies in Arts and Sciences No. 803
Linköping Studies in Behavioural Science No. 226
Faculty of Arts and Sciences
Linköping 2021

Linköping Studies in Arts and Sciences • No. 803
Linköping Studies in Behavioural Science • No. 226

At the Faculty of Arts and Sciences at Linköping University, research and doctoral studies are carried out within broad problem areas. Research is organized in interdisciplinary research environments and doctoral studies mainly in graduate schools. Jointly, they publish the series Linköping Studies in Arts and Sciences. Division of Psychology at the Department of Behavioural Sciences and Learning.

Distributed by:
Department of Behavioural Sciences and Learning
Linköping University
581 83 Linköping

Carl-Johan Uckelstam
Looking into the Future:
How to Use Advanced Statistical Methods for Predicting Psychotherapy
Outcomes in Routine Care

Edition 1:1
ISBN 978-91-7929-709-1
ISSN 0282-9800
ISSN 1654-2029

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

<https://creativecommons.org/licenses/by-nc/4.0/>

©Carl-Johan Uckelstam
Department of Behavioural Sciences and Learning, 2021

Cover illustration by: Jacob Stålhammar
Printed by: LiU-Tryck, Linköping 2021

“The ideal scientist thinks like a poet and works like a bookkeeper...”

(Wilson, 1998, p. 62)

ABSTRACT

Background: Psychotherapy research has shifted from mainly focusing on the average effects of different treatments to concentrating more on questions related to the individual patient. This aligns with the goals of precision medicine and patient-focused research and might include studying predictors to forecast important patient behaviors, such as premature termination or expected change rates.

When research attention shifts, it often gives rise to the implementation of new statistical methods that, in turn, can illuminate new challenges that must be addressed. For example, the fields of classical statistics and data science have merged in interesting ways in recent times, which has led to an expansion of the methodological toolbox available for psychotherapy research. However, some strengths and limitations of these new approaches should be recognized as they come from slightly different research traditions. For instance, the broader goals of scientific exploration, such as description, prediction, and explanation, might be addressed and given different importance under these approaches.

Studying these matters in the context of routine care leads to specific considerations that must be reflected upon, such as how to define relevant outcomes and draw sound conclusions from observational data collected in a naturalistic setting.

Aim: The aim of the thesis was to concretize these matters by contrasting how traditional methods for predicting certain psychotherapy outcomes have been studied in the past, and how more advanced statistical methods might be used to enhance knowledge of how to predict these outcomes today.

Studies/Results: Three studies were performed: Paper I focused on how Multi Level Modeling (MLM) can be used to study aspects of the relationship between working alliance and treatment outcome that has been overlooked by most earlier studies. The classical way of studying this question has been by using simple correlation analysis to explore the association between *patient*-rated alliance and treatment outcome. However, with the help of MLM and ratings from both parties of the therapeutic dyad, it is possible to analyze how these ratings relate to each other at different levels in the data (i.e., from patient to patient within the same therapist, denoted *within therapist* herein, and from therapist to therapist, denoted *between therapists* herein). MLM also makes it possible to explore the degree of correlation between working alliance and treatment outcome when studied between compared to within therapists, which might have importance for how to direct further studies of this association.

Latent Profile Analysis (LPA) and item-level analysis were used in Paper II to give nuance to the relationship between psychological distress at baseline and change rate during treatment. By using LPA and exploring all 34 items in the patient-rated self-assessment Clinical Outcomes in Routine Evaluation Outcome Measure (CORE-OM), it was possible to show that specific compositions of symptom distress at the start of treatment, and not just high or low levels of distress (which is the conventional way of studying this relationship), can be associated with change rate.

Machine Learning (ML) was used in Paper III to extend the practice of using item-level analysis from CORE-OM ratings. This was done to detect dropout patients in the early phase of treatment. An advanced ML algorithm called Random Forest (RF) was used to explore complex patterns of symptom distress and change in item response to predict dropout during the early phase of treatment. Explored with common practices for ML (e.g., supervised learning, cross validation, model training and testing), results indicated that dropout cases could be detected with high accuracy in the sample. This is an extension of the majority of earlier studies on dropout that have explored more or less static predictors collected before the treatment had started (e.g., age, education).

Conclusions: The papers included in the thesis showed how advanced statistical methods can be used to enhance our understanding of how to predict important psychotherapy outcomes that have been overlooked with the use of traditional methods. The thesis also exemplified how different goals of scientific exploration can be studied in the context of routine care with the use of different statistical frameworks and discussed some of the challenges and opportunities worth noting when entering this line of research.

Keywords: Statistics, machine learning, multilevel modeling, latent profile analysis, prediction, psychotherapy, routine care, outcome

ACKNOWLEDGEMENTS

I wish to acknowledge **Västra Götalandsregionen, Kungälv's sjukhus** and **Forskningsrådet för Hälsa, Arbetsliv och Velfärd (FORTE)**. Without their generous funding, this project would not have been possible.

I wish to express my deepest gratitude to my supervisors.

Fredrik Falkenström, my main supervisor during the final parts of my doctoral studies and co-supervisor during the early parts. He sparked my interest in advanced statistics and defined the path of my research. He truly has the elements of a genius and his warm, humble, and precise advice has been inspirational and essential at each step of this project.

Björn Philips, my main supervisor during the early parts of my doctoral studies and co-supervisor during the final parts. He is the perfect match between a top researcher and a skillful clinician. His ability of connecting the dots between abstract statistics and the realities of the clinician and patient has been crucial for highlighting the essence of this project. His sympathetic ear, sound advice and creative thinking have been specifically valuable throughout this process.

Rolf Holmqvist, my co-supervisor during the whole doctoral program. He is the ultimate role model and mentor. His positive energy, empathic and humble ways combined with a sharp mind and precise feedback has been particularly valuable to me. Without the generosity in sharing data from his study of psychological treatment in primary care in Sweden this thesis project would not have been possible. Rolf also supported the initial part of my doctoral studies with funding he received from FORTE (dnr 2012-0238). Rolf and his wife **Bodil Holmqvist** also generously housed me on numerous occasions on my visits to Linköping. I have truly enjoyed the stimulating discussion we have had during these times.

I would like to pay my special regards to **Katarina Andersson** the former head of the psychiatric unit at Kungälv's sjukhus, **Åsa Gustavsson** the former head of psykiatrimottagning Kungälv as well as **Mari Andersson** the current head of psykiatrimottagning Kungälv. Your pragmatic problem-solving abilities have been essential for making this thesis a reality. Your warm support and assistance in helping me devote the necessary time for finishing this thesis has been essential.

I would also like to acknowledge my former and current colleagues in my research group at IBL: **Paul Bergman, Malin Bäck, Annika Ekeblad, My Frankl, Ylva Gidhagen, Mattias Holmqvist Larsson, Niclas Kullgard, Börje Lech, Clara Möller, Tommy Skjulsvik, Erika Viklund** and **Anna-Karin Åkerman**. Thank you for supporting me and giving me valuable feedback over the years.

Finally, on a personal note, I wish to recognize the invaluable support of my true love **Märta Bergman** and our children **Tage** and **Ester**. Your warm encouragements and happy distractions have kept me smiling and focused on what is important in life. Märta has also provided insightful comments on the thesis. I love you all so much!

To my mom and dad, **Hasse** and **Ewa Uckelstam** my greatest's supporters, thank you for always believing in me and supporting me in every aspect of this project and other important parts of my life.

To my sister **Joanna Uckelstam** and her partner **Gerald Riessland** for your loving support and for cheering me on and helping me find time outside work to boost my energy.

To my mother-in-law **Lena Bergman** for providing an excellent writing environment and for your support and interesting discussions regarding clinical implications of the thesis over a nice cup of coffee.

.

LIST OF PAPERS

- I. **Uckelstam, C.-J.**, Holmqvist, R., Philips, B., & Falkenström, F.
(2018). A relational perspective on the association between working alliance and treatment outcome. *Psychotherapy Research*, 30(1), 13–22. <https://doi.org/10.1080/10503307.2018.1516306>
- II. **Uckelstam, C.-J.**, Philips, B., Holmqvist, R., & Falkenström, F.
(2019). Prediction of treatment outcome in psychotherapy by patient initial symptom distress profiles. *Journal of Counseling Psychology* 66(6), 736–746. <http://dx.doi.org/10.1037/cou0000345>
- III. **Uckelstam, C.-J.**, Philips, B., Holmqvist, R., & Falkenström, F.
(manuscript). Predicting psychotherapy dropout with patient initial symptom change patterns: A machine learning approach.

ABBREVIATIONS

AI	Artificial Intelligence
ANOVA	Analysis of Variance
AUC	Area Under the receiver operating characteristics Curve
BAI	Becks Anxiety Inventory
BDI	Becks Depression Inventory
CBT	Cognitive Behavioral Therapy
CDC	Center for Disease Control and Prevention
CORE-OM	Clinical Outcomes in Routine Evaluation Outcome Measure
CT	Cognitive Therapy
FIML	Full Information Maximum Likelihood
GLM	Generalized Linear regression Model
HRSD	Hamilton Rating Scale for Depression
ICBT	Internet Delivered Cognitive Behavioral Therapy
IDD	Independent and Identically Distributed
IPT	Interpersonal Psychotherapy
LCA	Latent Class Analysis
LGCM	Latent Growth Curve Model
LOCF	Last Observation Carried Forward
LPA	Latent Profile Analysis
MADRS	Montgomery-Åsberg Depression Rating Scale
MANOVA	Multivariate Analysis of Variance
MAR	Missing At Random
ML	Machine Learning
MLM	Multi Level Modeling
OQ-45	Outcome Questionnaire-45
PAI	Personalized Advantage Index
PDT	Psychodynamic Therapy
RCT	Randomized Controlled Trial (RCT)
RF	Random Forest
ROM	Routine Outcome Monitoring
SEM	Structural Equation Modeling
SMR	Social Relations Model
SVM	Support Vector Machines
WAI	Working Alliance Inventory

CONTENTS

INTRODUCTION.....	1
FROM AVERAGE CAUSAL EFFECTS TO PRECISION MEDICINE.....	1
<i>Precision medicine</i>	2
DESCRIPTION, PREDICTION, AND EXPLANATION.....	5
<i>The link between scientific goal and research design</i>	7
<i>The link between research design and study setting</i>	7
<i>Internal validity</i>	8
<i>External validity</i>	8
<i>Moderation analysis</i>	10
<i>Occam's razor vs Hickam's dictum</i>	11
DEFINITIONS OF OUTCOMES	12
<i>Self-assessments</i>	12
Core batteries	13
Change between two time points.....	14
Growth rate.....	15
<i>Dropout as an outcome of psychotherapy</i>	17
TRADITIONAL METHODS FOR PREDICTING THE THESIS OUTCOMES.....	20
<i>Working alliance and treatment outcome</i>	20
<i>Initial symptom distress and change rate</i>	21
<i>Change in symptom distress and dropout</i>	23
ADVANCED METHODS FOR STUDYING THE THESIS OUTCOMES.....	25
<i>Multilevel modeling</i>	25
<i>Latent Profile Analysis</i>	26
<i>Machine learning</i>	28
AIMS OF THE THESIS.....	35
SUMMARY OF INCLUDED STUDIES	37
PAPER I	37
PAPER II	39
PAPER III.....	41
GENERAL DISCUSSION	45
DISCUSSION OF FINDINGS	45
STRENGTHS AND LIMITATIONS.....	49
CLINICAL IMPLICATIONS AND FUTURE RESEARCH.....	51
CONCLUSIONS	55
REFERENCES.....	57

INTRODUCTION

In the late 20th century, Howard, Moras, Brill, Martinovich and Lutz (1996) summarized three central and still relevant questions that could be asked about the effectiveness of any psychotherapy intervention: “(a) does it work under special, experimental conditions? (b) does it work in practice? and (c) is it working for this patient.” (Howard et al., 1996, p. 1059). The same authors also wrote that most studies at that time focused on the two first questions (a, b) but considerably less research has been dedicated to the third question (c) which is generally the most important factor for clinicians.

However, over the past decades, with the rise of practice-based research, Routine Outcome Monitoring (ROM), and precision medicine, studying matters relating to the individual patient (i.e., the third question) have become an international endeavor commonly referred to as patient-focused research. When attention shifts, it often leads to the development of innovative ways to study new questions. In turn, this shift also enables old questions to be readdressed with new approaches (e.g., via updated statistical frameworks and/or data collection procedures).

As psychotherapy research stands on the brink of a technological revolution that could lead to a paradigm shift as we enter the arena of big data, Machine Learning (ML), and Artificial Intelligence (AI), it is important to recognize some opportunities and challenges that arise when applying advance statistical modeling to explore psychotherapy outcomes in routine care.

From average causal effects to precision medicine

A drawback with most classical group-based studies, such as Randomized Controlled Trials (RCT), is that they only provide knowledge about average effects, which say very little about a specific patient. In other words, comparing groups of patients does not straightforwardly reflect the practical importance of treatment effects even when reporting the size of an effect. For example, patients who started treatment severely distressed might still have a lot of problems despite having made major improvements. Also, some patients might be unaffected by a treatment even though the treatment showed an average effect in a sample.

Studying predictors that might influence treatment outcome at the individual patient level is an important next step for psychotherapy research

to consider. Such studies might for example lead to the development of data-driven approaches to optimize recommendations and decisions that could be beneficial to the individual patient. However, these questions can become very complex from a statistical point of view compared to exploring the average effects of different treatments. Psychotherapy research is at the beginning of this development and has up to now predominantly applied classical statistical frameworks, such as regression analysis, to these types of studies.

Precision medicine

Using a patient's unique characteristics to tailor treatment is often denoted *personalized* medicine or *precision* medicine (Khoury & Galea, 2017). The phrase personalized medicine stems from the 1940's but is slowly being replaced with the more current term precision medicine out of concern that personalized might be misinterpreted to imply that treatments are being developed uniquely for an individual (DeRubeis, 2019)

A fundamental notion in precision medicine is that the average response to treatment does not always apply to the individual patient, and therefore, matching patients to the best treatment for their specific features might result in better treatment outcomes overall. The concept might also involve studying predictors to forecast important patient behaviors such as premature termination or expected change rates regardless of treatment type.

According to Fraguas et al. (2017), psychotherapy might be an especially suitable target for precision medicine because of the many factors that could influence treatment outcome (e.g., therapist and patient characteristics as well as other external factors). However, the complexity of the subject can also make statistical modeling challenging.

A contemporary example of precision medicine in psychotherapy is the so-called *Personalized Advantage Index* (PAI; DeRubeis et al., 2014), which was first introduced to aid treatment selection between Cognitive Behavioral Therapy (CBT) and the antidepressant drug paroxetine. This study used data from an RCT that examined differences in treatment outcome between CBT and paroxetine ($N = 154$) assessed using the Hamilton Rating Scale for Depression (HRSD; Hamilton, 1960)(DeRubeis et al., 2005). Guided by three earlier studies that examined predictors for treatment outcome using this data, the authors selected four so called *prognostic* (Intake-HRSD, chronic subtype, age & IQ), and five *prescriptive* predictors ("marital status, employment status, life events, comorbid personality disorder, & prior medication trials" DeRubeis et al., 2014, p.1) to forecast post-treatment HRSD scores for each patient in the sample. Prognostic predictors are

variables used to forecast how a patient will do at some future time compared to other patients regardless of treatment. Prescriptive variables are basically the same as moderators (which will be explained further in the chapter Moderation analysis) and can be used to explore the best treatment for a specific patient by studying interactions between different variables and treatments (Fournier et al., 2009).

The next step of the development of PAI was to include all prognostic and prescriptive predictors and type of treatment (CBT or paroxetine) in separate Generalized Linear regression Models (GLM) to generate regression coefficients for each patient. All nine predictors and treatment type were entered as main effects terms, and the interaction between treatment type and the five prescriptive variables were entered as interaction terms¹. From these coefficients, the authors were then able to create so-called *factual* and *counterfactual* predictions of each patient's end-HRSD score.

The factual prediction represents the end-HRSD score for a patient while receiving the treatment he or she actually got, and the counterfactual prediction represents the prediction for the same patient if he or she would have received the other treatment. The PAI score was then calculated for each patient as the difference between these two predictions.

As patients were randomized to either CBT or medication, it was possible to calculate whether patients who by chance received their optimal treatment (based on their PAI scores) fared better (i.e., had a lower end-HRSD) than patients who did not.

Results revealed, however, no significant difference ($p = 0.09$) between the conditions. Yet, the authors stated that the PAI-score difference was quite small and below clinical significance for approximately 40% of the sample (three points on the HRSD). Thus, they argued that, where there was no clear guidance from the PAI score for these patients, other factors might be more relevant to use for treatment selection (e.g., patient preferences or costs). A reanalysis including only 60% of the sample with a PAI score above three points showed that patients randomized to their most suitable treatment fared better than patients who were not randomized to their most suitable treatment ($M = 1.78$, $SD = 6.38$, $p = .006$) with an effect size of $d = .58$.

If these results hold, the PAI method shows promise for use as a practical tool to identify patients with obvious differences in outcomes between treatments (i.e., with high PAI scores). However, as stated by DeRubeis et al.

¹ The authors used leave-one-out cross-validation, which means that when parameters were generated for a specific patient in the sample, that patient was excluded from the analysis, which left 153 patients to base the prediction on.

(2014), treatment selection is probably best guided by other factors when the PAI score approaches zero.

The PAI method has been replicated for treatment selection between Cognitive Therapy (CT) or Interpersonal Psychotherapy (IPT) (Huibers et al., 2015) and for CBT versus CBT with integrated exposure and emotion-focused elements, with promising results (Friedl, Berger, Krieger, Caspar, & Grosse Holtforth, 2019). However, to the best of my knowledge, there have been no prospective tests of PAI performance with a naturalistic sample, which means that the applicability of the method for new patients in a real clinical context is still unclear.

Another approach to precision medicine is to alter the intervention during treatment to better fit individual patient needs. An example of this is a study conducted by Forsell, Isacsson, et al. (2019). Those authors identified patients suffering from insomnia and at risk of treatment failure during ongoing Internet-delivered Cognitive-Behavioral Therapy (ICBT) and altered the treatment based on this information.

Patients in the study were identified as either “red” (at risk of treatment failure) or “green” (not at risk of treatment failure) with the help of a semiautomated algorithm during week four (of nine) of treatment. The procedures for allocating patients have not yet been published in detail but were primarily based on the literature on predictors and moderators of change in CBT for insomnia, and a regular regression analysis framework was used.

After week four, patients identified as “red” ($N=102$) were randomly assigned to either continue standard ICBT ($N=51$) or adapted ICBT ($N=51$), and the rest ($N=149$) were classified as not at risk (i.e., green) and continued standard ICBT.

Patients randomized to adapted ICBT were contacted by their therapist for a telephone interview to evaluate the patient’s difficulties with treatment and to explore how the treatment might be altered to help the patient. The most common adaptations were: “telephone support, printed materials via regular mail, and/or increased text-message reminders” (Forsell et al., 2019, p. 318).

Results showed that patients classified as “green” (i.e., not at risk) improved significantly more on the primary outcome (Insomnia Severity Index) than at-risk patients (i.e., red) given standard ICBT. In addition, at-risk patients who received adapted ICBT improved significantly more than at-risk patients who received standard ICBT.

These results highlight the advantage of providing therapists with an updated estimation of risk of treatment failure during treatment, an estimation that could be used to alter the treatment in more favorable ways.

The studies mentioned above used classical regression analysis in innovative ways. However, the application of more advanced statistical frameworks might be a way to gain additional insights from the data. DeRubeis (2019) argues that one problem with analyzing single (or very few) predictor variables is that they hardly account for sufficient variance to reach significance (often due to the shortage of power to detect small interaction effects), which can lead to a lack of clinical relevance for such studies. This is not surprising since single predictor variables would have to be sufficiently powerful to influence most patients to change clinical practice but, at the same, time sufficiently subtle to have escaped previous discovery by clinicians and researchers. Nonetheless, even if single variables do not account for much of the relevant variance, it might still be possible that information from many variables can be used to inform clinical practice.

Machine Learning (ML) and Latent Profile Analysis (LPA), which will be discussed in the chapter Advanced methods for studying the thesis outcomes, might be suitable methods for extracting useful information from a large collection of variables. Mayer et al. (2020) have also shown how Structural Equation Modeling (SEM) can be used to gain additional information from approaches similar to the PAI method (e.g., by including latent variables in the analysis).

Studies examining different types of outcomes in naturalistic settings (e.g., rates of change or dropout) and different types of predictors (e.g., baseline distress, quality of working alliance) are also needed to understand the broad range of factors that might influence psychotherapy outcomes at the individual patient level. However, it is important to clarify some statistical boundaries and practices that are necessary in order to draw sound conclusions from such studies.

Description, prediction, and explanation

Scientific discovery can be divided into three central goals: *description*, *prediction*, and *explanation*. Description involves summarizing data in parsimonious ways, which can be done by simple analysis such as using the mean and standard deviation to describe the proportion of depressed patients in a population, or with more advance methods, such as cluster analysis, to explore similarities in a sample.

Prediction involves using variables that appear *before* an outcome to forecast new future observations. This might include simple tasks, such as quantifying the association between the level of psychological distress at baseline and the rate of premature termination from treatment, to more

complex tasks such as using hundreds of variables to predict the risk of dropout during treatment.

Finally, explanation focuses on understanding driving forces and casual mechanisms for certain phenomena (Hamaker, Mulder, & van IJzendoorn, 2020; Hernán, 2018; Shmueli, 2010)

Scholars from different fields of research have devoted much time to clarifying the relationship between these constructs by asking questions such as: is it possible to predict what will occur in a specific situation without being able to explain the phenomenon? And, is it possible to explain a phenomenon without being able to predict it? (Pedhazur, 1975). An example of this is Darwinian evolution theory, which can be used to explain a phenomenon but probably not to predict future events (Shmueli, 2010).

According to Hamaker et al. (2020), there has been a tendency in the general research community to view description and prediction as subordinate motives, or merely as springboards to the higher goal of explanation. Yet, if applied in adequate ways, prediction and explanation are vital parts for scientific development.

The boundaries between these goals are not always clear, and they are often linked to one another. For example, theory (which is derived from explanatory research) is probably the best guide for selecting useful predictors, but predictive research might also lead to new theoretical formulations. If we can make successful predictions based on a certain explanation, we have very good reasons for accepting the explanation (Pedhazur, 1975). For instance, some complex relationships found in high dimensional data (i.e., complex data with many variables) might not be possible to hypothesize. However, with the use of advanced prediction models, it could be possible to discover new patterns related to potential causal mechanisms, which in turn could generate new hypotheses. The discovery of underlying complex patterns might also be used to advance or alter already existing explanatory models (Shmueli, 2010).

Nevertheless, according to Pedhazur (1975), it is not uncommon for researchers to fall into the trap of using “purely predictive studies for the purpose of explaining phenomena” (Pedhazur, 1975, p. 197). But the reverse is also true, for example, Shmueli (2010) has written that “models that possess high explanatory power are often assumed to inherently possess predictive power” (Shmueli, 2010, p. 289).

To overcome such ambiguity, it is critical for the researcher to clearly state the scientific goal of the study and align it with an appropriate research design. Other important factors are to conduct the study in a suitable research setting and to analyze the data with an adequate statistical model.

The link between scientific goal and research design

Depending on whether a study leans more towards predictive, explanatory, or descriptive research, it influences what variables to include and how to interpret the results. For example, if forecasting is the primary goal, it doesn't really matter what type of predictors the researcher includes as long as they contribute to the prediction at hand. Sometimes predictors might even seem contradictory, for example, Pedhazur (1975) has argued that although psychotherapy probably decreases a depressed individual's probability of suicide, for prediction purposes, more psychotherapy is probably predictive of an increased risk of suicide. Furthermore, if possessing a set of household items correlated strongly with a patient's success in psychotherapy, it would be valid to treat those items as useful *predictors*, but it would not be very wise to recommend every patient to buy these items as they are most likely not the cause of the increase in success rate.

There are more or less optimal matches among the goals of description, prediction, and explanation with different types of research designs. For example, patients in cross-sectional studies are assessed at a single point in time. Such data has a natural fit for descriptive research and can for instance be used to assess the degree of correlation between therapist and patient ratings of the working alliance at a certain point in time. This type of data is, however, not the best fit for prediction or explanation purposes as these goals imply a sequential chain of events (Hamaker et al., 2020).

Longitudinal research, on the other hand, usually consists of repeated measures collected over time. This design has a more natural fit for prediction and can be used to forecast individual patient development over time. Finally, experimental research involves manipulating variables in a controlled environment, which is closely linked to the concept of explanation.

The link between research design and study setting

Non-experimental studies typically utilize observational data where patients are studied without the researcher trying to control the environment or manipulate the dependent variables. Cronbach (1957) has written that: "The correlator's mission is to observe and organize the data from Nature's experiments" (Cronbach, 1957, p. 672) (i.e., observing) while the "...the experimenter is interested only in the variation he himself creates..." (Cronbach, 1957, p. 671) (i.e., manipulation).

These types of studies focus on different types of validities. In naturalistic studies, for instance, the emphasis is typically on increasing external validity, whereas experimental studies focus on increasing internal

validity. In turn, these validities dictate what variation is possible to explain or account for in a study (see the next section).

Internal validity

In relation to psychotherapy, Cronbach's experimenter could be a researcher interested in exploring differences in treatment outcome between groups of patients given different treatments. Internal validity is an important factor in such an example. This type of validity concerns whether the study can eliminate alternative explanations for the observed variation in outcome and answer the research questions in an unbiased way.

The Randomized Controlled Trial (RCT) is a procedure specifically developed to increase internal validity. The logic behind RCT is to use concepts from experimental science to address the question of causality. In other words, to distinguish treatment effects from other confounding variables such as spontaneous improvement, regression towards the mean (explained later) and/or placebo effects. A fundamental procedure in RCTs is to randomly assign patients to a treatment or control condition to balance different patient variables (observed and unobserved) between the groups. Also, to minimize bias, there should be no prior knowledge of which condition patients will be allocated to and, ideally, both patients and clinicians should be unaware of who receives the active treatment, which is quite difficult in psychotherapy research. Other important procedures of the RCT are: to use a representative sample, to stringently operationalize the treatment intervention, to ensure that patients not receiving the treatment are more or less treated in the same way (to maintain group comparability) and to ensure that the number of patients needed to determine if an observed difference between the groups is reliable or not, i.e., a power analysis (Carey & Stiles, 2016; Hariton & Locascio, 2018; Seligman, 1995; Shean, 2014).

Thus, by increasing internal validity it is possible to show, or at least come close to showing, a causal relationship between a treatment and differences in treatment outcome between groups of patients. However, it might remain unclear whether similar effects would be observed in less controlled settings such as routine care, and the regular RCT say nothing about the *cause* of individual patients results.

External validity

External validity concerns generalizability, i.e., the degree to which findings can be extended to other contexts. Studies aiming at strong external validity are usually less controlled than RCTs. For example, in studies conducted in first-line mental health care services the term "treatment" might include a

mix of different methods conducted by a broad range of therapists with different backgrounds (e.g., education, experience level, etc.). In addition, patients are not normally selected with predefined inclusion/exclusion criteria or randomized to different treatment conditions. Thus, under such circumstances, it is not possible to disentangle treatment effects from other factors that might have caused the patient to change.

Nevertheless, by monitoring variation in variables, it is still possible to study the proportion of variance in an outcome that was accounted for by variation in another variable. It is important to note that this is not to say that the variation was necessarily *caused* by the other variable or vice versa.

A strength with analyses conducted in naturalistic settings is that in order to break through the noise that comes from the lack of control in these studies, the observed effects must be quite robust to reach significance. Therefore, it is also more likely that the effects will generalize to other similar contexts. Obvious limitations are that more subtle relationships might be missed, and that it is not possible to reject other explanations for the observed effect between two variables as they could both be confounded by a third variable unknown to the researcher.

In some situations, it is impractical or unethical to conduct an experiment even though the underlying question might still be to study causation. An interesting opinion on this problem is presented by Hernán (2018). He argues that the practice of completely excluding casual language from non-experimental studies, as suggested by many journal editors, might lead to negative consequences. He claims that the way researchers by-pass such limitations are normally simply to use different words for causation, such as “association” or “link.” However, this type of word choice might lead to ambiguity about the underlying goal of the study. Even though results from an observational study will always be a biased *estimate* of a casual effect, simply sweeping the problem under the rug by calling the association something else is not helpful. In fact, if the goals are not clearly stated, the critical evaluation of studies can become more difficult, which in turn could hamper scientific progress (Hamaker et al., 2020).

Hernán, Hsu, and Healy (2019) argue that there is a need to develop formal methods for quantifying causal effects from observational studies in order to properly adjust for confounding variables in such an analysis. This progress has already been initiated in some parts of statistics and data science (e.g, Johanson, 2020).

Another initiative to bridge the gap between internal and external validity is the so-called pragmatic RCT. This method can be viewed as a compromise between the regular RCT and the observational study design by adapting certain aspects of the RCT to better fit a clinical context. Adaptations could

for example include broaden the inclusion criteria to better represent the heterogeneity of the clinical patient sample, or choosing outcome measures that reflect “real world concerns” such as return to work or cost-effectiveness etc. (Hotopf, 2002).

Moderation analysis

Moderation analysis is a common method for analyzing interactions between variables in psychotherapy research. For example, when attention shifts from studying the average effects of different treatments to the exploration of individual heterogeneity, moderation analysis becomes an important tool (Mayer et al., 2020)

Moderators can be categorical (e.g. class or gender) or continuous (e.g. level of social support) and influence the “...direction and/or strengths of the relation between the independent and dependent variable” (Baron & Kenny, 1986, p. 1174). For example, there might be a strong relationship between stress and quality of life. However, depending on a person’s level of social support, the association between stress and quality of life might be stronger or weaker, or even change direction. For instance, a person with a high level of social support might still report decent quality of life despite experiencing a lot of stress. In other words, social support has a moderating effect on the relationship between quality of life and stress.

Moderator analysis can be said to operate on the borderline between prediction and explanation, depending on how it is used. For example, if the aim is to find moderators to confirm or create a theory of why certain patients are suitable for a particular treatment, the analysis leans more towards explanation. If moderators are used to purely forecast purposes, they correspond more to the goal of prediction (Kazdin, 2007).

Some statistical notations are occasionally used for the same, or almost the same, concept. An example of this is the use of prescriptive predictors (e.g., variables related to treatment outcome due to different treatments) and prognostic predictors (e.g., variables related to treatment outcome regardless of treatment type) mentioned in the chapter Precision medicine above. In mathematical terms, there is no difference between how a moderator or prescriptive variable is calculated or entered into a model, but there could be interpretive differences.

The development of new statistical models for prediction and moderation analyses is growing at a fast rate and ranges from simpler models, such as classical ordinary least square regression, to different types of ML and SEM that incorporate latent variables (Mayer et al., 2020). A relevant question to

ask when a researcher is exposed to such an abundance of different models is how to decide on the type of model to select for a particular problem?

Occam's razor vs Hickam's dictum

William of Occam (1285-1349) asserted in the 14th century: “Plurality must not be posited without necessity.”² The essence of this statement called Occam's razor is often translated into something like “the simplest solution to a problem is usually the best one”. John Hickam, a graduate from Harvard University School of Medicine in the 1940's who later became head of medicine at Indiana University (Mani, Slevin, & Hudon, 2011), presented a counter argument to this maxim called Hickam's dictum. The motto is usually summarized with the somewhat peculiar statement: “A patient can have as many diagnoses as he darn well pleases.” (R. B. Miller, 1998, p. 213)³. By this he means that when a patient displays many different symptoms, a single diagnosis might not always be adequate to explain all the symptoms (as suggested by Occam's razor). Occasionally, different rare diseases might occur at the same time in a single patient. The essence of Hickam's dictum is sometimes extended to the notion “there could be many causes for multiple outcomes” (Edwards, Anderson, & Stranges, 2019; Tierney, Midthun, & Saint, 2004).

In relation to statistics and prediction, a similar but not entirely equal dilemma can be described with the *bias-variance trade-off* (e.g., Mitchell, 1997). More concretely, a statistical model that is too simplistic might *underfit* the data (i.e., miss relevant relationships), rendering predictions with low variance but high bias (i.e., performance might not be exceptional with the available data, but it would probably be similar with new data). A model that is too complex, on the other hand, might *overfit* the data (i.e., correspond too closely to the available data), rendering high variance but low bias (i.e., it performs well with the observed data but not with new data).

It is important to keep this trade-off in mind when moving from less complex to more advanced statistical modeling. However, increasing the complexity of a model can sometimes create the potential to explore completely new relationships that would not have been possible to study with the use of traditional models.

² The exact statement has not been found in any of his writings, but a close approximation according to Kneale and Kneale (1962) can be found in (Occam, 1495).

³ A common citation for Hickam's dictum is Tierney et al. (2004) published in the New England Journal of Medicine. However, the authors in turn cite Miller (1998), who unfortunately did not reference the exact saying.

Even though parsimony is an important principle in statistics, it can fail us if we use models that do not capture important relationships in the data. Nevertheless, it is important to note that a model that is too complex might not be applicable on a universal level as it might overfit the data.

With the considerations discussed in this chapter in mind, I will now outline how these matters can be studied in the context of routine care. First, however, it is important to clarify the outcomes being studied in the thesis and discuss some of the strengths and limitations of these definitions.

Definitions of outcomes

This thesis focusses primarily on studying outcomes in routine care derived from patient-rated self-assessments of psychological distress and dropout evaluated by the therapist. The mechanisms that *caused* the observed variations in these outcomes to occur are treated as unknown with an underlying assumption that they were due to a mix of different factors, such as patient and therapist variables, treatment, and other external factors

Self-assessments

One of the most central aspects in quantitative research is the dependency of this approach on measurement. Psychology researchers are generally aware of the many weaknesses and obstacles that arise when trying to assign numbers to complex phenomenon such as psychological distress. At the same time, the benefits that come with it, such as the potential to use statistical methods for analyzing and understanding change, do often far outweigh the costs.

Standardized patient-rated self-assessments are currently the most common way to quantify psychological distress in psychotherapy research, and these assessments generally consist of validated and reliable items with established population norms. The most common areas to measure are probably depression and anxiety with measures, such as Becks Depression Inventory (BDI; Beck, Steer, & Garbin, 1988), Becks Anxiety Inventory (BAI; Beck, Epstein, Brown, & Steer, 1988), or the Montgomery-Åsberg Depression Rating Scale (MADRS; Montgomery & Åsberg, 1979). Other common areas to assess are for example functionality, quality of life, and subjective wellbeing.

The decision to use self-assessment as an outcome might influence the observed effects of a study, for example, a meta-analysis by Cuijpers, Li, Hofmann, and Andersson (2010) found that when psychotherapy for adult depression was assessed with patient-rated self-assessments, it indicated less effect than when measured with clinician-rated assessments.

Nevertheless, Ogles (2013) presents some arguments for why patient-rated self-assessments have become so popular for measuring changes in psychological distress in psychotherapy. For instance, self-assessments are especially appealing in practice-oriented studies as they take less time from the clinician to administer, and funders of mental health usually want proof of effectiveness from their “customers,” i.e., the patients, rather than the providers, i.e., therapist-rated assessments.

Important questions to ask when assessing treatment outcome with patient-rated self-assessments are “who will use the measure”, “in what context”, and for “what purpose?” For instance, there is often a tradeoff between global versus more specific properties of a measure. In a naturalistic setting, such as a mental health care service, patients seek treatment for multifaceted and different types of problems. In this context, a broader measure might be a good fit to capture global aspects of patient problems. In more controlled studies where certain diagnostic groups are studied, specific scales intended to measure specific problems are likely to be a better fit. In addition, some measures used in research might be too time consuming for the patient to complete in routine care (Thornicroft & Slade, 2000).

Core batteries

Measures have been developed specifically for practice-oriented research such as the Clinical Outcomes in Routine Evaluation Outcome Measure (CORE-OM; Evans, Connell, Barkham, Mellor-Clark, & Audin, 2002) and the Outcome Questionnaire-45 (OQ-45; Lambert & Finch, 1999). These types of measures are constructed to fit any patient regardless of psychological distress and are usually administered repeatedly during treatment to measure change. The most important motivation behind these so called core batteries has been to capture global aspects of patient distress and to be able to compare results from studies conducted in different contexts.

The CORE-OM was used in all three papers included in this thesis. This measure was developed to meet three broad goals: standardization, acceptability by clinicians, and suitability for routine use (Evans et al., 2000). It consists of 34 items intended to measure psychological distress experienced during the last calendar week. The questions encompass four broad problem areas: “well-being”, “problems”, “functioning”, and “risk” (to self or others). Items are scored by the patient on a five-point scale ranging from *Not at all* to *Most or all the time*. The score ranges from 0-40 points, where higher scores indicate greater distress. The Swedish version of the measure, which was used in the included studies, has shown satisfactory reliability (Elfström et al., 2013).

Change between two time points

An intuitive way of quantifying change in psychological distress derived from self-assessments collected before and after treatment is to calculate so called *change scores* by simply deducting the pretest score from the posttest score. However, some drawbacks with this practice have been discussed by psychometricians since the 1950s.

Lord (1956) noted, for example, that the reliability of the pretest score and the posttest score inevitably influences the reliability of the change score so that “differences between scores tend to be much more unreliable than the scores themselves” (Lord, 1956, p. 449). Cronbach and Furby (1970) agreed with this view and added that “scores formed by subtracting pretest scores from posttest scores lead to fallacious conclusions, primarily because such scores are systematically related to any random error of measurement” (Cronbach & Furby, 1970, p. 68).

Another issue with change scores is that they are susceptible to regression towards the mean, which can be described as “a statistical phenomenon that can make natural variation in repeated data look like real change” (Barnett, Pols, & Dobson, 2005, p. 215). This almost universal phenomenon stipulates that if a first measure is extreme, then it is more likely that the next measure will be closer to the mean of the population. For example, if a measure of psychological distress collected at the first treatment session indicates severe distress, then the rating collected at the last treatment session for the same patient will tend to indicate less distress.

A proposed way to control for regression towards the mean between two time points is to calculate residualized change scores. There are different ways for computing such scores, but a common method is to first fit a linear regression model to the sample with the posttest score as the dependent variable and the pretest score as the independent variable. It is then possible to predict the posttest score based on the pretest score for each patient. The last step is to subtract the predicted posttest score from the observed posttest score for each individual to rule out the pretest-posttest correlation (Kisbu-Sakarya, MacKinnon, & Aiken, 2013). This procedure was used to define change in psychological distress in Paper I.

Nevertheless, according to Castro-Schilo and Grimm (2018), it is important to note that the limitations of using simple change scores has become somewhat more nuanced in recent times. For example, the authors point out that when correlation between the pretest and posttest scores is low, change scores are not as biased as previously thought. Also, the decision to use change scores versus residualized change scores should not be influenced by the reliability of the measure since measurement error will obscure true change, regardless of which of the two models the researcher chooses. To

complicate matters further, Allison (1990) argues that when choosing between change scores or residualized scores: “purely mathematical or statistical consideration will not suffice because the models are incommensurable” (Allison, 1990, p. 106). He means that none of the models are suitable for all situations, and in some cases, none of the models are appropriate. The choice between models depends on the design of the study and the hypothesis about the expected change. For example, the author asks the hypothetical question: “does regression towards the mean within groups lead to regression towards the mean between groups?” (Allison, 1990, p. 93). In conclusion, choosing among models for capturing change between two time points is notoriously difficult, and it is beyond the scope of this thesis to summarize all aspects related to this issue. However, one way to attain a more complete analysis of change is to analyze ratings collected over the course of treatment (e.g., at every session).

Growth rate

As noted above, in comparison to analyzing change with only pre- and post-measures (i.e., at two time points), a more informative approach is to assess ratings collected over the course of therapy.

The classical way of analyzing such data has been with methods such as the Multivariate Analysis of Variance (MANOVA) or the repeated measures Analysis of Variance (ANOVA) (Curran & Hussong, 2003). However, there are several drawbacks with these methods for studying repeated measures and most of these limitations are linked to ignoring different assumptions that are made about the data.

Firstly, ANOVA and MANOVA assume that observations (e.g., data points from patients) are independent of each other. This assumption, often called *non-independence*, is violated when some observations are more similar to other observations due to clustering. Data is often hierarchical in nature in psychotherapy research; for instance, patients are nested (grouped) within different therapists. In this case, the patients exist at Level 1, and the therapist (the grouping factor) exists at Level 2. This means that patient-level observations (e.g., ratings made by specific patients) are not independent because patients treated by a specific therapist might have more similar outcomes than patients treated by another therapist. In such a case, the estimates from the model may not correctly reflect how the variables in the data relate to one another. Violation of non-independence can bias statistical significance testing of repeated measures and/or inflate/deflate estimated relationships (Grwitsch & Munz, 2004). In fact, Baldwin, Murray, and Shadish (2005) have conducted an analysis of group-administered treatments

and were able to show that only about 20-60% of the studies included in the analysis ($k=33$) remained significant after correcting for improper handling of the non-independence assumption.

Traditional models for analyzing repeated measures also make very restrictive assumptions about the error variances for observations that are collected over time. This assumption, often denoted *sphericity*, specifies that error variance must be the same across time, and that correlations between measurements collected over time must be equal (Tasca & Gallop, 2009). However, according to the same authors, these assumptions are unlikely to hold for the typical psychotherapy study. For example, imagine a study where symptom distress is assessed at three time-points: pre-, post-treatment, and at follow-up, and patients are randomly allocated to either a treatment or a control group. In general, patients who begin treatment do not report the same levels of symptom distress, so variability in the measure between individuals can be expected. In addition, when treatment begins, one would expect symptom distress to decline, especially for those who started treatment with high levels of distress. This means that variability at post-treatment and follow-up can be expected to be lower than at pretreatment. Also, it is likely that patients allocated to the control condition would report less change in symptom distress than patients allocated to the treatment condition, which makes the assumption of equality of error variances even more unlikely. Measures of symptom distress for an individual taken close in time can be expected to be more strongly correlated than measurements taken further apart in time. In other words, the sphericity assumptions for both variances and covariances are likely to be violated in the typical psychotherapy study. Violating the assumption of sphericity in ANOVA and MANOVA could result in an inflated Type I error (i.e. the rejection of a true null hypothesis) (Tasca & Gallop, 2009).

Another drawback with these traditional models is that they treat time as a fixed variable. This means that they do not model individual change and variability in change over time but rather assume that all patients have the same average improvement rate. Patients who attend psychotherapy, however, do not usually have the same rate of improvement. There might be situations where it would be interesting to assess individual variability in patients' rates of change and predict these individual differences. Another drawback with only emphasizing group means and variances (i.e., by treating time as a fixed variable) is that each individual must have complete data at all time points to be included in the analysis. Missing data is a common problem, especially in naturalistic settings. There seems to be no satisfactory way of handling missing data in these traditional models. For example, one way is simply to delete all patients who did not complete all measurements

(listwise deletion), which might result in smaller sample sizes (power), and a potentially biased sample. Another way is to impute the missing values with a method, such as the Last Observation Carried Forward (LOCF), which can bias the results in many ways, or with multiple imputations, which can be a very complicated process (Tasca & Gallop, 2009).

Another way of analyzing repeated ratings of psychological distress is with Latent Growth Curve Modeling (LGCM), which was used to quantify the change rate in Paper II. There is no need to make any assumptions about sphericity using this framework. It includes random effects that make it possible to model individual change and variances. Missing data is possible to handle in more adequate ways such as with Full Information Maximum Likelihood estimation (FIML). This means, in simple terms, that if data for variable “X” is Missing At Random (MAR), the FIML algorithm will calculate the most likely value of “X” by using the current parameters derived from all valid data points in the sample (Allison, 2003; Enders, 2001). In addition, unbalanced time scores can be used (under the MAR assumption), which means that every patient does not need to be measured at the same time points. It is also possible to model nonlinear change. A common form for improvement in psychotherapy has turned out to be logarithmic rather than linear (e.g., Howard et al., 1996). Nevertheless, it remains important to note that estimates of patients with fewer measurements in the LGCM will probably be regressed to the grand mean of the sample, and generally, patients with more measurements will have a higher weight in the model.

Another, more general concern when analyzing longitudinal data in psychotherapy is that it is likely that some therapists are more effective with their patients than others. This concept is commonly referred to as *therapist effects*. If such effects are ignored in the longitudinal analysis, the risk of Type I error can increase (leading to the conclusion that the treatment had an effect when, in fact, it had no effect), especially when comparing different interventions with very few therapists per treatment (Magnusson, Andersson, & Carlbring, 2018). However, there are ways to account for this by including a third level in the model. This would allow therapists to have varying intercepts (baseline levels of distress) and slopes (rates of change).

Dropout as an outcome of psychotherapy

Another important outcome to study in psychotherapy is the proportion of patients who end treatment prematurely (or drop out of therapy). However, there is no consensus in the field of how to define dropout. This is

problematic because different definitions might significantly influence the rate of dropout reported in studies.

A broad definition of this concept that probably most researchers could agree on is given by Swift, Greenberg, Tompkins, and Parkin (2017) who have stated that: “dropout occurs whenever a client begins an intervention, but then unilaterally terminates it against provider recommendations and prior to recovering from the problems that led him or her to seek treatment in the first place” (Swift et al., 2017, p. 48). However, as will be discussed in this section, there are many variations of how to quantify and measure this phenomenon in psychotherapy research.

According to an early meta-analysis by Wierzbicki and Pekarik (1993), the most common way to define dropout historically has been by means of “...treatment duration or number of treatment sessions” (Wierzbicki & Pekarik, 1993, p. 190) such as failure to attend a fixed number of sessions or to complete a treatment protocol. A strength with this way of measuring dropout is that it is easy to assess by the researcher (i.e., it has high reliability) but an obvious limitation is that dropout might occur at any time in therapy (rendering low validity).

Hatchett and Park (2003) have proposed that dropout might also be defined by changes in psychological distress. Using the OQ-45 as an example, their definition stipulates that: “clients whose last scores fell within the normal range on the inventory would be classified as appropriate terminators, whereas clients whose last scores fell within the clinical range would be classified as premature terminators” (Hatchett & Park, 2003, p. 230). There are several limitations worth noting with this approach which was also raised by Hatchett and Park (2003). For example, *i*) the return to normal criteria might not be suitable for all patients (e.g., for patients with comorbid psychiatric diagnoses), *ii*) some patients might be in the normal range when they start treatment, and *iii*) some types of distress might not be captured with self-assessment forms. A more general concern (not raised by Hatchett and Park) is that this definition assumes that if patients go long enough in therapy, they will inevitably be cured. However, all forms of therapy are not suitable for all patients, and some patients may be suffering from concerns that are not curable with psychotherapy.

A more meaningful definition of dropout might be to let the therapist determine whether dropout has occurred or not. Still, such a definition is not without limitations. The most apparent one is that therapists might use different standards for deciding whether a patient should be defined as a dropout or not (i.e., rendering low reliability).

The meta-analysis by Wierzbicki and Pekarik (1993) shows that different definitions significantly influence the rate of dropout reported in their sample

of studies ($k = 125$). More specifically, “studies that defined dropout in terms of termination because of failure to attend a scheduled session reported lower dropout rates than did studies that defined dropout in terms of either therapist judgment or the number of sessions attended.” (Wierzbicki & Pekarik, 1993, p. 193).

In another study, Hatchett and Park (2003) used a sample of 85 students from a university counseling center to calculate four different dropout rates based on different definitions. These definitions were based on therapist judgement, failure to attend the last scheduled appointment, median split,⁴ and failure to return after the intake appointment. Results showed that rates of dropout varied substantially depending on which of these definitions the authors used. For example, therapist judgement and the missed last appointment definitions yielded the same dropout rate of 41 %, the median split definition yielded a 53 % dropout rate, and the intake-only definition yielded a 17 % dropout rate.

Swift Callahan and Levine (2009) conducted a similar study with 135 patients from a university-based training clinic and included the method used for defining dropout based on change in symptom distress. Their results showed dropout rates of 77 % for the clinically significant change method, 50 % for the median split method, 8 % for the intake only definition, 48 % for the missed-appointment definition, and 74 % for the therapist judgment method.

Finally, a more recent meta-analysis by Swift and Greenberg (2012) also showed that dropout rate was significantly moderated by the definition of their sample ($k = 669$, $N = 83,834$ patients). More specifically, “dropout rates were the highest when determined by therapist judgment (37.6%) and were the lowest when determined by completion of a set number of sessions (18.3%) or a treatment protocol (18.4%)” (Swift & Greenberg, 2012, p. 555).

It is evident from the results above that observed dropout rates vary substantially depending on different definitions. Dropout rates also seem to vary within the same cluster of definitions. For example, therapist-judged dropout rates ranged from 38% to 74% in different studies. Consequently, there is no easy way to define dropout. An adequate definition of this concept includes finding a balance between using a simple measure to increase reliability and, at the same time, using a meaningful measure to increase validity. It is likely that different predictors might also be more or less important, depending on how dropout is defined.

⁴ All patients that attended fewer than four sessions, which was the median of attended sessions in the sample, were classified as dropout cases.

One way to increase the reliability of therapist-rated dropout might be to incorporate clear descriptions of what counts as dropout and letting therapists assess it using identical forms. A similar procedure was used in Paper III.

This chapter discussed strengths and limitations with the definitions of outcomes used in the three papers of this thesis. The next chapter will be devoted to how specific predictors for these outcomes have been studied in the past and how more advanced statistical modeling might be used to enhance understanding of these relationships.

Traditional methods for predicting the thesis outcomes

Working alliance and treatment outcome

The association between working alliance and treatment outcome is one of the most replicated findings in psychotherapy research. The latest meta-analysis by Flückiger, Del Re, Wampold, and Horvath (2018) included 295 studies with 30 000 patients and found that the “overall alliance–outcome association for face-to-face psychotherapy was $r = .278$ (95% confidence intervals [.256, .299], $p < .0001$; equivalent of $d = .579$).”

Even though there are many different ways of defining working alliance, almost 70 % of the more recent studies included in the meta-analysis above have used some variation of the Working Alliance Inventory (WAI; Horvath & Greenberg, 1989).

WAI was founded on Bordin's (1979) definition of working alliance consisting of three related components: *goals*, *tasks*, and *bonds*. This definition stipulates that a strong working alliance emerges when both parties (i.e., patient and therapist) reach an agreement about the goals of the treatment, how to achieve these goals, and develop a strong personal bond between each other. Different variations of the WAI measure exist; for instance, WAI can be rated by the patient (Hatcher & Gillaspay, 2006) or the therapist (Tracey & Kokotovic, 1989), but they all include scoring questions associated to the three domains mentioned above, on a 7-point *Likert scale*.

The classical way of studying the relationship between working alliance and treatment outcome has been to analyze simple correlations between *patient*-rated alliance and treatment outcome. For example, 85% of the studies in Flückiger et al. (2018) meta-analysis only included information from the patient about the quality of the working alliance.

A limitation with studying working alliance from only one point on the dyad is that it is not possible to explore how ratings from both parties might

influence each other. If the patient has a positive view about the quality of the working alliance, it is likely that this will influence the therapist's view and vice versa. It would be interesting to explore questions, such as: "if a therapist consistently rates strong alliances with his or her patients, does that correspond to strong alliance ratings from that therapist's patients?" Or: "if a therapist rates a specifically strong working alliance with a specific patient, does that correspond to a particularly high rating of the alliance from that patient"?

Other questions that might be interesting to explore are how the relationship between the quality of the working alliance and treatment outcome varies depending on whether it is studied *between* or *within* therapists. Ratings analyzed from patient to patient treated by the same therapist is denoted as *within* therapist variation, and average ratings analyzed from therapist to therapist is denoted as *between* therapist's variation herein. For example, "do some therapists receive higher alliance ratings from patients than other therapists, and if so, do these therapists have better outcomes on average"? Or is the general trend more that therapists tend to have strong alliances and good outcomes with some of their patients, and weak alliances and poor outcomes with others"?

Questions like these have important clinical implications. For example, such matters might point to interesting ways of directing further study of the topic. If, for instance, there are obvious differences between therapists in their ability to foster strong alliances, it would be interesting to explore common characteristics of these therapists. However, if it was more likely that therapists tended to have different alliances from patient to patient, other research strategies might be more constructive (e.g., how to match therapists and patients, or how to handle specifically demanding therapeutic relationships).

Important to note, however, is that these types of questions are not possible to study with the use of classical correlation analysis as it would involve untangling the nested structure of the data (i.e., the fact that patients are grouped by different therapists).

Initial symptom distress and change rate

Some authors suggest that the patient is the single most influential factor for successful treatment (Norcross & Lambert, 2011; Wampold & Imel, 2015). However, according to Bohart and Wade (2013), demographic variables, such as patient age, gender, or socioeconomic status, have failed to demonstrate strong correlations with treatment outcome. Likewise, variables, such as patient motivation, attachment style and coping style have only

accounted for small parts of the total variance in outcomes, or yielded mixed findings. However, according to the same authors: "...there is evidence that a higher level of distress at the start of therapy is the best predictor of outcome, more so than the client's diagnosis, problem chronicity, or treatment population" (Bohart & Wade, 2013, p. 226). It is important to note, however, that studies primarily have shown that higher distress at baseline correlates with greater *change* in therapy but not necessarily with better final outcome (e.g., return to the range of the normal population (Brown, Burlingame, Lambert, Jones, & Vaccaro, 2001; Hansen & Lambert, 2003).

Possible explanations for the relationship between distress at baseline and change rate are that patients that are more distressed might have more room for improvement, the phenomenon of regression towards the mean (discussed in the chapter Change between two time points, above), and/or mathematical coupling.

Mathematical coupling can be described as: "...the effect that occurs when one variable directly or indirectly contains the whole or part of another, and the two variables are analyzed using standard correlation or regression techniques" (Y. K. Tu, Maddick, Griffiths, & Gilthorpe, 2004, p. 134). In simple terms and at the conceptual level, mathematical coupling is problematic because change and baseline scores are united algebraically, which could influence statistical testing of the null hypothesis (i.e., whether or not the growth curve in the regression is zero).

Nevertheless, using psychological distress at baseline as a predictor for change rate could be an accessible way for therapists to estimate treatment development over time. In fact, one of the most popular systems for monitoring and predicting rates of change for psychotherapy patients in routine care, the so called *OQ-analyst* (Lambert & Shimokawa, 2011), is based on this idea.

In the *OQ-analyst* system the OQ-45 form is administered to the patient at every session and scores are plotted on a graph over time. The system then compares the patient's development over time with normative data (from 11,492 patients) and gives a "signal" if the patient deviates from his or hers estimated course. The estimated course is entirely based on the patient's psychological distress at baseline. The total range of the intake score is divided into 50 distinct groups, and specific trajectories for these clusters are estimated with the use of mixed linear models. Thus, the expected trajectory for a specific patient is chosen based on the first OQ-45 rating and which of those 50 groups the patient ends up in (Lambert et al., 2002).

A potential problem with only using the total summarized baseline score as a predictor of change rate is that other important information in the rating related to change rate might be overlooked. To investigate these matters in

more detail, more advanced statistical modeling that takes all items into account can be used to enhance knowledge about this relationship.

Change in symptom distress and dropout

As discussed in the chapter Dropout as an outcome of psychotherapy, above, there is no consensus on how to define psychotherapy dropout. In addition, different definitions of dropout might be related to specific predictors (Warnick, Gonzalez, Robin Weersing, Scahill, & Woolston, 2012). The lack of a coherent definition of this concept might also be a contributing factor for the ambiguous conclusions reached by many studies about moderators and predictors for dropout, which will be discussed in the following section.

The most comprehensive meta-analysis that examines moderators for dropout rate was conducted by Swift and Greenberg (2012) and included 669 studies with 83,834 patients. In this analysis “dropout rate was coded as the percentage of clients who started the intervention and who were identified as dropouts according to each author’s method of operationalization” (Swift & Greenberg, 2012, p. 551). Five broad categories of moderators for dropout rate were investigated: treatment (e.g., orientation, format, time limitations, manualization, and setting), patient (e.g., diagnosis, age, race, gender, marital status, employment, education) therapist (e.g., experience level, gender, race, age), and study design (e.g., definition of dropout, efficacy, or effectiveness study).

Results showed that higher dropout rates were found in studies that examined non-manualized psychotherapy with no predetermined time limit conducted at university-based clinics. Higher dropout rates were also found for younger patients with less education and for therapies conducted by clinicians with less experience (e.g., trainees working towards a degree). Finally, higher dropout rates were also found in studies using therapist-judged definitions of dropout (as mentioned in the chapter Dropout as an outcome of psychotherapy) and in studies conducted in naturalistic settings (i.e., effectiveness studies). Nevertheless, it is important to note that the observed effect sizes for these moderators were quite small and ranged from $d = 0.16$ to 0.29 .

Other studies have tried to replicate these findings. For example, Rubin, Dolev and Zilcha-Mano (2018) conducted a study of 413 patients treated at a university consulting center and examined four of the seven patient predictors analyzed in the Swift and Greenberg study (age, gender, education, & marital status). Patients in the study received non-limited psychodynamic therapy, and dropout was defined as all patients who terminated therapy before the 10th treatment session. Results showed that the relationship between lower

education and higher risk of dropout was replicated. However, age had the *opposite* relationship to dropout compared to the findings by Swift and Greenberg (i.e., younger age yielded less likelihood for dropout). Both studies found non-significant correlations between dropout rate and patient gender and marital status.

A literature review conducted by Roos and Werbart (2013) replicated the finding from Swift and Greenberg that more experienced therapists had fewer dropout patients than less experienced therapists. However, it is important to note that this study also included the results from Swift and Greenberg's meta-analysis, which could lead to the risk of circular reasoning.

Karyotaki et al. (2015) conducted a meta-analysis examining predictors for dropout in self-guided web-based interventions for depression ($k = 10$, $N = 2705$ patients). Dropout was defined as "...a completion rate of less than seventy five percent of the intervention modules..." (Karyotaki et al., 2015, p. 2720). Both findings for age and education level (observed in the Swift and Greenberg study) were replicated (i.e., lower age and less education yielded a higher risk for dropout). However, contrary to the findings in Swift and Greenberg, male gender significantly increased the risk of dropout in this study.

Finally, a study by Keefe, Chambless, Barber, and Milrod (2020) was conducted to explore "predictors and moderators for dropout in cognitive-behavioral and psychodynamic therapies for panic disorder" (Keefe et al., 2020, p. 1). This study included 200 patients, and "dropout was defined as not completing at least 16 sessions of treatment within a period of 12 weeks plus a 4-week grace period for scheduling complications" (Keefe et al., 2020, p. 3). Of the significant predictors found in Swift and Greenberg's review, this study examined patient education level and age but was not able to replicate any correlation between these variables and rates of dropout.

It is clear from these results that most studies examining predictors of dropout have focused on relatively static demographic variables collected before treatment had started. It is also evident that these predictors seem to be dependent on type of study and dropout definition, which lead to mixed findings. Also, investigating a static predictor in relation to dropout does not say anything about *when* the risk of dropout might increase or decrease for a specific patient.

Considering these challenges, it is clear that other ways of predicting dropout are needed. One innovative way to complement the literature might be to explore if local prediction systems can be developed to detect patients at risk of dropout during the course of treatment.

However, to train and tune models to fit specific contexts and find complex prediction patterns might not be an optimal task for most traditional

statistical methods used in psychotherapy research as they are primarily developed to find simpler, more interpretative, and generalizable patterns in data.

Advanced methods for studying the thesis outcomes

Multilevel modeling

Data from psychotherapy studies usually consists of ratings collected from patients that are clustered within different therapists (as mentioned in the chapter Growth rate above). There are many reasons for trying to disentangle this type of dependency in data. For example, methodological motives, such as handling non-independence in repeated measures, and more substantive motives, such as analyzing relationships in the data that vary at more than one level (e.g., at the therapist and patient levels).

This type of grouping in data is simply ignored in regular correlation analysis. However, the implementation of MLM as demonstrated by for example Raudenbush and Bryk (2002) has made it possible to separate these different levels in data.

An extension of the regular MLM is the *multivariate* MLM, which can include two or more outcomes, such as alliance ratings from both therapists and patients (Baldwin, Imel, Braithwaite, & Atkins, 2014; Raudenbush, Brennan, & Barnett, 1995). Separate intercepts (i.e., averaged ratings across therapist) and residuals (i.e., variations in ratings within each therapist) in the multivariate MLM are created for both rater perspectives. This procedure enables the exploration of how ratings from both raters relate to each other at different levels in the data (i.e., within and between therapists).

For example, assume that a fictional therapist “Märta” rated stronger alliances with her fictional patients “Ester”, “Tage”, and “Carl-Johan” than the typical therapist in the sample did (i.e., “Märta” had a higher *therapist*-rated intercept than the grand mean of the sample). Then it would be interesting to explore if “Märta” also received higher alliance scores on average from her patients than the typical therapist in the sample (i.e., whether “Märta’s” *patient*-rated intercept was higher than the grand mean of the sample). This association is possible to explore by correlating the deviation of the intercepts from their grand means.

However, it would also be interesting to explore how the alliance ratings related to each other *within* each therapist. For example, say that “Märta” rated her alliance with “Carl-Johan” higher than what she did with her two

other patients (i.e., the residual for “Carl-Johan’s” rating is above “Märta’s” *therapist*-rated intercept), does that mean that “Carl-Johan” also rated his alliance with “Märta” higher than what “Ester” and “Tage” did (i.e., “Carl-Johan”’s residual is above “Märtas” *patient*-rated intercept)? This would be possible to explore by correlating the residuals with each other.

Finally, MLM also makes it possible to study the relationship between the quality of the working alliance and treatment outcome in a more direct way. It might be used to investigate whether the correlation between working alliance and treatment outcome is stronger when studied within or between therapists. Such questions could have important implications for how to direct further studies (as discussed in the chapter Working alliance and treatment outcome, above).

The relationship between alliance and treatment outcome explored within and between patients was addressed in a recent meta-analysis (Flückiger et al., 2020). In this study, the time-lagged effects of session-by-session patient-rated alliance and outcome were analyzed (between sessions 1-7). The between patient associations that can be explored with this type of data is whether patients with stronger alliances in the early phase of treatment have better outcomes at the end of treatment. A within patient effect that can be explored is whether sessions with unusually strong alliances (in relation to the patient’s other ratings) are followed by unusually low symptom ratings (in comparison to the other symptom ratings for that patient) or vice versa.

Results showed that “better average early alliance scores had a significant positive association with patients’ posttreatment outcome” (Flückiger et al., 2020, p. 837). Similarly, the within patient associations showed that “higher than usual alliance scores in one session were followed by lower-than-usual symptoms in the following session” (Flückiger et al., 2020, p. 837). However, the reverse effect was also true; unusually high symptom ratings at a particular session were followed by an unusually low alliance rating the following session.

It is evident that MLM can nuance the association between the quality of the working alliance and treatment outcome in meaningful ways beyond the capacity of traditional correlation analysis.

Latent Profile Analysis

Baseline psychological distress might be one of the strongest predictors for change rate discovered in psychotherapy research (see the chapter Initial symptom distress and change rate, above). However, most studies have only examined whether *high* or *low* ratings of psychological distress are related to change rate and not whether specific answer patterns of the rating could

contain valuable information. Exploring such questions involves analyzing all items that make up the rating and not just sums of the scores (e.g., mean values).

Most traditional statistical methods (e.g., factor analysis, MLM, & regression) are so called *variable-centered* approaches that emphasize associations between variables. However, many psychotherapy questions are best answered with methods that take a person-centered approach where the emphasis is on relationships among individuals (Muthén & Muthén, 2000).

Variable-centered approaches rest on the assumption that all cases must be drawn from a single population from which average parameters can be estimated. Person-centered approaches, in contrast, consider the option that the sample might contain subpopulations with different sets of parameters (Wang & Hanges, 2011).

An example of such a model is Latent Profile Analysis (LPA; Lazarsfeld & Henry, 1968), which is a type of mixture model. A mixture model is a general term for probabilistic models for “...representing the presence of subpopulations within an overall population...” (Ge, Song, Ding, & Huang, 2017, p. 8). Mixture models were first initiated more than 100 years ago by the statistician Carl Pearson who fitted a mixture of two normal populations with different means and variances on crab data provided by his colleague Raphael Weldon. Since then, several models have evolved that incorporate this method such as the LPA (McLachlan, Lee, & Rathnayake, 2019).

The general aim of the LPA is to “recover hidden groups in data based on the means of continuous observed variables” (Oberski, 2016, p. 280). Latent Class Analysis (LCA) does the same thing but for categorical variables⁵. In LPA, patients who are similar to each other on the observed variables (e.g., ratings of symptom distress) are grouped together, and patients who diverge from this pattern are allocated to different groups. These unobserved groups of individuals are called latent classes. The objective is to find the least number of latent classes that fit the data in an optimal way. This is done by stepwise adding more classes until a good fit to the data has been reached.

The parameters in the LPA model are probabilities that an individual belongs to a specific class. This model allows researchers to explore patterns in data and how these patterns relate to other important variables (Berlin, Williams, & Parra, 2014). LPA can be used to investigate if groups of

⁵ Another difference between LCA and LPA is that the continuous indicators in LPA are allowed to be correlated, and these “...correlations (as well as covariance and variances) can be set to be freely estimated or fixed to be equal across classes” (Berlin et al., 2014, p. 3).

patients answer items in a self-assessment form in similar ways, and whether these groups have different change rates.

There are several advantages to using LPA and LCA over other methods used for clustering, such as k -mean clustering. For one, LPA and LCA are model based, which makes it possible to obtain an optimal solution by comparing fit indices across models (Nylund, Asparouhov, & Muthén, 2007). Also, because latent classes are grounded on probabilities, each individual is allowed to have some membership to all latent classes, which makes it possible to model how certain and/or precise the classifications are (Asparouhov & Muthén, 2014; Berlin et al., 2014). Another advantage of the LPA/LCA methods is that they are flexible in terms of the type of data that can be used in the model (e.g., binary, or continuous variables).

However, there are also limitations worth noting. For example, there are no established rules for deciding on the optimal number of classes. There can be difficulties with model convergence⁶ and the need for quite large sample sizes. Also, there is always the risk that there are other explanations beyond the presence of latent classes that was detected with the model. It is possible that latent classes were due to the nonnormality of the data or because of a misspecification of the model (Bauer & Curran, 2004).

Nevertheless, LPA offers an interesting opportunity to further explore the association between baseline psychological distress and change rate that goes beyond just level of distress and focuses on hidden patterns in the rating.

Machine learning

Dropout has generally been studied by examining relatively static predictors collected at the start of treatment and analyzed with the use of traditional statistical models directed towards the goal of explanation rather than the goal of prediction. However, other statistical frameworks developed specifically for prediction might be used to increase the ability to detect patients at risk of dropout. I will return to this more specific question in later parts of this chapter.

Arthur Samuel invented the term Machine Learning (ML) in the 1950's by using the game of checkers to illustrate the concept. In a ground-breaking article (Samuel, 1959), he was able to show that a computer could improve at the game the more it played by learning to incorporate moves that made up a winning strategy. Subsequently, Tom Mitchell has provided a more formal definition of ML as: "A computer program is said to learn from experience E

⁶ New estimates of parameters in the maximum likelihood equation do not improve when more iterations are run so that it gets stuck at a local minima (usually resolved by running the algorithm with different starting values).

with respect to some class of tasks T and performance measure P , if its performance at tasks in T , as measured by P , improves with experience E ." (Mitchell, 1997, p. 2). In broader terms, the essence of ML is that a machine (i.e., a computer) can learn from data without being explicitly programmed to come up with a fixed solution. Artificial Intelligence (AI) is closely linked to ML but can be viewed as a broader term where ML is one of its branches.

The purpose of an ML analysis can be split into two wide-ranging categories: *supervised* and *unsupervised* learning. Labeled data is needed in supervised learning. For example, the algorithm would need to know the "right" label for each patient (e.g., dropout or completer) to learn patterns associated between the labels and a set of predicting variables (or features). In contrast, unlabeled sample data is used in unsupervised learning to group patients based on similarities over specific variables (Brézillon & Gonzalez, 2014). An example of unsupervised learning is exploring groups of patients with similar response patterns on a self-assessment form (similar to the aim of Paper II), and an example of supervised learning is training an algorithm to find patterns in symptoms ratings to separate dropout patients from completers (the aim of Paper III).

ML is used in our daily lives in many different ways; for example, in Google searches and personalized marketing as well as in self-driving cars (Jordan & Mitchell, 2015). ML has until recently predominantly been applied by data scientists and engineering groups. However, when researchers from other fields of research begin to work in this arena it is important to recognize some fundamental differences between research traditions. For example, according to Mohr, Zhang, and Schueller (2017), data science is often more exploratory in nature than clinical psychology, which usually is more hypothesis-driven. In addition, ML comes from a tradition of practical application with more emphasis on prediction than on explanation.

ML algorithms can be separated into parametric and non-parametric models. In simple terms, parametric models rest on the assumption that the data follows a certain probability distribution that can be defined with a predefined number of parameters, such as the mean and standard deviation of a normal distribution. Non-parametric models, in contrast, treat the data mechanism as unknown and can contain an infinite number of parameters that usually gets more complex as the number of samples increases (Breiman, 2001b).

Breiman (2001b) denotes these two approaches as the *data* modeling culture (i.e., parametric models) and the *algorithmic* modeling culture (i.e., non-parametric models). Even though models from both cultures are applied under the ML-umbrella, ML has a stronger tilt towards the algorithmic modeling culture. The rise of this community started in the mid 1980's with

the development of some powerful algorithms intended to increase prediction accuracy (e.g., neural nets & decision trees). The field at this time consisted primarily of data scientists, physicists, and engineers who started working on complex problems for which regular parametric models did not work as well, such as handwriting-, image- and speech-recognition tasks.

Under the algorithmic modeling umbrella, most statistical problems are viewed as a set of X s (predictor variables, e.g., items on a self-assessment form) that goes into a model and outputs a set of Y s (predictions, e.g., “dropout” or “completer”) at the other end. The task is to find an algorithm (i.e., a model) that maps the road from the X s to the Y s in an optimal way. Performance is usually measured by concluding whether future X s will be a good predictor of Y s (Breiman, 2001b).

A common procedure to solve to this problem is to first divide the data into a training set and a test set (80 % of the data are typically used for training and 20 % for testing). The training set is used to finetune the algorithm to find the best patterns to apply for prediction, and the test set is used to validate the model with unseen data in an unbiased way. To find a stable solution and counter overfitting, a method called k -fold cross validation is often used during training. In short, this means that the training set is further randomly split into k equal parts or folds. As an example, let’s set $k = 10$ (which is often used). The model is then given nine of these parts ($k-1$) to train on, and when an optimal solution has been reached, it is tested on the remaining part (1/10). This procedure is then repeated 10 times so that each part of the training set has served as a cross-validation set one time. It is also possible to run k -fold cross validation several times with different settings to optimize the model for better performance (Cucker & Smale, 2002). The final step is to pick the model with the best performance and run it on the test set (the 20 % part of the data that was not included during training). A robust model usually has similar performance in the training and test parts of the analysis (Mitchell, 1997).

Very few assumptions are made about the data under the algorithmic modeling umbrella except that it is drawn Independent and Identically Distributed (IID) from an unknown distribution. In simple terms, this means that all samples are thought to be created by the same process with no memory of past generated samples, much like tossing dice. If we throw a die a few times, all the values we observe can be said to be IID random variables. Nevertheless, in the typical psychotherapy study, some patients are usually more likely to be included than others, which makes it difficult to guarantee a perfectly random sample. Thus, the samples are often treated as a “reasonable approximation” of IID. It is hard to estimate the effects of this practice on psychotherapy research, however, one study indicates that some ML

algorithms, such as Support Vector Machines (SVM), can be successfully applied to data that does not support the IID assumption (Steinwart, Hush, & Scovel, 2009).

Breiman (2001b) states that one limitation with models that make too many assumptions about the data is that it might lead to biased conclusions if the available data does not fit these assumptions. Common ways to check how well the underlying assumptions of a model fit the data are with goodness-of-fit tests or residual analysis. Nevertheless, one drawback with most of these tests is that they usually do not reject the misspecification of the model until the lack of fit is extreme (Bickel, Ritov, & Stoker, 2006).

However, both cultures have pros and cons for reaching conclusions from data. For example, parametric models are usually more parsimonious and easier to interpret and can be a better fit for simpler problems (which leans more towards the goal of explanation). Non-parametric models, on the other hand, are generally more flexible and might achieve better performance for prediction tasks but are usually much harder to interpret and can have higher risks of overfitting (Graham et al., 2019).

The aim of traditional parametric models is usually to draw inferences about a population from a sample. There are some important differences between the concepts of inference and prediction. Inference is usually concerned with characterizing a relationship between the data and an outcome variable and not *per se* to make predictions about future data (even though these models can still be used to make predictions). In contrast, prediction is specifically aimed at forecasting unseen outcomes or future behavior (Glasser, 2014).

An advantage of some ML approaches compared to traditional statistical approaches is that they can detect very complex interactions in data that can be used for prediction, but these may be too complex to interpret in any meaningful way (Graham et al., 2019). If the number of predictor variables is greater than the number of patients, ML methods can be particularly helpful (Bzdok, Altman, & Krzywinski, 2018). This is an especially important benefit because most other classical statistical models will usually breakdown when too many variables are added to the model.

In psychotherapy research, as well as in most other fields of research, the concepts of inference/explanation, and prediction are complementary as they could point to meaningful, although somewhat different conclusions. For example, we might want to infer which psychological processes are associated with patient dropout from treatment (a suitable aim for classical parametric statistics), as well as build a program for detecting these patients before they terminate treatment (a suitable aim for an ML algorithm). In the latter case, however, understanding the underlying mechanisms is not the

primary goal, the main thing is which variables are the most useful for detecting dropout cases based on a set of predictor variables (Bzdok et al., 2018).

For example, a meta-analysis by Reich and Berman (2018) has demonstrated an association between greater psychological distress at baseline and higher risk of dropout. Deterioration or no change in psychological distress during the early phase of treatment has also been shown to be related to a higher likelihood of dropout (Lutz et al., 2014). ML and item-level analysis make it possible to explore very complex patterns of this association and study if they can be used to detect dropout cases during the course of treatment. Such an aim leans very close to “raw” prediction. Nevertheless, such a study might lead to better insights into potential predictors for dropout and how to build local prediction systems that make use of data collected in routine care.

Despite some great advantages, it is apparent that ML techniques are not a universal solution for all situations, and they come with their own limitations. For example, ML models often sacrifice interpretability over predictive power as they usually give limited information about why a subject was classified in a specific way (Hart & Wyatt, 1990). This raises issues when ML algorithms are implemented in clinical settings. A researcher may experience difficulties explaining the rationale behind a specific classification if it was guided by a complex ML-algorithm (Hariman, Ventriglio, & Bhugra, 2019). Because interpretability is a critical limitation of ML, it has been a prioritized research area for some time (e.g., Li et al., 2009). It should be noted, however, that most models do allow some degree of interpretation at a global level with tools such as variable importance measures (e.g. Koutsouleris et al., 2016).

Another important issue noted by Mohr et al. (2017) is that researchers know very little about the “shelf lives” of the predictive models generated by some ML algorithms. When commonly used data collection technologies, such as smartphones and other devices, are updated, the associated raw data that these technologies collect will also change. The way people use their devices changes over time, too. For example, the way people communicate via smartphones has in recent years shifted away from SMS texting and phone calls to communication through messaging apps and social media. ML algorithms can become increasingly inaccurate when people change how they use their devices and, in extension, the data collected by the device technologies.

An early example of this was the so- called Google *FluTrends* platform that used “flu-related” internet search terms to predict outbreaks of influenza in different geographical regions. The platform (launched in 2008) initially

showed promising results with prediction accuracies as good as those provided by the Centers for Disease Control and Prevention (CDC) but with a faster delivery rate to the public. Therefore, the system was implemented in 29 countries and extended to include other diseases. It performed remarkably well for a short period of time until it stopped working. Evidently people's search behaviors changed, which in turn weakened the Google Flu Trends' models (Butler, 2013).

The most recent development in ML seems to be building bridges between the goals of prediction and explanation. ML algorithms typically excel at quite explicit tasks, such as detecting cancer in medical scans or facial recognition tasks, but they cannot take what they have learned and use it to adapt to a new but related environment. For example, "a robot that understands that dropping things causes them to break would not need to toss dozens of vases onto the floor to see what happens to them" (Knight, 2019, para. 5). Bengio et al. (2019) have initiated work in this area by developing algorithms that can create hypotheses about which variables might be causally related to each other and then examine how alterations to different variables fit that theory. This might be an interesting topic for psychotherapy research to consider in the future.

ML is an exciting new arena for psychotherapy research to explore. It is, however, important to note that ML comes with its own limitations and advantages and stems from a slightly different research tradition compared to conventional methods. In addition, ML is not a universal solution to all research questions but could point to meaningful, although somewhat different, conclusions.

As the MLM, LPA, and ML approaches are becoming important tools for developing psychotherapy research, there is a need to compile how they can be applied to real clinical data, to illustrate some of their strengths and limitations, as well as to study how they can be combined to act in complementary ways to enhance the field of psychotherapy research.

AIMS OF THE THESIS

The overarching aim of this thesis project was to study how advanced statistical methods can be used to enhance understanding of how to predict different types of psychotherapy outcomes in routine care. To fulfill these aims, three studies were conducted using a multi-method approach comprising Multilevel Modeling (MLM) (Paper I), Latent Profile Analysis (LPA) (Paper II), and Machine Learning (ML) (Paper III). Underlying this broader scope were three more explicit research objectives: to study

- how MLM can be used to explore aspects of the relationship between working alliance and treatment outcome overlooked with the use of traditional methods.
- how LPA and item level analysis can be used to nuance the relationship between psychological distress at baseline and change rate during treatment.
- how ML can be used to predict dropout cases based on complex patterns of symptom distress undetectable with traditional methods.

SUMMARY OF INCLUDED STUDIES

Paper I

Aims: The first aim of this study was to investigate if variability in alliance scores can be best explained by differences *between* or *within* therapists. If variability is mainly due to differences between therapists, this would indicate that patients seeing the same therapist would tend to rate their alliances in similar ways (i.e., some therapists receive high alliance ratings from most of their patients, and some therapists receive low alliance ratings from most of their patients). In contrast, if variability is best explained by differences *within* therapists, this would indicate that therapists tend to have strong alliances with some of their patients, and weak alliances with others, leading to small average differences between therapists. The same type of within and between therapist variability was also studied in the therapist's *own* ratings of the alliance (i.e., do some therapists rate on average higher alliances with their patients compared to other therapist? or do therapists in general rate very different alliances from patient to patient?).

The second aim was to explore whether this type of variability in alliance ratings can be used to predict treatment outcome. For example, how strong was the correlation between therapist's *average* alliance scores with their *average* treatment outcomes? And how strong was the correlation between therapist's *specific* alliance score with a certain patient and treatment outcome for that patient?

Methods: Data from a naturalistic setting collected from primary care units in Sweden was used in the study. The Working Alliance Inventory Short version (WAI-S; Tracey & Kokotovic, 1989) was used to measure therapist ratings of the working alliance, and the Working Alliance Inventory Short version Revised (WAI-SR; Falkenström, Hatcher, & Holmqvist, 2015) was used to measure patient ratings of the working alliance. Clinical Outcomes in Routine Evaluation Outcome Measure (CORE-OM; Evans, Connell, Barkham, Mellor-Clark, & Audin, 2002), was used to measure psychological distress pre to post treatment. Therapists and patients had to have filled out the WAI rating at the third treatment session to be included in the analysis, and patients also had to have complete the CORE-OM form pre- and post-treatment. Using these criteria 439 patients treated by 50 therapists were selected for analysis.

An analytical design called the “One-With-Many” was used to distinguish the different sources of variability observed in the data (Marcus, Kashy, & Baldwin, 2009). This approach was guided by a modification of the so-called “Social Relations Model” described by Kenny (SMR; Kenny, 1994) where variability in people’s views of one another in a group format is thought to be influenced by three factors: the *perceiver*, the *partner*, and the *relationship*. Marcus et al. (2009) has named their transformed design One-With-Many to reflect the specific situation in individual therapy where the therapist (the one) might rate their relationship with many patients (the many), but the patient can only rate his or her relationship with one therapist.

A multivariate multilevel model based on the two-intercept approach was used to disentangle the different sources of variability described in the One-With-Many analytic design (Raudenbush et al., 1995). Wald-z tests were used to test the significance of variability in the alliance scores as well as their correlations. Treatment outcome was examined with the use of residualized change scores, which were computed using the CORE-OM scores from the first treatment session to predict CORE-OM scores at the last session.

Results: Patients who saw the same therapist had very different opinions about the quality of the working alliance. In line with this, there were no significant differences in the aggregated scores of patient-rated alliances between therapists. Similarly, therapists rated significantly different alliances from patient to patient, however, some therapists also rated higher alliances on average than other therapists. In addition, patient’s average alliance score for a particular therapist correlated in general with the therapist’s own average alliance score. However, there was an even stronger association between therapist- and patient-specific alliance scores within each therapist. The results displayed no significant relation between patient’s average alliance scores (for particular therapists) and therapist’s average treatment outcomes. Neither were there any significant relationships between the therapist’s own average alliance ratings and their average treatment outcomes. In contrast, both patient- and therapist-specific ratings of the alliance were related to outcome. In other words, there was a strong correlation between alliance and outcome when analyzed *within* each therapist, but the average alliance scores for a specific therapist did not predict outcome. In other words, the general trend in the sample was that therapists had strong alliances and good outcomes with some of their patients, and weak alliances and limited outcomes with some of their patients. This is in contrast to one group of therapists with strong alliances and good outcomes with most of their patients.

Discussion: This study demonstrated the type of questions researchers might ask when they can separate variances into piles of *between* and *within* therapist variability. These matters would not have been possible to investigate without the use of MLM.

The results of the study stand in contrast to some earlier findings. For example, some studies have demonstrated stronger relationships between working alliance and treatment outcome when they were analyzed *between* therapists compared to *within*, and some have shown that both between- and within-therapist variations were associated with outcome (Baldwin, Wampold, & Imel, 2007; Del Re, Flückiger, Horvath, Symonds, & Wampold, 2012; Kivlighan, Gelso, Ain, Hummel, & Markin, 2015; Koszycki, Kelly, & Leybman, 2010; Marcus et al., 2009; Marcus, Kashy, Wintersteen, & Diamond, 2011). However, the present study is the first to indicate that *only* within alliance variability was associated with treatment outcome.

Why is it important to study whether variability in alliance *between* or *within* therapists is more associated with treatment outcome? For example, questions like this conveys something about how to direct further studies. If, for instance, there were obvious differences between therapists in their ability to foster strong alliances with their patients, and these therapists also had better treatment outcomes, it would be interesting to explore common characteristics of these therapists. If, on the other hand, it was more likely that therapists generally tended to have different alliances and outcomes from patient to patient (as suggested in the present study) other research strategies might be more constructive (e.g., how to better match therapists and patients or how to handle specifically demanding therapeutic relationships).

Paper II

Aims: The aim of this study was to explore patterns of psychological distress at the first treatment session, and whether these patterns could be used to predict treatment response.

Methods: Data from 1095 first-line mental health service patients in Sweden was used (the same database as in Paper I). Treatment consisted primarily of supportive therapy, Psychodynamic Therapy (PDT) and CBT. All participants completed the CORE-OM at each session during the course of treatment. CORE-OM consists of 34 items related to psychological distress experienced over the last week divided among four major problem areas: “functioning”, “risk” (of harm to self or others), “subjective well-being”, and “problems/symptoms”.

LPA was used to explore different distress profiles based on patient responses to the 34 CORE-OM items from the first treatment session. Patients with similar distress profiles were grouped together by the LPA algorithm, and the optimal number of such clusters (or latent classes) were investigated. Two analyses were conducted: one using patient raw scores (for each of the 34 items), and one using scaled scores. The scaled scores were computed for each patient by deducting each item score from the patient's total score (i.e., the mean of the 34 items). The reasons for exploring scaled scores was to identify groups of patients that might have similar items that stand out, or deviate, from their typical distress score. With this procedure, it was possible to control for the general distress factor and thus better characterize patients based on their most pronounced symptoms, as opposed to levels of psychological distress for each item. The inclusion of the raw scores analysis was motivated to compare profiles based on different scoring methods since the scaled-scores approach is not a typical practice in LPA.

The next step of the analysis was to use Latent Growth Curve Modeling (LGCM) to investigate differences in treatment response (i.e., change rates) between the profiles. The latent classes derived from the most appropriate LPA model were entered as time invariant covariates in the LGM. Baseline psychological distress (i.e., intercepts) as well as change rate per session were tested for group differences.

Results: Four profiles with significantly different degrees of distress at the start of treatment and with different change rates were discovered (with all p 's < .05 using Holm's correction). The most obvious dissimilarity between the profiles was their degree of endorsement of the risk area in relation to the total score. Profile 1 included 15% of the sample ($n = 161$) and was characterized by low initial distress (8.91 points), a very slow improvement rate (-0.25 points per session), and an average of 20% endorsement of the risk area (calculated by taking the average risk area score for the profile, divided by the average total score of the profile). Profile 2 included 46 % of the sample ($n = 509$), and showed average initial distress (15.01 points), a slow improvement rate (-0.62 points per session), and a 13 % endorsement of the risk area. Profile 3 comprised 25% of the sample ($n = 275$) and showed high levels of initial distress (20.76), a fast improvement rate (-1.58 points per session), and low endorsement of the risk area (11 %). Finally, profile 4 was characterized by very high initial distress (22.31), and a moderate improvement rate (-1.17 points per session) coupled with high endorsement of the risk area (51 %).

As predicted, the general trend in the raw-score analysis was that more distressed patients were grouped together on all items, however, this analysis

also marked the level of endorsement of the risk area as the most evident distinction between the profiles.

Discussion: By using the statistical procedures described above, it was possible to distinguish groups of patients with similar initial distress patterns and their relation to treatment response. Earlier studies have recognized the importance of baseline initial distress as a predictor of change rate (e.g., Y.-K. Tu & Gilthorpe, 2007). The general trend is usually that more distressed patients show faster recovery rates (due to, e.g., regression to the mean and/or because these patients have more room for improvement). However, even though this pattern was replicated in the study, results also suggest that endorsement of specific clusters of items might have a significant influence on change rate. For example, even if patients resembling profile 3 had significantly lower initial distress than patients resembling profile 4, they still had significantly faster recovery rates on average. The most obvious distinction between these profiles was how much of their total score was influenced by the endorsement of the risk area items (11 % compared to 51 %). In other words, a high total score in combination with low scores on the risk area might predict faster recovery rates. In addition, different change rates might be expected if two patients showed the same initial distress level but varied substantially in their score on the risk area items. If replicated, this information might be used by the clinician at the start of treatment to give a rough estimate of possible change rates for different patients.

This study focused on how to predict different treatment responses based on pretreatment characteristics. Such knowledge might be used to adapt treatment to the individual patient to fulfill the intentions of precision medicine

Paper III

Aims: The aim of this paper was to investigate whether specific patterns of symptom distress in the early phase of treatment could be used to detect patients at risk of dropping out of treatment.

Methods: The same dataset as in Paper I and Paper II was used in the study. To be included in the analysis patients had to have completed the CORE-OM form at sessions one to three, and their therapists needed to report whether the end of treatment was *planned* or *unplanned*. An unplanned ending was marked if a patient dropped out without giving any reason for it, or because they perceived the treatment as ineffective. Planned endings occurred if the termination was scheduled beforehand, or if the therapist and patient had come to a mutual agreement about ending treatment. 260 patients

were selected for analysis, 31(12 %) were considered dropouts, and 229 (88 %) as completers.

The rationale for exploring symptom distress to predict dropout was based on two earlier studies that suggested that dropout patients are generally more distressed when they enter treatment than completers (Reich & Berman, 2018), and that symptom change in the early phase of treatment might influence the likelihood of patients dropping out of treatment (Lutz et al., 2014)

Supervised learning together with a ML algorithm called Random Forest (RF) were used to explore fine-grained patterns of psychological distress to predict dropout. All 34 items from the CORE-OM form collected at each treatment session (one to three) for each patient were added to the model. To capture change over time, item change between sessions was also included as a predictor. Lastly, scaled scores were calculated in the same manner as in Paper II and added as predictors to capture a patient's relative, as in contrast to absolute, psychological distress.

Common ML procedures were used to train and validate the model. For example, the data was randomly split into a training set comprising 75% of the sample, and a test set comprising the remaining 25% of the sample. The training set was used to finetune the algorithm to find relevant patterns for prediction, and the test set was used to validate the model. During the training step, 10-fold cross validation was applied to counter overfitting and attain robust results. Area Under the receiver operating characteristics Curve (AUC) was chosen as a performance measure to evaluate different models. An AUC of > 0.7 was considered good performance in accordance with Rice and Harris (2005).

Results: The best fitting RF model achieved an AUC of 0.72 in the training part of the analysis and 0.76 when applied to the test set. The final model correctly identified 86% of the dropout patients and 72 % of the completers in the test set.

Discussion: This study showed how to use response to specific items from a common patient-rated self-assessment form to detect patients at risk of dropping out of treatment. Using item response as opposed to summary scores for prediction can be seen as an extension of the majority of earlier studies. The RF algorithm in combination with the supervised learning framework might be an especially suitable strategy for this task. RF is a good fit for data containing excessive variables or variables with high correlations, outliers, and/or missing values. RF is also robust against overfitting and can automatically consider interactions between variables. In addition, RF usually performs well with data of reasonable size and is specifically suitable for

datasets with more variables than patients (Breiman, 2001a; Breiman, Friedman, Stone, & Olshen, 1984)

Most earlier studies on dropout have only considered variables collected before treatment had started. Yet, if risk of dropout can be assessed continually during the early phase of treatment, clinicians can act on this information before termination has occurred. A similar model could, e.g., be added to a monitoring system (where a patient's progress in treatment is tracked by frequently administering the CORE-OM form), to aid clinicians in detecting patients at risk of dropout. The results of this study might thus be viewed as proof of concept of how to make better use of data collected as part of routine care. Even though the model results are unlikely to be universally applicable, the methods described herein might be used to develop a tailored model to fit a specific local setting, such as a mental health service clinic.

The study supports the possibility of using complex patterns of item response in the early phase of psychotherapy treatment to predict dropout. However, it is important to note that future research is needed to confirm these findings.

GENERAL DISCUSSION

Discussion of findings

The aim of this thesis project was to study how application of complex statistical methods can be used to gain knowledge about how to predict different types of psychotherapy outcomes in routine care. The three studies included in the thesis approached this question from different angles by incorporating three separate statistical frameworks: Multilevel Modeling (MLM), Latent Profile Analysis (LPA), and Machine Learning (ML).

The first paper explored how MLM can be used to separate within and between therapist correlations of the quality of the working alliance and treatment outcome. Such questions can have important clinical implications (as discussed in Introduction) but are not possible to explore with the use of traditional correlational analysis. Results showed that the association was stronger when studied within therapists than between therapists. In other words, there were no strong correlation between a therapist's average quality of working alliance and average treatment outcome. However, there was a clear correlation between working alliance and treatment outcome when studied within each therapist. For example, a higher than usual alliance with a specific patient for a therapist was generally associated with better than usual treatment outcome for that therapist with that patient. The study showed how MLM can be used to nuance the relationship between working alliance and treatment outcome.

The second study investigated how LPA and item-level analysis can be used to explore specific compositions of patient symptom distress at the first treatment session and whether such patterns can be associated with change rate. LPA enabled the exploration of latent constructs⁷ in the CORE-OM rating, which is an extension of using summary scores of a rating to predict change rate. Results showed that the degree of risk items endorsed by the patient compared to other questions in the rating can influence change rate. Higher scores on questions related to the risk of harm to self or others compared to other questions in the rating was generally associated with slower change rates.

⁷ A latent variable is not directly observable in the data but can be inferred from observable data, in this case from item responses of the CORE-OM rating collected at the first treatment session.

Finally, the third paper studied whether patterns of symptom distress in the early phase of treatment could be used to predict risk of dropout from treatment. This study exemplified how techniques from data science and ML can be implemented in psychotherapy research for prediction. For example, ML has specific procedures to counter overfitting (e.g., *k*-fold cross validation) and some algorithms, such as the Random Forest (RF) algorithm (used in Paper III), are especially suitable for data containing more variables than cases and can automatically consider interactions between variables. Traditional models, such as logistic regression applied in classical ways (i.e., without the ML procedures training, testing, and cross validation), cannot be used to explore as complex patterns in the data, and it would have been difficult to assess how the model would have performed with new data that was not included when fitting the model (e.g., to evaluate potential problems with overfitting).

The thesis project explored how advanced statistical methods can be used to refine questions related to for whom therapy will most likely be effective. The next step to consider in this line of research is how to implement and use this knowledge for clinical decision making at the individual patient level.

A natural springboard for applying these ideas is to add them to already existing frameworks that can be implemented at the clinic, such as Routine Outcome Monitoring (ROM) systems. ROM has attracted a lot of attention during recent decades and started as a means for clinicians to become more aware of self-assessment bias⁸ and to detect patients who were not responding to treatment (Lambert & Shimokawa, 2011). These systems commonly contain some sort of self-assessment that is provided to the patient at every session with scores plotted on a graph over time (for a summary see, e.g., Wampold, 2015). According to Wampold (2015): "...the use of ROM is the most noteworthy advance in psychotherapy in the last 25 years..." (Wampold, 2015, p. 458).

However, Lambert (2017) has stated that the utility of information derived from ROM may "...vary as a function of the degree of discrepancy between clinicians' views of progress and measured progress..." (Lambert (2017, p. 82). In other words, if a system is too simplistic, it might only predict what is already known and be of little use to the clinician. In fact, there is growing evidence that many clinicians do not use feedback systems even when they are implemented at the clinic (Ionita & Fitzpatrick, 2014; Jensen-Doss et al., 2018). Challenges to implementing feedback systems probably depend on a mix of factors ranging from organizational to

⁸ A common notion where people in general estimate their character, abilities, or future prospects in a more optimistic way than reality admits (Karpen, 2018).

attitudinal issues (Boswell, 2020; Hovland & Moltu, 2019; Solstad, Castonguay, & Moltu, 2019). However, a contributing factor might be that these systems must detect information that would not have been apparent to the therapist in order to add value for them. As noted by Wampold (2015), ROM as we know it might be coming of age and needs to be upgraded to the next generation. Good prediction is a prerequisite for good feedback.

As shown in Paper I within- therapist variation in working alliance might be included as a predictor of outcome. Also, as demonstrated in Paper II, it might be possible to include a more fine-grained analysis of a patient's symptom distress profile at pre-treatment as a predictor of change rate, in contrast to just using the total score as a predictor. Finally, as indicated in Paper III, ML has great potential to detect very complex (non-linear) interactions in this type of data, which can be used for prediction. These findings alone are not enough to revolutionize ROM, but studies like these might be a step in the right direction. I will elaborate more on potential future advances of clinical decision supports systems such as ROM in the coming chapter Clinical implications and future research.

A detailed description of the methods used in the three articles was provided in Introduction by moving from less complex models that have been around in the field for some time (MLM) to methods adapted in the past decade (ML). When MLM was first introduced to psychotherapy research, it provided a theoretical framework that could be used to reformulate earlier research topics and to generate new research ideas. Paper I demonstrated how MLM can be used to broaden the horizon of the relationship between working alliance and treatment outcome beyond a simple correlation by properly accounting for therapist and patient factors related to the correlation. According to Kahn (2011), this type of application of MLM has been specifically beneficial for psychotherapy research.

As described in Introduction, there are several advantages of using MLM over traditional models, especially for longitudinal analysis. Despite these advantages, there has been a substantial time lag in implementing this framework in the field of psychotherapy research, and it is only recently that it has become the standard for conducting repeated measures analysis.

Although three different statistical frameworks were used in this project, it is also important to note that they can be combined to function in complementary ways. This was shown in Paper II by combining LPA and LGCM which is a SEM framework similar to MLM. The first part of this study incorporated what in ML terminology is called unsupervised learning by fitting an LPA model to an unlabeled sample of patients to divide them into groups of related cases. The next step was to investigate whether these different groups varied in their rates of change, which was assessed with

LGCM. This study also demonstrated the possibility to incorporate all items from a self-assessment form instead of solely summary scores. A more fine-grained analysis like this has the potential to detect important complex associations that would have otherwise been hard to find in the data.

Paper III expanded the fine-grained analysis even more by exploring item-level symptom patterns and changes in these patterns over time to predict psychotherapy dropout. This study demonstrated a shift away from classical inference/explanation to a focus more on prediction. In applied fields of psychotherapy research, such as precision medicine, prediction can be a very important tool to forecast different patient behaviors during the course of treatment. As these models are becoming increasingly accurate, it is likely that they will be implemented in computer systems and used by clinicians in everyday clinical work in the future. Similar systems, commonly referred to as *clinical decision support systems*, are currently being implemented in other areas of healthcare (e.g., for medical diagnostics). These systems typically apply different types of algorithms to search for patterns in large datasets from multiple sources to provide useful information to clinicians (Lysaght, Lim, Xafis, & Ngiam, 2019). However, in psychotherapy studies, this work is still in early development and consists primarily of attempts at the concept level, such as the PAI method mentioned in Introduction. Nevertheless, computer-aided decision making has the potential to become an important advancement in the field of psychotherapy research in the future, and Paper III reports on the possible application of such a system to detect patients at risk of dropping out of treatment.

As noted in Introduction, ML originated in data science, which is primarily focused on finding solutions to specific problems. An ML algorithm can be trained and finetuned on data from a particular clinic or hospital (given enough data) and be customized to make predictions for that context. Clinical research, in comparison, is traditionally more focused on finding generalizable answers to broader questions, which in this example, would translate into finding predictors that relate to *most* psychotherapy patients or in other words explanation/inference. However, these traditions might work in complementary ways. For example, as discussed in the introduction, knowledge gained from broader clinical studies is essential to guide the search for the most likely predictors for a specific context and to understand a particular phenomenon. It is important for the psychotherapy research community to build bridges between these traditions, and this thesis is an attempt to move in that direction. In fact, studies of the relationship between explanation and prediction seem to be the most recent focus of AI and ML of today. In line with this, Shmueli (2010) has stated that “statisticians who see the uniqueness and importance of predictive modeling

alongside explanatory modeling have the capability of contributing to scientific advancement as well as achieving meaningful practical results...” (p. 302).

Woo, Chang, Lindquist, and Wager (2017) argue that the first phase in implementing new techniques or methods, such as ML, is primarily testing and discovery. The aim in this phase is typically to demonstrate that the new method works (proof-of-concept) and to explore different domains of application. This is also evident in respect to clinical psychology where many ML studies have focused on questions that are not immediately useful for clinicians. Woo et al. (2017) argue that in the next step, it is vital to focus more on the real-world applicability of ML techniques instead of theoretical possibilities in an undefined future. However, a barrier to the development of using ML methods in clinical psychology is that these techniques can be quite difficult to grasp and require a substantial investment of time to master for researchers operating in the social sciences. Mooney and Pejaver (2018) suggest that training in ML should be accessible for researchers regardless of their field of practice but should not be made mandatory. For example, whereas data scientists need strong programming and algorithmic skills, psychology researchers might need more of a conceptual understanding where the focus lies on interpretive and translational skills. This is important because responsible innovation might be necessary to further advance the field since the use of traditional approaches cannot answer all types of questions.

Even though these new methodologies come with new challenges, the opportunities far outweigh the costs. The questions discussed above will probably become of the utmost importance for the field of clinical psychology and related fields of psychology in the near future.

Strengths and limitations

The thesis had a broad scope, which could be viewed both as a strength and a limitation. The main goal was to study how to apply different types of statistical frameworks to data from routine care to advance the field of psychotherapy research. To best accomplish this goal, questions related to different areas of psychotherapy were explored, for example: the working alliance, rates of change, and dropout.

There were several limitations in each paper. A limitation in Paper I was that an MLM approach with repeated measures could have been used for assessing treatment outcome instead of pre- to post-measures. However, in order to be able to compare the results of this study with results from Marcus et al. (2009) (who used the same model), the analysis design was not altered.

Alliance is usually assessed in the early phase of treatment (Horvath, Del Re, Flückiger, & Symonds, 2011), and we decided to choose the same session for assessing alliance as Marcus et al. (2009) to simplify comparison between the studies.

A limitation of Paper II was that it was difficult to exactly pinpoint the meaning of the different latent profiles, and it was difficult to make a clear-cut decision regarding the optimal number of classes. Also, only a linear change rate was considered when assessing differences in rates of change in the classes. However, limiting the change rate to a linear one enabled easy comparison between profiles, which was the major goal of the analysis. Another limitation was that only a limited number of patients attended all the treatment sessions, which could imply that some patients had disproportionate weight in the model. However, this issue was addressed by investigating Cook's distance, which displayed no apparent outliers. It would also have been possible to implement common procedures from ML in the study, such as validating the model with the use of a cross-validation sample.

The major limitation of Paper III was that it had a low sample size combined with an unbalanced class distribution (i.e., more completers than dropouts). Even though it was possible to show that the model preformed equally well on training data (with 10-fold cross validation) compared to test data, there is still a risk that the model might not perform as well with new data (i.e., overfitting). Another limitation with this study was that it did not use a more simplistic model to compare the potential advantage of the complex model. For example, a model such as logistic regression applied to fewer predictor variables in the data could have been used as a base model to compare the performance. However, the major goal for this analysis was not to attain generalizable results or explain a phenomenon but to show how to build a local prediction system that could make use of data collected as part of routine care. Nevertheless, this study might help to generate hypotheses of how to predict psychotherapy dropout in more general ways, such as indicating novel possible predictors.

To be able to advance clinical practice and open the door to personalized medicine, it is important to draw on knowledge gained from broader clinical research and to finetune algorithms to fit specific contexts. An example of a good synergic relationship is if systems can be implemented that, on the one hand, inform the clinician about treatment progress and, on the other hand, collect data that can be used to improve predictions about relevant patient behaviors.

Clinical implications and future research

According to Iniesta, Stahl, and McGuffin (2016), there are a number of obstacles that must be overcome for complex statistical models to be beneficial for clinicians in a practical sense. For example, they must be validated with data from the context where they will be implemented. Models trained on samples with very specific characteristics, e.g., with patients from RCTs, might not generalize without problems to clinical practice. The performance of the model should also be superior to the clinician's conclusions, otherwise there is no advantage to applying it, which is a potential problem with ROM use currently. The model should also preferably work with data that is easy to obtain and possible to collect from a sufficiently large sample size to make adequate predictions, which is partly true for the data used in Paper III. DeRubeis (2019) also points out that implementation of these models in clinical practice must be thoughtful and consider the preferences and goals of clinicians, patients, and the health care systems.

Looking into the future of how to develop the coming decision support systems for psychotherapy such as ROM, a fundamental question to ask is how to build systems that are truly beneficial for the patient. For example, ROM initially showed promising results as a tool for enhancing treatment outcome (e.g., Shimokawa, Lambert, & Smart, 2010), but later studies have shown mixed results. A meta-analysis by Kendrick et al. (2016) ($k = 12$, $N = 3696$) for instance found no significant difference in treatment outcome between therapies conducted with ROM, compared to therapies conducted without the use of ROM. However, some studies suggest that ROM might be more effective in certain situations. For instance, de Jong, van Sluis, Nugter, Heiser and Spinhoven (2012) found that therapists characteristics (such as commitment of using the feedback) moderated the effects of ROM in their sample. Another study found that the benefits of using ROM might depend on the severity of the patient's problems (Davidson, Perry, & Bell, 2015) and ROM might be more beneficial in short-term therapies (De Jong et al., 2014).

Lambert and Harmon (2018) suggests that more research is needed to understand the mechanism of change of ROM. They also propose that ROM might be enhanced by improving the prediction of risk of treatment failure. This is important because a fundamental value of ROM is to supply clinicians with valid information about how the patient might develop over time.

One way of closing in on this goal could be to implement ideas from data science and ML. It would for example be interesting to explore dynamic systems that can change a prediction based on how the patient progresses in treatment. Different predictors might for instance be more useful at certain

phases of treatment. Separate models could be developed to fit specific stages so that each model uses information collected up to a certain point in time to predict the next phase, and so on. Such procedures might for example be implemented with the help of a ML ‘pipeline’ where the workflow of the analysis is divided into different parts that can be ‘pipelined together’ to create a prediction model (see e.g., Zhou, 2018).

Another interesting approach would be to implement a self-learning system that could make changes to the model while simultaneously monitoring the results of the changes in order to improve the prediction as more data is collected over time. Such an approach might be especially suitable in contexts where data is collected as part of routine care.

Increased prediction performance might also be reached by exploring a wider range of variables collected during the course of treatment. To hypothesize, session by session ratings of the working alliance might be useful predictors in the early phase of treatment and other variables such as for example engagements in social activities might be more important in later stages. It is also possible that specific predictors might become more or less important depending on classifications made by the model in earlier phases. Data collected from different sources, such as for example electronic health records, smart phones, therapists as well as patients should be further explored as they might be combined to work in complementary ways.

Vast amounts of resources have been invested in the private sector for developing ML algorithms to work on tasks such as face recognition, language processing, and product recommendations (e.g., Murray, 2019). There might be possibilities for psychotherapy research to draw on some of these advances. For example, product recommendations methods such as collaborative filtering algorithms (e.g., Alyari & Jafari Navimipour, 2018) could perhaps be altered and applied for matching therapists and patients with each other to increase the likelihood of a favorable treatment outcome. In addition, language processing techniques (e.g., the use of specific words in therapy and/or intonation etc.) as well as facial expressions might be possible to use to predict different psychotherapy outcomes in the future.

It is important to note however that the amount and quality of the data available are usually more important factors for successful predictions than finding an optimal ML-algorithm among hundreds of different models (e.g., Banko & Brill, 2001). This means that in order to reach the goals discussed above, data collection procedures would need to be upgraded as well.

To come back to the question stated above regarding how to build systems that can be truly beneficial to the patient, “more of the same” might not be the best way forward. In other words, improving performance of forecasting patient rated symptom distress during the course of treatment may

not be enough. Future systems would probably need to be upgraded to help clinicians forecast other types of patient behaviors as well such as for example risk of dropout, time needed in therapy to reach sufficient improvements, or risk of suicide and self-harm. It would also be an important development if future systems could advise the clinician how to proceed to increase the probability of a successful outcome given a certain prediction.

Another obstacle with forecasting psychotherapy treatment outcome might be linked to limitations in the current biopsychosocial model for classifying mental health problems. However, the British Psychological Society's Division of Clinical Psychology has proposed an alternative to psychiatric diagnosis and classification called the Power Threat Meaning (PTM) framework to identify patterns in "...emotional distress, unusual experiences and troubling behaviour..." (Johnstone & Boyle, 2018, p. 5). A central aim of this project is to implement a fundamentally different approach to pattern identification in relation to emotional and behavioral distress. According to the authors, an important difference between this approach and earlier models is that it does not place social and relational factors subordinate to underlying biological factors but rather sees biology as a mediator and an enabler of all human experience. A new way of conceptualizing mental health problems might be a fruitful way to guide the search for novel predictors to clinically relevant questions in the future.

What are the clinical implications of the studies in this thesis project? MLM was used in Paper I to show that variations in working alliance ratings between patients treated by the same therapist were related to outcome. In contrast, there was no strong evidence that some therapists had stronger alliances and better outcomes overall than other therapists. In other words, therapists tended to have strong alliances and good outcome with some patients, and weaker alliances and outcomes with others. This might not be a major revelation, but says something about how to best direct further research. If the opposite was true, implying that one group of therapists had strong alliances and outcomes with most of their patients, and another group had weak alliances and outcomes with most of their patients, it would be interesting to compare these groups for the characteristics associated with this difference. However, our results suggested that therapists generally tend to foster strong alliances with some of their patients and weak ones with others. Therefore, it might be important to encourage therapists to be attentive to such variability and learn how to cope with particularly challenging therapeutic relationships. More specifically, by letting the patient continuously assess the alliance over the course of treatment the therapist might more easily detect whether a certain patient deviates in comparison to his or her other patients. If ratings are lower than usual, the therapists can

take measures to try to improve the alliance with the patient in question. For example, by consulting a supervisor, or try to repair the alliance rupture through therapeutic metacommunication etc.

It is, however, also important to note that studies have indicated that there might be larger differences between therapists than what was reported in Paper I (e.g., Baldwin, Wampold, & Imel, 2007; Del Re, Flückiger, Horvath, Symonds, & Wampold, 2012; Kivlighan, Gelso, Ain, Hummel, & Markin, 2015; Koszycki, Kelly, & Leybman, 2010; Marcus et al., 2009; Marcus, Kashy, Wintersteen, & Diamond, 2011). In other words, this question is open for debate and in need of further research.

Paper II demonstrated a relationship between certain structures of patient-rated psychological distress at baseline and responses to treatment. The study was a step towards precision medicine as it aimed to increase awareness of baseline factors associated with outcome. Prognostic information of this type can be used to plan treatment or to monitor response in ongoing treatments. This paper also introduced a way of thinking about how to group patients based on items that stand out after adjusting for the general distress factor. Such information might have important clinical implications in other areas as well.

Paper III was an extension of this analysis and demonstrated a novel approach to detecting patients likely to drop out of therapy by analyzing complex symptom distress patterns during the early phase of treatment. If clinicians could use information like this collected during treatment to predict the risk of dropout, they might act on the information before termination has happened. An example of the practical implementation of such a model would be to include it in a ROM system where patient progress is tracked with symptom and problem ratings at each session. If the patient demonstrates patterns associated with the risk of dropout, the system could signal this to the clinician. A clinician would then be able to obtain valuable information about the treatment in a practical way as part of routine care.

CONCLUSIONS

The papers in this thesis showed how psychotherapy outcomes, such as changes in symptom distress, proportion of dropout, and rates of change, can be predicted with the help of modern statistical procedures with different foundations (MLM, LPA, & ML). A broad set of variables were used to produce these predictions, e.g., cross-sectional data on patient symptom distress, process variables such as ratings of the working alliance, and repeated ratings of symptom distress during treatment. The scientific goals of the studies touched upon description, prediction, and explanation but with a greater leaning towards prediction.

This work might be a springboard to further exploration in this line of research where the overarching goal is to use data to help the individual patient reach an optimal treatment outcome, in line with the goals of precision medicine and patient-focused research.

However, a quote by the British author Arthur C. Clarke expresses the caution required with these new approaches: “Any sufficiently advanced technology is indistinguishable from magic” (Clarke, 1972, p. 182). When statistical models are viewed as “black boxes” without any knowledge of their inner workings, they could easily attract this kind of magical thinking. Nevertheless the “garbage in, garbage out” maxim does apply to all statistical modeling, which means that if the data that goes into a model is biased or faulty in some way, then the outputs will reflect those issues.

There remain many obstacles that must be overcome for the ideas discussed in this thesis to be implemented in everyday clinical practice. Nevertheless, it is possible that novel statistical modeling and computer-aided decision making might be the paradigm shift needed to respond to the fundamental question posed by Gordon Paul in the late 1960’s “What treatment, by whom, is most effective for this individual with that specific problem, and under which set of circumstances?” (Paul, 1967, p. 111).

REFERENCES

- Allison, P. D. (1990). Change scores as dependent variables in regression analysis. *Sociological Methodology*, 20, 93-114.
- Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of Abnormal Psychology*, 112(4), 545-557. <https://doi.org/10.1037/0021-843X.112.4.545>
- Alyari, F., & Navimipour, J. N. (2018). Recommender systems: A systematic review of the state of the art literature and suggestions for future research. *Kybernetes*, 47(5), 985-1017. <https://doi.org/10.1108/K-06-2017-0196>
- Asparouhov, T., & Muthén, B. (2014). Auxiliary variables in mixture modeling: Three-step approaches using Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(3), 329-341. <https://doi.org/10.1080/10705511.2014.915181>
- Baldwin, S. A., Imel, Z. E., Braithwaite, S. R., & Atkins, D. C. (2014). Analyzing multiple outcomes in clinical research using multivariate multilevel models. *Journal of Consulting and Clinical Psychology*, 82(5), 920-930. <https://doi.org/10.1038/jid.2014.371>
- Baldwin, S. A., Murray, D. M., & Shadish, W. R. (2005). Empirically supported treatments or type I errors? Problems with the analysis of data from group-administered treatments. *Journal of Consulting and Clinical Psychology*, 73(5), 924-935. <https://doi.org/10.1037/0022-006X.73.5.924>
- Baldwin, S. A., Wampold, B. E., & Imel, Z. E. (2007). Untangling the alliance-outcome correlation: Exploring the relative importance of therapist and patient variability in the alliance. *Journal of Consulting and Clinical Psychology*, 75(6), 842-852. <https://doi.org/10.1037/0022-006X.75.6.842>
- Banko, M., & Brill, E. (2001, July). *Scaling to very very large corpora for natural language disambiguation*. Paper presented at the 39th Annual Meeting on Association for Computational Linguistics, Morristown, NJ, USA. Retrieved from <https://www.microsoft.com/en-us/research/publication/scaling-to-very-very-large-corpora-for-natural-language-disambiguation/>
- Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: What it is and how to deal with it. *International Journal of Epidemiology*, 34, 215-220. <https://doi.org/10.1093/ije/dyh299>
- Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and

- statistical considerations. *Journal of Personality and Social Psychology*, 51(6), 1173–1182. <https://doi.org/10.1007/BF02512353>
- Bauer, D. J., & Curran, P. J. (2004). The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychological Methods*, 9(1), 3–29. <https://doi.org/10.1037/1082-989X.9.1.3>
- Beck, A. T., Epstein, N., Brown, G., & Steer, R. A. (1988). An inventory for measuring clinical anxiety: Psychometric properties. *Journal of Consulting and Clinical Psychology*, 56(6), 893–897. <https://doi.org/10.1037/0022-006X.56.6.893>
- Beck, A. T., Steer, R. A., & Garbin, M. G. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review*, 8(1), 77–100.
- Bengio, Y., Deleu, T., Rahaman, N., Ke, N. R., Lachapelle, S., Bilaniuk, O., Goyal, A., & Pal, C. (2020, April) *A meta-transfer objective for learning to disentangle causal mechanism*. Paper presented at the 8th International Conference on Learning Representations, Virtual Conference, Formerly Addis Ababa, Ethiopia. Retrieved from <https://arxiv.org/abs/1901.10912>
- Berlin, K. S., Williams, N. A., & Parra, G. R. (2014). An introduction to latent variable mixture modeling (Part 1): Overview and cross-sectional latent class and latent profile analyses. *Journal of Pediatric Psychology*, 39(2), 174–187. <https://doi.org/10.1093/jpepsy/jst084>
- Bickel, P. J., Ritov, Y., & Stoker, T. M. (2006). Tailor-made tests for goodness of fit to semiparametric hypotheses. *The Annals of Statistics*, 34(2), 721–741. <https://doi.org/10.1214/009053606000000137>
- Bohart, A. C., & Wade, A. G. (2013). The client in psychotherapy. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (pp. 219–257). New York, NY: Wiley.
- Bordin, E. S. (1979). Working alliance. *Psychotherapy: Theory, Research, & Practice*, 16(3), 252–260. <https://doi.org/10.1016/B978-0-12-809324-5.05450-X>
- Boswell, J. F. (2020). Monitoring processes and outcomes in routine clinical practice: A promising approach to plugging the holes of the practice-based evidence colander. *Psychotherapy Research*, 30(7), 829–842. <https://doi.org/10.1080/10503307.2019.1686192>
- Breiman, L. (2001a). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Breiman, L. (2001b). Statistical modeling : The two cultures. *Statistical Science*, 16(3), 199–231.
- Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees*. London: Routledge Taylor &

- Francis Group.
- Brézillon, P., & Gonzalez, A. J. (2014). *Context in computing*. New York, NY: Springer.
- Brown, G. S., Burlingame, G. M., Lambert, M. J., Jones, E., & Vaccaro, J. (2001). Pushing the quality envelope: A new outcomes management system. *Psychiatric Services*, 52(7), 925–934. <https://doi.org/10.1176/appi.ps.52.7.925>
- Butler, D. (2013). When google got flu wrong. *Nature*, 494, 155–156.
- Bzdok, D., Altman, N., & Krzywinski, M. (2018). Statistics versus machine learning. *Nature Methods*, 15(4), 233–234. <https://doi.org/10.1038/nmeth.4642>
- Carey, T. A., & Stiles, W. B. (2016). Some problems with randomized controlled trials and some viable alternatives. *Clinical Psychology and Psychotherapy*, 23(1), 87–95. <https://doi.org/10.1002/cpp.1942>
- Castro-Schilo, L., & Grimm, K. J. (2018). Using residualized change versus difference scores for longitudinal research. *Journal of Social and Personal Relationships*, 35(1), 32–58. <https://doi.org/10.1177/0265407517718387>
- Clarke, A. C. (1972). *The lost worlds of 2001*. New York, NY: New American Library.
- Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist*, 12, 671–684.
- Cronbach, L. J., & Furby, L. (1970). How we should measure “change”: Or should we? *Psychological Bulletin*, 74(1), 68–80. <https://doi.org/10.1037/h0029382>
- Cucker, F., & Smale, S. (2002). Best choices for regularization parameters in learning theory: On the bias-variance problem. *Foundations of Computational Mathematics*, 2(4), 413–428. <https://doi.org/10.1007/s102080010030>
- Cuijpers, P., Li, J., Hofmann, S. G., & Andersson, G. (2010). Self-reported versus clinician-rated symptoms of depression as outcome measures in psychotherapy research on depression: A meta-analysis. *Clinical Psychology Review*, 30(6), 768–778. <https://doi.org/10.1016/j.cpr.2010.06.001>
- Curran, P. J., & Hussong, A. M. (2003). The use of latent trajectory models in psychopathology research. *Journal of Abnormal Psychology*, 112(4), 526–544. <https://doi.org/10.1037/0021-843X.112.4.526>
- Davidson, K., Perry, A., & Bell, L. (2015). Would continuous feedback of patient’s clinical outcomes to practitioners improve NHS psychological therapy services? Critical analysis and assessment of quality of existing studies. *Psychology and Psychotherapy: Theory, Research and Practice*, 88(1), 21–37. <https://doi.org/10.1111/papt.12032>

- De Jong, K., Timman, R., Hakkaart-Van Roijen, L., Vermeulen, P., Kooiman, K., Passchier, J., & Busschbach, J. Van. (2014). The effect of outcome monitoring feedback to clinicians and patients in short and long-term psychotherapy: A randomized controlled trial. *Psychotherapy Research*, 24(6), 629–639.
<https://doi.org/10.1080/10503307.2013.871079>
- de Jong, K., van Sluis, P., Nugter, M. A., Heiser, W. J., & Spinhoven, P. (2012). Understanding the differential impact of outcome monitoring: Therapist variables that moderate feedback effects in a randomized clinical trial. *Psychotherapy Research*, 22(4), 464–474.
<https://doi.org/10.1080/10503307.2012.673023>
- Del Re, A. C., Flückiger, C., Horvath, A. O., Symonds, D., & Wampold, B. E. (2012). Therapist effects in the therapeutic alliance-outcome relationship: A restricted-maximum likelihood meta-analysis. *Clinical Psychology Review*, 32, 642–649.
<https://doi.org/10.1016/j.cpr.2012.07.002>
- DeRubeis, R. J. (2019). The history, current status, and possible future of precision mental health. *Behaviour Research and Therapy*, 123, 1–6.
<https://doi.org/10.1016/j.brat.2019.103506>
- DeRubeis, R. J., Cohen, Z. D., Forand, N. R., Fournier, J. C., Gelfand, L. A., & Lorenzo-Luaces, L. (2014). The personalized advantage Index: Translating research on prediction into individualized treatment recommendations. A demonstration. *PLoS ONE*, 9(1), 1–9.
<https://doi.org/10.1371/journal.pone.0083875>
- DeRubeis, R. J., Hollon, S. D., Amsterdam, J. D., Shelton, R. C., Young, P. R., Salomon, R. M., ... Gallop, R. (2005). Cognitive therapy vs medications in the treatment of moderate to severe depression. *Archives of General Psychiatry*, 62(4), 409–416.
<https://doi.org/10.1001/archpsyc.62.4.409>
- Edwards, J., Anderson, K. K., & Stranges, S. (2019). Migrant mental health, Hickam's dictum, and the dangers of oversimplification. *International Journal of Public Health*, 64(4), 477–478.
<https://doi.org/10.1007/s00038-018-1194-3>
- Elfström, M. L., Evans, C., Lundgren, J., Johansson, B., Hakeberg, M., & Carlsson, S. G. (2013). Validation of the Swedish version of the Clinical Outcomes in Routine Evaluation Outcome Measure (CORE-OM). *Clinical Psychology & Psychotherapy*, 20(5), 447–455.
<https://doi.org/10.1002/cpp.1788>
- Enders, C. K. (2001). The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychological Methods*, 6(4), 352–370.
<https://doi.org/10.1037/1082-989x.6.4.352>

- Evans, C., Connell, J., Barkham, M., Mellor-Clark, J., & Audin, K. (2002). Towards a standardised brief outcome measure: Psychometric properties and utility of the CORE-OM. *British Journal of Psychiatry*, (180), 51–61.
- Evans, C., Mellor-Clark, J., Margison, F., Barkham, M., Audin, K., Connell, J., & McGrath, G. (2000). CORE: Clinical Outcomes in Routine Evaluation. *Journal of Mental Health*, 9(3), 247–255.
<https://doi.org/10.1080/713680250>
- Falkenström, F., Hatcher, R. L., & Holmqvist, R. (2015). Confirmatory factor analysis of the patient version of the working alliance inventory–short form revised. *Assessment*, 22, 581–593.
<https://doi.org/10.1177/1073191114552472>
- Flückiger, C., Del Re, A. C., Wampold, B. E., & Horvath, A. O. (2018). The alliance in adult psychotherapy: A meta-analytic synthesis. *Psychotherapy*, 55(4), 316–340. <https://doi.org/10.1037/pst0000172>
- Flückiger, C., Rubel, J., Del Re, A. C., Horvath, A. O., Wampold, B. E., Crits-Christoph, P., ... Barber, J. P. (2020). The reciprocal relationship between alliance and early treatment symptoms: A two-stage individual participant data meta-analysis. *Journal of Consulting and Clinical Psychology*, 88(9), 829–843. <https://doi.org/10.1037/ccp0000594>
- Forsell, E., Isacsson, N., Blom, K., Jernelöv, S., Abdesslem, F. Ben, Lindefors, N., ... Lindefors, N. (2019). Predicting Treatment Failure in Regular Care Internet-Delivered Cognitive Behavior Therapy for Depression and Anxiety Using Only Weekly Symptom Measures. *Journal of Consulting and Clinical Psychology*.
- Fournier, J. C., Derubeis, R. J., Shelton, R. C., Gallop, R., Amsterdam, J. D., & Hollon, S. D. (2009). Antidepressant medications versus cognitive therapy in depressed patients with or without personality disorder. *British Journal of Psychiatry*, 192(2), 124–129.
<https://doi.org/10.1192/bjp.bp.107.037234>. Antidepressant
- Fraguas, D., Díaz-Caneja, C. M., State, M. W., O'Donovan, M. C., Gur, R. E., & Arango, C. (2017). Mental disorders of known aetiology and precision medicine in psychiatry: A promising but neglected alliance. *Psychological Medicine*, 47(2), 193–197.
<https://doi.org/10.1016/j.physbeh.2017.03.040>
- Friedl, N., Berger, T., Krieger, T., Caspar, F., & Grosse Holtforth, M. (2019). Using the personalized advantage index for individual treatment allocation to cognitive behavioral therapy (cbt) or a cbt with integrated exposure and emotion-focused elements (cbt-ee). *Psychotherapy Research*, 30(6), 763–775.
<https://doi.org/10.1080/10503307.2019.1664782>
- Ge, Z., Song, Z., Ding, S. X., & Huang, B. (2017). Data mining and analytics

- in the process industry: The role of machine learning. *IEEE Access*, 5, 20590–20616. <https://doi.org/10.1109/ACCESS.2017.2756872>
- Glasser, S. P. (2014). *Essentials of clinical research*. Cham: Springer.
- Graham, S., Depp, C., Lee, E. E., Nebeker, C., Tu, X., Kim, H.-C., & Jeste, D. V. (2019). Artificial Intelligence for mental health and mental illnesses: An overview. *Current Psychiatry Reports*, 21(116), 1–18. <https://doi.org/10.1007/s11920-019-1094-0>
- Grwitch, M. J., & Munz, D. C. (2004). Are your data nonindependent? A practical guide to evaluating nonindependence and within-group agreement. *Understanding Statistics*, 3(4), 231–257. <https://doi.org/10.1207/s15328031us0304>
- Hamaker, E. L., Mulder, J. D., & van IJzendoorn, M. H. (2020). Description, prediction and causation: Methodological challenges of studying child and adolescent development. *Developmental Cognitive Neuroscience*, 46, 1–14. <https://doi.org/10.1016/j.dcn.2020.100867>
- Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery & Psychiatry*, 23(1), 56–62.
- Hansen, N. B., & Lambert, M. J. (2003). An evaluation of the dose-response relationship in naturalistic treatment settings using survival analysis. *Mental Health Services Research*, 5(1), 1–12. <https://doi.org/10.1023/A:1021751307358>
- Hariman, K., Ventriglio, A., & Bhugra, D. (2019). The future of digital psychiatry. *Current Psychiatry Reports*, 21(88), 1–8. <https://doi.org/10.1007/s11920-019-1074-4>
- Hariton, E., & Locascio, J. J. (2018). Randomised controlled trials—the gold standard for effectiveness. *BJOG*, 125(13), 1–4. <https://doi.org/10.1111/1471-0528.15199>
- Hart, A., & Wyatt, J. (1990). Evaluating black-boxes as medical decision aids: Issues arising from a study of neural networks. *Medical Informatics*, 15(3), 229–236. <https://doi.org/10.3109/14639239009025270>
- Hatcher, R. L., & Gillaspie, J. A. (2006). Development and validation of a revised short version of the Working Alliance Inventory. *Psychotherapy Research*, 16(1), 12–25. <https://doi.org/10.1080/10503300500352500>
- Hatchett, G. T., & Park, H. L. (2003). Comparison of four operational definitions of premature termination. *Psychotherapy*, 40(3), 226–231. <https://doi.org/10.1037/0033-3204.40.3.226>
- Hernán, M. A. (2018). The C-word: Scientific euphemisms do not improve causal inference from observational data. *American Journal of Public Health*, 108(5), 616–619. <https://doi.org/10.2105/AJPH.2018.304337>
- Hernán, M. A., Hsu, J., & Healy, B. (2019). A second chance to get causal inference right: A classification of data science tasks. *Chance*, 32(1),

- 42–49. <https://doi.org/10.1080/09332480.2019.1579578>
- Horvath, A. O., Del Re, A. C., Flückiger, C., & Symonds, D. (2011). Alliance in Individual Psychotherapy. *Psychotherapy*, 48(1), 9–16. <https://doi.org/10.1037/a0022186>
- Horvath, A. O., & Greenberg, L. S. (1989). Development and validation of the working alliance inventory. *Journal of Counseling Psychology*, 36(2), 223–233. <https://doi.org/10.1016/j.jcin.2014.01.154>
- Hotopf, M. (2002). The pragmatic randomised controlled trial. *Advances in Psychiatric Treatment*, 8(5), 326–333. <https://doi.org/10.1192/apt.8.5.326>
- Hovland, R. T., & Moltu, C. (2019). Making way for a clinical feedback system in the narrow space between sessions: Navigating competing demands in complex healthcare settings. *International Journal of Mental Health Systems*, 13(1), 1–11. <https://doi.org/10.1186/s13033-019-0324-5>
- Howard, K. I., Moras, K., Brill, P. L., Martinovich, Z., & Lutz, W. (1996). Evaluation of psychotherapy efficacy, effectiveness, and patient progress. *American Psychologist*, 51(10), 1059–1064.
- Huibers, M. J. H., Cohen, Z. D., Lemmens, L. H. J. M., Arntz, A., Peeters, F. P. M. L., Cuijpers, P., & DeRubeis, R. J. (2015). Predicting optimal outcomes in cognitive therapy or interpersonal psychotherapy for depressed individuals using the personalized advantage index approach. *PLoS ONE*, 10(11), 1–17. <https://doi.org/10.1371/journal.pone.0140771>
- Iniesta, R., Stahl, D., & McGuffin, P. (2016). Machine learning, statistical learning and the future of biological research in psychiatry. *Psychological Medicine*, 46, 2455–2465. <https://doi.org/10.1017/S0033291716001367>
- Ionita, G., & Fitzpatrick, M. (2014). Bringing science to clinical practice: A Canadian survey of psychological practice and usage of progress monitoring measures. *Canadian Psychology/Psychologie Canadienne*, 55(3), 187–196. <https://doi.org/10.1037/a0037355>
- Jensen-Doss, A., Haimes, E. M. B., Smith, A. M., Lyon, A. R., Lewis, C. C., Stanick, C. F., & Hawley, K. M. (2018). Monitoring treatment progress and providing feedback is viewed favorably but rarely used in practice. *Administration and Policy in Mental Health and Mental Health Services Research*, 45(1), 48–61. <https://doi.org/10.1007/s10488-016-0763-0>
- Johanson, F. (2020). Machine Learning for Causal Inference from Observational Data with Applications in Healthcare. Retrieved from <https://www.chalmers.se/en/projects/Pages/Machine-Learning-for-Causal-Inference-from-Observational-Data.aspx>
- Johnstone, L., & Boyle, M. (2018). *The power threat meaning framework*.

- British Psychological Society* (Vol. 3). Leicester: British Psychological Society. <https://doi.org/10.1108/17465641211279798>
- Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255–260. <https://doi.org/10.1126/science.aaa8415>
- Kahn, J. H. (2011). Multilevel modeling: Overview and applications to research in counseling psychology. *Journal of Counseling Psychology*, 58(2), 257–271. <https://doi.org/10.1037/a0022680>
- Karpen, S. C. (2018). The social psychology of biased self-assessment. *American Journal of Pharmaceutical Education*, 82(5), 441–448. <https://doi.org/10.5688/ajpe6299>
- Karyotaki, E., Kleiboer, A., Smit, F., Turner, D. T., Pastor, A. M., Andersson, G., ... Cuijpers, P. (2015). Predictors of treatment dropout in self-guided web-based interventions for depression: An “individual patient data” meta-analysis. *Psychological Medicine*, 45(13), 2717–2726. <https://doi.org/10.1017/S0033291715000665>
- Kazdin, A. E. (2007). Mediators and mechanisms of change in psychotherapy research. *Annual Review of Clinical Psychology*, 3(1), 1–27. <https://doi.org/10.1146/annurev.clinpsy.3.022806.091432>
- Keefe, J. R., Chambless, D. L., Barber, J. P., & Milrod, B. L. (2020). Predictors and moderators of treatment dropout in cognitive-behavioral and psychodynamic therapies for panic disorder. *Psychotherapy Research*, 25, 1–11. <https://doi.org/10.1080/10503307.2020.1784487>
- Kendrick, T., El-Gohary, M., Stuart, B., Gilbody, S., Churchill, R., Aiken, L., ... Moore, M. (2016). Routine use of patient reported outcome measures (PROMs) for improving treatment of common mental health disorders in adults. *Cochrane Database of Systematic Reviews*, (7). <https://doi.org/10.1002/14651858.CD011119>
- Kenny, D. A. (1994). *Interpersonal perception: A social relations analysis*. New York, NY: Guilford Press.
- Khoury, M. J., & Galea, S. (2017). Will precision medicine improve population health? *JAMA*, 316(13), 1357–1358. <https://doi.org/10.1097/01.ogx.0000511937.49229.57>
- Knight, W. (2019, August, 10) An AI pioneer wants his algorithms to understand the 'why'. *Wired*. Retrieved from <https://www.wired.com/story/ai-pioneer-algorithms-understand-why/>
- Kisbu-Sakarya, Y., MacKinnon, D. P., & Aiken, L. S. (2013). A Monte Carlo comparison study of the power of the analysis of covariance, simple difference, and residual change scores in testing two-wave data. *Educational and Psychological Measurement*, 73(1), 47–62. <https://doi.org/10.1016/j.physbeh.2017.03.040>
- Kivlighan, D. M., Gelso, C. J., Ain, S., Hummel, A. M., & Markin, R. D.

- (2015). The therapist, the client, and the real relationship: An actor–partner interdependence analysis of treatment outcome. *Journal of Counseling Psychology*, 62(2), 314–320.
<https://doi.org/10.1037/cou0000012>
- Kneale, W., & Kneale, M. (1962). *The development of logic*. Oxford: Oxford University Press.
- Koszycski, D. C., Kelly, A. C., & Leybman, M. J. (2010). Between-therapist and within-therapist differences in the quality of the therapeutic relationship: Effects on maladjustment and self-critical perfectionism. *Journal of Clinical Psychology*, 66(7), 681–697.
<https://doi.org/10.1002/jclp>
- Koutsouleris, N., Kahn, R. S., Chekroud, A. M., Leucht, S., Falkai, P., Wobrock, T., ... Hasan, A. (2016). Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: A machine learning approach. *The Lancet Psychiatry*, 3(10), 935–946.
[https://doi.org/10.1016/S2215-0366\(16\)30171-7](https://doi.org/10.1016/S2215-0366(16)30171-7)
- Lambert, M. J. (2017). Maximizing psychotherapy outcome beyond evidence-based medicine. *Psychotherapy and Psychosomatics*, 86, 80–89. <https://doi.org/10.1159/000455170>
- Lambert, M. J., & Finch, A. E. (1999). The outcome questionnaire. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcome assessment* (2ed ed., pp. 831–870). Mahwah, NJ: Earlbaum.
- Lambert, M. J., & Harmon, K. L. (2018). The merits of implementing routine outcome monitoring in clinical practice. *Clinical Psychology: Science and Practice*, 25(4), 1–13. <https://doi.org/10.1111/cpsp.12268>
- Lambert, M. J., & Shimokawa, K. (2011). Collecting client feedback. In J. C. Norcross (Ed.), *Psychotherapy relationships that work: Evidence-based responsiveness* (pp. 203–223). New York, NY: Oxford University Press.
- Lambert, M. J., Whipple, J. L., Bishop, M. J., Vermeersch, D. a, Gray, G. V, & Finch, A. E. (2002). Comparison of empirically-derived and rationally-derived methods for identifying patients at risk for treatment failure. *Clinical Psychology and Psychotherapy*, 9(May), 149–164.
 Retrieved from
<http://onlinelibrary.wiley.com/doi/10.1002/cpp.333/abstract>
- Lazarsfeld, P., & Henry, N. (1968). *Latent structure analysis*. New York, NY: Houghton-Mifflin.
- Li, B., Krishnan, V. G., Mort, M. E., Xin, F., Kamati, K. K., Cooper, D. N., ... Radivojac, P. (2009). Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics*, 25(21), 2744–2750. <https://doi.org/10.1093/bioinformatics/btp528>

- Lord, F. M. (1956). The measurement of growth. *Educational and Psychological Measurement*, 16, 437–459.
- Lutz, W., Hofmann, S. G., Rubel, J., Boswell, J. F., Shear, M. K., Gorman, J. M., ... Barlow, D. H. (2014). Patterns of Early Change and Their Relationship to Outcome and Early Treatment Termination in Patients With Panic Disorder. *Journal of Consulting and Clinical Psychology*, 82(2), 287–297. <https://doi.org/10.1037/a0035535>
- Lysaght, T., Lim, H. Y., Xafis, V., & Ngiam, K. Y. (2019). AI-assisted decision-making in healthcare. *Asian Bioethics Review*, 11, 299–314. <https://doi.org/10.1007/s41649-019-00096-0>
- Magnusson, K., Andersson, G., & Carlbring, P. (2018). The consequences of ignoring therapist effects in trials with longitudinal data: A simulation study. *Journal of Consulting and Clinical Psychology*, 86(9), 711–725. <https://doi.org/10.1037/ccp0000333>
- Mani, N., Slevin, N., & Hudon, A. (2011). What three wise men have to say about diagnosis. *BMJ (Online)*, 343, 1–4. <https://doi.org/10.1136/bmj.d7769>
- Marcus, D. K., Kashy, D. A., & Baldwin, S. A. (2009). Studying psychotherapy using the one-with-many design : The therapeutic alliance as an exemplar. *Journal of Counseling Psychology*, 56(4), 537–548. <https://doi.org/10.1037/a0017291>
- Marcus, D. K., Kashy, D. A., Wintersteen, M. B., & Diamond, G. S. (2011). The therapeutic alliance in adolescent substance abuse treatment: A one-with-many analysis. *Journal of Counseling Psychology*, 58(3), 449–455. <https://doi.org/10.1037/a0023196>
- Mayer, A., Zimmermann, J., Hoyer, J., Salzer, S., Wiltink, J., Leibing, E., & Leichsenring, F. (2020). Interindividual differences in treatment effects based on structural equation models with latent variables: An EffectLiteR tutorial. *Structural Equation Modeling: A Multidisciplinary Journal*, 27(5), 798–816. <https://doi.org/10.1080/10705511.2019.1671196>
- McLachlan, G. J., Lee, S. X., & Rathnayake, S. I. (2019). Finite mixture models. *Annual Review of Statistics and Its Application*, 6, 355–378. <https://doi.org/10.1198/tech.2002.s651>
- Miller, R. B. (1998). Epistemology and psychotherapy data: The unspeakable, unbearable, horrible truth. *Clinical Psychology: Science and Practice*, 5(2), 242–250. <https://doi.org/10.1111/j.1468-2850.1998.tb00147.x>
- Mitchell, T. M. (1997). *Machine learning*. New York, NY: McGraw-Hill Science/Engineering/Math. https://doi.org/10.1007/978-3-642-21004-4_10
- Mohr, D. C., Zhang, M., & Schueller, S. M. (2017). Personal sensing:

- Understanding mental health using ubiquitous sensors and machine learning. *Annual Review of Clinical Psychology*, 13, 23–47.
<https://doi.org/10.1146/annurev-clinpsy-032816-044949>
- Montgomery, S. A., & Åsberg, M. (1979). A new depression scale designed to be sensitive to change. *British Journal of Psychiatry*, 134, 382–389.
<https://doi.org/10.1192/bjp.134.4.382>
- Mooney, S. J., & Pejaver, V. (2018). Big data in public health: Terminology, machine learning, and privacy. *Annual Review of Public Health*, 39, 95–112. <https://doi.org/10.1146/annurev-publhealth-040617-014208>
- Murray, J. (2019, Mars, 23) Machine learning academics are flocking to tech giants and that might be a problem. *Binary District Journal*. Retrieved from <https://thenextweb.com/syndication/2019/03/23/machine-learning-academics-are-flocking-to-tech-giants-and-that-might-be-a-problem/>
- Muthén, B., & Muthén, L. (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research*, 24(6), 882–891. Retrieved from <https://static1.squarespace.com/static/57309137ab48de6f423b3eec/t/5a00cc548165f5a597e7db13/1510001751444/Muthen%26Muthen2000.pdf>
- Norcross, J. C., & Lambert, M. J. (2011). Evidence- based therapy relationships. In J. C. Norcross (Ed.), *Psychotherapy relationships that work: Evidence- based responsiveness* (pp. 3–24). New York, NY: Oxford University Press.
- Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study. *Structural Equation Modeling*, 14(4), 535–569. <https://doi.org/10.1080/10705510701575396>
- Oberski, D. (2016). Mixture models: Latent profile and latent class analysis. In J. Robertson & M. Kaptein (Eds.), *Modern statistical methods for HCI. Human–Computer Interaction Series* (pp. 275–287). Cham: Springer.
- Occham, W. (1495). *Quaestiones et decisiones in quattuor libros Sententiarum Petri Lombardi. Centilogium theologicum*. Ed: Augustinus de Ratisbona, Jodocus Badius Ascensius. Lyon: Johannes Trechsel.
- Ogles, B. M. (2013). Measuring change in psychotherapy reserach. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (6th ed., pp. 134–166). New York, NY: Wiley.
- Paul, G. L. (1967). Strategy of outcome research in psychotherapy. *Journal of Consulting Psychology*, 31(2), 109–118.
<https://doi.org/10.1037/h0024436>
- Pedhazur, E. J. (1975). *Multiple regression in behavioral research:*

- Explanation and prediction*. Belmont, CA: Wadsworth.
<https://doi.org/10.2307/2285468>
- Raudenbush, S. W., Brennan, R. T., & Barnett, R. C. (1995). A multivariate hierarchical model for studying psychological change within married couples. *Journal of Family Psychology*, 9(2), 161–174.
<https://doi.org/10.1037/0893-3200.9.2.161>
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.
- Reich, C. M., & Berman, J. S. (2018). Are psychotherapies with more dropouts less effective? *Psychotherapy Research*, 30(1), 1–18.
<https://doi.org/10.1080/10503307.2018.1534018>
- Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's d, and r. *Law and Human Behavior*, 29(5), 615–620. <https://doi.org/10.1007/s10979-005-6832-7>
- Roos, J., & Werbart, A. (2013). Therapist and relationship factors influencing dropout from individual psychotherapy: A literature review. *Psychotherapy Research*, 23(4), 394–418.
<https://doi.org/10.1080/10503307.2013.775528>
- Rubin, A., Dolev, T., & Zilcha-Mano, S. (2018). Patient demographics and psychological functioning as predictors of unilateral termination of psychodynamic therapy. *Psychotherapy Research*, 28(5), 672–684.
<https://doi.org/10.1080/10503307.2016.1241910>
- Samuel, A. L. (1959). Some studies in machine learning. *IBM Journal of Research and Development*, 3(3), 210–229.
- Seligman, M. E. P. (1995). The effectiveness of psychotherapy the consumer reports study. *American Psychologist*, 50(12), 965–974.
- Shean, G. (2014). Limitations of randomized control designs in psychotherapy research. *Advances in Psychiatry*, 2014, 1–5.
<https://doi.org/10.1155/2014/561452>
- Shimokawa, K., Lambert, M. J., & Smart, D. W. (2010). Enhancing treatment outcome of patients at risk of treatment failure: meta-analytic and mega-analytic review of a psychotherapy quality assurance system. *Journal of Consulting and Clinical Psychology*, 78(3), 298–311.
<https://doi.org/10.1037/a0019247>
- Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3), 289–310. <https://doi.org/10.1214/10-STS330>
- Solstad, S. M., Castonguay, L. G., & Moltu, C. (2019). Patients' experiences with routine outcome monitoring and clinical feedback systems: A systematic review and synthesis of qualitative empirical literature. *Psychotherapy Research*, 29(2), 157–170.
<https://doi.org/10.1080/10503307.2017.1326645>

- Steinwart, I., Hush, D., & Scovel, C. (2009). Learning from dependent observations. *Journal of Multivariate Analysis*, 100(1), 175–194. <https://doi.org/10.1016/j.jmva.2008.04.001>
- Swift, J. K., Callahan, J., & Levine, J. C. (2009). Using clinically significant change to identify premature termination. *Psychotherapy*, 46(3), 328–335. <https://doi.org/10.1037/a0017003>
- Swift, J. K., & Greenberg, R. P. (2012). Premature discontinuation in adult psychotherapy: A meta-analysis. *Journal of Consulting and Clinical Psychology*, 80(4), 547–559. <https://doi.org/10.1037/a0028226>
- Swift, J. K., Greenberg, R. P., Tompkins, K. A., & Parkin, S. R. (2017). Treatment refusal and premature termination in psychotherapy, pharmacotherapy, and their combination: A meta-analysis of head-to-head comparisons. *Psychotherapy*, 54(1), 47–57. <https://doi.org/10.1037/pst0000104>
- Tasca, G. A., & Gallop, R. (2009). Multilevel modeling of longitudinal data for psychotherapy researchers: I. The basics. *Psychotherapy Research*, 19(4–5), 429–437. <https://doi.org/10.1080/10503300902849475>
- Thornicroft, G., & Slade, M. (2000). Are routine outcome measures feasible in mental health? *Quality in Health Care*, 9(2), 84. <https://doi.org/10.1136/qhc.9.2.84>
- Tierney, L. M., Midthun, D. E., & Saint, S. (2004). Occam's Razor versus Saint's Triad. *New England Journal of Medicine*, 350, 599–603.
- Tracey, T. J., & Kokotovic, A. M. (1989). The factor structure of the shortened working alliance inventory. *Psychological Assessment*, 1(3), 207–210.
- Tu, Y.-K., & Gilthorpe, M. S. (2007). Revisiting the relation between change and initial value: A review and evaluation. *Statistics in Medicine*, 26, 443–457. <https://doi.org/10.1002/sim.2538>
- Tu, Y.-K., Maddick, I. H., Griffiths, G. S., & Gilthorpe, M. S. (2004). Mathematical coupling can undermine the statistical assessment of clinical research: Illustration from the treatment of guided tissue regeneration. *Journal of Dentistry*, 32(2), 133–142. <https://doi.org/10.1016/j.jdent.2003.10.001>
- Wampold, B. E. (2015). Routine outcome monitoring: Coming of age—with the usual developmental challenges. *Psychotherapy*, 52(4), 458–462. <https://doi.org/10.1037/pst0000037>
- Wampold, B. E., & Imel, Z. E. (2015). *The great psychotherapy debate* (Second Edi). New York: Routledge Taylor & Francis Group.
- Wang, M., & Hanges, P. J. (2011). Latent class procedures: Applications to organizational research. *Organizational Research Methods*, 14(1), 24–31. <https://doi.org/10.1177/1094428110383988>
- Warnick, E. M., Gonzalez, A., Robin Weersing, V., Scahill, L., & Woolston,

- J. (2012). Defining dropout from youth psychotherapy: How definitions shape the prevalence and predictors of attrition. *Child and Adolescent Mental Health*, 17(2), 76–85. <https://doi.org/10.1111/j.1475-3588.2011.00606.x>
- Wierzbicki, M., & Pekarik, G. (1993). A meta-analysis of psychotherapy dropout. *Professional Psychology: Research and Practice*, 24(2), 190–195.
- Wilson, E. O. (1998). *Consilience: The unity of knowledge*. *Annals of Emergency Medicine*. New York, NY: Vintage books.
<https://doi.org/10.1016/j.annemergmed.2010.05.024>
- Woo, C.-W., Chang, L. J., Lindquist, M. A., & Wager, T. D. (2017). Building better biomarkers: brain models in translational neuroimaging. *Nature Neuroscience*, 20(3), 365–377.
<https://doi.org/10.1016/j.physbeh.2017.03.040>
- Zhou, L. (2018, September, 5) How to build a better machine learning pipeline. *Datanami*. Retrieved from
<https://www.datanami.com/2018/09/05/how-to-build-a-better-machine-learning-pipeline/>

Papers

The papers associated with this thesis have been removed for copyright reasons. For more details about these see:

<http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-172754>

LINKÖPING STUDIES IN BEHAVIOURAL SCIENCE

200. ÅHLFELDT, EMANUEL. Hållbart utvecklingsarbete i vård och omsorg. Ett institutionellt perspektiv på projekt i en professionell och byråkratisk kontext. 2017. ISBN: 978-91-7685-445-7
201. VERNMARK, KRISTOFER. Therapeutic alliance and different treatment formats when delivering internet-based CBT for depression. 2017. ISBN: 978-91-7685-436-5
202. FRANKL, MY. Psychotherapy for Substance Use Disorder – the importance of affects. 2017. ISBN: 978-91-7685-429-7
203. ABDULLA, AFRAH. Readiness or resistance? – Newly arrived adult migrants' experiences, meaning making, and learning in Sweden. 2017. ISBN: 978-91-7685-428-0
204. OLSSON, LINDA. "Count on me!". Mathematical development, developmental dyscalculia and computer-based intervention. 2018. ISBN: 978-91-7685-409-9
205. COLLIANDER, HELENA, Being and Becoming a Teacher in Initial Literacy and Second Language Education for Adults. 2018. ISBN: 978-91-7685-304-7
206. TOPOOCO, NAIRA. Blended Cognitive Behavior Therapy. Efficacy and Acceptability for Treating Depression in the Adult and Adolescent Population. 2018. ISBN: 978-91-7685-297-2
207. SOLÍS MARCOS, IGNACIO. Challenges in Partially Automated Driving: A Human Factors Perspective. 2018. ISBN: 978-91-7685-296-5
208. THORSTEN, ANJA. Berättelseskivande i skolan. Att studera, beskriva och utveckla ett kunnande. 2018. ISBN: 978-91-7685-258-3
209. ELEONOR, BREDLÖV EKNOR. Shaped for beauty. Vocational and gendered subjectivities in private education for the beauty industry. 2018. ISBN: 978-91-7685-248-4
210. VESTERGREN, SARA. Psychological change as an outcome of participation in collective action. 2018. ISBN: 978-91-7685-242-2
211. MÖLLER, CLARA. Mentalizing. Competence and process. 2018. ISBN: 978-91-7685-189-0
212. SÖDERBERG GIDHAGEN, YLVA. Psychological treatment of outpatients with substance use disorders in routine care – attachment style, alliance, and treatment outcome. 2018. ISBN: 978-91-7685-197-5

213. HAGMAN, WILLIAM. When are nudges acceptable? Influences of beneficiaries, techniques, alternatives and choice architects. 2018. ISBN: 978-91-7685-160-9
214. RAHM, LINA. Educational imaginaries: a genealogy of the digital citizen. 2019. ISBN: 978-91-7685-158-6
215. HALVARSSON LUNDQVIST, AGNETA. Learning Dynamics of Workplace Development Programmes. Studies in Swedish national programmes. 2019. ISBN: 978-91-7685-124-1
216. HOLMQVIST LARSSON, MATTIAS. Rupture and Repair in the Working Alliance: Relation to Psychotherapy Outcome and Within-Session Interaction. 2019. ISBN: 978-91-7685-111-1
217. LINDQVIST, HENRIK. Student teachers' and beginning teachers' coping with emotionally challenging situations. 2019. ISBN: 978-91-7685-078-7
218. GRÖNLUND, AGNETA. Återkoppling i analoga och digitala klassrum. Spänningsfyllda verksamheter i samhällskunskapsundervisning. 2019. ISBN: 978-91-7685-074-9
219. ÅKERBLOM, ERIKA. Utbildning och hälsa i nationens intresse. Styrningsteknologier och formering av en förädlad befolkning. 2019. ISBN: 978-91-7685-053-4220.
220. ÖSTERBORG WIKLUND, SOFIA. Folkbildning i global (o)rättvisa. Makt och motstånd i folkhögskolans internationalisering och transnationella kurser. 2019. ISBN: 978-91-7519-002-0
221. DAHLIN, MATS. Development and evaluation of an internet-based treatment for generalized anxiety disorder. An acceptance-based approach. 2020. ISBN: 978-91-7929-793-0
222. ERIKSSON, ELISABETH. Återkoppling i lågstadielklassrum. 2020. ISBN: 978-91-7929-769-5
223. KÖPSÉN, JOHANNA. Knowledge in VET curricula and power in society and labour market: Policy and practice: demands-based and employer-driven Swedish higher vocational education. 2020. ISBN: 978-91-7929-768-8
224. SJÖGREN, BJÖRN. Bystander behaviors in pre victimization: Associations with moral disengagement, efficacy beliefs, and student-teacher relationship quality 2021. ISBN: 978-91-7929-728-2
225. PERSSON-ASPLUND, ROBERT. Learning how to recover from stress-related disorders via internet-based interventions. 2021. ISBN: 978-91-7929-722-0

FACULTY OF ARTS AND SCIENCES

Linköping Studies in Arts and Sciences No. 802, 2021.
Linköping Studies in Behavioural Science No. 226, 2021.
Department of Behavioural Sciences and Learning (IBL)

Linköping University
SE-581 83 Linköping, Sweden

www.liu.se