Improving Predictive Modeling With Machine Learning Techniques

Inauguraldissertation zur Erlangung der Doktorwürde (Dr. rer. nat.) im Fach Psychologie, Fachbereich I an der Universität Trier

UNIVERSITÄT
TRIER

Vorgelegt im
April 2023

von

Björn Bennemann

Gutachter:

Prof. Dr. Wolfgang Lutz

Prof. Dr. Julian A. Rubel

Fachbereich I – Psychologie
Klinische Psychologie und Psychotherapie

Improving Predictive Modeling With Machine Learning Techniques

Björn Bennemann

Inauguraldissertation zur Erlangung der
Doktorwürde (Dr. rer. nat.)

Gutachter:
Prof. Dr. Wolfgang Lutz
Prof. Dr. Julian Rubel

2023

Dissertationsort: Trier

Danksagung

Mein Dank gilt an erster Stelle meinem Doktorvater Prof. Dr. Wolfgang Lutz, der mich während meiner gesamten Promotionsphase begleitet, unterstützt und gefördert hat. Ohne seinen kontinuierlichen Input und unsere anregenden Diskussionen, die vielen Möglichkeiten, die er mir geboten, und Türen, die er mir geöffnet hat, wäre diese Arbeit in der Form nicht entstanden.

Darüber hinaus möchte ich Prof. Dr. Julian Rubel danken, der mich vom Beginn meiner wissenschaftlichen Karriere bis zur Promotion immer fachlich unterstützt hat. Trotz seines Weggangs zwecks eigener beruflicher Ziele war er stets ansprechbar und hat für mich immer eine Rolle als fachlicher Mentor eigenommen. Auch die gemeinsamen Unternehmungen bei Konferenzen und Tagungen werden mir in Erinnerung bleiben.

Des Weiteren möchte ich mich bei Dr. Brian Schwartz bedanken, der mich ebenfalls von Anfang an fachlich und freundschaftlich unterstützt hat. Egal ob über neue Methodik oder neue Serien, ich habe die gemeinsamen Diskussionen immer als sehr bereichernd erlebt.

Daneben möchte ich mich bei Dr. Anne-Katharina Deisenhofer bedanken. Unsere gemeinsamen Laufeinheiten erlebte ich immer als den nötigen Ausgleich zum Arbeitsstress und als große Ressource.

Mein Dank gilt weiterhin allen Kolleginnen und Kollegen, die immer für mich da waren und einen unschätzbaren Beitrag zu meinem Durchhaltevermögen geleistet haben. Ein spezieller Dank geht dabei an Dr. Kaitlyn Poster, die unseren Publikationen und zuletzt auch dieser Arbeit den so wertvollen sprachlichen Feinschliff verliehen hat. Vielen Dank an alle für die vergangenen und die noch kommenden Jahre!

Da die Grundidee der gesamten Dissertation mit den Daten der Psychotherapieambulanz der Universität Trier steht und fällt, müssen auch das Leitungsteam, allen voran Dr. Birgit Weinmann-Lutz, und alle Therapeutinnen und Therapeuten Erwähnung finden, ohne deren organisatorische und klinische Arbeit diese Forschung nicht möglich gewesen wäre.

Ein ganz besonderer Dank gilt meiner Lebensgefährtin Kerstin Schenke, die ebenso großes Durchhaltevermögen über die Dauer meiner Promotion bewiesen hat, indem sie meine Launen stets ausgehalten hat und immer an mich geglaubt hat, sogar dann, wenn ich es selbst am wenigsten tat. Ich liebe dich sehr.

Table of Contents

Abstract vi
List of All Authored and Co-Authored Publications viii
List of Publications Included in the Cumulative Dissertation x
List of Tables xi
List of Figures xii
List of Abbreviations xiii
1 Introduction 1
2 Theoretical Background 3
2.1 Shifts and Developments in Psychotherapy Research 3
2.2 Patient-Focused Research in Psychotherapy 5
2.3 From Personalized to Precision Mental Health 5
2.4 Prediction 7
2.4.1 Data Requirements 8
2.4.2 Statistical and Analytical Requirements 9
2.4.3 Requirements for Generalizability and Clinical Utility 11
3 Research Questions 13
3.1 Study I 13
3.2 Study II 13
3.3 Study III 13
4 Methodological Aspects 14
4.1 Machine Learning 14
4.1.1 Random Forest 15
4.1.2 Elastic Net 15
4.1.3 Nearest Neighbor 16
4.1.4 Ensembles 16
4.1.5 Optimization of Algorithm Parameters 17
5 Study I: Predicting Patients who Will Drop out of Outpatient Psychotherapy Using Machine Learning Algorithms 18
5.1 Introduction 18
5.2 Methods 18
5.2.1 Patients and Treatment 18

5.2.2 Measures ... 19

5.2.3 Selection of ML Algorithms ... 19

5.2.4 Data Analytic Strategy ... 19

5.3 Results ... 20

5.4 Conclusions ... 23

6 Study II: Therapist Interventions and Skills as Predictors of Dropout in Outpatient Psychotherapy ... 25

6.1 Introduction ... 25

6.2 Methods ... 26

6.2.1 Patients and Treatment ... 26

6.2.2 Measures ... 26

6.2.3 Video Selection and Rating Procedure ... 26

6.2.4 Data Analytic Strategy ... 26

6.3 Results ... 27

6.4 Conclusion ... 29

7 Study III: Fostering the Upward Spiral After a Sudden Gain in Routine Care Cognitive Behavioral Therapy ... 31

7.1 Introduction ... 31

7.2 Methods ... 32

7.2.1 Patients and Treatment ... 32

7.2.2 Measures ... 32

7.2.3 Data Preparation ... 32

7.2.4 Data Analytic Strategy ... 33

7.3 Results ... 34

7.4 Conclusion ... 35

8 General Discussion ... 37

8.1 General Conclusions and Future Research ... 38

8.2 General Limitations ... 40

8.3 Concluding Remarks ... 42

9 References ... 43

10 Original Publications ... 68

Abstract

There is no longer any doubt about the general effectiveness of psychotherapy. However, up to 40% of patients do not respond to treatment. Despite efforts to develop new treatments, overall effectiveness has not improved. Consequently, practice-oriented research has emerged to make research results more relevant to practitioners. Within this context, patient-focused research (PFR) focuses on the question of whether a particular treatment works for a specific patient. Finally, PFR gave rise to the precision mental health research movement that is trying to tailor treatments to individual patients by making data-driven and algorithm-based predictions. These predictions are intended to support therapists in their clinical decisions, such as the selection of treatment strategies and adaptation of treatment. The present work summarizes three studies that aim to generate different prediction models for treatment personalization that can be applied to practice. The goal of Study I was to develop a model for dropout prediction using data assessed prior to the first session (*N* = 2543). The usefulness of various machine learning (ML) algorithms and ensembles was assessed. The best model was an ensemble utilizing random forest and nearest neighbor modeling. It significantly outperformed generalized linear modeling, correctly identifying 63.4% of all cases and uncovering seven key predictors. The findings illustrated the potential of ML to enhance dropout predictions, but also highlighted that not all ML algorithms are equally suitable for this purpose. Study II utilized Study I's findings to enhance the prediction of dropout rates. Data from the initial two sessions and observer ratings of therapist interventions and skills were employed to develop a model using an elastic net (EN) algorithm. The findings demonstrated that the model was significantly more effective at predicting dropout when using observer ratings with a Cohen's *d* of up to .65 and more effective than the model in Study I, despite the smaller sample (*N* = 259). These results indicated that generating models could be improved by employing various data sources, which provide better foundations for model development. Finally, Study III generated a model to predict therapy outcome after a sudden gain (SG) in order to identify crucial predictors of the upward spiral. EN was used to generate the model using data from 794 cases that experienced a SG. A control group of the same size was also used to quantify and relativize the identified predictors by their general influence on therapy outcomes. The results indicated that there are seven key predictors that have varying effect sizes on therapy outcome, with Cohen's *d* ranging from 1.08 to 12.48. The findings suggested that a directive approach is more likely to lead to better outcomes after an SG, and that alliance ruptures can be effectively compensated for. However, these effects were reversed in the control group. The results of the three studies are discussed regarding their usefulness to support clinical decision-making and their implications for the implementation of precision mental health.

List of All Authored and Co-Authored Publications

Bennemann, B., Caspar, F., Evers, O., Laireiter, A.-R., Lutz, W., Poster, K., Rief, W., Strauß, B., & Taubner, S. (2021). Förderung persönlicher Kompetenzen in der psychotherapeutischen Aus- und Weiterbildung. In Rief, W., Schramm, E., & Strauß, B. (Eds.), *Psychotherapie – Ein kompetenzorientiertes Lehrbuch* (1st ed., pp. 31–52). Elsevir GmbH.

Bennemann, B., Schwartz, B., Giesemann, J., & Lutz, W. (2022). Predicting patients who will drop out of out-patient psychotherapy using machine learning algorithms. *The British Journal of Psychiatry*, *220*(4), 192–201. https://doi.org/10.1192/bjp.2022.17

Bennemann, B., Schwartz, B., & Lutz, W. (2023). Fostering the upward spiral after a sudden gain in routine care cognitive behavioral therapy. [Manuscript submitted for publication].

Boyle, K., Deisenhofer, A. K., Rubel, J. A., Bennemann, B., Weinmann-Lutz, B., & Lutz, W. (2020). Assessing treatment integrity in personalized CBT: the inventory of therapeutic interventions and skills. *Cognitive Behaviour Therapy, 49*(3), 210–227. https://doi.org/10.1080/16506073.2019.1625945

Deisenhofer, A. K., Meyer, L., Schaffrath, J., Weinmann-Lutz, B., Poster, K., Bennemann, B., & Lutz, W. (2022). Vermittlung psychotherapeutischer Fertigkeiten durch die Arbeit mit Simulationspatientinnen. Die Evaluation einer innovativen Lehrmethode im universitären Kontext. *Zeitschrift für Klinische Psychologie und Psychotherapie, 50*(3-4), 145–158. https://doi.org/10.1026/1616-3443/a000634

Deisenhofer, A. K., Rubel, J. A., Bennemann, B., Aderka, I. M., & Lutz, W. (2022). Are some therapists better at facilitating and consolidating sudden gains than others? *Psychotherapy Research*, *32*(3), 343–357. https://doi.org/10.1080/10503307.2021.1921302

Moggia, D., Bennemann, B., Schwartz, B., Hehlmann, M. I., Driver, C. D., & Lutz, W. (in review). Process-Based Psychotherapy Personalization: Considering Causality with Continuous-Time Dynamic Modeling. *Psychotherapy Research.*

Giesemann, J., Delgadillo, J., Schwartz, B., Bennemann, B., & Lutz, W. (2023). Predicting dropout from treatment using different machine learning algorithms, resampling methods, and sample sizes. *Psychotherapy Research*, 1–13. https://doi.org/10.1080/10503307.2022.2161432

Lutz, W., Clausen, S. A., Bennemann, B., Zimmermann, D., Prinz, J., Rubel, J. A., & Deisenhofer, A. K. (2019). Chancen von E-Mental-Health und eProzessdiagnostik in der ambulanten Psychotherapie: Der Trierer Therapie Navigator. *Verhaltenstherapie, 29*(3), 145–154. https://doi.org/10.1159/000501026

Lutz, W., Deisenhofer, A. K., Rubel, J., Bennemann, B., Giesemann, J., Poster, K., & Schwartz, B. (2022). Prospective evaluation of a clinical decision support system in psychological therapy. *Journal of Consulting and Clinical Psychology, 90*(1), 90–106. https://doi.org/10.1037/ccp0000642

Poster, K., Bennemann, B., Hofmann, S. G., & Lutz, W. (2021). Therapist Interventions and Skills as Predictors of Dropout in Outpatient Psychotherapy. *Behavior therapy*, 52(6), 1489–1501. https://doi.org/10.1016/j.beth.2021.05.001

List of Publications Included in the Cumulative Dissertation

Study I

Bennemann, B., Schwartz, B., Giesemann, J., & Lutz, W. (2022). Predicting patients who will drop out of out-patient psychotherapy using machine learning algorithms. *The British Journal of Psychiatry*, *220*(4), 192–201. https://doi.org/10.1192/bjp.2022.17

Study II

Poster, K., Bennemann, B., Hofmann, S. G., & Lutz, W. (2021). Therapist Interventions and Skills as Predictors of Dropout in Outpatient Psychotherapy. *Behavior therapy*, 52(6), 1489–1501. https://doi.org/10.1016/j.beth.2021.05.001

Study III

Bennemann, B., Schwartz, B., & Lutz, W. (2023). Fostering the upward spiral after a sudden gain in routine care cognitive behavioral therapy. [Manuscript submitted for publication].

List of Tables

Table 1: Predictors used for model generation. Predictors were routinely collected at intake..21

Table 2: Mean scores across all 100 confusion matrices for the models excluding and including ITIS variables...29

Table 3: Mean regression weights for all relevant variables across the three sessions after a sudden (pseudo) gain and the mean differences between these two groups.............................35

List of Figures

Figure 1: Distribution of the 20 outer cross-validation (CV) models generated by each algorithm and ensemble ranked from best to worst using the area under the curve (AUC).....23

Figure 2: Means of the relative importance of the pre-selected predictor variables across all 100 models..28

Figure 3: Number of occurrences of each variable in each outer CV run for both groups (i.e., sudden gainers and non-gainers)..34

List of Abbreviations

ADA	Adapted boosted classification trees
AUC	Area under the curve
ASC	Assessment for signal clients
BPSR	Bern post session reports
BSI	Brief symptom inventory
CBT	Cognitive behavioral therapy
CoV	Coefficient of variation
CV	Cross-validation
DASK	Dysfunctional attitudes scale – short form
EMA	Ecological momentary assessment
EMI	Emotionality inventory
EN	Elastic net
FEP	Questionnaire for the evaluation of psychotherapeutic progress
GAF	Global assessment of functioning
GBM	Generalized boosted regression model
GLM	Generalized linear model
GLMAIC	Generalized linear model with stepwise feature selection using Akaike information criterion
GLMBOOST	Boosted generalized linear model
GSI	Global severity index
HAQP	Helping alliance questionnaire – patient version
HAQT	Helping alliance questionnaire – therapist version
HSCL-11	Hopkins symptom checklist short form
KODAP	Coordination of data collection and analysis at research and training outpatient clinics for psychotherapy
LASSO	Least absolute shrinkage and selection operator
LDA	Linear discriminant analysis
ML	Machine learning
NN	Nearest neighbor
IAPT	Improving access to psychological therapies

IDCL-P	International diagnostic checklist for personality disorders
IIP	Interpersonal problems
INK	Incongruence questionnaire
ILE	Inventory of stressful life events
ITIS	Inventory of therapeutic interventions and skills
OQ	Outcome questionnaire
PFR	Patient-focused research
PG	Pseudo gain
POR	Practice-oriented research
PROGRESS	Prognosis research strategy
PSM	Propensity score matching
PSSI	Personality style and disorder inventory
PSTB	Bern postsession reports – patient perspective
RCI	Reliable change index
RCT	Randomized controlled trial
RF	Random forest
RMSE	Root-mean-square error
SCID-I	Structured clinical interview for DSM-IV
SG	Sudden gain
SMART	Stratified medicine approaches for treatment selection
SRS	Session rating scale
TH	Therapist expectations
TK	Techniker Krankenkasse
TSIL	Treatment selection idea lab
TSTB	Bern postsession reports – therapist perspective
TreatEx	Treatment expectations
XGB	Extreme gradient boosting

1 Introduction

Psychological disorders are prevalent in all cultures worldwide and account for a large proportion of the disease burden. In 2010, mental and substance use disorders accounted for over 183 million (7.4%) of all disability adjusted life years worldwide. Such disorders accounted for 8.6 million years (0.5%) of life lost to premature mortality and over 175 million (22.9%) of all years lived with disability, illustrating the impact of psychological disorders on society (Whiteford et al., 2013). These findings are further supported by the World Health Organization, which predicts that depressive disorders will be the most common cause of illness by 2030 (World Health Organization, 2008). In order to reduce these alarming numbers, psychotherapy plays a fundamental role in the care and treatment of psychological disorders being an effective and cost-efficient form of treatment (e.g., Barkham & Lambert, 2021; Ewbank et al., 2020; Lazar, 2014).

Psychotherapy is to be understood as a group of scientifically based procedures and methods for the treatment of mental illnesses and relevant psychological factors in physical illnesses (Strauß, 2021). Further, it is an interpersonal treatment with psychological methods based on empirically founded psychological concepts. Hereby, the goal is to change behavior, cognitions, emotions and other characteristics in accordance with the patient's wishes (Wampold & Imel, 2015). Over the past decades, many studies have demonstrated the efficiency and effectiveness of psychotherapy for the treatment of psychological disorders (Barkham & Lambert, 2021; Mohr et al., 2014; Wampold, 2019). However, only about 60–69% of all patients have a positive outcome, meaning that about 31–40% of all patients have no improvement or even deteriorate during therapy (Barkham & Lambert, 2021; McAleavey et al., 2019). Furthermore, approximately 20% of all patients drop out of therapy (Swift et al., 2017), often due to deterioration or non-response (Lutz et al., 2014).

As a result of these numbers, the practical relevance of *treatment-focused* research has been increasingly criticized (Kazdin, 2008; Rothwell, 2005), leading to the development of *practice-oriented* research (POR; Castonguay et al., 2021). One branch of this research focuses on the prediction of complex therapeutic events such as treatment discontinuation, sudden gains (SGs), or treatment outcomes in the context of *patient-focused* research (PFR; Lutz, de Jong, et al., 2021). One goal here is to generate useful predictions that can be applied directly in clinical practice to improve treatment outcomes by providing the therapist with clinical support tools (e.g., Boswell, 2020; de Jong et al., 2021; Lutz, de Jong, & Rubel, 2015; Lutz, Deisenhofer, et al., 2022; Lutz, Zimmermann, et al., 2017). Since therapists are often inaccurate in their predictions (Ægisdóttir et al., 2006; Chapman et al., 2012; Macdonald & Mellor-Clark, 2015), comprehensive data acquisition and methodology is necessary (Kessler et al., 2019) to generate individual predictions for the individual patient in the sense of a *precision mental health* procedure (e.g., Z. D. Cohen & DeRubeis, 2018; Kessler et al., 2017; Lutz, de Jong, et al., 2021; Lutz, Saunders, et al., 2006).

For this reason, new methodologies have emerged in PFR, namely machine learning (ML) algorithms, that can handle large data sets to further improve individualized predictions (e.g., Delgadillo, 2021; Delgadillo & Lutz, 2020). Nevertheless, in addition to numerous positive opinions (Aafjes-van Doorn et al., 2021; Chekroud et al., 2021; Kessler et al., 2019; Taubitz et al., 2022), there is also criticism that questions the usefulness of ML or calls for more research (Liu et al., 2019; Makridakis et al., 2018; Wilkinson et al., 2020).

This umbrella therefore summarizes three studies that used advanced methodology and machine learning to make individualized predictions in the sense of precision mental health for common and highly relevant therapeutic events.

Study I (Chapter 5) described the generation process of a predictive dropout model using ML and self-report data assessed prior to the first session that can be implemented into clinical practice, also revealing critical factors that increase the risk of dropout. In addition, it reviewed numerous ML algorithms in terms of their predictive performance in a naturalistic psychotherapy setting, identifying those that are appropriate for naturalistic data.

Study II (Chapter 6) then used and extended these findings to generate another dropout model, using data from the first sessions as well as observer ratings regarding therapist competence and adherence (Boyle et al., 2020) to investigate whether these ratings improve the prediction of subsequent dropout.

Study III (Chapter 7) extended the prediction process to therapy outcome. Here, ML was used to examine which factors fostered symptom improvement after a sudden gain (SG). Additionally, a control group that did not experience a SG was used to quantify the effects of the identified outcome predictors after a SG.

Before the three studies are described in more detail, a short theoretical overview is provided from which the specific research questions were derived. After formulating the three main research questions, the basic methodology of the three studies is presented. The main section of this umbrella provides a summary of the studies and is followed by a general discussion that addresses future research directions and clinical applications.

2 Theoretical Background

Psychotherapy as a form of treatment for mental illness first became popular at the beginning of the 20th century in the form of psychoanalysis (Norcross et al., 2011). Today there are several hundred different psychotherapy orientations, which differ more or less from each other with regard to certain aspects (McAleavey & Castonguay, 2015). Nevertheless this umbrella understands psychotherapy as “a class of treatments defined by overlapping techniques, mechanisms, and proposed outcomes” (Castonguay et al., 2013; p. 87). This definition clarifies the focus of this work, as the results pertain to applications for clinical practice independent of the treatment approach. Furthermore, the definition implies the paradigm shift that psychotherapy research has undergone. In the past, the primary focus of research in psychotherapy was to determine the average efficacy and effectiveness of specific approaches. However, the current focus has shifted towards identifying therapeutic strategies and techniques that are effective for specific patients, as well as understanding the factors within routine care that contribute to positive outcomes (Lutz, Castonguay, et al., 2021). This characterizes the shift from a treatment-focused to a patient-focused research paradigm.

2.1 Shifts and Developments in Psychotherapy Research

Eysenck (1952) began the first major psychotherapy research debate by claiming that psychotherapy was no more effective than spontaneous remission. After countless studies, reviews and meta-analyses, this claim can be impressively refuted (Barkham & Lambert, 2021; Carpenter et al., 2018; Munder et al., 2019; Smith & Glass, 1977; Wampold, 2019).

After the debate about the general efficacy of psychotherapeutic treatment, research interest shifted to the relative efficacy of different treatment approaches (Luborsky & Singer, 1975; Lutz, Castonguay, et al., 2021). This led to the emergence of two camps, one trying to prove that their treatment approach is the most effective one and the other assuming that there are no significant differences in effectiveness between different therapeutic approaches (Wampold & Imel, 2015). Through this debate, the latter camp developed the *common factors* model (Wampold, 2001, 2015), which posits that all treatments are influenced by certain shared factors. Although current studies tend to support the *common factors* theory, as that there are no differences in average effectiveness (American Psychological Association, 2019; Elkin et al., 1989; Laska et al., 2014; Nahum et al., 2019), this debate has yet to be fully concluded (Cuijpers et al., 2019; Hofmann & Barlow, 2014; Laska et al., 2014).

The two debates described above can be seen within the context of treatment-focused research, which includes two different study approaches, namely *efficacy* trials and *effectiveness* trials. Both concepts refer to the ability of therapeutic interventions to produce positive outcomes for patients. Efficacy is often defined as the demonstrated impact of a treatment under ideal conditions. Here, randomized controlled trials (RCTs) are considered the gold standard for ensuring internal validity and are often used to establish various treatment guidelines (British Psychological Society, 2010; Härter et al., 2017; Wampold & Imel, 2015). Nevertheless, RCTs have been criticized for their lack of practical relevance (Rothwell, 2005; Shean, 2014) and for expanding the scientist-practitioner gap (e.g., Barkham & Mellor-Clark, 2003; Kazdin, 2016; Lutz, 2002). As a result, the number of naturalistic studies testing the effectiveness of treatments under routine care conditions have increased. This effort to narrow the scientist-practitioner gap and disseminate research findings in the sense of *evidence-based practice*, led to the emergence of so-called empirically supported treatments (Chambless & Ollendick, 2001; Herbert, 2003). Subsequently, various therapy manuals and guidelines for practitioners have been developed and made available to clinicians (e.g., Christophersen & VanScoyoc, 2013; Foa & Rothbaum, 2001; Hayes et al., 2011; Linehan, 1993; Mufson, 2011; Shapiro, 2001; Yalom & Leszcz, 2020). Despite efforts to bridge the gap between research and clinical practice by conducting effectiveness trials and developing therapy manuals, a disconnection between scientists and practitioners remains. This could be because it takes a significant amount of time for treatment-focused research findings to be disseminated to clinicians through guidelines and manuals (Lutz, de Jong, et al., 2021). Furthermore, treatment-focused research typically only focuses on group averages and fails to address individual patient needs (Howard et al., 1996). Therefore, clinicians cannot rely solely on treatment-focused research to provide answers to basic questions, making treatment-focused research inapplicable to individual cases (Barkham & Margison, 2007).

To address these issues and narrow the scientist-practitioner gap, a new research direction, namely POR, was established. This research direction “reflects a bottom-up approach to building and using scientific knowledge” (Castonguay et al., 2021; p. 192). POR is intended to increase clinicians’ integration into research and is “complementary to studies that are conducted in controlled settings” (Castonguay et al., 2021; p. 191), whereby Castonguay et al. (2013) distinguishes three approaches to POR. First, there is *practice-based evidence*, which primarily focuses on factors that foster change in psychotherapy. Second, *practice research networks* (PRNs) that refer to active groups of clinicians that cooperate to collect data and conduct research studies. Third, *patient-focused research* that examines patients’ patterns of change and supports therapists in their decision-making based on scientific methods. The focus of this field of research is the individualization of therapy for the patient at hand (Lutz, de Jong, et al., 2021). The studies summarized in this umbrella were conducted in the context of PFR (i.e., individualization).

2.2 Patient-Focused Research in Psychotherapy

As described above, approximately 60–69% improve in therapy, which means that approximately 31–40% do not benefit from therapy (Barkham & Lambert, 2021). Furthermore, there are indications that about 5–10% deteriorate in therapy (Hansen et al., 2002). These figures demonstrate the great variability in therapeutic success and that treatment-focused research alone is not sufficient to improve therapy benefits. In the context of PFR, Howard et al. (1996) and Lutz (2002) argued for idiographic research to address the shortcomings of treatment-focused research and work to close the scientist-practitioner gap. One goal is to gather data from individual patients prior to and during treatment to support clinical decision-making and predict their progress.

These predictions are derived from extensive data collected from prior studies or routine clinical care that has led to the development of empirically-derived statistical models. These models are made accessible to clinicians and aim to enable psychotherapists to monitor and predict treatment progress and to adjust psychotherapy in order to improve treatment outcomes (Lutz, de Jong, et al., 2021).

This results in a data-driven decision support system for clinicians that considers the individual needs of each patient and follows empirically supported rules (Lutz, Deisenhofer, et al., 2022; e.g., Lutz, Zimmermann, et al., 2017; Norcross & Wampold, 2011). Furthermore, these predictions can help identify patients who are at risk of dropping out or deteriorating during treatment (Lutz et al., 2019; e.g., Lutz et al., 2018). In this regard, this research paradigm provides a sophisticated connection between the nomothetic and idiographic approaches in psychological evaluation and treatment and shares many similarities with current ideas in *precision medicine* or *precision mental health* (Lutz, de Jong, et al., 2021; National Research Council, 2011).

2.3 From Personalized to Precision Mental Health

Personalized mental health is an approach to psychotherapeutic treatment in the sense of patient-focused research. It aims to optimize treatment trajectories and outcomes by personalizing assessment, diagnosis, and treatment strategies based on patients’ unique factors. This approach to therapy replaces the conventional method of searching for a one-size-fits-all treatment. Instead, it focuses on tailoring therapy to meet the specific needs of each individual patient (e.g., Z. D. Cohen et al., 2021; Z. D. Cohen & DeRubeis, 2018; DeRubeis et al., 2014; Lutz, Saunders, et al., 2006).

Via these personalized treatment approaches, researchers have now begun picking up on what clinicians have been doing for years. However, clinicians' decisions were based on intuition rather than empirical evidence (Perlis, 2016). Z. D. Cohen et al. (2021) outlined four sources/levels in a developmental continuum of personalized treatment. The first level, as described above, relies on *clinicians' intuition* rather than theory- or data-driven sources, making it highly error-prone. The second level is *theory-driven*, which is more systematic and mindful of research evidence, but still relies heavily on clinical opinion and indirect evidence (cf. Ong et al., 2020). The third level is characterized by an *evidence-driven* approach, where clinical decision are informed by data and evidence, but without solely relying on it. The use of data serves as a tool or benchmark and is combined with clinician's insight and experience (e.g., Demir et al., 2022; Lutz, Zimmermann, et al., 2017; Schaffrath et al., 2022). The last and most advanced level/source is *data-driven*, where therapy is heavily guided by objective data to inform the treatment plan. The goal of data-driven personalization is to make the treatment process more objective and less reliant on intuition or personal experience. Data-driven approaches are based on statistical algorithms and are thus highly standardized and reproducible.

In recent years, many studies have shown the superiority of data-driven clinical decision making, for example in selecting the treatment intensity (Delgadillo et al., 2022; Delgadillo, Appleby, et al., 2020; Lorenzo-Luaces et al., 2017; Saunders et al., 2020), the treatment package (Deisenhofer et al., 2018; DeRubeis et al., 2014; Huibers et al., 2015; Lutz, Saunders, et al., 2006; Schwartz et al., 2021), and the treatment components or techniques (Lutz, Deisenhofer, et al., 2022). Furthermore, clinical decision-making is often outperformed by data-driven statistical prediction models (Ægisdóttir et al., 2006), especially for less experienced clinicians (Spengler et al., 2009).

In addition to determining outcomes, data-driven clinical decision making can help identify patients who have an increased risk of dropout, so that interventions to increase the likelihood of retention can be initiated (Lutz et al., 2019; Lutz et al., 2020). These studies have typically employed large datasets and ML algorithms, as they are capable of analyzing complex patterns in data. Given the superiority of these data-driven approaches, this final level/source of personalized treatment or personalized mental health is often referred to as precision mental health, as they use extensive data and advanced methodology (Z. S. Chen et al., 2022; Insel, 2014). Precision mental health can thus be allocated to the research tradition of PFR, which aims, among other things, to maximize each patient's likelihood of improvement by implementing research findings directly into clinical practice.

According to Lutz et al. (2019), two branches of precision mental health can be distinguished: prediction and adaptation. Prediction involves constructing data-driven prognostic models to identify patients who are at risk (e.g., deterioration or dropout). Adaptation involves using data collected during treatment to assess a patient's progress in "real-time" and adjust the treatment plan accordingly. As the studies within this umbrella focus on prediction, this stage will be examined in greater detail.

2.4 Prediction

In clinical contexts, prediction or prognostic modeling refers to the process of using data and statistical models to estimate the likelihood of a certain course of treatment or outcome for a patient. This can include, for example, predicting the likelihood of a patient's response to a particular treatment or the risk of dropout. Therefore, prognostic models can be defined "as any models that are constructed in a way that aims to predict an outcome of interest for a specific population" (Z. D. Cohen et al., 2021; p. 690). Usually, adequate prediction of prognostic endpoints (e.g., end of therapy), generally requires multiple prognostic factors (Steyerberg et al., 2013). The PROGnosis RESearch Strategy (PROGRESS) group defined a prognostic factor as "any measure that, among people with a given health condition (that is, a start point), is associated with a subsequent clinical outcome (an endpoint)" (Hemingway et al., 2013). Therefore, prognostic factors help inform clinical and therapeutic decisions either directly or as part of prognostic models for individualized risk or outcome prediction (Riley et al., 2013).

Such prognostic models have been able to predict treatment outcome in obsessive-compulsive disorders (Hilbert et al., 2021), psychotic disorders (Koutsouleris et al., 2018), depressive disorders (Lee et al., 2018) and in disorder heterogeneous samples (Bone et al., 2021; Hilbert et al., 2020). Further, Kessler et al. (2016) were able to predict the persistence and severity of depressive disorders. Also, prediction models have been developed for internet therapy (Pearson et al., 2019) and long-term effects on depression treatment (van Bronswijk et al., 2019). However, the literature on predictors of psychotherapy outcome is diverse due to differences in the data used for the models. Replicated predictors include initial impairment, disorder chronicity, treatment expectation, and number of prior treatments (Constantino et al., 2018; Lutz et al., 2007; Lutz et al., 1999; Lutz et al., 2019). Besides treatment outcome, early response during the first sessions of therapy and the probability of treatment dropout have also been predicted (Lutz, Arndt, et al., 2017; Lutz et al., 2014; Lutz et al., 2019). Reliable prediction models can aid therapists in their treatment planning, allowing them to effectively identify patients at risk and allocate resources more effectively.

In order to generate such prognostic models, sophisticated procedures are necessary that derive the important information from the data. Steyerberg et al. (2013) formulate three points for the generation of reliable prognostic models for clinical practice. First, the dataset used must be of high quality, meaning that it contains much representative data (e.g., naturalistic patient data from routine care) and reliable predictors that have an association with the target variable of interest. Second, there must be a sound statistical analysis plan that captures the interactions between all predictors and their influence on the target variable. Third, the model must be validated in independent datasets to prevent overfitting and to evaluate generalizability. These three points will be further described below.

2.4.1 Data Requirements

For the prediction of complex therapeutic phenomena, large amounts of data are needed, which has been referred to as big data (N > 1000; Delgadillo, 2021). Despite existing effects in psychological studies, these are often not detected due to statistical underpowering (Stanley et al., 2018). Therefore, sample size calculation can help improve the detection of existing effects and improve the generalizability of results (Archer et al., 2021; Riley et al., 2019). Further, a large amount of data is needed to apply complex statistical methods (i.e., machine learning) that are able to assess complex associations and high-level interactions between predictors.

With large collaborative projects in routine care, the assessment of big data in psychotherapy research is becoming increasingly feasible. For instance, in the UK, data is regularly gathered from 200 outpatient psychotherapy services as part of the Improving Access to Psychological Therapies (IAPT) program (Clark, 2018; Wakefield et al., 2021), while in Germany, a significant project with the health insurance company Techniker Krankenkasse (TK) has collected data from private practice psychotherapists (Strauss et al., 2015). Additionally, university clinics have collaborated to collect their data in a project that coordinates data collection and analysis at research and training outpatient clinics for psychotherapy (KODAP; Velten et al., 2017). In addition to assessing more patients to increase the size of data sets, more data can also be obtained by measuring individual patients more frequently, for example, before and after each session (Lutz et al., 2019). This improves data resolution, possibly revealing effects that would otherwise remain hidden.

In addition to data quantity, data quality is equally important for prediction purposes. A key consideration is whether the data set, regardless of its size, contains appropriate predictors that are associated with the target variable (e.g., the outcome of a treatment). For example, psychological disorders are not homogeneous, which makes it difficult to capture them using unidimensional (e.g., PHQ-9) instruments (Lutz, de Jong, et al., 2021). Nevertheless such scales are often used in research despite the fact that unidimensional measures with the same label do no not necessarily measure the same construct (Böhnke & Croudace, 2016; Fried, 2017). Therefore, new approaches like modern item response theory address this problem by enabling the development of common metrics (see Lutz, de Jong, et al., 2021; Wahl et al., 2014). Further, to improve data quality for prediction, multidimensional measurements (e.g., OQ-45) are more suitable, because of their ability to cover a wider range of problems, as is common in clinical studies (Hill & Lambert, 2004).

Additionally, all moderating predictors should be assessed, if possible. For example, patients from different ethnic backgrounds may understand and answer items differently (e.g., Wetzel & Böhnke, 2017), making it clear that predictors must be included, which moderate other predictive variables. Moreover, different assessment sources should be used (e.g. self-report measurements and observer ratings), as they do not necessarily overlap in their findings, even when they target the same construct (Stepankova Georgi et al., 2019).

In summary, for a reliable prediction model, it is essential to have sufficient and appropriate data, to include as many relevant predictors as possible, and to consider different sources of data assessment.

2.4.2 Statistical and Analytical Requirements

In order to analyze such big data, new methodological approaches are needed that can take the size and complexity of the data into account. One important development is the integration of ML algorithms into psychotherapy research (Chekroud et al., 2021; Delgadillo, 2021; Lee et al., 2018). ML refers to a collection of modeling techniques that have the abilityto automatically learn from data, while also being able to effectively capture complex, non-linear relationships and high-level interactions (Brownlee, 2016).

The advantages of ML over classic regression in mental health prediction have often been demonstrated (Kessler et al., 2016; Rosellini et al., 2018; Symons et al., 2020; Wardenaar et al., 2014; Webb et al., 2020). However, it can be challenging to select a specific modeling approach, because of the abundance of recommendations (Adibi et al., 2020). For this reason, it is useful to build several prediction models with different approaches and compare them using cross-validation, a holdout test sample or bootstrapping (e.g., Delgadillo, Rubel, & Barkham, 2020; Hilbert et al., 2020; Lutz et al., 2019; Webb et al., 2020). Another approach to face this problem is the development of ensembles (i.e., pooling the information provided by different algorithms) to attain strong predictions (Kessler et al., 2019).

In order to make ML algorithms work properly, adequate data pre-processing is necessary. Although ML is able to handle different scales of predictors, skewed distributions etc., appropriate data preparation can improve predictions. Normally, pre-processing includes transformation, reduction, imputation, variable selection, and balancing of data (Z. D. Cohen et al., 2021; Delgadillo, 2021).

While transformation may be suboptimal for models that aim to discover complex or nonlinear patterns (e.g., decision trees), it can be helpful in certain contexts, depending on the data (Brownlee, 2016). The process of reduction includes making choices regarding the consolidation of several infrequent categories into a smaller number of more prevalent categories, with the aim of reducing the impact of biased results stemming from small sample sizes. Imputation is often used when a significant proportion of data is missing at random. However, if the missing data is systematic, an alternative method is to consider the missing data as a predictor. Variable selection is an important step and the basis of the modeling process. Ideally, only predictive variables are in the dataset, eliminating those that have no benefit for the prediction of the target variable. Because of varying recommendations and the difficulty to identify these predictors before the actual modeling process, Sauerbrei et al. (2020) published an extensive review on variable selection, outlining the strengths and weaknesses of different approaches. Last, balancing refers to the process of adjusting a dataset in cases where there is a class imbalance, such as when predicting an event with a low occurrence rate (e.g., dropout). This imbalance can negatively impact prediction accuracy (Japkowicz & Stephen, 2002). Various techniques can be employed to address this issue, including down-sampling, up-sampling, and synthetic sampling methods (Chawla et al., 2002).

Even with appropriate pre-processing, ML does not always have an advantage over more traditional methods (Bone et al., 2021; Christodoulou et al., 2019; Makridakis et al., 2018; Rubel et al., 2020). Wilkinson et al. (2020) and Liu et al. (2019) argue that personalized medical care faces serious challenges that cannot be addressed by algorithmic complexity alone and that many studies have evaluated ML models under unrealistic conditions with little relevance to routine clinical practice. Kessler et al. (2019) therefore proposed that large observational datasets rather than randomized controlled trials are needed to derive rules for routine clinical practice. In contrast to Wilkinson et al. (2020), the authors propose that ML methods, especially ML ensembles are very suited to this task.

Current findings have demonstrated that machine learning (ML) has significant potential for enhancing predictive accuracy. However, the circumstances under which this approach is most effective require further exploration. Additionally, previous studies have primarily focused on predicting outcomes, with less attention devoted to examining other therapeutic events like dropout. It therefore remains unclear whether such binary events can be predicted effectively.

2.4.3 Requirements for Generalizability and Clinical Utility

In order for a predictive model to be utilized in clinical practice, it is necessary to test its generalizability (Steyerberg et al., 2013). The quality of the model must be assessed using performance parameters, and it is recommended to use multiple parameters to obtain a thorough evaluation (Delgadillo, 2021; Handelman et al., 2019). Additionally, it is crucial to ensure that the model is not overfitted. Overfitting is a problem that can occur when a model becomes too complex and starts to fit the training data too closely, to the point where it begins to capture noise and random fluctuations in the data rather than the underlying patterns (Rudin & Carlson, 2019). This means that the model performs very well on the training data, but when it is used to make predictions on new, unseen data, its performance is often poor. Cross-validation (CV) is an effective tool to avoid overfitting, in which the data set is split into a training and a test set several times (Cawley & Talbot, 2010). While the training sets are used to generate the model, the test sets are used to evaluate them in order to reduce double dipping (i.e., performing variable selection and model fitting in the same sample, Fiedler, 2011; Vul et al., 2009).

According to Delgadillo (2021), there are various levels of CV quality that determine clinical utility. On the first level, an internal CV is performed on the entire dataset. On the second level, the final test set is completely independent of the modeling process. On the third level, the test sample is obtained from a different study or context (i.e., a holdout sample), while on the final level, the model is prospectively tested on a new sample. Good protection against overfitting is necessary to ensure and evaluate the practicability of a model.

Next, based on the theoretical background described above, the dissertation’s main research questions are described. In the following, the associated basic methodological features are introduced. Subsequently, the three research studies that address the presented research questions are summarized and discussed.

3 Research Questions

As described above, to date there are comparatively few prediction models addressing dropout. Furthermore, there are various recommendations regarding the use of the huge number of different algorithms and ensembles. Therefore, under consideration of past research and recommendations on prediction and personalization in naturalistic contexts, Study I pursued the following questions:

3.1 Study I

1. Do machine learning algorithms and ensembles generate prediction models that have comparable abilities to predict dropout in a naturalistic CBT setting, or are there significant differences between them?
2. Can the most effective prediction model validly predict dropout prior to the first session, and to what extent does it outperform a generalized linear model?

As specified in the theoretical framework, incorporating data from different assessment sources can be useful for prediction modeling. Thus, Study II intended to leverage the findings of Study I to address the subsequent research question:

3.2 Study II

1. Can the inclusion of observer video ratings of therapist interventions and skills from early therapy sessions improve the prediction of dropout beyond intake assessments?

In Study III, prediction modeling shifted to another context. Machine learning and findings from Study I were used to investigate the processes after a sudden gain and identify which factors foster a better outcome. Therefore, the following research questions were pivotal:

3.3 Study III

1. Can previous results regarding outcome predictors after a sudden gain be replicated with a large sample and machine learning algorithms?
2. What factors influence therapy outcome after a sudden gain and how large are the effects compared with a control group?

4 Methodological Aspects

All three studies outlined in this dissertation apply machine learning (ML) algorithms to develop predictive models in order to investigate the research questions. Consequently, this umbrella will initially provide a methodical overview of machine learning in general, before delving into the specifics of individual algorithms. Given the extensive use of algorithms in Study I, a comprehensive description of all of them is beyond the scope of this umbrella. Hence, only the most pertinent ones will be discussed in detail below.

4.1 Machine Learning

Machine learning is a subfield of artificial intelligence that involves using algorithms and statistical models to allow computer systems to automatically learn and improve from experience without being explicitly programmed (Alpaydin, 2010). The goal of machine learning is to enable computers to recognize patterns, make predictions, and make decisions based on data (Goodfellow et al., 2016). Here, two types of ML can be distinguished: supervised and unsupervised learning. Unsupervised machine learning methods reveal new relationships in data, while supervised methods learn established relationships from known data.

While inferential statistics focus on testing hypotheses, ML algorithms explore and model unknown relationships in a more open-ended way, without strict assumptions. Rather than confirming specific relationships between predictor X and criterion Y, their goal is to construct the most accurate prediction model possible (Yarkoni & Westfall, 2017). Consequently, evaluating ML models emphasizes overall predictive accuracy, rather than the significance of individual predictors. The ideal approach to assess accuracy is to test the model on new data, on which it has not been trained, i.e., validation or test data. Such an approach enables the model’s accuracy to be estimated without overfitting to the original training or development data (Delgadillo, 2021).

Overall, ML offers an effective method for reducing dimensionality by identifying inherent data patterns and selecting pertinent features. These abilities will be used to build a dropout model with pre-treatment data, comparing the usefulness of different algorithms and ensembles (Study I). These findings will then be used to generate another dropout model with additional data from observer ratings to see if such data can improve dropout prediction (Study II). Further, ML will be used to identify predictors that foster the upward spiral after a sudden gain by generating an outcome prediction model with data after a sudden gain, which is also compared to a control group (Study III)

4.1.1 Random Forest

Random forest (RF) is a machine learning algorithm that builds an ensemble of decision trees to improve the predictive accuracy and generalizability of the model (Breiman, 2001). The algorithm randomly selects a subset of features and data points to build each tree in the forest, and then aggregates the predictions of all trees to make a final prediction. The strength of RF is that it can reduce overfitting and improve generalization performance by building multiple trees on random subsets of the data, each biased towards a different subset of features and data points, and combine the results by averaging or majority voting (Breiman, 2001).

RF has been shown to improve variable selection in different settings by reducing under- and overfitting (Altman & Krzywinski, 2017; W. Chen et al., 2018; Menze et al., 2009; Ok et al., 2012). Because of these advantages, a RF algorithm was used to generate a prediction model in Study I. Further, RF algorithms were used to impute missing data in all studies, as simulation studies have shown that they perform well up for up to 30% missing values (Stekhoven & Bühlmann, 2012).

4.1.2 Elastic Net

The elastic net (EN) algorithm is a regularized linear regression model that combines both the least absolute shrinkage and selection operator (LASSO) method and the ridge regression method. The LASSO method adds a penalty term equal to the absolute value of the coefficients, which can lead to sparsity in the solution, i.e., some coefficients become zero. The ridge regression method does the same with the difference that the penalty term is equal to the squared value of the coefficients and is not able to fully shrink the coefficients to zero. The EN algorithm combines both LASSO and ridge regularization methods by adding a penalty term that is a linear combination of both penalty terms. This allows the algorithm to have both the advantages of LASSO (sparsity) and Ridge (stability), and to select the most important features, while keeping correlated features together (Zou & Hastie, 2005). Friedman et al. (2010) state that the EN algorithm is a versatile method that can be used for both regression and classification tasks, and can handle a variety of data and a wide range of data types, including continuous, binary, and categorical variables. They also note that regularization is particularly important in high-dimensional data to avoid overfitting and to improve generalization performance on new data. Further, Hastie et al. (2015) discuss that the EN algorithm is often used in ML and data analysis applications where regularization and feature selection are important.

Especially because of the ability to select crucial predictors and to exclude redundant variables, EN was used for variable selection in all the three studies. Also, EN has often been applied in psychotherapeutic research contexts and seems to handle clinical data well (e.g., Fisher & Bosley, 2020; Lutz et al., 2019; Webb et al., 2020).

4.1.3 Nearest Neighbor

The nearest neighbor (NN) algorithm works by finding the closest training data points in the data set to a new data point and then predicting the value of the new data point based on the values of its NNs (Hastie et al., 2009). In the case of classification tasks, the algorithm selects the label of the most frequent class among the k-NNs, where k is a hyperparameter specified by the user that defines the amount of NNs for classification (Alpaydin, 2010). In the case of regression tasks, the algorithm predicts the value of the new data point based on the average of the values of its k-NNs. For example, if a new data point has three neighbors with values 2, 4, and 3, the algorithm will predict the value of the new data point to be 3, since 3 is the average value of its three NNs (Müller & Guido, 2016). The NN algorithm is easy to implement and can work well on small datasets, but can become computationally expensive for larger datasets as it needs to calculate the distances between the new data point and all the training data points (Kotsiantis et al., 2007). The NN technique has found application in varying contexts (Cai, 2022; Jabbar et al., 2013; Laios et al., 2020; Lutz, Lambert, et al., 2006; Lutz et al., 2005; Lutz, Zimmermann, et al., 2017; Rubel et al., 2020; Xiao et al., 2010). The NN algorithm was used for prediction modeling in Study I and for propensity score matching in Study III to generate a comparable control group.

4.1.4 Ensembles

In ML, an ensemble is a technique used to improve the accuracy and robustness of a model by combining the predictions of multiple models. Ensemble methods are often used when a single model is unable to capture the complexity of the data or when multiple models can provide complementary information. Brownlee (2021) distinguishes three different ensemble techniques that are most commonly used in practice, namely bagging, boosting, and stacking.

Bagging involves training multiple models on different bootstrap samples of the data and combining their predictions. This approach can help to reduce overfitting and improve generalization (Zhou, 2012).

Boosting involves training a sequence of models that are optimized to correct the errors of the previous models. The objective is to convert many weak models to a strong one (Zhou, 2012). This approach can help to improve the accuracy of the final model, but it can also be more computationally expensive than other ensemble methods.

Stacking is an ensemble technique in machine learning, where multiple base models are trained and their predictions are combined to train a meta-model that aims to make the best predictions (Rokach, 2010). Initially, multiple base models are trained on the training dataset, with each algorithm learning different aspects of the dataset. Then the base models' predictions are used as input for the meta-model, which produces the final predictions. The meta-model can be any algorithm trained on the base models' predictions. The idea behind stacking is that the meta-model can leverage the strengths of the base models and compensate for their weaknesses to make better predictions (Burkov, 2019; Zhou, 2012).

In Study I, different ensemble learners as well as manually generated ensembles were used and compared regarding the generation of a dropout model. Because large datasets are required (Dietterich, 2000), no ensembles were used in the other studies.

4.1.5 Optimization of Algorithm Parameters

When using ML algorithms, the parameter settings are of crucial importance. For example, with the elastic net algorithm, it is important how large the penalty should be and with which ratio the LASSO portion should be taken into account. These parameters are usually labeled with a λ and α, respectively. Such parameters exist for all algorithms and are crucial, as they can significantly impact a model's performance (Géron, 2019).

In order to avoid poor model performance due to inappropriate parameter settings, different settings were tested during the model generation process with the help of the R-packages caret (v6.0-84; Kuhn, 2019) and caretEnsemble (v2.0.0; Deane-Mayer & Knowles, 2016). Especially in Study I, in which many algorithms were compared, caret and caretEnsemble were used to make a fair comparison between the algorithms and to find the optimal model. In addition, caret was used in the other two studies to find the optimal parameters for model generation.

5 Study I: Predicting Patients who Will Drop out of Outpatient Psychotherapy Using Machine Learning Algorithms

Bennemann, B., Schwartz, B., Giesemann, J., & Lutz, W. (2022). Predicting patients who will drop out of out-patient psychotherapy using machine learning algorithms. The British Journal of Psychiatry, 220(4), 192–201. https://doi.org/10.1192/bjp.2022.17

5.1 Introduction

Study I focused on dropout prediction. Approximately one in five patients drop out of treatment (Swift et al., 2017), leading to various negative consequences for the patient and society (Barrett et al., 2008; Björk et al., 2009; Delgadillo et al., 2014; Wells et al., 2013). Therefore, identifying potential dropout patients could lead to the development of clinical support tools that minimize the risk of dropout (Lutz, Rubel, et al., 2015; Lutz et al., 2019). However, findings from past studies have been heterogeneous, with only younger age and lower education level being consistently associated with dropout (Swift & Greenberg, 2012). Thus, dropout is a complex phenomenon needing advanced methodology to make reliable predictions (e.g., Lutz et al., 2019).

In this context, ML algorithms offer the possibility of making reliable predictions due to their ability to capture complex relationships (Brownlee, 2016; Z. D. Cohen et al., 2021; Lutz, Deisenhofer, et al., 2022). However, ML is not always advantageous over traditional methods (Christodoulou et al., 2019; Rubel et al., 2020; Wilkinson et al., 2020). Further, predictions using ML to address binary events like dropout are rare. For this purpose, this study aimed, besides predicting dropout, to investigate the use of various ML algorithms for the prediction of a binary event in a naturalistic setting.

5.2 Methods

5.2.1 Patients and Treatment

The study used a sample of 2543 patients treated at the University of Trier outpatient CBT Clinic between 2007 and 2021. Patients were included if they had completed a battery of questionnaires at intake, had begun therapy after the diagnostic phase, and completed or dropped out of treatment. The patient data until 2017 was used for model training purposes, the remaining data for testing purposes (i.e., holdout sample). Diagnoses were based on the German version of the structured clinical interview for axis I DSM-IV Disorders (SCID-I;

Wittchen et al., 1997) and the International Diagnostic Checklist for Personality Disorders (IDCL-P; Bronisch et al., 1996). The mean scores of the Outcome Questionnaire (OQ; Ellsworth et al., 2006) and the Brief Symptom Inventory (BSI; Franke, 2000) indicated a moderate to severe general level of distress in all samples. Patients were treated by 220 therapists (79.5% female).

5.2.2 Measures

The study assessed dropout in therapy and defined it as the patient stopping therapy despite the therapist's recommendation to continue. Further, 77 routinely collected self-report intake variables that were assessed before the first session were analyzed (e.g., BSI scales, demographic variables).

5.2.3 Selection of ML Algorithms

Only algorithms that had already been used in relevant literature were applied for the analyses. Specifically, algorithm selection was based on the Stratified Medicine Approaches for Treatment Selection (SMART) mental health prediction tournament at the 2019 Treatment Selection Idea Lab (TSIL) conference (Z. D. Cohen et al., 2018). Using this information, as well as an examination of the literature provided by the tournament organizer, 21 algorithms were examined more closely. Both, linear and non-linear machine learning algorithms were considered ML, as suggested by Brownlee (2016).

5.2.4 Data Analytic Strategy

Data with more than 10% missing values were excluded, while the missing values for other variables were imputed using a trained random forest for the training and holdout sample separately (Stekhoven & Bühlmann, 2012). For all modeling processes, the R-packages caret (v6.0-84; Kuhn, 2019) and caretEnsemble (v2.0.0; Deane-Mayer & Knowles, 2016) were used to train and tune the hyperparameters of all algorithms and ensembles, selecting the best model based on the area under the curve (AUC).

First, all 21 individual algorithms were ranked based on their Brier score and AUC. Further, the correlations of the predictions of all algorithms during the model-building process were compared. For model generation, a nested CV approach with 20 outer and 10 inner loops was used in the training sample, while the synthetic minority oversampling technique (SMOTE; Chawla et al., 2002) was applied to address the problem of class imbalance. For data pre-processing, all categorical variables were dichotomized and all continuous variables were centered with the mean of the training sample for each outer CV loop.

After the ranking and correlation comparison, algorithm ensembles were generated using five different types of ensembles (two best, three best, two least correlated, three least correlated, best algorithm with least correlating one). These ensembles were generated via stacking, either with a generalized linear model (GLM) or the best algorithm, leading to 10 ensembles.

This entire procedure was repeated twice, first using only the significant predictors from a previous study (Zimmermann et al., 2017). Second, a preceding EN regularization analysis was conducted for each training part of the outer loops inside the nested CV framework. Therefore, only those variables that had predictive power for model generation in EN regularization were used. In summary, 30 ensembles (10 ensembles x 3 procedures) were generated. All ensembles as well as the five best single algorithms from each procedure (i.e., 15 single algorithms overall) were then compared using a nested CV with 20 outer loops and 10 inner loops including 3 repetitions with the data pre-processing described above. Evaluation parameters were the mean Brier score and mean AUC across all CVs.

Further, *t*-tests between the best and worst model as well as between the best model and a GLM were conducted for each parameter, using the distributions obtained by the values of the 20 outer loops for each of both parameters to assess the robustness of the models and to prevent sampling artifacts. The best ensemble/algorithm was then tested in the still unused and independent holdout sample to assess the generalizability of the model, to prevent overfitting, and to identify the most relevant predictors of dropout.

5.3 Results

The best algorithm for predicting outcomes was generalized boosted regression modeling when using all variables, RF when using selected predictors, and adapted boosted classification trees when using only significant predictors. The ensemble of the best and least correlated algorithms (i.e., RF and NN) stacked with a GLM including a preceding EN regularization generated the best model (see Table 1 for all models).

Table 1

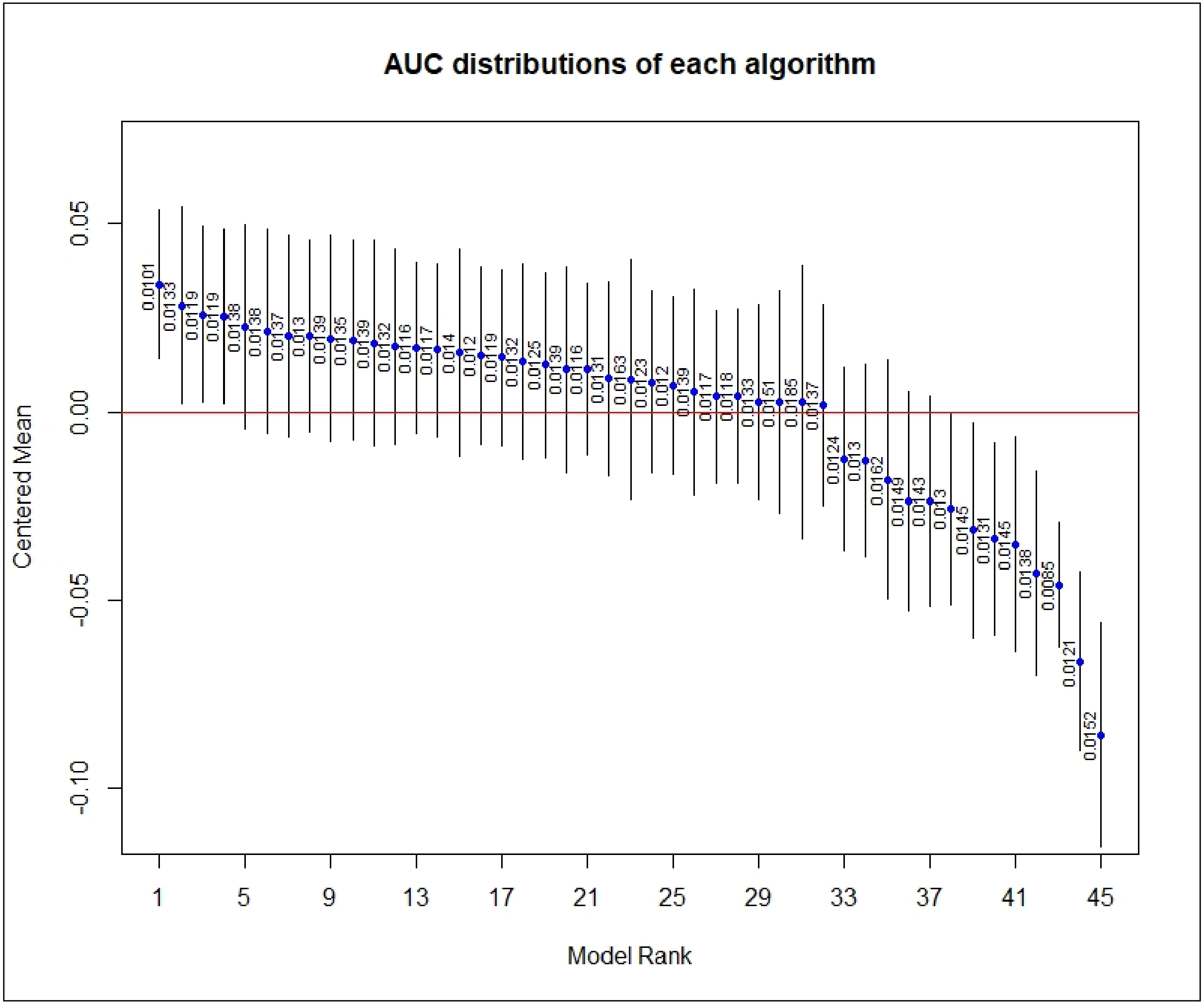
Mean scores of the models generated by all 45 algorithms and ensembles.

Algorithm/Ensemble	Stacking Method	Variables used	Brier score	AUC	Training AUC
Best with lowest correlation	GLM	Selected with EN	.1983	.6581	.6617
Two best	GLM	Selected with EN	.1983	.6577	.6674
Three best	GLM	Selected with EN	.1985	.6535	.6673
Two best	GBM	All	.1989	.6550	.6515
Three best	GLM	All	.1994	.6513	.6549
Best with lowest correlation	GLM	All	.1992	.6497	.6492
Two best	GLM	All	.1995	.6518	.6530
Three best	GBM	All	.1998	.6523	.6557
GBM	-	Selected with EN	.2022	.6661	.6608
Two best	GLM	Manually selected	.1995	.6493	.6464
RF	-	Selected with EN	.2041	.6605	.6602
Best with lowest correlation	GLM	Manually selected	.1997	.6488	.6430
Best with lowest correlation	GBM	All	.2004	.6506	.6483
Three best	GLM	Manually selected	.1998	.6468	.6461
Three least correlating	GLM	All	.2010	.6435	.6494
Three least correlating	GBM	All	.2006	.6412	.6485
Three best	ADA	Manually selected	.2011	.6435	.6488
ADA	-	All	.2071	.6525	.6485
GBM	-	All	.2055	.6457	.6497
Best with lowest correlation	ADA	Manually selected	.2017	.6403	.6424
XGB	-	Selected with EN	.2069	.6480	.6584
Two best	ADA	Manually selected	.2014	.6349	.6482
RF	-	All	.2058	.6392	.6428
ADA	-	Selected with EN	.2099	.6475	.6591
XGB	-	All	.2075	.6448	.6451
Two least correlating	GLM	All	.2049	.6197	.6129
Three least correlating	GLM	Manually selected	.2053	.6193	.6095
Two least correlating	GLM	Manually selected	.2056	.6143	.6060
GBM	-	Manually selected	.2208	.6525	.6369
GLMBOOST	-	Selected with EN	.2309	.6408	.6516
Three least correlating	ADA	Manually selected	.2059	.6066	.6150
Two least correlating	ADA	Manually selected	.2064	.6087	.6092
Two least correlating	GBM	All	.2060	.6010	.6121
ADA	-	Manually selected	.2240	.6349	.6440
GLMBOOST	-	Manually selected	.2342	.6364	.6379
GLMBOOST	-	All	.2306	.6349	.6487
Two least correlating	GLM	Selected with EN	.2064	.5971	.5872
LDA	-	Manually selected	.2347	.6364	.6377
Three least correlating	GLM	Selected with EN	.2074	.5986	.6058
Three best	RF	Selected with EN	.2180	.6085	.6143
GLMAIC	-	Manually Selected	.2342	.6342	.6392
Two best	RF	Selected with EN	.2376	.5893	.5902
Three least correlating	RF	Selected with EN	.2490	.5661	.5607
Best with lowest correlation	RF	Selected with EN	.2586	.5864	.5838
Two least correlating	RF	Selected with EN	.2859	.5465	.5489

Note: EN = Elastic net; GLM = Generalized linear model; RF = Random forest; GLMBOOST = Boosted generalized linear model; ADA = Adapted boosted classification trees; GLMAIC = Generalized linear model with stepwise feature selection using Akaike information criterion; XGB = Extreme gradient boosting; LDA = Linear discriminant analysis; GBM = Generalized boosted regression model; All ensembles and algorithms are ranked.

The best model was able to identify 63.4% of holdout cases correctly before the first session occurred, while a GLM was able to identify 46.2% correctly. Lower education level, younger age, lower scores on the compulsive scale of the personality style and disorder inventory (PSSI), higher scores on the negativistic and antisocial scale of the PSSI, and higher scores on the BSI overall score as well as on the additional scale of the BSI (i.e., mean of the four additional items) were the main predictors of dropout (i.e., relative importance > 90%). A paired one-sided *t*-test revealed a highly significant effect between the overall best and overall worst models concerning the AUC score (AUC_{best} = 0.6581; AUC_{worst} = 0.5465; *t*(19) = 8.30, *p* < .001, Cohen's *d* = 1.86 [0.11; 2.58]). Comparing the overall best model with the GLM using all variables, the effect was still significant (AUC_{best} = 0.6581; AUC_{GLM} = 0.6253; *t*(19) = 2.63, *p* < .01, Cohen's *d* = 0.59 [0.11; 1.06]). The differences stayed significant when comparing the Brier score distributions. In summary, the distributions of each algorithm/ensemble revealed that the best ones hardly differed from one another (see Figure 1). Nevertheless, some models seemed to make significantly worse predictions. For the Brier score, the pattern was very similar.

Figure 1
Distribution of the 20 outer cross-validation (CV) models generated by each algorithm and ensemble ranked from best to worst using the area under the curve (AUC).



Note: Each value was grand-mean centered; the horizontal line represents the total average of all models. The numbers on the graphs are the standard deviations.

5.4 Conclusions

Study I generated a model that correctly identified 63.4% of all cases in an independent holdout sample, which was of high clinical value compared to the identification percentage of 46.2% by the GLM. This superiority was further supported by the significant differences between the distributions (Cohen's *d* = 0.59 [0.11; 1.06]). Although 63.4% does not seem very precise at first, it must be acknowledged that this prediction was made before the first session had occurred and therefore helps to identify patients who may drop out after the first session. Knowing the important predictors (e.g., lower educational level) for the generation process, clinicians could use the information to generate a more precise case concept for the individual patient before the first session to help reduce the risk of dropout. It is important for clinicians to pay attention to the complementarity of the relationship in order to establish a good alliance, which is in line with previous research (Lutz et al., 2019; Zimmermann et al., 2017). With regard to the BSI, it seems reasonable to first treat symptoms such as sleep problems, loss of appetite, suicidal thoughts and feelings of guilt, which cause high levels of distress. Lower education and younger age also seem to increase the probability of dropout, a finding also identified in other studies (Lutz et al., 2019; Swift & Greenberg, 2012; Zimmermann et al., 2017). One possibility is that such patients need more time to understand how therapy could help them concretely. Therefore, this should be part of the treatment strategy, especially in the first sessions. Nevertheless, to ensure that Study I's final model can be applied in different settings, it is important to test its efficacy in various contexts to further evaluate its generalizability, preferably using prospective data (Delgadillo, 2021).

Concerning Study I's modeling strategy, there are numerous other possibilities to build models using ML algorithms. However, this study still extensively reviewed the utility of various algorithms and ensembles. No other study has compared different algorithms with such a large dataset. Study I thus contributes to the growing body of research on precision mental health and the closely related use of ML algorithms in psychotherapy research (Z. D. Cohen et al., 2021). It indicates that it is possible to predict potential dropout patients prior to treatment using ML approaches, but also that some advanced ensembles perform even worse than a GLM. Tree-based and boosted algorithms that include variable selection performed better compared to more advanced algorithms like neural networks. Ensembles with weakly correlated algorithms or many algorithms tended to underperform, unless a strong algorithm with good predictive power was included, which is consistent with prior recommendations (Mayer, 2019; Ramzai, 2019). Thus, effective use of machine learning entails analyzing the data structure beforehand to identify how to extract key information and enhance predictions.

6 Study II: Therapist Interventions and Skills as Predictors of Dropout in Outpatient Psychotherapy

Poster, K., Bennemann, B., Hofmann, S. G., & Lutz, W. (2021). Therapist Interventions and Skills as Predictors of Dropout in Outpatient Psychotherapy. *Behavior Therapy*, 52(6), 1489–1501. https://doi.org/10.1016/j.beth.2021.05.001

6.1 Introduction

This study also focused on generating a dropout model, this time by examining whether video ratings can improve dropout predictions alongside self-report instruments from patients and therapists.

As described in Study I, the issue of dropout in psychological interventions is a significant concern, with one in five patients terminating treatments prematurely (Swift et al., 2017). Besides patient factors, therapist variables also contribute to the likelihood of patients terminating treatment prematurely. Between 6.2% and 12.9% of dropout variance can be attributed to individual therapists, with therapists' individual dropout rates ranging between 1.2–73.2% (Saxon et al., 2017; Zimmermann et al., 2017). Professional background and experience level have been associated with dropout, but findings remain inconsistent (Hamilton et al., 2011; Swift & Greenberg, 2012). However, a positive therapeutic alliance with therapist behaviors such as empathy, positive regard, and collaboration is associated with treatment completion (Crits-Christoph et al., 2006; Sharf et al., 2010). Treatment factors such as time-limitation, manualization, and treatment setting also have been associated with dropout, but not treatment orientation or format (Swift & Greenberg, 2012).

These findings illustrate that dropout cannot be solely attributed to patient characteristics. For this reason, the aim of Study II was to generate another dropout model using the findings from Study I and to add therapist and patient variables from the first sessions as well as therapist interventions and skills assessed via the Inventory of Therapeutic Interventions and Skills (ITIS; Boyle et al., 2020). Because data from different assessment sources are normally more reliable (Lutz, de Jong, et al., 2021), it was hypothesized that a model with observer ratings from the ITIS predicts dropout better than a model without them.

6.2 Methods

6.2.1 Patients and Treatment

Therapies were conducted at the University of Trier outpatient CBT clinic from 2017 to 2019. A total of 259 patients (61.8% female, average age 36.3 years) were diagnosed by independent clinicians using the German version of the SCID-I. Process and outcome data were routinely collected and therapy sessions were videotaped with consent from patients and therapists. Treatment length averaged 24.9 sessions and 69.1% of patients completed treatment. Dropout patients had significantly shorter treatment lengths. Treatments were conducted by 65 therapists (83.1% female).

6.2.2 Measures

Dropout was defined as in Study I. A total of 95 variables measured at intake or after the first session were examined as potential predictors of dropout. Besides the 77 variables from Study I, 18 clinician-rated variables such as the alliance and expected improvement were assessed. Additionally, 35 items from the ITIS (Boyle et al., 2020), observer-based video ratings assessing therapist adherence and competence, were used for the modeling process.

6.2.3 Video Selection and Rating Procedure

One early therapy session per case was rated using the ITIS (*N*=263). Typically, the third session was rated, but if the video was unavailable, had poor quality, or was too long, the next available session was used. On average, session 3.2 was rated by a total of eight extensively trained raters, who received regular supervision and had good inter-rater reliability for the skills items and excellent reliability for the interventions items of the ITIS. Raters were blind to diagnoses, termination status, and outcome.

6.2.4 Data Analytic Strategy

First, missing values were imputed as in Study I and variables were screened as possible predictors via significant bivariate correlations to achieve a more favorable balance between the number of data points and variables (Hastie et al., 2015). After that, EN was used to generate dropout models, because of its ability to handle clinical data and prevent overfitting, which was also confirmed in Study I (Fisher & Bosley, 2020; Lutz et al., 2019; Pavlou et al., 2016; Webb et al., 2020). Additionally, EN was chosen, because it is a linear algorithm and therefore less demanding regarding data quantity (Domingos, 2012). In order to find the best hyperparameters, 100 overall combinations of α and λ were defined to find the best fitting parameters using the AUC as the model evaluation parameter.

The dataset was split into a training (70%) and test set (30%) to prevent overfitting (Rudin & Carlson, 2019), using a repeated CV in the training data with 10 folds and three repetitions. Up-sampling was used to minimize the impact of class imbalance (Hand & Vinciotti, 2003; Japkowicz & Stephen, 2002). The training-test split and model generation were repeated 100 times with varying training and test sets to minimize the influence of one training set's specific sample characteristics. After each split, all continuous variables were centered and all categorical variables were dichotomized for training and test set separately.

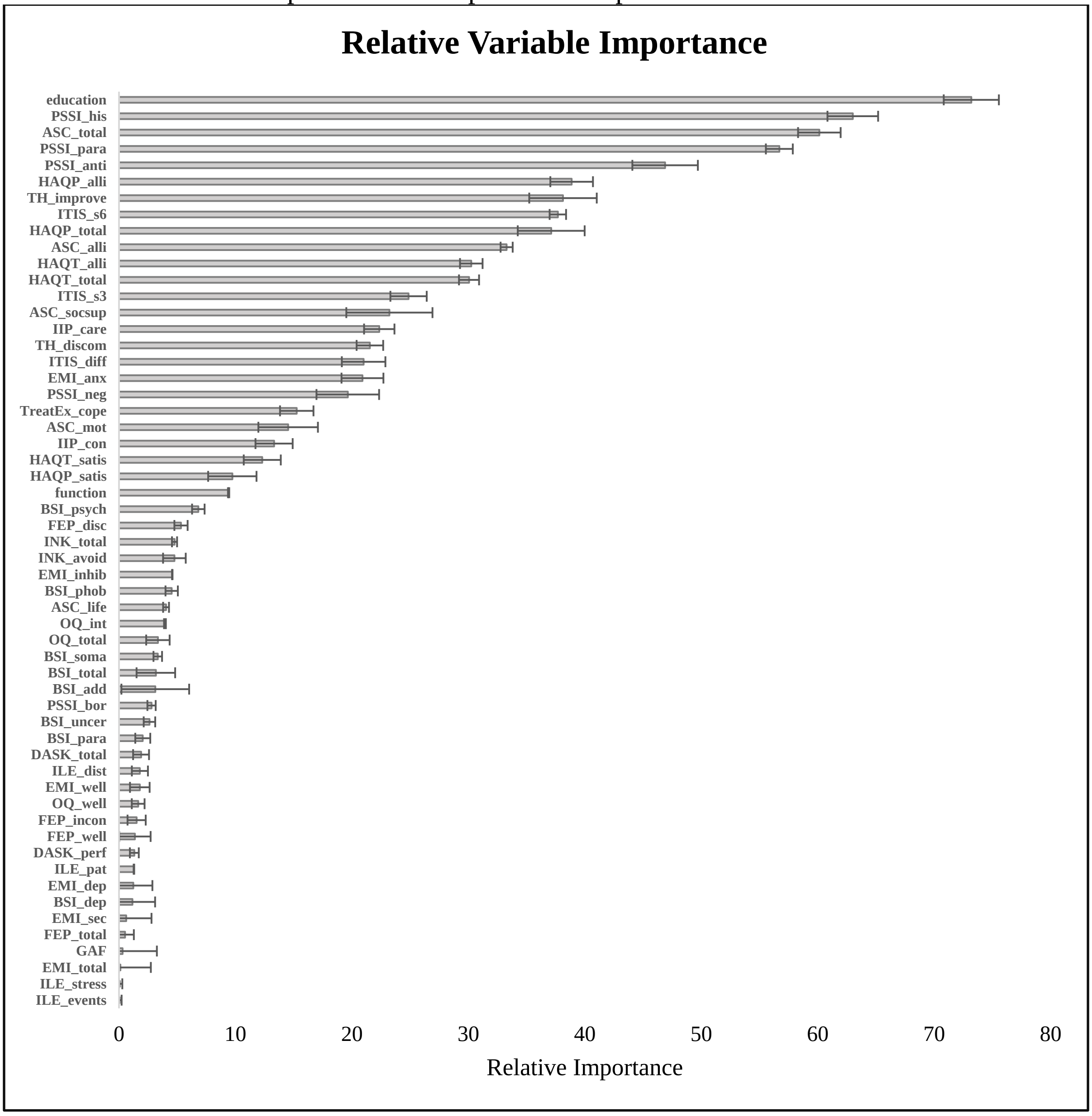
This procedure was conducted twice, once under inclusion and once under exclusion of the ITIS variables to examine the impact of the ITIS variables on model generation. The means of the Brier score, accuracy and the AUC across all 100 models generated were examined for the set with and the set without the ITIS variables to identify the best overall model. Finally, the distributions of each parameter (i.e., Brier score, accuracy, AUC) were compared via three one-sided pairwise *t*-tests, one for each parameter. The impact of the ITIS variables was assessed by calculating Cohen's *d* for each model parameter.

6.3 Results

Fifty-six of 130 variables were significantly correlated with dropout, including three ITIS items: use of cognitive techniques was linked to lower dropout rates, while use of feedback/summaries and treatment difficulty were associated with higher dropout rates. Figure 2 shows the importance of all pre-selected predictors across all 100 EN generations.

Figure 2

Means of the relative importance of the pre-selected predictor variables across all 100 models.



Note: PSSI = Personality style and disorder inventory; ASC = Assessment for signal clients; HAQP = Helping alliance questionnaire – patient version; TH = Therapist expectations; ITIS = Inventory of therapeutic interventions and skills; HAQT = Helping alliance questionnaire – therapist version; IIP = Interpersonal problems; EMI = Emotionality inventory; TreatEx = Treatment expectations; BSI = Brief symptom inventory; FEP = Questionnaire for the evaluation of psychotherapeutic progress; INK = Incongruence questionnaire; OQ = Outcome questionnaire; DASK = Dysfunctional attitudes scale – short form; ILE = Inventory of stressful life events; GAF = Global assessment of functioning.

The three ITIS variables were included in most of the 100 model generations (range: 73% - 97%), indicating the importance of these items for dropout model generation. The best models were generated when using the dataset including the ITIS variables. This was the case for all three parameters (Brier score: 0.2096 vs. 0.2138; accuracy: 0.6758 vs. 0.6687; AUC: 0.7226 vs. 0.7130). For the confusion matrices, see Table 2.

Table 2
Mean scores across all 100 confusion matrices for the models excluding and including ITIS variables.

Model including intake variables only	**Observed**		
Predicted	Regular	Dropout	Total
Regular	37.63	9.13	46.76
Dropout	16.71	14.53	31.24
Total	54.34	23.66	78

Model including intake and ITIS variables	**Observed**		
Predicted	Regular	Dropout	Total
Regular	38.00	8.95	49.58
Dropout	16.34	14.71	28.42
Total	54.34	23.66	78

Note: ITIS = Intervention of therapeutic interventions and skills

The *t*-tests revealed a significant effect between the dropout model including the ITIS variables as predictors and the model based on intake variables alone for the Brier score ($t(99) = 6.49$; $p < .001$; Cohen's $d = 0.65$), accuracy ($t(99) = 2.10$; $p < .05$; Cohen's $d = 0.21$) and the AUC ($t(99) = 3.93$; $p < .001$; Cohen's $d = 0.39$). Predictors that were in at least 95% of the models generated by elastic net were ITIS treatment difficulty, ITIS use of feedback/summaries, a paranoid or histrionic personality style, a high total score in the assessment for signal clients (ASC; Lambert et al., 2007) and a high education level, with the last two predictors having a negative association with dropout.

6.4 Conclusion

Study II demonstrated that video ratings improve dropout predictions compared to those that use self-report assessments only. The model including the ITIS variables outperformed the model without them on all three parameters with small to medium effects (J. Cohen, 1988). Although the added predictive benefit of the observer-rated variables may seem descriptively small, it is important to remember that these variables made a significant contribution to dropout prediction beyond intake variables that already covered a very wide range of potentially influencing factors. Overall, the model showed good prediction performance with a brier score of .2096, an accuracy score of .6758 and an AUC score of .7226.

In line with previous findings (Lutz et al., 2019; Swift & Greenberg, 2012; Zimmermann et al., 2017) and findings from Study I, several intake variables, including histrionic personality style and lower education level, significantly predicted dropout. Besides, three ITIS variables improved dropout prediction, namely *therapist application of cognitive techniques, therapist use of feedback and summaries,* and *treatment difficulty*.

The use of cognitive techniques was a protective factor against dropout. When cognitive techniques are applied effectively early in treatment, they may facilitate symptom relief (Lorenzo-Luaces et al., 2015), in turn strengthening the patient's confidence in the treatment's effectiveness and increasing commitment to continue treatment. Therapists' competent use of feedback and summaries was associated with a higher dropout risk. Although these skills are considered important for therapy outcome, it is possible that the intensified use of these strategies are not necessarily a cause of dropout, but rather a therapeutic response to an at-risk patient. When therapists perceive the patient as difficult or at risk of dropping out of treatment, they may react by making a stronger effort to gain feedback, which is in line with Cooper et al. (2016). Last, higher observer-rated treatment difficulty was also predictive of dropout. This result is consistent with findings showing patient intake symptom severity, interpersonal impairment, and personality disorders to be associated with a higher probability of dropout (Lutz et al., 2019; Lutz et al., 2018; Swift & Greenberg, 2012).

Study II used the findings from Study I and extended them with data from the first sessions and other sources, namely video ratings. Once again, a model was generated that can be used in clinical practice, demonstrating that data from various sources do improve predictions (Lutz, de Jong, et al., 2021). Clinicians could obtain information from this model about which patients are at risk of dropping out of therapy and adjust their strategy accordingly, e.g., by increasing the use of cognitive techniques to reduce the level of distress. Like Study I, this study adds to the growing body of research on precision mental health, but needs to be further tested in new data for use in other contexts.

7 Study III: Fostering the Upward Spiral After a Sudden Gain in Routine Care Cognitive Behavioral Therapy

Bennemann, B., Schwartz, B., & Lutz, W. (2023). Fostering the upward spiral after a sudden gain in routine care cognitive behavioral therapy. [Manuscript submitted for publication].

7.1 Introduction

Study III shifted the focus from predicting dropout to predicting outcome following sudden gains (SGs). The phenomenon of a SG in psychotherapy refers to a disproportionate improvement in symptoms between two therapy sessions (Tang & DeRubeis, 1999). Recent research has shown that trajectories of patients in therapy often appear non-linear and have interindividual differences (Lutz, de Jong, et al., 2021; Lutz et al., 2013). A common phenomenon that reflects this varying responsiveness is the SG (Shalom & Aderka, 2020). SGs have been found across a variety of therapies (Gaynor et al., 2003; Kelly et al., 2007; Kelly et al., 2005; Lemmens et al., 2016; Tang et al., 2005; Vittengl et al., 2015) and for a range of disorders (Aderka, Anholt, et al., 2012; Aderka et al., 2011; Clerkin et al., 2008; Hofmann et al., 2006; Wiedemann et al., 2020). Further, SGs are found in both randomized controlled trials and naturalistic samples (Wucherpfennig, Rubel, Hollon, & Lutz, 2017), also showing a therapist effect in naturalistic samples (Deisenhofer et al., 2022).

Since approximately one third of all patients experience a SG, but not all benefit equally from it, understanding the process that occurs after a SG is of high clinical relevance (Shalom & Aderka, 2020). Studies that have investigated the processes after a SG are rare, but show findings mostly consistent with Tang and DeRubeis’ (1999) theory that there is an improvement in the therapeutic alliance that leads to an upward spiral and thus to a better outcome (Lutz et al., 2013; Wucherpfennig, Rubel, Hollon, & Lutz, 2017; Zilcha-Mano et al., 2019). The objective of this study was therefore to replicate the results of past studies using advanced methodology that ensures better generalizability (i.e., elastic net) and to identify additional predictors of outcome after a SG. Additionally, the impact of these predictors on outcome was assessed by comparing and quantifying the effects with a control group that did not experience SGs.

7.2 Methods

7.2.1 Patients and Treatment

The sample comprised a total of 3626 patients treated at the University of Trier outpatient CBT Clinic between 2007 and 2022. Patients were included in the analyses if they completed a questionnaire battery before therapy, completed therapy (i.e., regular termination or dropout), and their therapy lasted at least 6 sessions. Of the remaining 2840 patients, 794 SG patients were identified and included. From the group of remaining patients, a control group of equal size was selected via NN matching, resulting in a total sample of 1588. Treatments were conducted by 218 therapists; each therapist treated 7.3 patients on average.

Diagnoses were based on the German version of the SCID-I and the IDCL-P. The mean score of the BSI indicated a moderate to severe general level of distress and did not differ significantly between groups.

7.2.2 Measures

The Hopkins Symptom Checklist short form (HSCL-11; Lutz, Tholen, et al., 2006) was assessed at the beginning of each session to identify sudden gains. The Global Severity Index (GSI) of the BSI was calculated to capture symptomatic distress at the beginning and end of therapy. The scales *therapeutic relationship*, *problem solving*, *problem actualization*, *motivational clarification*, and *resource activation* of the Bern Post Session Reports (BPSR; Flückiger et al., 2010) from both the patient (PSTB) and therapist (TSTB) perspectives were assessed after each session. Further, all items from the Session Rating Scale (SRS; Duncan et al., 2003) were assessed after each session. Additionally, the Global Assessment of Functioning Scale (GAF; American Psychiatric Association, 2005) and the coping item *How well is your patient coping emotionally and psychologically?* was rated by the therapist after each session. Last, alliance ruptures were assessed by the therapist as well as the patient after each session via the item *During today's session, did you perceive any tension, misunderstandings, or inconsistencies in the relationship with your patient/therapist?*.

7.2.3 Data Preparation

In a first step, missing data were imputed. First, Level 1 data were imputed, with the mean values subsequently used for Level 2 imputations. The randomForest method of the mice package in R (v3.14.0; van Buuren, 2021) was used for this purpose. No variable had more than 30% missingness, for which good algorithm performance has previously been shown (Stekhoven & Bühlmann, 2012).

Next, all sudden gain patients were identified, using the criteria by Tang and DeRubeis (1999). However, the first criterion had to be adapted, as the HSCL-11 was used. The reliable change index (RCI; Jacobson & Truax, 1991) was used for the first criterion, as suggested by Stiles et al. (2003). The RCI for the HSCL-11 was 0.61.

To create a control group, propensity score matching (PSM; see Stuart et al., 2004 for an overview) was used, which balances two groups for comparability. The nearest neighbor method suggested by Ho et al. (2007) was used and the following variables were included for matching, based on the work of Delgadillo et al. (2016) and Wucherpfennig, Rubel, Hofmann, and Lutz (2017): pre-treatment BSI, age, therapy expectations, medication at intake, sex, education status, marital status, and employment status. Each sudden gainer was matched with the most similar non-gainer based on these variables. To ensure comparability, χ^2-tests and *t*-tests were conducted.

After matching, a comparable *pseudo gain* (PG) session was selected for each patient in the non-gainer group. The PG session was chosen based on the time point in relation to the total number of sessions. This allowed for effect sizes to be calculated for predictors to provide more precise information about the processes after a SG.

7.2.4 Data Analytic Strategy

To analyze sudden gainers' outcome predictors independent of their general influence on outcome, two comparable prediction models were generated using the mean scores and coefficients of variation (CoV) of each variable (except the GSI pre score) over the first three sessions after a SG or PG and the GSI post score as the outcome. The CoV is defined as the standard deviation standardized by the mean.

A repeated nested cross-validation with ten outer and five inner loops using EN regularization as the prediction model generation algorithm (Friedman et al., 2010) was conducted separately for the SG and the PG group, repeating the inner cross-validations three times to avoid overfitting (Brownlee, 2016; Cawley & Talbot, 2010). One hundred overall combinations of α and λ were defined to find the best fitting parameters with the root-mean-square error (RMSE) as the model evaluation parameter for each outer run of the CV.

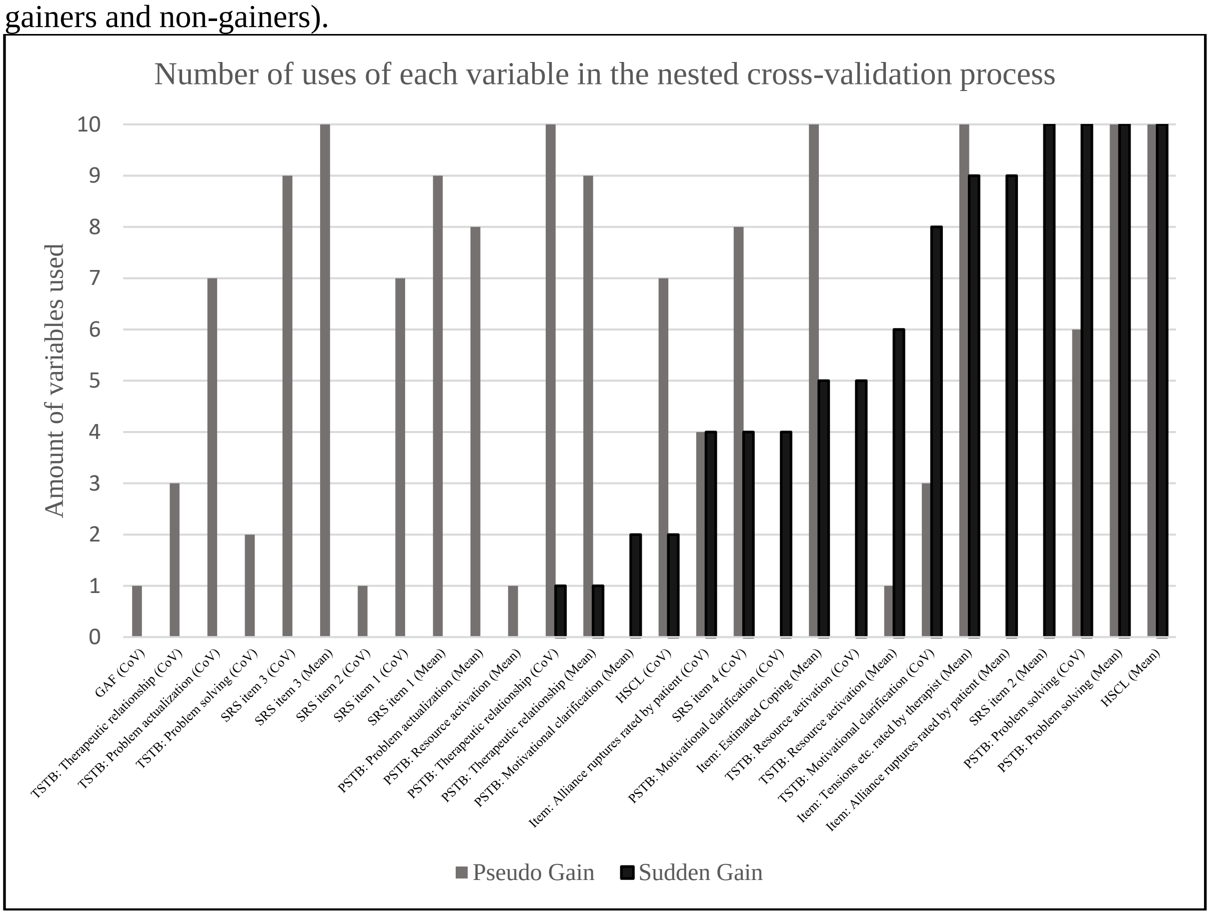
For each predictor, one RMSE value and one regression weight were obtained in each outer CV run. Predictors included in all 10 runs were used for further analysis. Predictor distributions between SG and PG were compared using *t*-tests to quantify the effects between the SG and the control group. In addition, model performance using the RMSE distributions were compared via a *t*-test.

7.3 Results

The two groups did not differ with respect to the matching variables, albeit the SG group had a significantly better therapy outcome (*t*(1586) = 3.59, *p* < .001, Cohen's *d* = 0.18). The two groups differed in terms of diagnoses, with the SG group having more affective disorders, while the PG had more anxiety disorders. All included predictors are shown in Figure 3.

Figure 3

Number of occurrences of each variable in each outer CV run for both groups (i.e., sudden gainers and non-gainers).



Note: HSCL = Hopkins symptom checklist short form; PSTB = Bern post-session reports for patients; TSTB = Bern post-session reports for therapists; SRS = Short rating scale; GAF = Global assessment of functioning; CoV = Coefficient of variation; All variables not included in the figure were excluded in each run in both groups.

Table 3 provides a comparison of the included variables and their mean values. Differences between the two groups were significant for all variables except for the mean score of the problem solving scale. Further, the model parameters did not differ significantly.

Table 3
Mean regression weights for all relevant variables across the three sessions after a sudden (pseudo) gain and the mean differences between these two groups.

	Sudden gainers (n = 794)	Non-gainers (n = 794)			
Variable	M (SD)	M (SD)	*t*-value	*p*-value	Cohen's *d*
Intercept	.000 (.000)†	.000 (.000)†	1.54	0.140	0.69
HSCL (mean)	.213 (.007)	.348 (.013)	27.91***	0.000	12.48
PSTB: Problem solving (mean)	-.013 (.010)	-.011 (.008)	0.55	0.589	0.25
PSTB: Problem solving (CoV)	.029 (.011)	.005 (.009)†	5.24***	0.000	2.34
PSTB: Therapeutic relationship (CoV)	.000 (.001)†	.025 (.001)	10.90***	0.000	4.88
SRS: Item 2 (mean)	-.017 (.010)	.000 (.000)†	5.37***	0.000	2.40
SRS: Item 3 (mean)	.000 (.000)†	-.018 (.005)	11.11***	0.000	4.97
Item: Alliance ruptures (mean)	.009 (.006)	.017 (.009)	2.42*	0.026	1.08
Item: Estimated coping (mean)	.003 (.005)†	-.013 (.008)	5.37***	0.000	2.40
Model parameter (RMSE)	.546 (.050)	.539 (.056)	0.33	0.748	0.15

Note: HSCL = Hopkins symptom checklist short form; PSTB = Bern post session reports for patients; SRS = Short rating scale; RMSE = Root mean square error; CoV = Coefficient of variation; * *p* > .05; ** *p* > .01; *** *p* > .001; † Mean value is not significantly different from 0. To determine significance, the Benjamini-Hochberg procedure was applied to prevent false discovery rates through multiple testing.

7.4 Conclusion

In Study III, a ML algorithm (i.e., EN) was utilized to detect predictors of outcome following a SG. A comparable control group was generated using PSM, whereby a pseudo gain session was assigned to this group. The same modeling process was applied to both groups and no significant differences in terms of RMSE were found. However, the predictor differences showed medium to large effects. As expected, the SG group had a lower HSCL mean value, indicating that this group has a better outcome, which is consistent with previous findings (Shalom & Aderka, 2020). Further, high average use of problem-solving strategies had a positive effect on both groups, whereas a high variation led to a worse outcome only in the SG group. This is probably due to the fact that the sample was treated mainly with a CBT focus and that sudden gains have higher effect sizes in CBT settings (Aderka, Nickerson, et al., 2012).

The CoV of the therapeutic alliance had the reversed effect, leading to a worse outcome only in the PG group, where alliance ruptures also had a larger effect. Because sudden gainers often showed reliable improvement, also regarding interpersonal problems, they were more able to handle interactional problems and disagreements. Further, patient coping led to a better outcome for non-gainers only. Consistent with this finding, for sudden gainers, goals and topics are crucial (SRS item 2), while for non-gainers the approach and method (SRS item 3) are vital, indicating that the *what* seems to be more important for sudden gainers than the *how*, which is reversed for the non-gainers. This also supports the finding that consistent use of problem-solving strategies is important for sudden gainers, as they have enough resources to apply concrete strategies and are better able to handle a directive approach.

Study III builds on the findings from the first study and puts them into a new context with a continuous outcome. In this way, the modeling process was used to identify crucial predictors. These findings could support clinicians with clear recommendations for their treatment plan after a sudden gain to extend the upward spiral to the end of therapy and beyond. Therefore, Study III fits well in the context of data-driven psychological therapy (Z. D. Cohen et al., 2021; Lutz, Schwartz, & Delgadillo, 2022). However, to apply these recommendations in other contexts, again, further testing is needed.

8 General Discussion

The three studies summarized in this umbrella represent three important contributions to the field of precision mental health and prediction modeling and can therefore be allocated to the patient-focused research paradigm. While Studies I and II concentrate on dropout prediction, Study III shifted the focus to another phenomenon, namely sudden gains. At the same time, all three studies used advanced methodology in the form of machine learning algorithms.

Study I extends the research on predicting dropout in outpatient psychotherapy. For the first time, a model was developed to predict dropout rates using data from a naturalistic sample prior to the first session, while comparing a large number of algorithms and ensembles. Only one study had a similar aim, but used a smaller sample, compared fewer algorithms and predicted outcome rather than dropout (Webb et al., 2020). The results of Study I indicated that not all algorithms and ensembles are equally effective for prediction purposes, revealing that some algorithms perform worse than a GLM. However, it was found that the most suitable algorithms and ensembles significantly outperformed average algorithms and that some algorithms appear to be better suited to the large amount of clinical naturalistic data. Finally, the best ensemble generated a model that could predict dropout rates, providing valuable information for clinicians.

In the spirit of precision mental health, Study II takes this approach further by using the findings from Study I to generate a dropout model with more potential predictors and additional data sources. Data from video ratings (Boyle et al., 2020) and information from the first two sessions were used to add more information to the generation process. No other study has used video ratings to predict dropout so far. Results highlight that two ITIS items were among the six best predictors of dropout, indicating that video ratings can improve dropout predictions beyond the level of self-report instruments.

Study III shifted the focus to sudden gains and used predictive modeling to identify and quantify predictors for outcome after a SG. Despite the importance of the topic for clinical practice, studies are rare compared to those investigating predictors of sudden gains. To date, there has been only one study that used a control group to examine the changes after a sudden gain and related them to outcome (Wucherpfennig, Rubel, Hofmann, & Lutz, 2017). The results highlighted that the constant use of problem-solving interventions after a SG is promising and the content of therapy is more important than how it is delivered.

In the following, some general conclusions that can be drawn from the three studies and their implications for future research are summarized.

8.1 General Conclusions and Future Research

Taking all three studies together, some general conclusions and future directions can be deduced. In general, all three studies demonstrated the usefulness of machine learning for prediction models and clinical utility.

Primarily, Study I revealed that not all ML algorithms are appropriate for naturalistic data in a clinical setting. By comparing more than 20 different algorithms and ensembles, this study showed that inappropriate ML algorithms perform even worse than a GLM. This finding is in line with previous arguments that personalized mental health is not only a matter of algorithmic complexity (Wilkinson et al., 2020). Therefore, prediction quality does not seem to be based on data quality and quantity alone, but also on a fitting methodology (Steyerberg et al., 2013). Further, Study I also demonstrated that ML has the ability to capture more complex relationships in the data and thus make more accurate predictions, leading to a reliable data-driven methodology (Z. D. Cohen et al., 2021). Hereby, the significance of feature selection in ensuring reliable prediction was highlighted. This is in line with previous research showing that variables that have no predictive power for the outcome can weaken the power of the model (Chowdhury & Turin, 2020). Ideally, therefore, from a large group of variables, those should be identified that make a significant contribution to the prediction of the target variables, e.g., using an EN analysis. Furthermore, depending on the amount of data, it should be investigated whether non-linear algorithms are suitable for the analysis in advance. Study I showed that not all data sets are suitable for this approach. Future studies should try to replicate these findings with another (continuous) outcome variable to determine whether these findings are transferable.

Further, Study I identified predictors crucial for dropout prediction, helping translate the findings into clinical practice. Study I's results indicate that a complementary alliance and a motivational approach are decisive for younger, less educated and/or interpersonally difficult patients, which is in line with other findings (Swift & Greenberg, 2012; Zimmermann et al., 2017). Additionally, higher initial impairment was a predictor of dropout, which is also consistent with previous studies (Lutz et al., 2018; Zimmermann et al., 2017). Study I extended these previous finding by identifying more concrete symptoms, namely *sleep problems, poor appetite*, *suicidal thoughts,* and *feelings of guilt* that predicted dropout and should therefore be treated first.

In Study II, the approach was enhanced by incorporating more variables from various data sources, as recommended by Lutz, de Jong, et al. (2021), which improved the predictions. Combining the unique information from each data source led to more reliable dropout prediction, as demonstrated by the higher AUC score compared to Study I results. Additionally, the model from Study II performed only slightly worse in terms of Brier score compared to the results in Study I, despite having a smaller sample size. The inclusion of additional variables from the first sessions, which provide valuable insight into the patient-therapist interaction and are enriched by video ratings, allowed for more reliable predictions. Besides the inclusion of two ITIS items, this fact is strengthened by the importance of the ASC predictor in Study II, which captured important interactional and motivational aspects from the first session that were not available in Study I. Nevertheless, the model from Study II also included predictors that assessed difficult personality styles in therapy and educational status, strengthening the importance of these variables in dropout prediction.

The improvement of the models from Studies I to II suggests that a dynamic modeling approach is appropriate. This means that models are provided with data from previous sessions and current progress throughout the course of therapy in order to update the predictions (e.g., Bone et al., 2021; Lutz et al., 2019). The model from Study I could therefore be seen as a starting point, which is augmented by additional data from different sources to improve predictions. This could lead to so-called clinical support tools, which can be applied directly into clinical practice to individualize therapy for the patient at hand in terms of precision mental health and patient-focused research (Lutz, de Jong, et al., 2021; Lutz, Deisenhofer, et al., 2022). However, acquiring many different sources of data, especially video ratings, is a costly and time-consuming process, limiting its feasibility to larger healthcare institutions rather than private practices. Future research should further investigate the use of such dynamic models in a prospective setting and how these can be implemented into clinical practice.

In Study III, the application of ML was switched to another context, namely sudden gains and outcome. It was demonstrated that constant use of problem-solving techniques is important to foster the upward spiral and that the content of the therapy session is more important than the way it is delivered. Further, problems regarding the alliance have a bigger impact for non-gainers. These findings are in line with the original theory for the upward spiral (Tang & DeRubeis, 1999) and previous findings (Bohn et al., 2013; Wucherpfennig, Rubel, Hofmann, & Lutz, 2017). In contrast to the other studies, this was the first to quantify these effects using a methodology that allows for generalizability, even though no holdout data was provided. The medium to large effect sizes highlight the importance of the right therapeutic strategy after a SG, which is crucial for the maintenance of the upward spiral and thus the therapeutic outcome. However, the identification of a SG is difficult in clinical practice, because according to Tang and DeRubeis' (1999) criteria, three sessions after a SG must be assessed. Nevertheless, this information is of great value to clinicians, because the results reveal which strategies were applied after a sudden symptom improvement to maintain it. Because there is evidence that the influence of sudden gains depends on the time point, at which they occur, future studies should take this into account (Stiles et al., 2003). It is possible that varying therapeutic strategies may be indicated after an early versus later sudden gain.

All three studies utilized machine learning algorithms for prediction in the field of precision mental health and patient-focused research. Results indicated that predictive modeling can offer considerable clinical value and can be directly applied in practice. While the models developed in these studies demonstrate significant utility and implement the suggestions by Steyerberg et al. (2013) presented in Chapter 2.4, further research is needed to validate them using a prospective holdout dataset to improve their generalizability. Therefore, it is important to note that despite the strengths of these studies, there are some limitations that need to be addressed. These limitations will be discussed in the following section.

8.2 General Limitations

All studies have a limitation due to the lack of prospective holdout samples. Although Study I used a holdout sample, it was not prospective and did not come from another outpatient clinic, which limits generalizability due to possible overfitting. In Studies II and III, creating a holdout dataset was not feasible due to the relatively small sample sizes. Nevertheless, precautions were taken to ensure generalizability, such as separating the data into a training set and a test set and using CV procedures (Ball et al., 2020). Further, all studies considered the recommendations for data pre-processing and statistical requirements and covered a wide range of potentially influencing factors (see Chapters 2.4.1 and 2.4.2; Z. D. Cohen et al., 2021; Delgadillo, 2021; Steyerberg et al., 2013). Therefore, all models are elaborated to the extent that they can be well applied in other data contexts to test their generalizability.

Especially in Studies II and III, the sample sizes should be discussed. Since ML algorithms are able to capture complex relationships, they also need correspondingly large data sets. Only Study I had a sample size of over 1000 patients, justifying the use of the term big data (Delgadillo, 2021). However, it remains uncertain whether the amount of data is sufficient for highly complex algorithms such as neural networks (Brownlee, 2016; Jacobucci & Grimm, 2020). In a similar study with a sample size of nearly 50000 cases, neural networks were able to produce better models (Giesemann et al., 2023). Due to these preconditions, an elastic net algorithm was used in Studies II and III, which can only capture linear correlations, but requires only smaller amounts of data (Domingos, 2012). In this context, data quality should also be mentioned, as self-report assessments and video ratings always have measurement errors, which can be large and make predictions more difficult compared to data from other disciplines such as neuroscience. Nevertheless, the datasets used in these studies are still comparatively large, have many variables and represent real-world settings, which further emphasizes the clinical utility of these models. However, larger data sets and intensive repeated measures, e.g., via ecological momentary assessment (EMA; Fisher & Bosley, 2020; Hehlmann et al., 2021), may further improve the models. EMA is more likely to reflect patients' real experiences and behavior in their daily lives than retrospective measurements, which are subject to several biases (e.g., retrospective biases; Ebner-Priemer & Trull, 2009), and should therefore be used in future studies.

Besides the sample constraints, it is important to take into account the composition of the samples used in the studies. All studies relied on data from the psychotherapy outpatient clinic at the University of Trier in Germany. Therapists had diverse cultural backgrounds and treated patients with migrant backgrounds as well as those with different gender identities and sexual orientations, addressing common mental illnesses among these populations. Nevertheless, some populations were underrepresented, which can be seen as a constraint, as building prediction models based on non-diverse samples may lead to algorithms that discriminate against minorities with rare disorders or of black ethnicity. This could happen due to a lack of diversity in the teams that developed the algorithms or in the training and test data used, highlighting the need for more inclusive and representative data collection practices. To address this concern, Forscher et al. (2020) suggest utilizing big team science approaches that involve the assessment of multi-cultural, multi-site data.

In addition to limitations due to sampling, the definition of dropout could be a limiting factor. In Studies I and II, dropout was defined as the patient stopping therapy despite the therapist's recommendation to continue. Because there are other definitions of dropout, transferring the results to contexts using other definitions should be made with caution.

Furthermore, Study III did not control for symptom severity after a SG or PG. For this reason, it cannot be distinguished whether the recommendations after a SG really apply to this phenomenon alone or to patients with lower symptom severity in general. Although this does not limit the results and recommendations, future studies should additionally control for symptom severity after a SG or PG to better understand and classify the phenomenon of sudden gains. Further, in Study III, the distribution of diagnoses in both groups was not identical, but there is no reason to assume that this interfered with the findings (Aderka & Shalom, 2021; Cuijpers et al., 2013). Last, in Study III crucial predictors may not have been assessed or longer-term processes may have been overlooked. Future studies should consider assessing more sessions and taking the timing of the SG into account in their analyses to investigate whether other predictors play an important role depending on the timing of SGs.

8.3 Concluding Remarks

Despite the limitations mentioned above, these studies offer valuable contributions to the field of precision mental health research and the use of new methodologies, specifically machine learning algorithms. The studies imply that it is possible to construct a model to predict an individual patient’s likelihood of dropping out based on data collected before the first session (Study I), and that this model can be enhanced by including additional data from early therapy sessions and data from various sources (Study II). Moreover, machine learning can be utilized to develop models that help to extract important information from crucial processes in therapy (Study III). Additionally, the modeling process can identify predictors that have a significant impact on the target variable, which can be used to tailor treatment plans in practice. By further developing these prognostic models, they can be further improved and applied in various contexts.

These three precision mental health approaches have the potential to improve already effective treatments by reducing the rate of treatment dropouts and adjusting treatment strategies after a sudden gain to ensure continued improvement and symptom relief. To optimize the use of these models, it is important to provide extensive training to therapists. Implementing data-driven and algorithm-based decision support in routine psychotherapy could potentially bridge the research-practice gap and establish a symbiosis of evidence-based practice and practice-based evidence (Barkham & Mellor-Clark, 2003; Kazdin, 2016; Lutz, 2002), an important goal in patient-focused research.

9 References

Aafjes-van Doorn, K., Kamsteeg, C., Bate, J., & Aafjes, M. (2021). A scoping review of machine learning in psychotherapy research. *Psychotherapy Research*, *31*(1), 92–116. https://doi.org/10.1080/10503307.2020.1808729

Aderka, I. M., Anholt, G. E., van Balkom, A. J. L. M., Smit, J. H., Hermesh, H., & van Oppen, P. (2012). Sudden gains in the treatment of obsessive-compulsive disorder. *Psychotherapy and Psychosomatics*, *81*(1), 44–51. https://doi.org/10.1159/000329995

Aderka, I. M., Appelbaum-Namdar, E., Shafran, N., & Gilboa-Schechtman, E. (2011). Sudden gains in prolonged exposure for children and adolescents with posttraumatic stress disorder. *Journal of Consulting and Clinical Psychology*, *79*(4), 441–446. https://doi.org/10.1037/a0024112

Aderka, I. M., Nickerson, A., Bøe, H. J., & Hofmann, S. G. (2012). Sudden gains during psychological treatments of anxiety and depression: A meta-analysis. *Journal of Consulting and Clinical Psychology*, *80*(1), 93–101. https://doi.org/10.1037/a0026455

Aderka, I. M., & Shalom, J. G. (2021). A Revised Theory of Sudden Gains in Psychological Treatments. *Behaviour Research and Therapy*, *139*, 103830. https://doi.org/10.1016/j.brat.2021.103830

Adibi, A., Sadatsafavi, M., & Ioannidis, J. P. A. (2020). Validation and Utility Testing of Clinical Prediction Models: Time to Change the Approach. *JAMA*, *324*(3), 235–236. https://doi.org/10.1001/jama.2020.1230

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., Nichols, C. N., Lampropoulos, G. K., Walker, B. S., Cohen, G., & Rush, J. D. (2006). The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction. *The Counseling Psychologist*, *34*(3), 341–382. https://doi.org/10.1177/0011000005285875

Alpaydin, E. (2010). *Introduction to machine learning* (2nd ed.). MIT Press.

Altman, N., & Krzywinski, M. (2017). Ensemble methods: bagging and random forests. *Nature Methods*, *14*(10), 933–934. https://doi.org/10.1038/nmeth.4438

American Psychiatric Association. (2005). *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association.

American Psychological Association. (2019). *Clinical practice guideline for the treatment of depression across three age cohorts.* APA Press. https://www.apa.org/depression-guideline/guideline.pdf

Archer, L., Snell, K. I. E., Ensor, J., Hudda, M. T., Collins, G. S., & Riley, R. D. (2021). Minimum sample size for external validation of a clinical prediction model with a continuous outcome. *Statistics in Medicine*, *40*(1), 133–146. https://doi.org/10.1002/sim.8766

Ball, T. M., Squeglia, L. M., Tapert, S. F., & Paulus, M. P. (2020). Double Dipping in Machine Learning: Problems and Solutions. *Biological Psychiatry. Cognitive Neuroscience and Neuroimaging*, *5*(3), 261–263. https://doi.org/10.1016/j.bpsc.2019.09.003

Barkham, M., & Lambert, M. J. (2021). The efficacy and effectiveness of psychological therapies. In M. Barkham, W. Lutz, & L. G. Castonguay (Eds.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (50th ed., pp. 135–189). John Wiley & Sons, Inc.

Barkham, M., & Margison, F. (2007). Practice-based Evidence as a Complement to Evidence-based Practice: From Dichotomy to Chiasmus. In C. Freeman & M. Power (Eds.), *Handbook of Evidence-based Psychotherapies* (pp. 443–476). John Wiley & Sons Ltd. https://doi.org/10.1002/9780470713242.ch25

Barkham, M., & Mellor-Clark, J. (2003). Bridging evidence-based practice and practice-based evidence: developing a rigorous and relevant knowledge for the psychological therapies. *Clinical Psychology & Psychotherapy*, *10*(6), 319–327. https://doi.org/10.1002/cpp.379

Barrett, M. S., Chua, W.-J., Crits-Christoph, P., Gibbons, M. B. C., & Thompson, D. (2008). Early withdrawal from mental health treatment: Implications for psychotherapy practice. *Psychotherapy: Theory, Research, Practice, Training*, *45*(2), 247. https://doi.org/10.1037/0033-3204.45.2.247

Björk, T., Björck, C., Clinton, D., Sohlberg, S., & Norring, C. (2009). What happened to the ones who dropped out? Outcome in eating disorder patients who complete or prematurely terminate treatment. *European Eating Disorders Review: The Professional Journal of the Eating Disorders Association*, *17*(2), 109–119. https://doi.org/10.1002/erv.911

Bohn, C., Aderka, I. M., Schreiber, F., Stangier, U., & Hofmann, S. G. (2013). Sudden gains in cognitive therapy and interpersonal therapy for social anxiety disorder. *Journal of Consulting and Clinical Psychology*, *81*(1), 177–182. https://doi.org/10.1037/a0031198

Böhnke, J. R., & Croudace, T. J. (2016). Calibrating well-being, quality of life and common mental disorder items: Psychometric epidemiology in public mental health research. *The British Journal of Psychiatry*, *209*(2), 162–168. https://doi.org/10.1192/bjp.bp.115.165530

Bone, C., Simmonds-Buckley, M., Thwaites, R., Sandford, D., Merzhvynska, M., Rubel, J., Deisenhofer, A.-K., Lutz, W., & Delgadillo, J. (2021). Dynamic prediction of psychological treatment outcomes: Development and validation of a prediction model using routinely collected symptom data. *The Lancet. Digital Health*, *3*(4), e231-e240. https://doi.org/10.1016/S2589-7500(21)00018-2

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

Boyle, K., Deisenhofer, A.-K., Rubel, J. A., Bennemann, B., Weinmann-Lutz, B., & Lutz, W. (2020). Assessing treatment integrity in personalized CBT: The inventory of therapeutic interventions and skills. *Cognitive Behaviour Therapy*, *49*(3), 210–227. https://doi.org/10.1080/16506073.2019.1625945

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

British Psychological Society. (2010). *Depression: The Treatment and Management of Depression in Adults (Updated Edition)*. British Psychological Society.

Bronisch, T., Hiller, W., Mombour, W., & Zaudig, M. (1996). *International diagnostic checklists for personality disorders according to ICD-10 and DSM-IV—IDCL-P*. Seattle, WA: Hogrefe and Huber Publishers.

Brownlee, J. (2016). *Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch*. Jason Brownlee.

Brownlee, J. (2021). *A Gentle Introduction to Ensemble Learning Algorithms*. https://machinelearningmastery.com/tour-of-ensemble-learning-algorithms/

Burkov, A. (2019). *The hundred-page machine learning book*. Andriy Burkov.

Cai, F. (2022). The Application of the Gestalt Theory in Music Psychotherapy for Piano. *Occupational Therapy International*, *2022*, 2119111. https://doi.org/10.1155/2022/2119111

Carpenter, J. K., Andrews, L. A., Witcraft, S. M., Powers, M. B., Smits, J. A. J., & Hofmann, S. G. (2018). Cognitive behavioral therapy for anxiety and related disorders: A meta-analysis of randomized placebo-controlled trials. *Depression and Anxiety*, *35*(6), 502–514. https://doi.org/10.1002/da.22728

Castonguay, L. G., Barkham, M., Lutz, W., & McAleavey, A. A. (2013). Practice-oriented research: Approaches and applications. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (6th ed., pp. 85–133). J. Wiley & Sons.

Castonguay, L. G., Barkham, M., Youn, S. J., & Page, A. C. (2021). Practice-based evidence - Findings from routine clinical settings. In M. Barkham, W. Lutz, & L. G. Castonguay (Eds.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (50th ed., pp. 191–222). John Wiley & Sons, Inc.

Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research, 11,* 2079–2107. https://www.jmlr.org/papers/volume11/cawley10a/cawley10a.pdf

Chambless, D. L., & Ollendick, T. H. (2001). Empirically supported psychological interventions: Controversies and evidence. *Annual Review of Psychology, 52,* 685–716. https://doi.org/10.1146/annurev.psych.52.1.685

Chapman, C. L., Burlingame, G. M., Gleave, R., Rees, F., Beecher, M., & Porter, G. S. (2012). Clinical prediction in group psychotherapy. *Psychotherapy Research, 22*(6), 673–681. https://doi.org/10.1080/10503307.2012.702512

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16,* 321–357. https://doi.org/10.1613/jair.953

Chekroud, A. M., Bondar, J., Delgadillo, J., Doherty, G., Wasil, A., Fokkema, M., Cohen, Z. D., Belgrave, D., DeRubeis, R. J., Iniesta, R., Dwyer, D., & Choi, K. (2021). The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry*, *20*(2), 154–170. https://doi.org/10.1002/wps.20882

Chen, W., Zhang, S., Li, R., & Shahabi, H. (2018). Performance evaluation of the GIS-based data mining techniques of best-first decision tree, random forest, and naïve Bayes tree for landslide susceptibility modeling. *The Science of the Total Environment*, *644*, 1006–1018. https://doi.org/10.1016/j.scitotenv.2018.06.389

Chen, Z. S., Kulkarni, P. P., Galatzer-Levy, I. R., Bigio, B., Nasca, C., & Zhang, Y. (2022). Modern views of machine learning for precision psychiatry. *Patterns*, *3*(11), 100602. https://doi.org/10.1016/j.patter.2022.100602

Chowdhury, M. Z. I., & Turin, T. C. (2020). Variable selection strategies and its importance in clinical prediction modelling. *Family Medicine and Community Health*, *8*(1), e000262. https://doi.org/10.1136/fmch-2019-000262

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, *110*, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

Christophersen, E. R., & VanScoyoc, S. M. (2013). *Treatments that work with children: Empirically supported strategies for managing childhood problems (2nd ed.)*. American Psychological Association. https://doi.org/10.1037/14137-000

Clark, D. M. (2018). Realizing the Mass Public Benefit of Evidence-Based Psychological Therapies: The IAPT Program. *Annual Review of Clinical Psychology*, *14*, 159–183. https://doi.org/10.1146/annurev-clinpsy-050817-084833

Clerkin, E. M., Teachman, B. A., & Smith-Janik, S. B. (2008). Sudden gains in group cognitive-behavioral therapy for panic disorder. *Behaviour Research and Therapy*, *46*(11), 1244–1250. https://doi.org/10.1016/j.brat.2008.08.002

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge.

Cohen, Z. D., Delgadillo, J., & DeRubeis, R. J. (2021). Personalized Treatment Approaches. In M. Barkham, W. Lutz, & L. G. Castonguay (Eds.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (50th ed., pp. 673–703). John Wiley & Sons, Inc.

Cohen, Z. D., & DeRubeis, R. J. (2018). Treatment selection in depression. *Annual Review of Clinical Psychology*, *14*, 209–236. https://doi.org/10.1146/annurev-clinpsy-050817-084746

Cohen, Z. D., Wiley, J. F., Lutz, W., Fisher, A. J., Kim, Thomas, Saunders, Rob, & Buckman, J. E. J. (2018). *SMART Mental Health Prediction Tournament.* osf.io/wxgzu

Constantino, M. J., Vîslă, A., Coyne, A. E., & Boswell, J. F. (2018). A meta-analysis of the association between patients' early treatment outcome expectation and their posttreatment outcomes. *Psychotherapy*, *55*(4), 473–485. https://doi.org/10.1037/pst0000169

Cooper, A. A., Strunk, D. R., Ryan, E. T., DeRubeis, R. J., Hollon, S. D., & Gallop, R. (2016). The therapeutic alliance and therapist adherence as predictors of dropout from cognitive therapy for depression when combined with antidepressant medication. *Journal of Behavior Therapy and Experimental Psychiatry*, *50*, 113–119. https://doi.org/10.1016/j.jbtep.2015.06.005

Crits-Christoph, P., Gibbons, M. B. C., Crits-Christoph, K., Narducci, J., Schamberger, M., & Gallop, R. (2006). Can therapists be trained to improve their alliances? A preliminary study of alliance-fostering psychotherapy. *Psychotherapy Research*, *16*(3), 268–281. https://doi.org/10.1080/10503300500268557

Cuijpers, P., Reijnders, M., & Huibers, M. J. H. (2019). The Role of Common Factors in Psychotherapy Outcomes. *Annual Review of Clinical Psychology*, *15*, 207–231. https://doi.org/10.1146/annurev-clinpsy-050718-095424

Cuijpers, P., Sijbrandij, M., Koole, S. L., Andersson, G., Beekman, A. T. F., & Reynolds, C. F. (2013). The efficacy of psychotherapy and pharmacotherapy in treating depressive and anxiety disorders: A meta-analysis of direct comparisons. *World Psychiatry*, *12*(2), 137–148. https://doi.org/10.1002/wps.20038

de Jong, K. de, Conijn, J. M., Gallagher, R. A. V., Reshetnikova, A. S., Heij, M., & Lutz, M. C. (2021). Using progress feedback to improve outcomes and reduce drop-out, treatment duration, and deterioration: A multilevel meta-analysis. *Clinical Psychology Review*, *85*, 102002. https://doi.org/10.1016/j.cpr.2021.102002

Deane-Mayer, Z. A., & Knowles, J. E. (2016). *caretEnsemble: Ensembles of Caret Models*. R package version 2.0.0.

Deisenhofer, A.-K., Delgadillo, J., Rubel, J. A., Böhnke, J. R., Zimmermann, D., Schwartz, B., & Lutz, W. (2018). Individual treatment selection for patients with posttraumatic stress disorder. *Depression and Anxiety*, *35*(6), 541–550. https://doi.org/10.1002/da.22755

Deisenhofer, A.-K., Rubel, J. A., Bennemann, B., Aderka, I. M., & Lutz, W. (2022). Are some therapists better at facilitating and consolidating sudden gains than others? *Psychotherapy Research*, *32*(3), 343–357. https://doi.org/10.1080/10503307.2021.1921302

Delgadillo, J. (2021). Machine learning: A primer for psychotherapy researchers. *Psychotherapy Research*, *31*(1), 1–4. https://doi.org/10.1080/10503307.2020.1859638

Delgadillo, J., Ali, S., Fleck, K., Agnew, C., Southgate, A., Parkhouse, L., Cohen, Z. D., DeRubeis, R. J., & Barkham, M. (2022). Stratified Care vs Stepped Care for Depression: A Cluster Randomized Clinical Trial. *JAMA Psychiatry*, *79*(2), 101–108. https://doi.org/10.1001/jamapsychiatry.2021.3539

Delgadillo, J., Appleby, S., Booth, S., Burnett, G., Carey, A., Edmeade, L., Green, S., Griffin, P., Johnson, E., Jones, R., Parker, P., Reeves-McLaren, L., & Lutz, W. (2020). The Leeds Risk Index: Field-Test of a Stratified Psychological Treatment Selection Algorithm. *Psychotherapy and Psychosomatics*, *89*(3), 189–190. https://doi.org/10.1159/000505193

Delgadillo, J., & Lutz, W. (2020). A Development Pathway Towards Precision Mental Health Care. *JAMA Psychiatry*, *77*(9), 889–890. https://doi.org/10.1001/jamapsychiatry.2020.1048

Delgadillo, J., McMillan, D., Lucock, M., Leach, C., Ali, S., & Gilbody, S. (2014). Early changes, attrition, and dose-response in low intensity psychological interventions. *British Journal of Clinical Psychology*, *53*(1), 114–130. https://doi.org/10.1111/bjc.12031

Delgadillo, J., Moreea, O., & Lutz, W. (2016). Different people respond differently to therapy: A demonstration using patient profiling and risk stratification. *Behaviour Research and Therapy*, *79*, 15–22. https://doi.org/10.1016/j.brat.2016.02.003

Delgadillo, J., Rubel, J., & Barkham, M. (2020). Towards personalized allocation of patients to therapists. *Journal of Consulting and Clinical Psychology*, *88*(9), 799–808. https://doi.org/10.1037/ccp0000507

Demir, S., Schwarz, F., & Kaiser, T. (2022). Therapy from my point of view: A case illustration of routine outcome monitoring and feedback in psychotherapeutic interventions. *Journal of Clinical Psychology*, *78*(10), 2029–2040. https://doi.org/10.1002/jclp.23408

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), e83875. https://doi.org/10.1371/journal.pone.0083875

Dietterich, T. G. (2000). An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. *Machine Learning*, *40*(2), 139–157. https://doi.org/10.1023/A:1007607513941

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78–87. https://doi.org/10.1145/2347736.2347755

Duncan, B. L., Miller, S. D., Sparks, J. A., Claud, D. A., Reynolds, L. R., Brown, J., & Johnson, L. D. (2003). The Session Rating Scale: Preliminary psychometric properties of a “working” alliance measure. *Journal of Brief Therapy*, *3*(1), 3–12.

Elkin, I., Shea, M. T., Watkins, J. T., Imber, S. D., Sotsky, S. M., Collins, J. F., Glass, D. R., Pilkonis, P. A., Leber, W. R., & Docherty, J. P. (1989). National Institute of Mental Health Treatment of Depression Collaborative Research Program. General effectiveness of treatments. *Archives of General Psychiatry*, *46*(11), 971-982. https://doi.org/10.1001/archpsyc.1989.01810110013002

Ellsworth, J. R., Lambert, M. J., & Johnson, J. (2006). A comparison of the Outcome Questionnaire-45 and Outcome Questionnaire-30 in classification and prediction of treatment outcome. *Clinical Psychology & Psychotherapy*, *13*(6), 380–391. https://doi.org/10.1002/cpp.503

Ewbank, M. P., Cummins, R., Tablan, V., Bateup, S., Catarino, A., Martin, A. J., & Blackwell, A. D. (2020). Quantifying the association between psychotherapy content and clinical outcomes using deep learning. *JAMA Psychiatry*, *77*(1), 35–43. https://doi.org/10.1001/jamapsychiatry.2019.2664

Eysenck, H. J. (1952). The effects of psychotherapy: An evaluation. *Journal of Consulting Psychology*, *16*(5), 319–324. https://doi.org/10.1037/h0063633

Fiedler, K. (2011). Voodoo Correlations Are Everywhere-Not Only in Neuroscience. *Perspectives on Psychological Science*, *6*(2), 163–171. https://doi.org/10.1177/1745691611400237

Fisher, A. J., & Bosley, H. G. (2020). Identifying the presence and timing of discrete mood states prior to therapy. *Behaviour Research and Therapy*, *128*, 103596. https://doi.org/10.1016/j.brat.2020.103596

Flückiger, C., Regli, D., Zwahlen, D., Hostettler, S., & Caspar, F. (2010). Der Berner Patienten- und Therapeutenstundenbogen 2000. *Zeitschrift Für Klinische Psychologie Und Psychotherapie*, *39*(2), 71–79. https://doi.org/10.1026/1616-3443/a000015

Foa, E. B., & Rothbaum, B. O. (2001). *Treating the trauma of rape: Cognitive-behavioral therapy for PTSD*. Guilford Press.

Forscher, P. S., Wagenmakers, E.-J., Coles, N. A., Silan, M. A. A., Dutra, N. B., Basnight-Brown, D., & IJzerman, H. (2020). *The Benefits, Barriers, and Risks of Big Team Science*. PsyArXiv. https://doi.org/10.31234/osf.io/2mdxh

Franke, G. H. (2000). *Brief symptom inventory (BSI) von LR Derogatis:(Kurzform der SCL-90-R)*. Beltz Test.

Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders, 208,* 191–197. https://doi.org/10.1016/j.jad.2016.10.019

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1). https://doi.org/10.18637/jss.v033.i01

Gaynor, S. T., Weersing, V. R., Kolko, D. J., Birmaher, B., Heo, J., & Brent, D. A. (2003). The prevalence and impact of large sudden improvements during adolescent therapy for depression: A comparison across cognitive-behavioral, family, and supportive therapy. *Journal of Consulting and Clinical Psychology*, *71*(2), 386–393. https://doi.org/10.1037/0022-006X.71.2.386

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc.

Giesemann, J., Delgadillo, J., Schwartz, B., Bennemann, B., & Lutz, W. (2023). Predicting dropout from psychological treatment using different machine learning algorithms, resampling methods, and sample sizes. *Psychotherapy Research*, 1–13. https://doi.org/10.1080/10503307.2022.2161432

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. *Adaptive computation and machine learning*. The MIT Press.

Hamilton, S., Moore, A. M., Crane, D. R., & Payne, S. H. (2011). Psychotherapy dropouts: Differences by modality, license, and DSM-IV diagnosis. *Journal of Marital and Family Therapy*, *37*(3), 333–343. https://doi.org/10.1111/j.1752-0606.2010.00204.x

Hand, D. J., & Vinciotti, V. (2003). Choosing k for two-class nearest neighbour classifiers with unbalanced classes. *Pattern Recognition Letters*, *24*(9-10), 1555–1562. https://doi.org/10.1016/S0167-8655(02)00394-X

Handelman, G. S., Kok, H. K., Chandra, R. V., Razavi, A. H., Huang, S., Brooks, M., Lee, M. J., & Asadi, H. (2019). Peering Into the Black Box of Artificial Intelligence: Evaluation Metrics of Machine Learning Methods. *American Journal of Roentgenology*, *212*(1), 38–43. https://doi.org/10.2214/AJR.18.20224

Hansen, N. B., Lambert, M. J., & Forman, E. M. (2002). The psychotherapy dose-response effect and its implications for treatment delivery services. *Clinical Psychology: Science and Practice*, *9*(3), 329–343. https://doi.org/10.1093/clipsy.9.3.329

Härter, M., Schorr, S., & Schneider, F. (2017). *S3-Leitlinie/Nationale VersorgungsLeitlinie Unipolare Depression*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-52906-5

Hastie, T., Friedman, F., & Robert T. (2009). *Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Scholars Portal.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical Learning with Sparsity*. Chapman and Hall/CRC. https://doi.org/10.1201/b18401

Hayes, S. C., Strosahl, K. D., & Wilson, K. G. (2011). *Acceptance and commitment therapy: The process and practice of mindful change*. Guilford Press.

Hehlmann, M. I., Schwartz, B., Lutz, T., Gómez Penedo, J. M., Rubel, J. A., & Lutz, W. (2021). The Use of Digitally Assessed Stress Levels to Model Change Processes in CBT - A Feasibility Study on Seven Case Examples. *Frontiers in Psychiatry, 12,* 613085. https://doi.org/10.3389/fpsyt.2021.613085

Hemingway, H., Croft, P., Perel, P., Hayden, J. A., Abrams, K., Timmis, A., Briggs, A., Udumyan, R., Moons, K. G. M., Steyerberg, E. W., Roberts, I., Schroter, S., Altman, D. G., & Riley, R. D. (2013). Prognosis research strategy (PROGRESS) 1: A framework for researching clinical outcomes. *BMJ (Clinical Research Ed.), 346,* e5595. https://doi.org/10.1136/bmj.e5595

Herbert, J. D. (2003). The science and practice of empirically supported treatments. *Behavior Modification*, *27*(3), 412–430. https://doi.org/10.1177/0145445503027003008

Hilbert, K., Jacobi, T., Kunas, S. L., Elsner, B., Reuter, B., Lueken, U., & Kathmann, N. (2021). Identifying CBT non-response among OCD outpatients: A machine-learning approach. *Psychotherapy Research, 31*(1), 52–62. https://doi.org/10.1080/10503307.2020.1839140

Hilbert, K., Kunas, S. L., Lueken, U., Kathmann, N., Fydrich, T., & Fehm, L. (2020). Predicting cognitive behavioral therapy outcome in the outpatient sector based on clinical routine data: A machine learning approach. *Behaviour Research and Therapy, 124,* 103530. https://doi.org/10.1016/j.brat.2019.103530

Hill, C. E., & Lambert, M. J. (2004). Methodological issues in studying psychotherapy processes and outcomes. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (5th ed.). Wiley.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis, 15*(3), 199–236. https://doi.org/10.1093/pan/mpl013

Hofmann, S. G., & Barlow, D. H. (2014). Evidence-based psychological interventions and the common factors approach: The beginnings of a rapprochement? *Psychotherapy, 51*(4), 510–513. https://doi.org/10.1037/a0037045

Hofmann, S. G., Schulz, S. M., Meuret, A. E., Moscovitch, D. A., & Suvak, M. (2006). Sudden gains during therapy of social phobia. *Journal of Consulting and Clinical Psychology*, *74*(4), 687–697. https://doi.org/10.1037/0022-006X.74.4.687

Howard, K. I., Moras, K., Brill, P. L., Martinovich, Z., & Lutz, W. (1996). Evaluation of psychotherapy. Efficacy, effectiveness, and patient progress. *The American Psychologist*, *51*(10), 1059–1064. https://doi.org/10.1037/0003-066X.51.10.1059

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), e0140771. https://doi.org/10.1371/journal.pone.0140771

Insel, T. R. (2014). The NIMH Research Domain Criteria (RDoC) Project: Precision medicine for psychiatry. *The American Journal of Psychiatry*, *171*(4), 395–397. https://doi.org/10.1176/appi.ajp.2014.14020138

Jabbar, M. A., Deekshatulu, B. L., & Chandra, P. (2013). Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm. *Procedia Technology*, *10*, 85–94. https://doi.org/10.1016/j.protcy.2013.12.340

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*(1), 12–19. https://doi.org/10.1037/0022-006X.59.1.12

Jacobucci, R., & Grimm, K. J. (2020). Machine Learning and Psychological Research: The Unexplored Effect of Measurement. *Perspectives on Psychological Science*, *15*(3), 809–816. https://doi.org/10.1177/1745691620902467

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, *6*(5), 429–449. https://doi.org/10.3233/IDA-2002-6504

Kazdin, A. E. (2008). Evidence-based treatment and practice: New opportunities to bridge clinical research and practice, enhance the knowledge base, and improve patient care. *The American Psychologist*, *63*(3), 146–159. https://doi.org/10.1037/0003-066X.63.3.146

Kazdin, A. E. (2016). Closing the research–practice gap: How, why, and whether. *Clinical Psychology: Science and Practice*, *23*(2), 201–206. https://doi.org/10.1111/cpsp.12155

Kelly, M. A. R., Cyranowski, J. M., & Frank, E. (2007). Sudden gains in interpersonal psychotherapy for depression. *Behaviour Research and Therapy*, *45*(11), 2563–2572. https://doi.org/10.1016/j.brat.2007.07.007

Kelly, M. A. R., Roberts, J. E., & Ciesla, J. A. (2005). Sudden gains in cognitive behavioral treatment for depression: When do they occur and do they matter? *Behaviour Research and Therapy*, *43*(6), 703–714. https://doi.org/10.1016/j.brat.2004.06.002

Kessler, R. C., Bossarte, R. M., Luedtke, A., Zaslavsky, A. M., & Zubizarreta, J. R. (2019). Machine learning methods for developing precision treatment rules with observational data. *Behaviour Research and Therapy*, *120*, 103412. https://doi.org/10.1016/j.brat.2019.103412

Kessler, R. C., van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Cai, T., Ebert, D. D., Hwang, I., Li, J., Jonge, P. de, Nierenberg, A. A., Petukhova, M. V., Rosellini, A. J., Sampson, N. A., Schoevers, R. A., Wilcox, M. A., & Zaslavsky, A. M. (2016). Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports. *Molecular Psychiatry*, *21*(10), 1366–1371. https://doi.org/10.1038/mp.2015.198

Kessler, R. C., van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., La Brenner, Ebert, D. D., Jonge, P. de, Nierenberg, A. A., Rosellini, A. J., & Sampson, N. A. (2017). Using patient self-reports to study heterogeneity of treatment effects in major depressive disorder. *Epidemiology and Psychiatric Sciences*, *26*(1), 22–36. https://doi.org/10.1017/S2045796016000020

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging Artificial Intelligence Applications in Computer Engineering*, *160*(1), 3–24.

Koutsouleris, N., Kambeitz-Ilankovic, L., Ruhrmann, S., Rosen, M., Ruef, A., Dwyer, D. B., Paolini, M., Chisholm, K., Kambeitz, J., Haidl, T., Schmidt, A., Gillam, J., Schultze-Lutter, F., Falkai, P., Reiser, M., Riecher-Rössler, A., Upthegrove, R., Hietala, J., Salokangas, R. K. R., . . . Borgwardt, S. (2018). Prediction Models of Functional Outcomes for Individuals in the Clinical High-Risk State for Psychosis or With Recent-Onset Depression: A Multimodal, Multisite Machine Learning Analysis. *JAMA Psychiatry*, *75*(11), 1156–1172. https://doi.org/10.1001/jamapsychiatry.2018.2165

Kuhn, M. (2019). *caret: Classification and Regression Training*. R package version 6.0-84.

Laios, A., Gryparis, A., DeJong, D., Hutson, R., Theophilou, G., & Leach, C. (2020). Predicting complete cytoreduction for advanced ovarian cancer patients using nearest-neighbor models. *Journal of Ovarian Research*, *13*(1), 117. https://doi.org/10.1186/s13048-020-00700-0

Lambert, M. J., Bailey, R., Kimball, K., Shimokawa, K., Harmon, S. C., & Slade, K. (2007). *Clinical support tools manual-brief version-40*. OQ Measures.

Laska, K. M., Gurman, A. S., & Wampold, B. E. (2014). Expanding the lens of evidence-based practice in psychotherapy: A common factors perspective. *Psychotherapy*, *51*(4), 467–481. https://doi.org/10.1037/a0034332

Lazar, S. G. (2014). The cost-effectiveness of psychotherapy for the major psychiatric diagnoses. *Psychodynamic Psychiatry*, *42*(3), 423–457. https://doi.org/10.1521/pdps.2014.42.3.423

Lee, Y., Ragguett, R.-M., Mansur, R. B., Boutilier, J. J., Rosenblat, J. D., Trevizol, A., Brietzke, E., Lin, K., Pan, Z., Subramaniapillai, M., Chan, T. C. Y., Fus, D., Park, C., Musial, N., Zuckerman, H., Chen, V. C.-H., Ho, R., Rong, C., & McIntyre, R. S. (2018). Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. *Journal of Affective Disorders*, *241*, 519–532. https://doi.org/10.1016/j.jad.2018.08.073

Lemmens, L. H. J. M., DeRubeis, R. J., Arntz, A., Peeters, F. P. M. L., & Huibers, M. J. H. (2016). Sudden gains in Cognitive Therapy and Interpersonal Psychotherapy for adult depression. *Behaviour Research and Therapy*, *77*, 170–176. https://doi.org/10.1016/j.brat.2015.12.014

Linehan, M. M. (1993). *Skills training manual for treating borderline personality disorder*. Guilford Press.

Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C., Ledsam, J. R., Schmid, M. K., Balaskas, K., Topol, E. J., Bachmann, L. M., Keane, P. A., & Denniston, A. K. (2019). A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. *The Lancet. Digital Health*, *1*(6), e271-e297. https://doi.org/10.1016/S2589-7500(19)30123-2

Lorenzo-Luaces, L., DeRubeis, R. J., van Straten, A., & Tiemens, B. (2017). A prognostic index (PI) as a moderator of outcomes in the treatment of depression: A proof of concept combining multiple variables to inform risk-stratified stepped care models. *Journal of Affective Disorders*, *213*, 78–85. https://doi.org/10.1016/j.jad.2017.02.010

Lorenzo-Luaces, L., German, R. E., & DeRubeis, R. J. (2015). It's complicated: The relation between cognitive change procedures, cognitive change, and symptom change in cognitive therapy for depression. *Clinical Psychology Review*, *41*, 3–15. https://doi.org/10.1016/j.cpr.2014.12.003

Luborsky, L., & Singer, B. (1975). Comparative studies of psychotherapies. Is it true that "everywon has one and all must have prizes"? *Archives of General Psychiatry*, *32*(8), 995–1008. https://doi.org/10.1001/archpsyc.1975.01760260059004

Lutz, W. (2002). Patient-Focused Psychotherapy Research and Individual Treatment Progress as Scientific Groundwork for an Empirically Based Clinical Practice. *Psychotherapy Research*, *12*(3), 251–272. https://doi.org/10.1080/713664389

Lutz, W., Arndt, A., Rubel, J., Berger, T., Schröder, J., Späth, C., Meyer, B., Greiner, W., Gräfe, V., Hautzinger, M., Fuhr, K., Rose, M., Nolte, S., Löwe, B., Hohagen, F., Klein, J. P., & Moritz, S. (2017). Defining and Predicting Patterns of Early Response in a Web-Based Intervention for Depression. *Journal of Medical Internet Research*, *19*(6), e206. https://doi.org/10.2196/jmir.7367

Lutz, W., Castonguay, L. G., Lambert, M. J., & Barkham, M. (2021). Traditions and new beginnings: Historical and current perspectives on research in psychotherapy and behavior change. In M. Barkham, W. Lutz, & L. G. Castonguay (Eds.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (50th ed., pp. 3–18). John Wiley & Sons, Inc.

Lutz, W., de Jong, K., & Rubel, J. (2015). Patient-focused and feedback research in psychotherapy: Where are we and where do we want to go? *Psychotherapy Research*, *25*(6), 625–632. https://doi.org/10.1080/10503307.2015.1079661

Lutz, W., de Jong, K., Rubel, J. A., & Delgadillo, J. (2021). Measuring, predicting, and tracking change in psychotherapy. In M. Barkham, W. Lutz, & L. G. Castonguay (Eds.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (50th ed., pp. 89–133). John Wiley & Sons, Inc.

Lutz, W., Deisenhofer, A.-K., Rubel, J., Bennemann, B., Giesemann, J., Poster, K., & Schwartz, B. (2022). Prospective evaluation of a clinical decision support system in psychological therapy. *Journal of Consulting and Clinical Psychology, 90*(1), 90-106. https://doi.org/10.1037/ccp0000642

Lutz, W., Ehrlich, T., Rubel, J., Hallwachs, N., Röttger, M.-A., Jorasz, C., Mocanu, S., Vocks, S., Schulte, D., & Tschitsaz-Stucki, A. (2013). The ups and downs of psychotherapy: Sudden gains and sudden losses identified with session reports. *Psychotherapy Research*, *23*(1), 14–24. https://doi.org/10.1080/10503307.2012.693837

Lutz, W., Hofmann, S. G., Rubel, J., Boswell, J. F., Shear, M. K., Gorman, J. M., Woods, S. W., & 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

Lutz, W., Lambert, M. J., Harmon, S. C., Tschitsaz, A., Schürch, E., & Stulz, N. (2006). The probability of treatment success, failure and duration—what can be learned from empirical data to support decision making in clinical practice? *Clinical Psychology & Psychotherapy, 13*(4), 223–232. https://doi.org/10.1002/cpp.496

Lutz, W., Leach, C., Barkham, M., Lucock, M., Stiles, W. B., Evans, C., Noble, R., & Iveson, S. (2005). Predicting change for individual psychotherapy clients on the basis of their nearest neighbors. *Journal of Consulting and Clinical Psychology, 73*(5), 904–913. https://doi.org/10.1037/0022-006X.73.5.904

Lutz, W., Leon, S. C., Martinovich, Z., Lyons, J. S., & Stiles, W. B. (2007). Therapist effects in outpatient psychotherapy: A three-level growth curve approach. *Journal of Counseling Psychology, 54*(1), 32–39. https://doi.org/10.1037/0022-0167.54.1.32

Lutz, W., Martinovich, Z., & Howard, K. I. (1999). Patient profiling: An application of random coefficient regression models to depicting the response of a patient to outpatient psychotherapy. *Journal of Consulting and Clinical Psychology, 67*(4), 571–577. https://doi.org/10.1037/0022-006X.67.4.571

Lutz, W., Rubel, J. A., Schiefele, A.-K., Zimmermann, D., Böhnke, J. R., & Wittmann, W. W. (2015). Feedback and therapist effects in the context of treatment outcome and treatment length. *Psychotherapy Research, 25*(6), 647–660. https://doi.org/10.1080/10503307.2015.1053553

Lutz, W., Rubel, J. A., Schwartz, B., Schilling, V., & Deisenhofer, A.-K. (2019). Towards integrating personalized feedback research into clinical practice: Development of the Trier Treatment Navigator (TTN). *Behaviour Research and Therapy*, *120*, 103438. https://doi.org/10.1016/j.brat.2019.103438

Lutz, W., Saunders, S. M., Leon, S. C., Martinovich, Z., Kosfelder, J., Schulte, D., Grawe, K., & Tholen, S. (2006). Empirically and clinically useful decision making in psychotherapy: Differential predictions with treatment response models. *Psychological Assessment, 18*(2), 133. https://doi.org/10.1037/1040-3590.18.2.133

Lutz, W., Schwartz, B., & Delgadillo, J. (2022). Measurement-Based and Data-Informed Psychological Therapy. *Annual Review of Clinical Psychology*, *18*, 71–98. https://doi.org/10.1146/annurev-clinpsy-071720-014821

Lutz, W., Schwartz, B., Hofmann, S. G., Fisher, A. J., Husen, K., & Rubel, J. A. (2018). Using network analysis for the prediction of treatment dropout in patients with mood and anxiety disorders: A methodological proof-of-concept study. *Scientific Reports*, *8*(1), 7819. https://doi.org/10.1038/s41598-018-25953-0

Lutz, W., Schwartz, B., Martín Gómez Penedo, J., Boyle, K., & Deisenhofer, A.-K. (2020). Working Towards the Development and Implementation of Precision Mental Healthcare: An Example. *Administration and Policy in Mental Health*, *47*(5), 856–861. https://doi.org/10.1007/s10488-020-01053-y

Lutz, W., Tholen, S., Schürch, E., & Berking, M. (2006). Reliabilität von Kurzformen gängiger psychometrischer Instrumente zur Evaluation des therapeutischen Fortschritts in Psychotherapie und Psychiatrie. *Diagnostica*, *52*(1), 11–25. https://doi.org/10.1026/0012-1924.52.1.11

Lutz, W., Zimmermann, D., Müller, V. N. L. S., Deisenhofer, A.-K., & Rubel, J. A. (2017). Randomized controlled trial to evaluate the effects of personalized prediction and adaptation tools on treatment outcome in outpatient psychotherapy: study protocol. *BMC Psychiatry*, *17*(1), 1–11. https://doi.org/10.1186/s12888-017-1464-2

Macdonald, J., & Mellor-Clark, J. (2015). Correcting Psychotherapists' Blindsidedness: Formal Feedback as a Means of Overcoming the Natural Limitations of Therapists. *Clinical Psychology & Psychotherapy*, *22*(3), 249–257. https://doi.org/10.1002/cpp.1887

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). Statistical and Machine Learning forecasting methods: Concerns and ways forward. *PloS One*, *13*(3), e0194889. https://doi.org/10.1371/journal.pone.0194889

Mayer, Z. (2019). *A Brief Introduction to caretEnsemble*. https://cran.r-project.org/web/packages/caretEnsemble/vignettes/caretEnsemble-intro.html

McAleavey, A. A., & Castonguay, L. G. (2015). The Process of Change in Psychotherapy: Common and Unique Factors. In O. C. Gelo, A. Pritz, & B. Rieken (Eds.), *Psychotherapy Research* (pp. 293–310). Springer Vienna. https://doi.org/10.1007/978-3-7091-1382-0_15

McAleavey, A. A., Youn, S. J., Xiao, H., Castonguay, L. G., Hayes, J. A., & Locke, B. D. (2019). Effectiveness of routine psychotherapy: Method matters. *Psychotherapy Research*, *29*(2), 139–156. https://doi.org/10.1080/10503307.2017.1395921

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., & Hamprecht, F. A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics*, *10*, 213. https://doi.org/10.1186/1471-2105-10-213

Mohr, D. C., Ho, J., Hart, T. L., Baron, K. G., Berendsen, M., Beckner, V., Cai, X., Cuijpers, P., Spring, B., & Kinsinger, S. W. (2014). Control condition design and implementation features in controlled trials: a meta-analysis of trials evaluating psychotherapy for depression. *Translational Behavioral Medicine*, *4*(4), 407–423. https://doi.org/10.1007/s13142-014-0262-3

Mufson, L. (2011). *Interpersonal psychotherapy for depressed adolescents*. Guilford Press.

Müller, A. C., & Guido, S. (2016). *Introduction to Machine Learnin with Python*. Oreilly & Associates Inc.

Munder, T., Flückiger, C., Leichsenring, F., Abbass, A. A., Hilsenroth, M. J., Luyten, P., Rabung, S., Steinert, C., & Wampold, B. E. (2019). Is psychotherapy effective? A re-analysis of treatments for depression. *Epidemiology and Psychiatric Sciences*, *28*(3), 268–274. https://doi.org/10.1017/S2045796018000355

Nahum, D., Alfonso, C. A., & Sönmez, E. (2019). Common Factors in Psychotherapy. In A. Javed & K. N. Fountoulakis (Eds.), *Advances in Psychiatry* (pp. 471–481). Springer International Publishing. https://doi.org/10.1007/978-3-319-70554-5_29

National Research Council. (2011). *Toward precision medicine: Building a knowledge network for biomedical research and a new taxonomy of disease*. National Academies Press.

Norcross, J. C., VandenBos, G. R., & Freedheim, D. K. (2011). *History of psychotherapy: Continuity and change*. American Psychological Association.

Norcross, J. C., & Wampold, B. E. (2011). What works for whom: Tailoring psychotherapy to the person. *Journal of Clinical Psychology*, *67*(2), 127–132. https://doi.org/10.1002/jclp.20764

Ok, A. O., Akar, O., & Gungor, O. (2012). Evaluation of random forest method for agricultural crop classification. *European Journal of Remote Sensing*, *45*(1), 421–432. https://doi.org/10.5721/EuJRS20124535

Ong, W. T., Murphy, D., & Joseph, S. (2020). Unnecessary and incompatible: a critical response to Cooper and McLeod’s conceptualization of a pluralistic framework for person-centered therapy. *Person-Centered & Experiential Psychotherapies*, *19*(2), 168–182. https://doi.org/10.1080/14779757.2020.1717987

Pavlou, M., Ambler, G., Seaman, S., Iorio, M. de, & Omar, R. Z. (2016). Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. *Statistics in Medicine*, *35*(7), 1159–1177. https://doi.org/10.1002/sim.6782

Pearson, R., Pisner, D., Meyer, B., Shumake, J., & Beevers, C. G. (2019). A machine learning ensemble to predict treatment outcomes following an Internet intervention for depression. *Psychological Medicine*, *49*(14), 2330–2341. https://doi.org/10.1017/S003329171800315X

Perlis, R. H. (2016). Abandoning personalization to get to precision in the pharmacotherapy of depression. *World Psychiatry*, *15*(3), 228–235. https://doi.org/10.1002/wps.20345

Ramzai, J. (2019). *Simple guide for ensemble learning methods*. https://towardsdatascience.com/simple-guide-for-ensemble-learning-methods-d87cc68705a2

Riley, R. D., Hayden, J. A., Steyerberg, E. W., Moons, K. G. M., Abrams, K., Kyzas, P. A., Malats, N., Briggs, A., Schroter, S., Altman, D. G., & Hemingway, H. (2013). Prognosis Research Strategy (PROGRESS) 2: Prognostic factor research. *PLoS Medicine*, *10*(2), e1001380. https://doi.org/10.1371/journal.pmed.1001380

Riley, R. D., Snell, K. I. E., Ensor, J., Burke, D. L., Harrell, F. E., Moons, K. G. M., & Collins, G. S. (2019). Minimum sample size for developing a multivariable prediction model: Part I - Continuous outcomes. *Statistics in Medicine*, *38*(7), 1262–1275. https://doi.org/10.1002/sim.7993

Rokach, L. (2010). *Pattern Classification Using Ensemble Methods (Series in machine perception and artificial intelligence v. 75)*. World Scientific.

Rosellini, A. J., Dussaillant, F., Zubizarreta, J. R., Kessler, R. C., & Rose, S. (2018). Predicting posttraumatic stress disorder following a natural disaster. *Journal of Psychiatric Research*, *96*, 15–22. https://doi.org/10.1016/j.jpsychires.2017.09.010

Rothwell, P. M. (2005). External validity of randomised controlled trials: “to whom do the results of this trial apply?”. *The Lancet*, *365*(9453), 82–93. https://doi.org/10.1016/S0140-6736(04)17670-8

Rubel, J. A., Zilcha-Mano, S., Giesemann, J., Prinz, J., & Lutz, W. (2020). Predicting personalized process-outcome associations in psychotherapy using machine learning approaches-A demonstration. *Psychotherapy Research*, *30*(3), 300–309. https://doi.org/10.1080/10503307.2019.1597994

Rudin, C., & Carlson, D. (2019, June 4). *The Secrets of Machine Learning: Ten Things You Wish You Had Known Earlier to be More Effective at Data Analysis*. http://arxiv.org/pdf/1906.01998v1

Sauerbrei, W., Perperoglou, A., Schmid, M., Abrahamowicz, M., Becher, H., Binder, H., Dunkler, D., Harrell, F. E., Royston, P., & Heinze, G. (2020). State of the art in selection of variables and functional forms in multivariable analysis-outstanding issues. *Diagnostic and Prognostic Research*, *4*, 3. https://doi.org/10.1186/s41512-020-00074-3

Saunders, R., Buckman, J. E. J., & Pilling, S. (2020). Latent variable mixture modelling and individual treatment prediction. *Behaviour Research and Therapy*, *124*, 103505. https://doi.org/10.1016/j.brat.2019.103505

Saxon, D., Barkham, M., Foster, A., & Parry, G. (2017). The Contribution of Therapist Effects to Patient Dropout and Deterioration in the Psychological Therapies. *Clinical Psychology & Psychotherapy*, *24*(3), 575–588. https://doi.org/10.1002/cpp.2028

Schaffrath, J., Weinmann-Lutz, B., & Lutz, W. (2022). The Trier Treatment Navigator (TTN) in action: Clinical case study on data-informed psychological therapy. *Journal of Clinical Psychology*, *78*(10), 2016–2028. https://doi.org/10.1002/jclp.23362

Schwartz, B., Cohen, Z. D., Rubel, J. A., Zimmermann, D., Wittmann, W. W., & Lutz, W. (2021). Personalized treatment selection in routine care: Integrating machine learning and statistical algorithms to recommend cognitive behavioral or psychodynamic therapy. *Psychotherapy Research*, *31*(1), 33–51. https://doi.org/10.1080/10503307.2020.1769219

Shalom, J. G., & Aderka, I. M. (2020). A meta-analysis of sudden gains in psychotherapy: Outcome and moderators. *Clinical Psychology Review*, *76*, 101827. https://doi.org/10.1016/j.cpr.2020.101827

Shapiro, F. (2001). *Eye movement desensitization and reprocessing (EMDR): Basic principles, protocols, and procedures*. Guilford Press.

Sharf, J., Primavera, L. H., & Diener, M. J. (2010). Dropout and therapeutic alliance: A meta-analysis of adult individual psychotherapy. *Psychotherapy: Theory, Research, Practice, Training*, *47*(4), 637. https://doi.org/10.1037/a0021175

Shean, G. (2014). Limitations of Randomized Control Designs in Psychotherapy Research. *Advances in Psychiatry*, *2014*, 1–5. https://doi.org/10.1155/2014/561452

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *The American Psychologist*, *32*(9), 752–760. https://doi.org/10.1037/0003-066X.32.9.752

Spengler, P. M., White, M. J., Ægisdóttir, S., Maugherman, A. S., Anderson, L. A., Cook, R. S., Nichols, C. N., Lampropoulos, G. K., Walker, B. S., Cohen, G. R., & Rush, J. D. (2009). The Meta-Analysis of Clinical Judgment Project. *The Counseling Psychologist*, *37*(3), 350–399. https://doi.org/10.1177/0011000006295149

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin*, *144*(12), 1325–1346. https://doi.org/10.1037/bul0000169

Stekhoven, D. J., & Bühlmann, P. (2012). Missforest--non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

Stepankova Georgi, H., Horakova Vlckova, K., Lukavsky, J., Kopecek, M., & Bares, M. (2019). Beck Depression Inventory-II: Self-report or interview-based administrations show different results in older persons. *International Psychogeriatrics*, *31*(5), 735–742. https://doi.org/10.1017/S1041610218001187

Steyerberg, E. W., Moons, K. G. M., van der Windt, D. A., Hayden, J. A., Perel, P., Schroter, S., Riley, R. D., Hemingway, H., & Altman, D. G. (2013). Prognosis Research Strategy (PROGRESS) 3: Prognostic model research. *PLoS Medicine, 10*(2), e1001381. https://doi.org/10.1371/journal.pmed.1001381

Stiles, W. B., Leach, C., Barkham, M., Lucock, M., Iveson, S., Shapiro, D. A., Iveson, M., & Hardy, G. E. (2003). Early sudden gains in psychotherapy under routine clinic conditions: Practice-based evidence. *Journal of Consulting and Clinical Psychology*, *71*(1), 14–21. https://doi.org/10.1037/0022-006X.71.1.14

Strauß, B. (2021). Was ist Psychotherapie? In W. Rief, E. Schramm, & B. Strauß (Eds.), *Psychotherapie: Ein kompetenzorientiertes Lehrbuch* (pp. 459–464). Elsevier Health Sciences.

Strauss, B. M., Lutz, W., Steffanowski, A., Wittmann, W. W., Böhnke, J. R., Rubel, J., Scheidt, C. E., Caspar, F., Vogel, H., Altmann, U., Steyer, R., Zimmermann, A., Bruckmayer, E., Heymann, F. von, Kramer, D., & Kirchmann, H. (2015). Benefits and challenges in practice-oriented psychotherapy research in Germany: The TK and the QS-PSY-BAY projects of quality assurance in outpatient psychotherapy. *Psychotherapy Research*, *25*(1), 32–51. https://doi.org/10.1080/10503307.2013.856046

Stuart, E. A., Rubin, D. B., & Osborne, J. (2004). Matching methods for causal inference: Designing observational studies. *Harvard University Department of Statistics Mimeo*.

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

Symons, M., Feeney, G. F. X., Gallagher, M. R., Young, R. M., & Connor, J. P. (2020). Predicting alcohol dependence treatment outcomes: A prospective comparative study of clinical psychologists versus 'trained' machine learning models. *Addiction*, *115*(11), 2164–2175. https://doi.org/10.1111/add.15038

Tang, T. Z., & DeRubeis, R. J. (1999). Sudden gains and critical sessions in cognitive-behavioral therapy for depression. *Journal of Consulting and Clinical Psychology*, *67*(6), 894–904. https://doi.org/10.1037/0022-006X.67.6.894

Tang, T. Z., DeRubeis, R. J., Beberman, R., & Pham, T. (2005). Cognitive changes, critical sessions, and sudden gains in cognitive-behavioral therapy for depression. *Journal of Consulting and Clinical Psychology*, *73*(1), 168–172. https://doi.org/10.1037/0022-006X.73.1.168

Taubitz, F.-S., Büdenbender, B., & Alpers, G. W. (2022). What the future holds: Machine learning to predict success in psychotherapy. *Behaviour Research and Therapy*, *156*, 104116. https://doi.org/10.1016/j.brat.2022.104116

van Bronswijk, S. C., Lemmens, L. H. J. M., Keefe, J. R., Huibers, M. J. H., DeRubeis, R. J., & Peeters, F. P. M. L. (2019). A prognostic index for long-term outcome after successful acute phase cognitive therapy and interpersonal psychotherapy for major depressive disorder. *Depression and Anxiety, 36*(3), 252–261. https://doi.org/10.1002/da.22868

van Buuren, S. (2021). *Multivariate Imputation by Chained Equations*. R package version 3.14.0.

Velten, J., Margraf, J., Benecke, C., Berking, M., In-Albon, T., Tania Lincoln, Lutz, W., Schlarb, A., Schöttke, H., Willutzki, U., & Jürgen Hoyer (2017). Methodenpapier zur Koordination der Datenerhebung und -auswertung an Hochschul- und Ausbildungsambulanzen für Psychotherapie (KODAP). *Zeitschrift Für Klinische Psychologie Und Psychotherapie*, *46*(3), 169–175. https://doi.org/10.1026/1616-3443/a000431

Vittengl, J. R., Clark, L. A., Thase, M. E., & Jarrett, R. B. (2015). Detecting Sudden Gains during Treatment of Major Depressive Disorder: Cautions from a Monte Carlo Analysis. *Current Psychiatry Reviews*, *11*(1), 19–31. https://doi.org/10.2174/1573400510666140929195441

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition. *Perspectives on Psychological Science*, *4*(3), 274–290. https://doi.org/10.1111/j.1745-6924.2009.01125.x

Wahl, I., Löwe, B., Bjorner, J. B., Fischer, F., Langs, G., Voderholzer, U., Aita, S. A., Bergemann, N., Brähler, E., & Rose, M. (2014). Standardization of depression measurement: A common metric was developed for 11 self-report depression measures. *Journal of Clinical Epidemiology*, *67*(1), 73–86. https://doi.org/10.1016/j.jclinepi.2013.04.019

Wakefield, S., Kellett, S., Simmonds-Buckley, M., Stockton, D., Bradbury, A., & Delgadillo, J. (2021). Improving Access to Psychological Therapies (IAPT) in the United Kingdom: A systematic review and meta-analysis of 10-years of practice-based evidence. *The British Journal of Clinical Psychology*, *60*(1), 1–37. https://doi.org/10.1111/bjc.12259

Wampold, B. E. (2001). *The Great Psychotherapy Debate: Models, methods, and findings.* Lawrence Erlbaum Associates Publishers.

Wampold, B. E. (2015). How important are the common factors in psychotherapy? An update. *World Psychiatry*, *14*(3), 270–277. https://doi.org/10.1002/wps.20238

Wampold, B. E. (2019). Research on the effectiveness of psychotherapy. In B. E. Wampold (Ed.), *The basics of psychotherapy: An introduction to theory and practice (2nd ed.)* (pp. 67–89). American Psychological Association. https://doi.org/10.1037/0000117-004

Wampold, B. E., & Imel, Z. E. (2015). *The Great Psychotherapy Debate: The evidence for what makes psychotherapy work.* Routledge. https://doi.org/10.4324/9780203582015

Wardenaar, K. J., van Loo, H. M., Cai, T., Fava, M., Gruber, M. J., Li, J., Jonge, P. de, Nierenberg, A. A., Petukhova, M. V., Rose, S., Sampson, N. A., Schoevers, R. A., Wilcox, M. A., Alonso, J., Bromet, E. J., Bunting, B., Florescu, S. E., Fukao, A., Gureje, O., . . . Kessler, R. C. (2014). The effects of co-morbidity in defining major depression subtypes associated with long-term course and severity. *Psychological Medicine*, *44*(15), 3289–3302. https://doi.org/10.1017/S0033291714000993

Webb, C. A., Cohen, Z. D., Beard, C., Forgeard, M., Peckham, A. D., & Björgvinsson, T. (2020). Personalized prognostic prediction of treatment outcome for depressed patients in a naturalistic psychiatric hospital setting: A comparison of machine learning approaches. *Journal of Consulting and Clinical Psychology*, *88*(1), 25–38. https://doi.org/10.1037/ccp0000451

Wells, J. E., Browne, M. O., Aguilar-Gaxiola, S., Al-Hamzawi, A., Alonso, J., Angermeyer, M. C., Bouzan, C., Bruffaerts, R., Bunting, B., Caldas-de-Almeida, J. M., Girolamo, G. de, Graaf, R. de, Florescu, S., Fukao, A., Gureje, O., Hinkov, H. R., Hu, C., Hwang, I., Karam, E. G., . . . Kessler, R. C. (2013). Drop out from out-patient mental healthcare in the World Health Organization's World Mental Health Survey initiative. *The British Journal of Psychiatry*, *202*(1), 42–49. https://doi.org/10.1192/bjp.bp.112.113134

Wetzel, E., & Böhnke, J. R. (2017). Differential Item Functioning. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *Encyclopedia of Personality and Individual Differences* (pp. 1–5). Springer International Publishing. https://doi.org/10.1007/978-3-319-28099-8_1297-1

Whiteford, H. A., Degenhardt, L., Rehm, J., Baxter, A. J., Ferrari, A. J., Erskine, H. E., Charlson, F. J., Norman, R. E., Flaxman, A. D., & Johns, N. (2013). Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *The Lancet*, *382*(9904), 1575–1586. https://doi.org/10.1016/S0140-6736(13)61611-6

Wiedemann, M., Stott, R., Nickless, A., Beierl, E. T., Wild, J., Warnock-Parkes, E., Grey, N., Clark, D. M., & Ehlers, A. (2020). Cognitive processes associated with sudden gains in cognitive therapy for posttraumatic stress disorder in routine care. *Journal of Consulting and Clinical Psychology*, *88*(5), 455.

Wilkinson, J., Arnold, K. F., Murray, E. J., van Smeden, M., Carr, K., Sippy, R., Kamps, M. de, Beam, A., Konigorski, S., Lippert, C., Gilthorpe, M. S., & Tennant, P. W. G. (2020). Time to reality check the promises of machine learning-powered precision medicine. *The Lancet. Digital Health*, *2*(12), e677-e680. https://doi.org/10.1016/S2589-7500(20)30200-4

Wittchen, H.-U., Wunderlich, U., Gruschwitz, S., & Zaudig, M. (1997). *SKID I. Strukturiertes Klinisches Interview für DSM-IV. Achse I: Psychische Störungen. Interviewheft und Beurteilungsheft. Eine deutschsprachige, erweiterte Bearb. d. amerikanischen Originalversion des SKID I*. Hogrefe.

World Health Organization. (2008). *The Global burden of disease: 2004 update*. World Health Organization.

Wucherpfennig, F., Rubel, J. A., Hofmann, S. G., & Lutz, W. (2017). Processes of change after a sudden gain and relation to treatment outcome-Evidence for an upward spiral. *Journal of Consulting and Clinical Psychology*, *85*(12), 1199–1210. https://doi.org/10.1037/ccp0000263

Wucherpfennig, F., Rubel, J. A., Hollon, S. D., & Lutz, W. (2017). Sudden gains in routine care cognitive behavioral therapy for depression: A replication with extensions. *Behaviour Research and Therapy*, *89*, 24–32. https://doi.org/10.1016/j.brat.2016.11.003

Xiao, Y., Griffin, M. P., Lake, D. E., & Moorman, J. R. (2010). Nearest-neighbor and logistic regression analyses of clinical and heart rate characteristics in the early diagnosis of neonatal sepsis. *Medical Decision Making*, *30*(2), 258–266. https://doi.org/10.1177/0272989X09337791

Yalom, I. D., & Leszcz, M. (2020). *The theory and practice of group psychotherapy*. Basic books.

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Zhou, Z.-H. (2012). *Ensemble methods: Foundations and algorithms*. *Chapman & Hall/CRC machine learning & pattern recognition series*. CRC Press.

Zilcha-Mano, S., Errázuriz, P., Yaffe-Herbst, L., German, R. E., & DeRubeis, R. J. (2019). Are there any robust predictors of "sudden gainers," and how is sustained improvement in treatment outcome achieved following a gain? *Journal of Consulting and Clinical Psychology*, *87*(6), 491–500. https://doi.org/10.1037/ccp0000401

Zimmermann, D., Rubel, J., Page, A. C., & Lutz, W. (2017). Therapist Effects on and Predictors of Non-Consensual Dropout in Psychotherapy. *Clinical Psychology & Psychotherapy*, *24*(2), 312–321. https://doi.org/10.1002/cpp.2022

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

10 Original Publications

Study I

Bennemann, B., Schwartz, B., Giesemann, J., & Lutz, W. (2022). Predicting patients who will drop out of out-patient psychotherapy using machine learning algorithms. *The British Journal of Psychiatry*, *220*(4), 192-201. https://doi.org/10.1192/bjp.2022.17

Author Contributions

W. L. provided the data set and advised on the structure and design of the study. B. B. developed the idea and performed the analyses. B. S., J. G. and W. L. advised on statistical issues. All authors were involved in interpreting data, drafting the work or critically revising it for important intellectual content. All authors provided final approval of the submitted version of the manuscript.

Predicting Patients who will Drop out of Outpatient Psychotherapy Using Machine Learning Algorithms

Björn Bennemann [a*], Brian Schwartz [a], Julia Giesemann [a], Wolfgang Lutz [a],

[a] University of Trier, Universitätsring 15, 54296 Trier, Germany

Reference:

Bennemann, B., Schwartz, B., Giesemann, J., & Lutz, W. (2022). Predicting patients who will drop out of out-patient psychotherapy using machine learning algorithms. *The British Journal of Psychiatry*. https://doi.org/10.1192/bjp.2022.17

Abstract

Background: About thirty percent of patients drop out of cognitive behavioral therapy (CBT), which has implications for psychiatric and psychological treatment. Findings concerning dropout remain heterogeneous. **Aims:** This paper aims to compare different machine learning (ML) algorithms using nested cross-validation, evaluate their benefit in naturalistic settings, and identify the best model as well as the most important variables. **Method:** The data set consisted of 2543 outpatients treated with CBT. Assessment took place before session one. Twenty-one algorithms and ensembles were compared. Two parameters (Brier score, area under the curve (AUC)) were used for evaluation. **Results:** The best model was an ensemble that used random forest and nearest neighbor modeling. During the training process, it was significantly better than generalized linear modeling (GLM) (Brier score: *d* = –2.93 [-3.95; -1.90]; AUC: *d* = 0.59 [0.33; 1.06]). In the holdout sample, the ensemble was able to correctly identify 63.4% of cases as dropout/regular, while the GLM only identified 46.2% correctly. The most important predictors were lower education, lower scores on the Personality Style and Disorder Inventory (PSSI) compulsive scale, younger age, higher scores on the PSSI negativistic and PSSI antisocial scale as well as on the Brief Symptom Inventory (BSI) additional scale (mean of the four additional items) and BSI overall scale. **Conclusions:** ML improves dropout predictions. However, not all algorithms are suited to naturalistic datasets and binary events. Tree-based and boosted algorithms including a variable selection process seem well-suited, while more advanced algorithms such as neural networks do not.

Keywords: dropout; machine learning; algorithms; ensembles; variable selection

Predicting Patients who will Drop out of Outpatient Psychotherapy Using Machine Learning Algorithms

Cognitive behavioral therapy (CBT) is an effective treatment for mental health problems.[1] However, approximately one in five patients drop out of treatment,[2] leading to many problems such as the lack of an adequate treatment.[3,4]

Because of these negative consequences, identifying patients at a high risk of dropping out could lead to the development of clinical support tools that minimize the risk of dropout in individual patients.[5,6] However, findings from past studies examining CBT treatments have been heterogeneous with only younger age and lower education level being consistently associated with dropout.[7–9] Most studies used small samples and heterogeneous methods. Therefore, an increase of statistical precision and large datasets are necessary to reliably identify patients at risk of dropping out of therapy.

Methodological Developments

Over the last years, machine learning (ML) approaches in particular have had a large impact on prediction modeling and on the most recent debate about the implementation of personalized or precision medicine concepts in mental health.[10,11] ML has been applied in various prediction contexts,[12–15] taking advantage of the ability to capture non-linear relations.[16]

Nevertheless, ML does not always have an advantage over more *traditional* methods,[17] indicating that personalized medical care faces serious challenges that cannot be addressed through algorithmic complexity alone.[18] It remains unclear which ML methods are most suited to data from an outpatient CBT setting and whether previous findings can be generalized to this context.[15,18] Further, to our knowledge, there is no study that investigated the use of ML algorithms for the prediction of a binary event in a naturalistic setting. For this reason, we pursued two aims in this study:

1.) Various ML algorithms will be systematically compared with regard to their personalized dropout predictions and under routine care outpatient CBT conditions.

2.) Findings from these comparisons will be used to generate a clinically useful dropout prediction model that can be used in clinical practice before the first session has occurred.

Method

Patients and Treatment

The analyses were based on a sample comprising 2543 patients treated at the University of Trier Outpatient CBT Clinic in Southwest Germany between 2007 and 2021. Patients were included when they had completed a battery of questionnaires at intake, had begun therapy after the diagnostic phase (i.e., completed at least three sessions) and completed (i.e., consensual termination) or dropped out of treatment (see supplemental materials 1 for a flowchart of selected patients). Written informed consent was obtained from all patients. The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008. All procedures involving human patients were approved by the ethics committee of the University of Trier.

All patient data collected from 2007 to 2017 were used for the model generating process (training sample) while the remaining data were used for testing purposes (holdout sample). Therapy took place once a week (range = 3–113 sessions). When patients dropped out, the number of sessions was significantly lower than when they completed therapy ($M_{dropout}$ = 17.2 sessions; $M_{completion}$ = 43.4 sessions; $t(2541) = 33.46$; $p < .001$; Cohen’s d = 1.49). This held true for the training sample ($M_{dropout}$ = 17.5 sessions; $M_{completion}$ = 44.3 sessions; $t(2041) = 31.36$; $p < .001$; Cohen’s d = 1.51) and for the holdout sample ($M_{dropout}$ = 16.4 sessions; $M_{completion}$ = 39.0 sessions; $t(498) = 12.51$; $p < .001$; Cohen’s d = 1.21).

Diagnoses were based on the German version of the Structured Clinical Interview for Axis I DSM–IV Disorders—Patient Edition (SCID-I)[19] and the International Diagnostic Checklist for Personality Disorders (IDCL-P).[20] Interviews were conducted by intensively trained independent clinicians before actual therapy began. All sessions were videotaped; interviews and diagnoses were discussed in expert consensus teams that included four senior clinicians. Final diagnoses were determined by consensual agreement of at least 75% of the team members. For an overview of patient characteristics and differences between the groups see supplemental materials 2.

The mean scores of the short-form of the Outcome Questionnaire[21] and the Brief Symptom Inventory (BSI[22]; German translation of Derogatis[23]) were 1.90 (SD = 0.56) and 1.30 (SD = 0.71) respectively, indicating a moderate to severe general level of distress.

Therapists

Patients were treated by 220 therapists (175 female, 41 male, 9 unknown) who participated in a three-year (full-time) or five-year (part-time) postgraduate training program with a CBT focus. Each therapist had at least 1 year of clinical training before beginning to treat patients. On average, therapists treated 11.6 patients each (*SD* = 6.2, range 1–26). Each therapist received one hour of group or individual supervision on a monthly basis. The session videos were used for supervision and research. Supervisors were senior clinicians with at least five years of clinical experience after completing training. In treatment, therapists scored a mean of 3.81 (*SD* = 0.89) on the "overall adherence item" of the Inventory of Therapeutic Interventions and Skills[24] (ITIS), which can be considered as moderately lege artis.

Measures

Dropout. Dropout was assessed via clinical judgment at the end of treatment. When the patient and therapist agreed on a consensual end of therapy, the treatment was considered regularly completed. In contrast, when the patient stopped coming to therapy, despite the therapist's appraisal that more sessions were necessary, the form of termination was considered a dropout. Examples of this operationalization of dropout include the patient stopped coming to sessions and was unable to be reached by phone or e-mail or the patient told the therapist that they will no longer bet coming to therapy anymore, despite the therapist's advice to continue therapy.

Intake variables. 77 variables measured at intake (i.e., before the first session) were included in the analyses. Table 1 shows all 77 variables as well as the mean differences. All variables were assessed via questionnaires.

Selection of ML Algorithms

In order to get an accurate picture of common algorithms used in sociological/ scientific/ medical contexts, we decided to use and compare only those algorithms that have already found application in the relevant literature. For this purpose, we particularly focused on the Stratified Medicine Approaches for Treatment Selection (SMART) Mental Health Prediction Tournament at the 2019 Treatment Selection Idea Lab (TSIL) conference, in which 13 different research groups developed different prediction models using the same dataset (for a further review, see Cohen et al.[25]). Using the information from this tournament, as well as an examination of the literature provided by the tournament organizer, we selected a total of

21 algorithms for closer examination (see Table 2). As we aimed to compare different algorithms regardless of them being linear or non-linear, we decided to include linear algorithms alongside the ML algorithms, as suggested by Brownlee.[16]

Table 1
Predictors used for model generation. Predictors were routinely collected at intake.

Variables	Training sample Dropout vs. regular		Holdout sample Dropout vs. regular		Training sample vs. Holdout sample	
	***t*-value / chi²-value**	***p*-value**	***t*-value / chi²-value**	***p*-value**	***t*-value / chi²-value**	***p*-value**
Male sex	-0.67	0.41	1.89	0.17	1.62	0.20
High education	-29.83	<.001	-12.16	<.001	2.11	0.15
Middle education	0.08	0.78	0.16	0.69	-0.62	0.43
Sick leave	6.03	<.05	0.01	0.93	-2.82	0.09
Children	0.31	0.58	-0.00	0.95	-5.40	<.05
Marital status	-8.54	<.01	-1.96	0.16	-3.06	0.08
Medication intake	-1.35	0.25	-0.00	1.00	-2.15	0.14
Age	-3.75	<.001	-1.70	0.09	-0.46	0.65
Outcome Questionnaire (OQ) -Total score	3.75	<.001	2.12	<.05	1.02	0.31
OQ - Symptom distress	3.14	<.01	2.02	<.05	1.49	0.14
OQ - Social role functioning	1.31	0.19	0.69	0.49	0.72	0.47
OQ - Interpersonal relationship	5.75	<.001	2.54	<.05	-0.70	0.48
Questionnaire for the Evaluation of Psychotherapy (FEP2) - Total score	3.20	<.01	2.38	<.05	0.08	0.93
FEP2 - Well-being	1.80	0.07	1.93	0.05	1.40	0.16
FEP2 - Discomfort	3.54	<.001	2.87	<.01	0.99	0.32

Table 1 continuation

Variables	**<u>Training sample</u> Dropout vs. regular**		**<u>Holdout sample</u> Dropout vs. regular**		**Training sample vs. Holdout sample**	
	***t*-value / chi²-value**	***p*-value**	***t*-value / chi²-value**	***p*-value**	***t*-value / chi²-value**	***p*-value**
FEP2 - Incongruence	3.60	<.001	2.66	<.01	0.03	0.97
FEP2 - Interpersonal	1.92	0.06	0.89	0.37	-1.66	0.10
Emotionality Inventory (EMI) - Total score	2.71	<.01	1.45	0.15	0.65	0.52
EMI - Anxiety	1.75	0.08	2.39	<.05	-0.85	0.40
EMI - Depression	3.42	<.001	1.63	0.10	1.71	0.09
EMI - Inhibition	0.81	0.42	0.38	0.71	-0.64	0.52
EMI - Security	3.09	<.01	1.58	0.12	-0.16	0.88
EMI - Well-being	2.71	<.01	0.23	0.41	0.68	0.50
Brief Symptom Inventory (BSI) - Total score	5.92	<.001	2.96	<.01	0.13	0.90
BSI - Somatic problem	4.12	<.001	2.61	<.01	-0.04	0.97
BSI - Obsessive compulsive	2.86	<.01	1.22	0.22	0.28	0.78
BSI - Uncertainty	4.32	<.001	1.94	0.05	-0.57	0.57
BSI - Depression	5.09	<.001	2.17	<.05	0.76	0.45
BSI - Anxiety	3.72	<.001	2.67	<.01	-0.31	0.76
BSI - Hostility	5.25	<.001	2.79	<.01	-0.92	0.36
BSI - Phobia	4.32	<.001	2.42	<.05	0.65	0.51
BSI - Paranoid	5.87	<.001	3.05	<.01	-1.58	0.11
BSI - Psychoticism	5.16	<.001	1.94	0.05	1.36	0.17
BSI - Additional	6.81	<.001	3.07	<.01	1.58	0.11

Table 1 continuation

Variables	**<u>Training sample</u> Dropout vs. regular**		**<u>Holdout sample</u> Dropout vs. regular**		**Training sample vs. Holdout sample**	
	***t*-value / chi²-value**	***p*-value**	***t*-value / chi²-value**	***p*-value**	***t*-value / chi²-value**	***p*-value**
Interpersonal Problems (IIP32) - Total score	1.86	0.06	0.97	0.33	-0.39	0.70
IIP - Autocratic/ dominant	4.82	<.001	2.60	<.01	-1.86	0.06
IIP - Confrontational	3.19	<.01	2.06	<.05	-0.65	0.51
IIP - Unapproachable	2.86	<.01	2.01	<.05	1.01	0.31
IIP - Introverted	1.58	0.11	0.81	0.42	-0.33	0.74
IIP - Submissive	-2.06	<.05	-1.87	0.06	0.19	0.85
IIP - Exploitable	-2.95	<.01	-1.91	0.05	-0.24	0.81
IIP - Caring	1.40	0.16	1.44	0.15	0.11	0.92
IIP - Expressive	1.16	0.25	-0.54	0.59	-0.28	0.78
Incongruence Questionnaire (INK23) - Total score	3.47	<.001	1.43	0.15	-1.08	0.28
INK - Approach	2.74	<.01	1.42	0.16	0.03	0.97
INK - Avoidance	3.78	<.001	1.27	0.21	-2.26	<.05
Dysfunctional Attitudes Scale - short form (DASK) - Total score	2.42	<.05	0.92	0.36	-0.38	0.70
DASK - Recognition	0.36	0.72	1.05	0.30	-1.37	0.17
DASK - Performance	2.91	<.01	0.72	0.47	-0.08	0.94
Inventory of Stressful Events (ILE) - Score for number of events	3.19	<.01	1.61	0.11	-2.31	<.05
ILE - Score for stress	3.18	<.01	1.41	0.16	-2.64	<.01

Table 1 continuation

Variables	**Training sample** **Dropout vs. regular**		**Holdout sample** **Dropout vs. regular**		**Training sample vs. Holdout sample**	
	***t*-value / chi²-value**	***p*-value**	***t*-value / chi²-value**	***p*-value**	***t*-value / chi²-value**	***p*-value**
ILE - Number in patient's life	3.54	<.001	2.24	<.05	-1.20	0.23
ILE - Number of events in close relationships	-1.55	0.12	-1.04	0.30	-0.22	0.83
ILE - Number of events in distant relationships	-3.86	<.001	-2.11	<.05	5.24	<.001
General Perceived Self-Efficacy Scale	-0.26	0.79	-0.80	0.43	0.90	0.37
Personality Style and Disorder Inventory - short form (PSSIK) - Antisocial	4.49	<.001	3.16	<.01	-0.08	0.93
PSSIK - Paranoid	6.07	<.001	2.47	<.05	-1.63	0.10
PSSIK - Schizoid	3.23	<.01	0.74	0.46	-0.17	0.87
PSSIK - Avoidant	0.34	0.74	-1.13	0.26	-0.67	0.50
PSSIK - Compulsive	-4.66	<.001	-1.52	0.13	-0.80	0.42
PSSIK - Schizotypal	1.42	0.16	-0.05	0.96	-1.71	0.09
PSSIK - Rhapsodic	-0.63	0.53	1.15	0.25	0.28	0.78
PSSIK - Narcissistic	1.44	0.15	0.94	0.35	0.79	0.43
PSSIK - Negativistic	6.06	<.001	2.85	<.01	-2.30	<.05
PSSIK - Dependent	4.23	<.001	1.93	0.05	-1.15	0.25
PSSIK - Borderline	4.51	<.001	2.19	<.05	0.33	0.74
PSSIK - Histrionic	3.43	<.001	2.28	<.05	-1.42	0.15
PSSIK - Depressive	4.23	<.001	1.63	0.10	-0.11	0.91
PSSIK - Altruistic	1.88	0.06	1.83	0.07	-0.04	0.97

Table 1 continuation

Variables	**Training sample** **Dropout vs. regular**		**Holdout sample** **Dropout vs. regular**		**Training sample vs. Holdout sample**	
	***t*-value / chi²-value**	***p*-value**	***t*-value / chi²-value**	***p*-value**	***t*-value / chi²-value**	***p*-value**
Patient-rated wellbeing	-2.91	<.01	-1.70	0.09	-0.90	0.37
Current emotional and psychological functioning	-3.05	<.01	-2.69	<.01	1.92	0.06
Therapy Expectations (TE) - Importance of psychotherapy	-1.70	0.09	0.64	0.53	0.03	0.97
TE - Difficulties attending psychotherapy	-1.93	0.05	1.58	0.11	-1.75	0.08
TE - Confidence in the helpfulness of psychotherapy	-2.82	<.01	-3.22	<.01	-0.64	0.52
TE - Amount of previous psychotherapy	-0.09	0.93	0.67	0.50	2.49	<.05
TE - Chronicity of the problem	1.37	0.17	2.10	<.05	-0.24	0.81
TE - Estimated future coping	-2.22	<.05	-1.47	0.14	-0.47	0.64

Note: Negative values indicate a negative correlation with the dropout variable or a higher value/ratio in the holdout sample. For dichotomous variables (first 7 variables) a chi² - test was used, for continuous variables, a *t*-test was used. High Education = University entrance qualification; Middle school = middle school graduation. The total score of the BSI additional scale is the mean of the four additional items of the BSI.

Data Analytic Strategy

Data preparation. All analyses were conducted using the free software environment R version 4.1.1.[26] No variables that had more than 10% missing values were included in the analyses. Therefore, we had to exclude a total of five variables (total scores of the Patient Health Questionnaire (PHQ-9), Affective Style Questionnaire (ASQ), Generalized Anxiety Disorder Assessment (GAD-7), Ten Item Personality Measure (TIPI) and Work and Social

Adjustment Scale (WSAS)). No patient was excluded from the analyses because of too many missing values. Variables with less than 10% missing values were imputed using a trained random forest in the R package missForest v1.4.[27] The imputations for the training and holdout samples were conducted separate from the cross-validation (CV) framework before the actual analyses. For model training, we used the R-packages caret v6.0-90[28] and caretEnsemble v2.0.1.[29] These packages tune the hyperparameters to their optimal settings depending on which one is being used. To ensure a fair comparison of the algorithms, we did not change the packages' default settings. Table 2 shows the algorithms used and the different tuning parameters that were tested (for a further review see Kuhn[30]). Identification of the best model was always based on the receiver operating characteristic (ROC) curve. The model with the largest area under the curve (AUC) was considered the best model. All models predicted dropout as a binary event (dropout vs. non-dropout).

Ranking and correlation of algorithms. First, we ranked all individual algorithms based on the two parameters (i.e., Brier, AUC) and compared the correlations of the predictions of all algorithms during the model building process using the corresponding function in the Caret package. For this purpose, we conducted a nested CV with 20 outer and 10 inner loops according to Brownlee's[31] recommendations. All continuous variables were centered separately for each outer CV loop of the training and test sets. Subtrahend was always the mean value from the training data of the respective variable to avoid data leakage and to ensure appropriate data preparation for the algorithms.[16] Dropout was dichotomized with 1 (dropout) and 0 (regular termination). Subsequently, each of the 21 ML algorithms generated a dropout prediction model based on each outer and inner CV training set that was then evaluated in the respective outer or inner CV test set to minimize overfitting[32] and the influence of sample characteristics. For the inner CV, we also applied a sampling method (Synthetic Minority Oversampling Technique; SMOTE)[33] to address the problem of class imbalance.[34,35] SMOTE is a hybrid method combining up and down sampling. It artificially generates new examples of the minority class using the nearest neighbors of these cases. Furthermore, the majority class examples are also under-sampled, leading to a more balanced dataset. Then, a performance ranking based on the Brier score and the AUC was generated as well as the correlation matrix for all algorithms.

Table 2
Classification of all machine learning algorithms (Brownlee, 2019) that were used in this study.

Regression	**Tuning Parameters**
Generalized Linear Model (GLM)	-
GLM with stepwise feature selection using AIC (GLMAIC)	-
Bayesian	
Bayesian GLM (BAYESGLM)	-
Naïve Bayes (NB)	usekernel[y/n]; laplace = 0; adjust = 1
Decision Tree	
C4.5-like Trees (C4.5)	C[0.01; 0.255; 0.5]; M[1; 2; 3]
Conditional Inference Trees (CTREE)	maxdepth[1; 2; 3]; mincriterion[0.01; 0.5; 0.99]
Artificial Neural Networks	
Feed-Forward Neural Network with single hidden layer (NNET)	size[1; 3; 5]; decay[0; 0.1; 0.0001]
Averaged feed-forward Neural Network with single hidden layer over different seeds (AVNN)	size[1; 3; 5]; decay[0; 0.1; 0.0001]; bag = FALSE
Monotone Multi-Layer Perceptron Neural Network (MONMLP)	hidden1[1; 3; 5]; n.ensemble = 1
Dimensionality Reduction	
Linear Discriminant Analysis (LDA)	-
Regularization	
Elastic Net (EN)	alpha[0.1; 0.55; 1]; lambda[0.0001; 0.001; 0.01]
Instance-based	
K-fold-Nearest Neighbors (kNN)	k[5; 7; 9]
Support Vector Machines (SVM)	cost[0.25; 0.5; 1];
Ensembles	
Generalized boosted regression modeling (GBM)	Interaction depth[1; 2; 3]; n.trees[50; 100; 150]; n.minobsinnode = 10; shrinkage = 0.1
Boosted Logistic Regression (LOGIT)	nIter[11; 21; 31]
Extreme Gradient Boosting (XGB)	Eta[0.3; 0.4]; max_depth[1; 2; 3]; colsample_bytree[0.6; 0.8]; subsample[0.5; 0.75; 1]; nrounds[50; 100; 150]; gamma = 0; min_child_weight = 1;
Random Forest (RF)	mtry[2; 4; 7]
Bagged Multivariate Adaptive Regression Splines (MARS)	nprune[2; 8; 14]; degree = 1
Bagged Classification and Regression Tree (CART)	-
Adapted boosted Classification Trees (ADA)	maxdepth[1; 2; 3]; iter[50; 100; 150]; nu = 0.1
Boosted Generalized Linear Model (GLMBOOST)	mstop[50; 100; 150]

Note: The numbers in the square brackets indicate the different tuning parameters tested using the R package caret. The words in bold are the categories of the respective algorithms.

Brier score. The Brier score ranges from 0 (best prediction) to 1 (worst prediction) by measuring probabilistic predictions.[36] Thus, it takes the certainty of the prediction into account. In effect, it is the mean squared error of the forecast

$$\frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2$$

Hereby, *N* is the total number of observations, *f* is the probability of the event (i.e., dropout) and *o* is the actual outcome (i.e., 0 or 1) of the event at instance t.

AUC. The AUC uses the sensitivity and specificity of a prediction and ranges from 0 (worst prediction) to 1 (best prediction). Based on signal detection theory[37], the AUC takes the base rate of the dependent variable into account.

Ensembles. We used the ranking of single algorithms and the correlation matrix to generate ensembles. Ensembles show better performance and greater robustness in certain contexts by reweighting the results of different algorithms, which can produce better overall results.[38] We decided to use five types of ensembles. The two and three best algorithms, the two and three least correlating algorithms and the best algorithm with the respective least correlating algorithm. The idea to merge algorithms with low correlations is that they probably assess different aspects of the dataset.[39] Therefore, it is possible that an ensemble of such algorithms improves the prediction significantly, even though one algorithm makes poor predictions on its own. These ensembles were merged either via a generalized linear modeling (GLM) algorithm or via the best algorithm across both parameters (i.e., Brier score and AUC) according to our ranking using the stacking method. Again we used Caret with its default settings to create an ensemble with the best parameters. In total, we generated ten ensembles (5 types of ensemble x 2 ways of merging).

Comparing ensembles and single algorithms. Next, we compared all 10 ensembles and the 5 best single algorithms. Again, we used a nested CV as described above. However, this time we used a 10-fold inner CV with 3 repetitions. Repeating the CV leads to a more precise result[40], so we conducted this procedure for a more adequate comparison.

Extending the procedure. In order to gain a more comprehensive picture, we repeated the entire procedure twice. For the first repetition, we only used the significant predictors (i.e., initial impairment, male sex, lower education status, more histrionic and less compulsive personality style and negative treatment expectations) from Zimmermann et al.[8] Thus, we evaluated the changes in the prediction when using these relevant predictors only. For the second repetition, we performed variable selection using an elastic net (EN) regularization with the Caret package for the training set after each split. As we examined a large set of variables, we evaluated whether some models improve, when preceded by variable selection. This was done 20 times for each training set of the outer loops inside the CV framework, preventing data leakage from the test set. For this EN selection after each split, we did not use Caret to choose the optimal setting, but set alpha to 0.1 for the first analysis and then altered alpha in increments of 0.1 until 1 was reached. An alpha of 0 is equal to a ridge regression, while an alpha of 1 equals a least absolute shrinkage and selection operator (LASSO) regression. We also defined lambda's range analogue to the alpha parameter. Lambda defines the magnitude of the regression penalty. This resulted in 100 different possible combinations of these two parameters (10 values for alpha x 10 values for lambda) to identify the best fitting model. Identification of the best model was always based on the AUC. The model with the highest value was considered the best model. At the end of the second repetition, we only included the predictors that had predictive power in the best model.

Conducting this entire procedure three times (with all variables, with only seven predictors, and with variables that had predictive power in the preceding EN analysis) led to a total of 30 ensembles (10 ensembles x 3 procedures) and 15 single algorithms (5 single algorithms x 3 procedures). Each ensemble and single algorithm generated a model via a nested CV with 20 outer loops and 10 inner loops with three repetitions. The model with the best mean prediction scores across all CVs and across both parameters was chosen. Generating 20 models via the outer CVs resulted in one distribution consisting of 20 Brier scores and one distribution consisting of 20 AUC scores for each algorithm/ensemble. In order to quantitatively compare the differences and distributions as well as the robustness against sampling artifacts, *t*-tests between the best and worst model as well as between the best model and a single GLM were conducted for each parameter.

For a final test we used the best ensemble/algorithm and let it generate a model with the whole training sample via a 10-fold CV with 3 repetitions. This model was then tested in the still unused and independent holdout sample to assess the generalizability of the model and to prevent overfitting.

Last, the holdout sample's confusion matrix was examined in order to assess the improvement of the prediction. Therefore, each case that had a higher risk than the mean of the training sample to drop out of therapy (i.e., 30.6%) was considered a predicted *dropout case*. Finally, the Caret package was used to determine the most important variables.

Results

After the first step, the algorithm with the best predictions when using all variables was Generalized Boosted Regression Modeling (GBM). When only using predictors that showed predictive power in a preceding EN analysis, Random Forest (RF) was the best algorithm. Adapted boosted Classification Trees (ADA) made the best predictions when only the seven significant predictors were used (see supplemental materials 3 for an overview of all algorithms). Especially boosting and tree-based approaches seemed to make the best predictions. Further, algorithms from different classes seemed to correlate the least with each other (see supplemental materials 4 for the low correlating algorithms).

Next, by using the rankings (supplemental materials 3) and correlations (supplemental materials 4), we generated the ensembles as described above for the final analyses. Comparing the different algorithms and ensembles, the best model across both parameters was generated by an ensemble with the best ML algorithm and its least correlating algorithm (i.e., RF and k-fold Nearest Neighbors (kNN)) that was merged via a GLM and had a preceding EN variable selection (Brier score = 0.1983; AUC = 0.6581). Table 3 provides an overview of all algorithms and ensembles.

The distributions of each algorithm/ensemble revealed that the best ones hardly differed from each other (see Figure 1). Nevertheless, some models seemed to make significantly worse predictions. For the Brier score, the pattern was very similar.

As a result of these distributions, we were able to compare model accuracy/robustness via t-tests. A paired one-sided t-test revealed a highly significant effect between the overall best and overall worst models concerning the AUC score (AUC_{best} = 0.6581; AUC_{worst} = 0.5465; t(19) = 8.30, p < .001, Cohen’s d = 1.86 [0.11; 2.58]). Comparing the overall best model with the models of a GLM using all variables, the effect was still significant (AUC_{best} = 0.6581; AUC_{GLM} = 0.6253; t(19) = 2.63, p < .01, Cohen’s d = 0.59 [0.11; 1.06]). For the Brier score, the effects were also significant when comparing the best with the worst model ($Brier_{best}$ = 0.1982; $Brier_{worst}$ = 0.2859; t(19) = –13.03, p < .001, Cohen’s d = –2.91 [–3.92; –1.89]) and when comparing the best model with the GLM model using all variables ($Brier_{best}$ = 0.1982; $Brier_{GLM}$ = 0.2384; t(19) = –13.11, p < .001, Cohen’s d = –2.93 [–3.95; –1.90]). All boxplots are shown in supplemental materials 5.

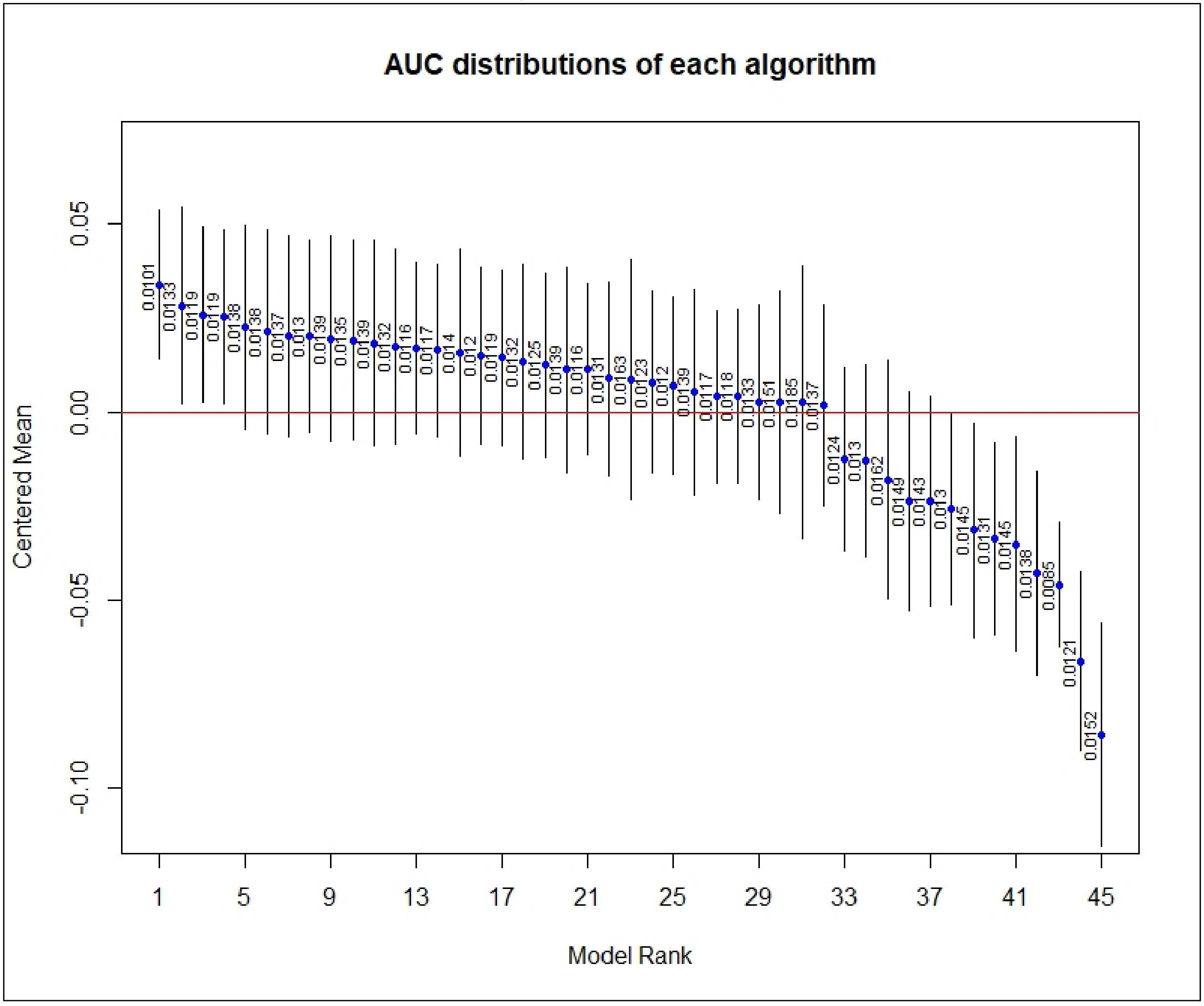
Table 3

Mean scores of the models generated by all 45 algorithms and ensembles.

Algorithm/Ensemble	Stacking Method	Variables used	Brier score	AUC	Training AUC
Best with lowest correlation	GLM	Selected with EN	.1983	.6581	.6617
Two best	GLM	Selected with EN	.1983	.6577	.6674
Three best	GLM	Selected with EN	.1985	.6535	.6673
Two best	GBM	All	.1989	.6550	.6515
Three best	GLM	All	.1994	.6513	.6549
Best with lowest correlation	GLM	All	.1992	.6497	.6492
Two Best	GLM	All	.1995	.6518	.6530
Three best	GBM	All	.1998	.6523	.6557
GBM	-	Selected with EN	.2022	.6661	.6608
Two best	GLM	Manually selected	.1995	.6493	.6464
RF	-	Selected with EN	.2041	.6605	.6602
Best with lowest correlation	GLM	Manually selected	.1997	.6488	.6430
Best with lowest correlation	GBM	All	.2004	.6506	.6483
Three best	GLM	Manually selected	.1998	.6468	.6461
Three least correlating	GLM	All	.2010	.6435	.6494
Three least correlating	GBM	All	.2006	.6412	.6485
Three best	ADA	Manually selected	.2011	.6435	.6488
ADA	-	All	.2071	.6525	.6485
GBM	-	All	.2055	.6457	.6497
Best with lowest correlation	ADA	Manually selected	.2017	.6403	.6424
XGB	-	Selected with EN	.2069	.6480	.6584
Two best	ADA	Manually selected	.2014	.6349	.6482
RF	-	All	.2058	.6392	.6428
ADA	-	Selected with EN	.2099	.6475	.6591
XGB	-	All	.2075	.6448	.6451
Two least correlating	GLM	All	.2049	.6197	.6129
Three least correlating	GLM	Manually selected	.2053	.6193	.6095
Two least correlating	GLM	Manually selected	.2056	.6143	.6060
GBM	-	Manually selected	.2208	.6525	.6369
GLMBOOST	-	Selected with EN	.2309	.6408	.6516
Three least correlating	ADA	Manually selected	.2059	.6066	.6150
Two least correlating	ADA	Manually selected	.2064	.6087	.6092
Two least correlating	GBM	All	.2060	.6010	.6121
ADA	-	Manually selected	.2240	.6349	.6440
GLMBOOST	-	Manually selected	.2342	.6364	.6379
GLMBOOST	-	All	.2306	.6349	.6487
Two least correlating	GLM	Selected with EN	.2064	.5971	.5872
LDA	-	Manually selected	.2347	.6364	.6377
Three least correlating	GLM	Selected with EN	.2074	.5986	.6058
Three best	RF	Selected with EN	.2180	.6085	.6143
GLMAIC	-	Manually Selected	.2342	.6342	.6392
Two best	RF	Selected with EN	.2376	.5893	.5902
Three least correlating	RF	Selected with EN	.2490	.5661	.5607
Best with lowest correlation	RF	Selected with EN	.2586	.5864	.5838
Two least correlating	RF	Selected with EN	.2859	.5465	.5489

Note: EN = Elastic net; GLM = Generalized linear model; RF = Random forest; GLMBOOST = Boosted generalized linear model; ADA = Adapted boosted classification trees; GLMAIC = Generalized linear model with stepwise feature selection using Akaike information criterion; XGB = Extreme gradient boosting; LDA = Linear discriminant analysis; GBM = Generalized boosted regression model; All ensembles and algorithms are ranked.

Figure 1
Distribution of the 20 outer cross-validation (CV) models generated by each algorithm and ensemble ranked from best to worst using the area under the curve (AUC).



Note: Each value was grand-mean centered; the horizontal line represents the total average of all models. The numbers on the graphs are the standard deviations.

The best model was able to identify 63.4% of all holdout cases correctly before the first session occurred (the confusion matrix is shown in supplemental material 6) having an AUC of 0.6694 and a Brier score of 0.1988. Thus, it achieved a substantial improvement over the model generated by a GLM using all variables (46.2 %). The main dropout predictors that made a substantial contribution (i.e., relative importance > 90%) to the model were lower education level, younger age, lower scores on the *compulsive* scale of the personality style and disorder inventory (PSSI), higher scores on the *negativistic* and *antisocial* scale of the PSSI and higher scores on the *additional* scale of the BSI as well as a higher total score (see supplemental materials 7 for an overview of all variables; see Liaw et al[41] for a description of how variable importance is calculated in a random forest model). The BSI additional scale is the mean of the four additional items not included in any of the dimension scores (“Poor appetite”, “Trouble falling asleep”, “Thoughts of death and dying”, “Feeling of guilt”).

Discussion

The aim was to evaluate the use of different ML algorithms in a naturalistic routine care setting by generating a predictive model to identify patients who are at risk of dropout. Two different indices were used to gain a more comprehensive picture of the results. We selected 21 algorithms for our study and used nested CV to compare them. We used the best algorithms and least correlating algorithms to generate ensembles that were also compared. The best model was an ensemble of the best algorithm with its least correlating algorithm (i.e., RF, kNN) that used only predictive variables and was merged via a GLM. Differences between the best ensemble and a single GLM as well as the worst algorithm were highly significant, independent of the examined parameters. When comparing the distributions of the best model and the GLM, a large effect size of up to $d = 2.93$ was found, indicating the superiority of the best model independent of the training sample used.

The best model was able to correctly identify 63.4 % of all cases in an independent holdout sample. Although this does not seem very precise at first, it must be acknowledged that this prediction was made *before* the first session of routine CBT and that a single GLM correctly identified only 46.2 %. Therefore, this model is of high clinical value and is able to identify patients who tend to drop out of therapy before the first session has occurred. The mostly identical values of the AUC and the Brier score in the holdout sample compared to the test set in the modeling process indicate good stability and generalizability of the model.

The most important variables used in the final prediction model also appear to differ significantly between dropout and consensual termination cases. Nevertheless, this is not true for all variables, suggesting that the model uses more than just the different mean values for prediction. Based on the relevant variables in the model, therapists should take time to build a complementary relationship with the patient and invest time in explaining how therapy can concretely help them. Particularly high levels of interpersonal variables that make it difficult to establish a functional therapeutic relationship (e.g., negativistic or antisocial personality style) appear to increase the risk of dropout. It is important for clinicians to pay attention to the complementarity of the relationship in order to establish a good alliance. This is especially crucial in the first session, as this is where the first impression is made. Here model predictions can be used to better prepare for potential interpersonal difficulties.

With regard to the BSI additional scale, it seems reasonable to first treat symptoms such as sleep problems, poor appetite, suicidal thoughts, and feelings of guilt. Although general symptom burden or functionality do not play a particularly important role, symptoms that are very obvious to the patient (e.g., sleep problems, distressing suicidal thoughts) appear to be important indicators. It is obvious that patients hope for a quick improvement of these symptoms in therapy, which has an important signal effect for therapists to focus on the treatment of these symptoms, especially at the beginning of therapy. Interestingly, lower education and younger age also seem to increase the probability of dropout. Other studies have also identified these variables[8], so these should be considered in therapy, even if they are invariant. Future studies should explore the underlying mechanisms of these on dropout probability to better understand the effects and to improve future models. Nevertheless, using the information from our model, clinicians could generate a more precise case concept for the individual patient before the first session to help patients gain confidence in therapy, facilitate the establishment of a functional therapeutic relationship, and thus reduce the risk of dropout. Therefore, the best model from our analyses could improve and further support measurement based care with regard to dropout prediction and prevention.

Further, the results indicate that ML algorithms/ensembles can have a true predictive advantage in naturalistic settings. However, this does not apply to all algorithms. Some produced significantly worse predictions, indicating that not all ML algorithms/ensembles are suited to naturalistic settings. Our results revealed that ensembles consisting of low correlating algorithms did not perform well except when a powerful algorithm that delivers good predictions on its own is included. The idea that low correlating algorithms assess different aspects of the dataset and thus should perform better than ensembles with more similar algorithms did not hold true. Mayer[39] states that an optimal ensemble of low correlating algorithms consists of those that perform similarly on their own. This could explain the worse prediction quality in our data, as this was not the case in our analyses.

Furthermore, the assumption *the more the better* also did not hold true. Ensembles that used more algorithms did not automatically perform better. This finding is in line with previous argumentation that selecting algorithms to create an ensemble does not follow easy rules like *the more the better*, but is a research topic of its own.[42] In addition, when using a large set of variables, a variable selection procedure should be part of model generation, either by using an algorithm that includes a selection procedure or a preceding variable selection. Our findings indicate that algorithms that had to handle many variables and did not include a variable selection procedure performed worse (e.g., linear discriminant analysis (LDA)). This finding is in line with the existing literature, stating that, in clinical settings, not every variable has predictive power for a certain outcome[43] and can thus weaken the power of the model.

Interestingly, tree-based and boosted algorithms seem to perform better compared to more advanced algorithms like neural networks. This finding appeared consistently, independent of the examined parameters. Therefore, for this kind of naturalistic binary data, boosted linear algorithms and tree-based approaches such as random forest seem very well-suited.

Limitations

Although this study has many strengths, several limitations must be mentioned. One reason for the poor performance of neural networks could be the data quality. Albeit naturalistic assessments include crucial predictive information, they are nowhere near perfect and always have measurement errors. Although this topic is not new[44], these errors prevent the algorithms from assessing the relevant relationships. These suggestions are in line with the existing literature.[43,45] A solution to this problem could be the usage of ecological momentary assessment (EMA) data, which provides more accurate descriptions of within-person processes at a higher resolution. For future studies, it is of great interest whether complex algorithms such as neural networks are more suitable for such data and are thus able to improve dropout predictions. Electrophysiological variables and neural imaging variables could also improve predictions,[46] but such assessments are expensive and time-consuming and therefore unlikely to be used in routine care. In addition, the amount of data could have played an important role. Complex algorithms that are able to assess high order interactions need a lot of data.[16] Thus, the size of our dataset limits the evaluation of these algorithms. Future studies should try to generate even larger datasets in order to evaluate the possible benefit of advanced algorithms.

Furthermore, it is possible that certain predictive variables were not collected. For example, we only collected whether patients were taking medication or not, regardless of what they were taking or for how long. Although this variable did not play a role in our model, it cannot be ruled out that more precise information could improve the model. The same applies to the variables that we had to exclude due to too many missing values (i.e., PHQ-9, ASQ, GAD-7, TIPI, WSAS). These variables contain important clinical information that could be important for prediction.

Moreover, we used only a small number of possible ML algorithms. Although we used many models that have already been applied in psychological studies to create a representative picture, it cannot be ruled out that an even more suitable approach for this kind of data exists. Also, as mentioned above, the use of ensembles requires a profound understanding of this topic. For our own ensembles, we used the stacking method. However, there are other options to create ensembles such as bagging or boosting.

Although the model is well protected against overfitting by the use of repeated and nested CV as well as a separate holdout sample, the possibility of overfitting cannot be completely ruled out. Furthermore, our holdout sample is quite similar to the training sample, which limits the generalizability of the results. Nevertheless, it must be noted that there are differences between the samples, especially with regard to the diagnosis, which is why a certain degree of generalizability can be assumed. Nevertheless, holdout samples from other institutes should be used in future studies to more robustly test generalizability.

In addition, although this model helps identifying patients who are at risk of dropping out of therapy, it does not reveal the reasons for this increased risk. No causal conclusions can be drawn from this model, which is a limitation of our model and of ML in general. Nevertheless, the identified predictors provide first clues as to which risk factors may be relevant to dropout. Moreover, the identification of patients at risk for treatment discontinuation is the first step to reducing the number of patients who drop out.

Conclusion

To our knowledge, this study is the first to use such a large naturalistic dataset to evaluate different ML algorithms and ensembles to identify a useful dropout prediction model. The current study compared several ML methods in order to evaluate the benefit of ML in naturalistic contexts and to generate a model that has high clinical value for identifying dropout risk on an individual level. The model identified over 60% of patients’ type of therapy termination correctly. This study’s findings highlight that it is possible to identify patients at risk of dropout before the first session has occurred and that ML algorithms provide an important contribution to model generation. Especially tree-based and boosted algorithms that include a variable selection procedure (e.g., EN) seem suited to building prediction models for psychotherapy dropout.

Future research should further explore treatment data to improve prediction models and use them to develop strategies to reduce the risk of dropout. By implementing these models into clinical support systems, the number of dropouts could be reduced, resulting in more effective therapy outcomes and less burden on patients and society.

Author details

Björn Bennemann, MSc, Department of Clinical Psychology and Psychotherapy, University of Trier, Germany; **Brian Schwartz**, MSc, Department of Clinical Psychology and Psychotherapy, University of Trier, Germany; **Julia Giesemann**, MSc, Department of Clinical Psychology and Psychotherapy, University of Trier, Germany; **Wolfgang Lutz**, Full Professor, Department of Clinical Psychology and Psychotherapy, University of Trier, Germany.

Björn Bennemann https://orcid.org/0000-0002-4085-2262
Brian Schwartz https://orcid.org/0000-0003-4695-4953
Julia Giesemann https://orcid.org/0000-0002-3370-793X
Wolfgang Lutz https://orcid.org/0000-0002-5141-3847

Author Contributions

W.L. provided the data set and advised on the structure and design of the study. B.B. developed the idea and performed the analyses. B.S., J.G. and W.L. advised on statistical issues. All authors were involved in interpreting data, drafting the work or critically revising it for important intellectual content. All authors provided final approval of the submitted version of the manuscript.

Additional Information

Correspondence: Björn Bennemann, E-mail: bennemann@uni-trier.de

Funding: This work was partly supported by the German Research Foundation (DFG) under Grant Nr. LU 660/10-1.

Declarations of interest: None.

Acknowledgements: We are thankful to Dr. Kaitlyn Poster for proofreading the manuscript.

Data availability: The data that support the findings of this study are available from the corresponding author, B.B., upon reasonable request.

References

1. Brachel R von, Hirschfeld G, Berner A, et al. Long-term effectiveness of cognitive behavioral therapy in routine outpatient care: a 5-to 20-year follow-up study. *Psychotherapy and psychosomatics* 2019; 88: 225–235.
2. Swift JK, Greenberg RP, Tompkins KA, et al. Treatment refusal and premature termination in psychotherapy, pharmacotherapy, and their combination: A meta-analysis of head-to-head comparisons. *Psychotherapy (Chic)* 2017; 54: 47–57.
3. Wells JE, Browne MO, Aguilar-Gaxiola S, et al. Drop out from out-patient mental healthcare in the World Health Organization’s World Mental Health Survey initiative. *Br J Psychiatry* 2013; 202: 42–49.
4. Rossi A, Amaddeo F, Bisoffi G, et al. Dropping out of care: inappropriate terminations of contact with community-based psychiatric services. *Br J Psychiatry* 2002; 181: 331–338.
5. Lutz W, Rubel JA, Schiefele A-K, et al. Feedback and therapist effects in the context of treatment outcome and treatment length. *Psychotherapy Research* 2015; 25: 647–660.
6. Kessler RC. The potential of predictive analytics to provide clinical decision support in depression treatment planning. *Curr Opin Psychiatry* 2018; 31: 32–39.
7. Swift JK and Greenberg RP. Premature discontinuation in adult psychotherapy: a meta-analysis. *J Consult Clin Psychol* 2012; 80: 547–559.
8. Zimmermann D, Rubel JA, Page AC, et al. Therapist Effects on and Predictors of Non-Consensual Dropout in Psychotherapy. *Clin Psychol Psychother* 2017; 24: 312–321.
9. Tehrani E, Krussel J, Borg L, et al. Dropping out of psychiatric treatment: a prospective study of a first-admission cohort. *Acta Psychiatr Scand* 1996; 94: 266–271.
10. Delgadillo J and Lutz W. A Development Pathway Towards Precision Mental Health Care. *JAMA Psychiatry* 2020; 77: 889–890.
11. Fabbri C, Kasper S, Kautzky A, et al. Genome-wide association study of treatment-resistance in depression and meta-analysis of three independent samples. *Br J Psychiatry* 2019; 214: 36–41.
12. Legge SE, Dennison CA, Pardiñas AF, et al. Clinical indicators of treatment-resistant psychosis. *Br J Psychiatry* 2020; 216: 259–266.
13. Lutz W, Deisenhofer A-K, Rubel J, et al. Prospective evaluation of a clinical decision support system in psychological therapy. *J Consult Clin Psychol* 2021.

14. Maj M, Stein DJ, Parker G, et al. The clinical characterization of the adult patient with depression aimed at personalization of management. *World Psychiatry* 2020; 19: 269–293.
15. Kessler RC, Bossarte RM, Luedtke A, et al. Machine learning methods for developing precision treatment rules with observational data. *Behav Res Ther* 2019; 120: 103412.
16. Brownlee J. Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch (2019).
17. Christodoulou E, Ma J, Collins GS, et al. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019; 110: 12–22.
18. Wilkinson J, Arnold KF, Murray EJ, et al. Time to reality check the promises of machine learning-powered precision medicine. *The Lancet Digital Health* 2020.
19. Wittchen H-U, Wunderlich U, Gruschwitz S, et al. SKID I. Strukturiertes Klinisches Interview für DSM-IV. Achse I: Psychische Störungen. Interviewheft und Beurteilungsheft. Eine deutschsprachige, erweiterte Bearb. d. amerikanischen Originalversion des SKID I 1997.
20. Bronisch T, Hiller W, Mombour W, et al. *International diagnostic checklists for personality disorders according to ICD-10 and DSM-IV—IDCL-P*: Seattle, WA: Hogrefe and Huber Publishers.
21. Ellsworth JR, Lambert MJ and Johnson J. A comparison of the Outcome Questionnaire-45 and Outcome Questionnaire-30 in classification and prediction of treatment outcome. *Clin Psychol Psychother* 2006; 13: 380–391.
22. Franke GH. *Brief symptom inventory (BSI) von LR Derogatis:(Kurzform der SCL-90-R)*: Beltz Test, 2000.
23. Derogatis LR. *SCL-90-R: Symptom Checklist-90-R Administration, Scoring, and Procedures Manual*: NCS Pearson, 1975.
24. Boyle K, Deisenhofer A-K, Rubel JA, et al. Assessing treatment integrity in personalized CBT: the inventory of therapeutic interventions and skills. *Cogn Behav Ther* 2020; 49: 210–227.
25. Cohen ZD, Wiley JF, Lutz W, et al. SMART Mental Health Prediction Tournament., osf.io/wxgzu (2018).
26. R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2021.

27. Stekhoven DJ. missForest: Nonparametric missing value imputation using random forest. *Astrophysics Source Code Library* 2015.
28. Kuhn M. *caret: Classification and Regression Training*: R package version 6.0-90, 2021.
29. Deane-Mayer ZA and Knowles JE. *caretEnsemble: Ensembles of Caret Models*: R package version 2.0.1, 2019.
30. Kuhn M. The caret Package: 6 Available Models, https://topepo.github.io/caret/available-models.html (2021, accessed 3 November 2021).
31. Brownlee J. Machine Learning Mastery: Nested Cross-Validation for Machine Learning with Python, https://machinelearningmastery.com/nested-cross-validation-for-machine-learning-with-python/ (2019).
32. Cawley GC and Talbot NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research* 2010; 11: 2079–2107.
33. Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: Synthetic Minority Over-sampling Technique. *jair* 2002; 16: 321–357.
34. Hand DJ and Vinciotti V. Choosing k for two-class nearest neighbour classifiers with unbalanced classes. *Pattern recognition letters* 2003; 24: 1555–1562.
35. Japkowicz N and Stephen S. The class imbalance problem: A systematic study. *Intelligent data analysis* 2002; 6: 429–449.
36. Brier GW. Verification of forecasts expressed in terms of probability. *Monthly weather review* 1950; 78: 1–3.
37. Green DM and Swets JA. *Signal detection theory and psychophysics*: Wiley New York, 1966.
38. Zhang C and Ma Y. *Ensemble machine learning: methods and applications*: Springer, 2012.
39. Mayer Z. A Brief Introduction to caretEnsemble, https://cran.r-project.org/web/packages/caretEnsemble/vignettes/caretEnsemble-intro.html (2019).
40. Kuhn M and Johnson K. *Applied predictive modeling*: Springer, 2013.
41. Liaw A and Wiener M. Classification and regression by randomForest. *R news* 2002; 2: 18–22.
42. Ramzai J. Simple guide for ensemble learning methods, https://towardsdatascience.com/simple-guide-for-ensemble-learning-methods-d87cc68705a2 (2019).

43. Chowdhury MZI and Turin TC. Variable selection strategies and its importance in clinical prediction modelling. *Fam Med Community Health* 2020; 8: e000262.
44. Lutz W, Jong K de, Rubel JA, et al. Measuring, Predicting and Tracking Change in Psychotherapy. In: Barkham M, Lutz W and Castonguay LG (eds) *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change*. 7 ed. New York: Wiley, 2021.
45. Jacobucci R and Grimm KJ. Machine Learning and Psychological Research: The Unexplored Effect of Measurement. *Perspect Psychol Sci* 2020; 15: 809–816.
46. Lueken U, Zierhut KC, Hahn T, et al. Neurobiological markers predicting treatment response in anxiety disorders: A systematic review and implications for clinical application. *Neurosci Biobehav Rev* 2016; 66: 143–162.

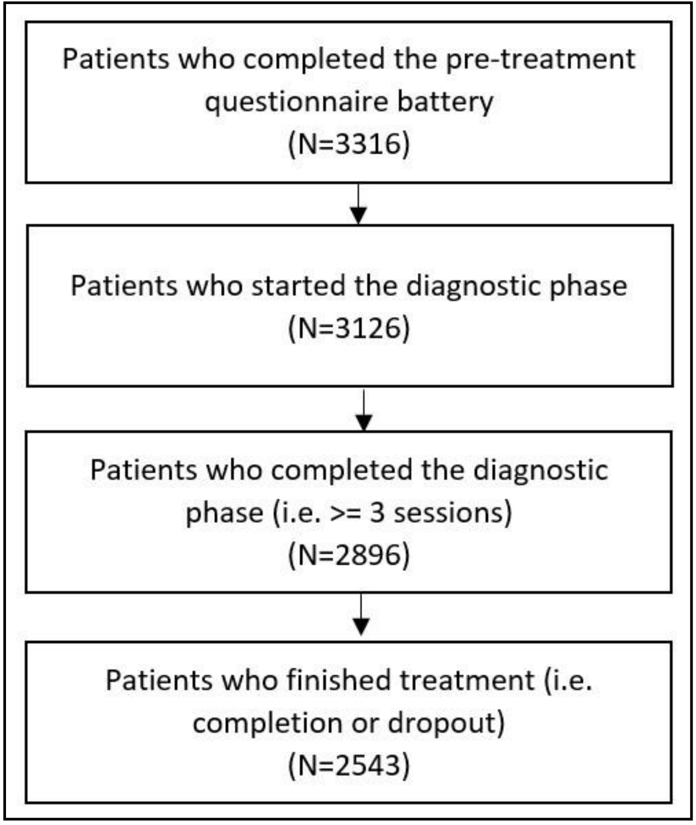
Supplementary Material

Predicting Patients who will Drop out of Outpatient Psychotherapy Using Machine Learning Algorithms

Björn Bennemann [a*], Brian Schwartz [a], Julia Giesemann [a], Wolfgang Lutz [a],

[a] University of Trier, Universitätsring 15, 54296 Trier, Germany

Supplemental Materials 1
Flowchart of selected patients for the training sample of this study.


Patients who completed the pre-treatment questionnaire battery (N=3316)
Patients who started the diagnostic phase (N=3126)
Patients who completed the diagnostic phase (i.e. >= 3 sessions) (N=2896)
Patients who finished treatment (i.e. completion or dropout) (N=2543)

Supplemental Materials 2

Patient characteristics of the training and holdout samples divided into dropout and regular completion.

Characteristics	**Training sample (*N* = 2043)**		***t*-value / chi²-value**	***p*-value**	**Holdout sample (*N* = 500)**		***t*-value / chi²-value**	***p*-value**
	Regular (*N* = 1418)	**Dropout (*N* = 625)**			**Regular (*N* = 346)**	**Dropout (*N* = 154)**		
Age (*M* years (*SD*))	36.7 (12.7)	34.4 (12.7)	3.75	<.001	36.4 (14.3)	34.2 (13.0)	1.70	0.09
Gender female (*n* (%))	1279 (62.6)	400 (64.0)	0.67	0.41	213 (61.6)	84 (55.5)	1.89	0.17
German nationality (*n* (%))	1352 (95.3)	590 (94.4)	0.64	0.43	321 (92.8)	137 (89.0)	1.55	0.21
Marital status (*n* married (%))	422 (29.7)	146 (23.4)	8.75	<.01	89 (25.7)	30 (19.5)	1.96	0.16
Education (*n* > 12 years (%))	662 (46.7)	210 (33.6)	29.83	<.001	179 (51.7)	53 (34.4)	12.17	<.001
Inability to work (*n* (%))	269 (19.0)	149 (24.3)	6.03	<.05	58 (16.8)	27 (17.5)	0.01	0.93
Intake of medication (*n* (%))	1011 (71.3)	429 (68.6)	1.35	0.25	232 (67.1)	103 (66.9)	0.00	1
Primary diagnosis (*n* (%))								
Affective disorder	667 (51.1)	287 (49.2)	0.50	0.48	170 (49.4)	59 (38.3)	4.85	<.05
Anxiety disorder	217 (16.6)	64 (11.0)	9.72	<.01	40 (11.6)	20 (13.0)	0.08	0.78
Adjustment disorder / PTSD	191 (14.6)	108 (18.5)	4.29	<.05	69 (20.1)	38 (24.7)	1.08	0.30
Other	343 (24.2)	166 (26.6)	1.18	0.28	67 (19.4)	37 (24.0)	1.14	0.29
Comorbidity (*n* (%))								
Two diagnoses	424 (30.0)	181 (29.0)	0.14	0.71	126 (36.4)	48 (31.2)	1.07	0.30
Three or more diagnoses	525 (37.0)	289 (47.7)	20.04	<.001	70 (20.2)	51 (33.1)	8.96	<.01

Note: Other diagnoses included e.g., obsessive-compulsive disorders, eating disorders, personality disorders, psychosis, and substance use disorders. For continuous variables, a *t* - test was used, for categorical variables, a chi² - test was used.

Patient characteristics of the training and holdout sample

Characteristics	**Training sample (*N* = 2043)**	**Holdout sample (*N* = 500)**	***t*-value / chi²-value**	***p*-value**
Age (*M* years (*SD*))	36.0 (12.8)	35.7 (14.0)	0.46	0.65
Gender female (*n* (%))	1279 (63.8)	297 (59.4)	1.62	0.20
German nationality (*n* (%))	1942 (95.1)	458 (91.6)	8.40	<.01
Marital status (*n* married (%))	568 (27.8)	119 (23.8)	3.06	0.08
Education (*n* > 12 years (%))	872 (42.7)	232 (46.4)	2.11	0.15
Inability to work (*n* (%))	418 (20.5)	85 (17.0)	2.82	0.09
Intake of medication (*n* (%))	1440 (70.0)	335 (66.7)	2.15	0.14
Number of dropouts (*n* (%))	625 (30.6)	154 (30.8)	0.00	0.97
Primary diagnosis (*n* (%))				
Affective disorder	954 (50.5)	229 (46.0)	3.08	0.08
Anxiety disorder	281 (14.9)	60 (12.0)	2.36	0.12
Adjustment disorder / PTSD*	299 (15.8)	107 (21.5)	8.51	<.01
Other	509 (24.9)	104 (20.8)	3.50	0.06
Comorbidity (*n* (%))				
Two diagnoses*	605 (29.6)	174 (34.8)	4.84	<.05
Three or more diagnoses*	823 (40.3)	121 (24.2)	43.83	<.001

Note: Other diagnoses included e.g., obsessive-compulsive disorders, eating disorders, personality disorders, psychosis, and substance use disorders. For continuous variables, a *t* - test was used, for categorical variables, a chi² - test was used.

Supplemental Materials 3

Mean scores of the models generated by each algorithm with all significant predictors from Zimmermann et al. (2017).

Overall rank	Algorithm	Brier score	AUC	Training AUC
1	ADA	.2259 (4)	.6393 (1)	.6456 (1)
2	GLMAIC	.2344 (5)	.6378 (2)	.6392 (2)
3	GLMBOOST	.2344 (6)	.6368 (3)	.6381 (7)
3	GBM	.2204 (2)	.6363 (7)	.6388 (5)
5	GLM	.2346 (9)	.6366 (4)	.6379 (8)
5	LDA	.2345 (7)	.6364 (6)	.6378 (10)
7	XGB	.2173 (1)	.6222 (13)	.6359 (12)
8	BAYESGLM	.2346 (10)	.6366 (5)	.6379 (9)
9	EN	.2345 (8)	.6325 (9)	.6389 (4)
10	SVM	.2350 (11)	.6330 (8)	.6345 (13)
10	RF	.2247 (3)	.5931 (16)	.6027 (16)
12	NNET	.2360 (13)	.6314 (11)	.6382 (6)
13	MONMLP	.2382 (16)	.6320 (10)	.6374 (11)
13	NB	.2361 (14)	.6239 (12)	.6282 (14)
15	AVNN	.2373 (15)	.6205 (14)	.6391 (3)
16	CART	.2352 (12)	.5783 (18)	.5881 (18)
17	MARS	.2635 (19)	.6033 (15)	.3995 (21)
17	CTREE	.2452 (17)	.5816 (17)	.6041 (15)
19	C4.5	.2649 (20)	.5723 (19)	.5902 (17)
19	LOGIT	.2527 (18)	.5572 (21)	.5625 (20)
21	kNN	.2743 (21)	.5663 (20)	.5794 (19)

Note: The digits in the brackets refer to the rank of the algorithm for the particular parameter. The overall rank is the sum of the single rankings concerning the two parameters without the training AUC. When sums were equal, the AUC was given priority. For the full names of the ML algorithms, see Table 2.

Mean scores of the models generated by each algorithm with all significant predictors identified with an elastic net analysis in the training sample.

Overall rank	Algorithm	Brier score	AUC	Training AUC
1	RF	.2037 (1)	.6584 (1)	.6610 (3)
2	GBM	.2090 (2)	.6567 (2)	.6637 (1)
3	ADA	.2093 (4)	.6515 (3)	.6607 (4)
4	XGB	.2090 (3)	.6442 (5)	.6624 (2)
5	GLMBOOST	.2306 (6)	.6469 (4)	.6531 (10)
6	EN	.2324 (8)	.6435 (6)	.6599 (5)
7	LDA	.2333 (9)	.6425 (7)	.6569 (7)
8	BAYESGLM	.2338 (10)	.6424 (8)	.6569 (8)
8	SVM	.2323 (7)	.6401 (11)	.6575 (6)
8	CART	.2113 (5)	.6256 (13)	.6278 (15)
11	GLM	.2339 (11)	.6422 (9)	.6568 (9)
12	GLMAIC	.2340 (12)	.6406 (10)	.6520 (11)
13	AVNN	.2374 (14)	.6289 (12)	.6489 (12)
14	NNET	.2459 (17)	.6247 (14)	.6436 (13)
14	MONMLP	.2381 (15)	.6079 (16)	.6217 (16)
16	CTREE	.2438 (16)	.5764 (17)	.5856 (17)
16	LOGIT	.2353 (13)	.5505 (20)	.5608 (19)
18	NB	.2856 (19)	.6214 (15)	.6325 (14)
19	kNN	.3052 (21)	.5608 (18)	.5564 (20)
19	C4.5	.2986 (20)	.5510 (19)	.5612 (18)
19	MARS	.2850 (18)	.5275 (21)	.4284 (21)

Note: The digits in the brackets refer to the rank of the algorithm for the particular parameter. The overall rank is the sum of the single rankings concerning the two parameters without the training AUC. When sums were equal, the AUC was given priority. For the full names of the ML algorithms, see Table 2.

Mean scores of the models generated by each algorithm with all variables.

Overall rank	Algorithm	Brier score	AUC	Training AUC
1	GBM	.2045 (1)	.6483 (2)	.6522 (1)
2	ADA	.2068 (3)	.6527 (1)	.6503 (2)
3	RF	.2053 (2)	.6423 (4)	.6443 (6)
4	GLMBOOST	.2311 (7)	.6426 (3)	.6492 (4)
4	XGB	.2096 (4)	.6319 (6)	.6489 (5)
6	EN	.2308 (6)	.6413 (5)	.6494 (3)
7	SVM	.2335 (8)	.6301 (7)	.6296 (8)
8	GLMAIC	.2375 (9)	.6291 (9)	.6287 (10)
9	CART	.2167 (5)	.6043 (14)	.6078 (14)
10	BAYESGLM	.2382 (12)	.6293 (8)	.6303 (7)
11	AVNN	.2376 (11)	.6241 (11)	.6296 (9)
12	LDA	.2389 (13)	.6243 (10)	.6265 (11)
13	MONMLP	.2376 (10)	.6008 (15)	.5965 (15)
14	GLM	.2396 (14)	.6234 (12)	.6259 (12)
15	NNET	.2498 (17)	.6226 (13)	.6131 (13)
16	CTREE	.2468 (16)	.5796 (17)	.5844 (17)
17	LOGIT	.2419 (15)	.5373 (19)	.5560 (18)
18	NB	.3680 (21)	.5957 (16)	.5962 (16)
19	C4.5	.3431 (20)	.5520 (18)	.5549 (19)
20	kNN	.3181 (19)	.5351 (20)	.5293 (20)
20	MARS	.2843 (18)	.5198 (21)	.4298 (21)

Note: The digits in the brackets refer to the rank of the algorithm for the particular parameter. The overall rank is the sum of the single rankings concerning the two parameters without the training AUC. When sums were equal, the AUC was given priority. For the full names of the ML algorithms, see Table 2.

Supplemental Materials 4

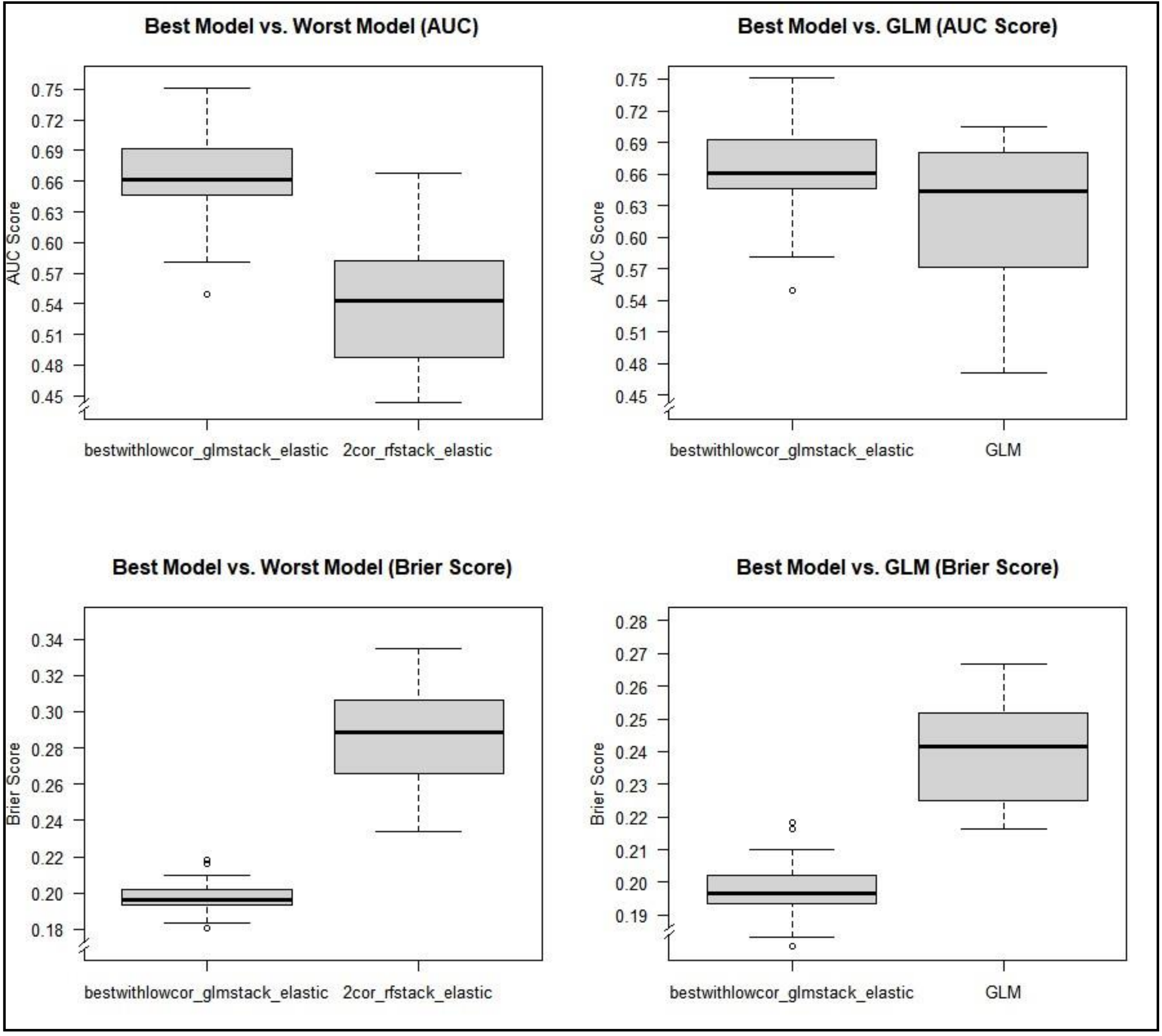
Overview of ensemble correlations by included predictors.

	Two least correlating	**Three least correlating**	**Best algorithm with least correlating**
All variables	CART + C4.5	kNN + C4.5 + GBM	GBM + kNN
Significant variables	MARS + C4.5	MARS + kNN + C4.5	ADA + LOGIT
Elastic net variables	LOGIT + MARS	LOGIT + MARS + C4.5	RF + kNN

Note: Significant variables = using significant variables from Zimmermann et al. (2017); Elastic net variables = using variables with predictive power in a preceding elastic net analysis; kNN = K-fold Nearest Neighbors; C4.5 = C4.5-like Trees; MARS = Bagged multivariate adaptive regression splines; LOGIT = Boosted logistic regression; CART = Bagged classification and regression tree; ADA = Adapted boosted classification tree; RF = Random forest; GBM = Generalized boosted regression model.

Supplemental Materials 5

Boxplots for the best models (overall), the worst models (overall) and the GLM models.



Note: GLM = Generalized linear model; bestwithlowcor_glmstack_elastic = ensemble of the best algorithm with its least correlating algorithm (i.e., random forest, nearest neighbor) using variables with predictive power in a preceding elastic net analysis stacked via GLM; 2cor_rfstack_elastic = ensemble of the two least correlating algorithms (i.e., boosted logistic regression, bagged multivariate adaptive regression splines) using variables with predictive power in a preceding elastic net analysis stacked via GLM; AUC = area under the curve.

Supplemental Materials 6

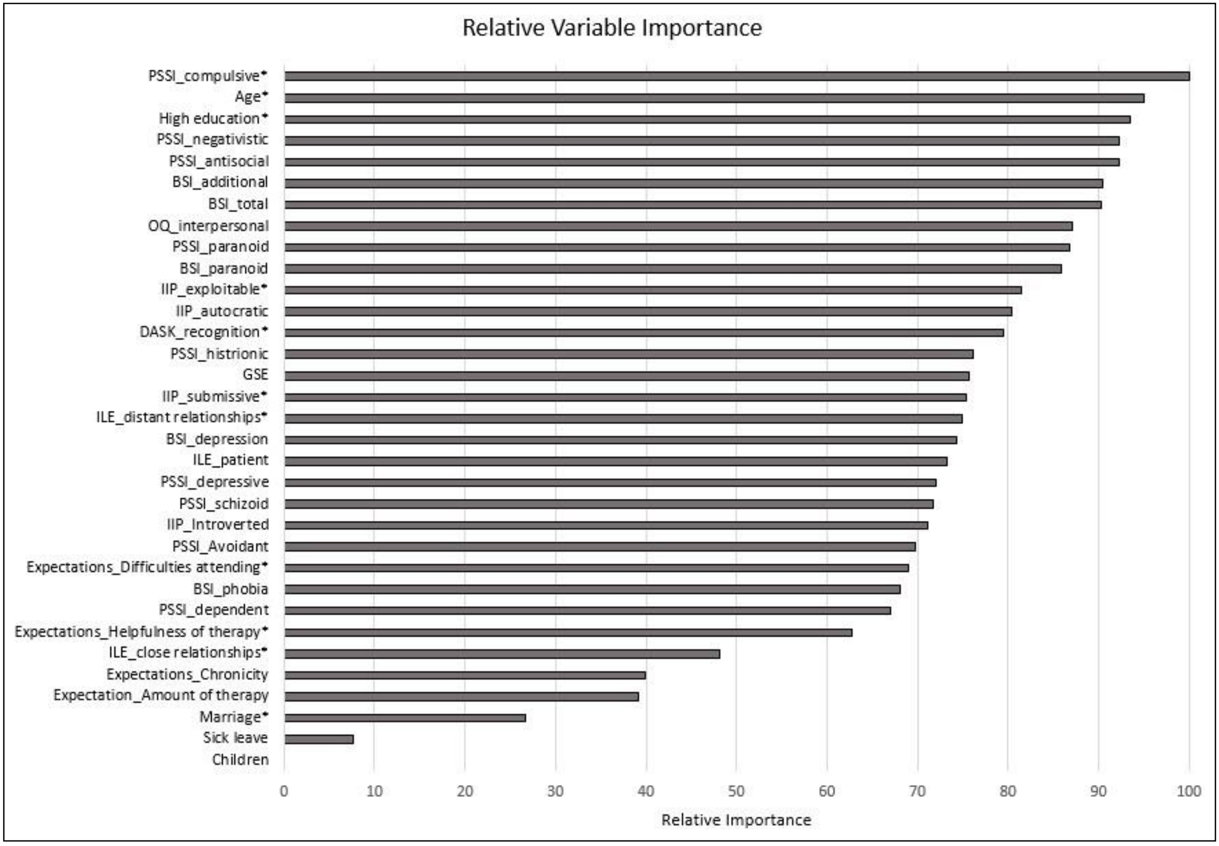
Confusion matrix of the best model generated by the best ensemble (i.e., best algorithm with its least correlating algorithm (RF, kNN) stacked via GLM using variables with predictive power in a preceding elastic net analysis)

	Observed		
Predicted	Regular	Dropout	Total
Regular	224	61	285
Dropout	122	93	215
Total	346	154	500

Note: kNN = K-fold nearest neighbors; RF = Random forest; GLM = Generalized linear model.

Supplemental Materials 7

Relative variable importance for the final model tested in the holdout sample.



Note: All variables not shown here were excluded from the ensemble by the preceding elastic net analysis (i.e., relative importance = 0). Since the nearest neighbor algorithm uses the entire data set and uses euclidean distance to determine predictions, no variable importance is available. Therefore, only the values from the random forest algorithm are shown here.

* The correlation between these variables and dropout is negative.

Study II

Poster, K., Bennemann, B., Hofmann, S. G., & Lutz, W. (2021). Therapist Interventions and Skills as Predictors of Dropout in Outpatient Psychotherapy. *Behavior therapy*, 52(6), 1489-1501. https://doi.org/10.1016/j.beth.2021.05.001

Author Contributions

All authors made substantial contributions to conception and design of the study. K. P. was responsible for conception, data preparation, data management, and writing of the manuscript. B. B. was responsible for data analyses and contributed to the writing of the results section. S. H. and W. L. critically evaluated the manuscript and added valuable comments. All authors contributed to and have approved the final manuscript.

Therapist Interventions and Skills as Predictors of Dropout in Outpatient Psychotherapy

Kaitlyn Poster[1], Björn Bennemann[1], Stefan G. Hofmann[2], Wolfgang Lutz[1]

[1]University of Trier

[2]Boston University

Author Note

Kaitlyn Poster https://orcid.org/0000-0002-1653-7772

Björn Bennemann https://orcid.org/0000-0002-4085-2262

Stefan G. Hofmann https://orcid.org/0000-0002-3548-9681

Wolfgang Lutz https://orcid.org/0000-0002-5141-3847

This work was supported by the German Research Foundation (DFG) under Grant Nr. LU 660/10-1. The authors report no conflict of interest.

Correspondence concerning this article should be addressed to Kaitlyn Poster, Department of Clinical Psychology and Psychotherapy, University of Trier, D-54296 Trier, Germany. E-mail: k.poster@uni-trier.de

Abstract

The current study employed machine learning to investigate whether the inclusion of observer-rated therapist interventions and skills in early sessions of psychotherapy improved dropout prediction beyond intake assessments. Patients were treated by post-graduate clinicians at a university outpatient clinic. Psychometric instruments were assessed at intake and therapeutic interventions and skills in the third session were routinely rated by independent observers. After variable pre-selection, an elastic net algorithm was used to build two dropout prediction models, one including and one excluding observer-rated session variables. The best model included observer-rated variables, and was significantly superior to the model including intake variables only. Alongside intake variables, two observer-rated variables significantly predicted dropout: therapist use of feedback and summaries and treatment difficulty. Observer ratings of therapist interventions and skills in early sessions of psychotherapy improve predictors of dropout from psychotherapy beyond intake variables alone. Future research could work toward personalizing dropout predictions to the specific dyad, thereby improving their validity and aiding therapists to recognize and react to increased dropout risk.

Keywords: dropout; machine learning; competence; video ratings; cognitive behavioral therapy

Therapist Interventions and Skills as Predictors of Dropout in Outpatient Psychotherapy

Dropout is a significant issue in psychological interventions, with a recent, comprehensive meta-analysis reporting an average of one in five patients terminating psychological treatments prematurely (Swift et al., 2017). Previous research has shown that patients who drop out of psychotherapy are characterized by less symptom improvement (Cahill et al., 2003), more dissatisfaction with therapy (Björk et al., 2009), persistent impairment (Barrett et al., 2008), and repeated utilization of the health care system (Carpenter et al., 1979). Therefore, dropout from psychological treatments enhances the burden of mental illness on society (Barrett et al., 2008).

Over the last several decades, several reviews and meta-analyses have identified a number of patient demographic variables associated with dropout, including younger age, lower socioeconomic status, lower education level, male sex, and ethnic minority status (Barrett et al., 2008; Garfield, 1994; Reis & Brown, 2006; Wierzbicki & Pekarik, 1993; Zimmermann et al., 2017). However, a more recent meta-analysis including 669 studies found only younger age and lower education level to be consistently associated with dropout. Further, eating and personality disorders were associated with higher dropout rates (Swift & Greenberg, 2012).

In recent years, the creation of large, naturalistic datasets and methodological advances have provided new opportunities to identify dropout predictors. For example, Lutz and colleagues (2018) used network analysis and machine learning algorithms to predict dropout based on patient intake variables and data from ecological momentary assessments. The authors were able to demonstrate that networks differed significantly between completers and dropouts and to identify six dropout predictors (initial impairment and sex, as well as four network parameters). Further, Lutz and colleagues (2019) used a machine learning approach (LASSO regression) to predict dropout based on a range of patient intake variables. The study identified seven significant predictors, including measures of initial general impairment, relationship impairment, personality traits, treatment expectations and level of education. These findings were used to develop personalized predictions of dropout for new patients. This information is provided to therapists at the beginning of treatment in the form of a visualization of the specific patient's dropout probability in comparison to that of the clinic's average patient. The utility of the comprehensive feedback system in which this dropout risk prediction tool is embedded is currently being evaluated (Lutz et al., 2017; Lutz et al., 2019).

While these are important efforts to better understand dropout and identify effective methods of minimizing its occurrence, it is important to remember that dropout does not only depend on the patient, but also on the therapist. Between 6.2% and 12.9 % of dropout variance can be attributed to the individual therapist, with therapists' individual dropout rates ranging between 1.2-73.2 % (Saxon et al., 2017; Zimmermann et al., 2017). Therefore, therapist variables seem to also make a substantial contribution to the likelihood of patients terminating treatment prematurely. However, therapist predictors of dropout have received much less research interest than patient predictors (Roos & Werbart, 2013). A few therapist variables have been associated with dropout in the literature, including therapist professional background (Hamilton et al., 2011), therapist experience level (Swift & Greenberg, 2012), ethnic match between patient and therapist, and therapist cultural competence (Owen et al., 2012; Sue, 1998). However, findings remain inconsistent and their generalizability questionable.

One rather consistent finding in the literature is the positive association between the therapeutic alliance and treatment completion (Sharf et al., 2010). A meta-analysis of 11 studies revealed a moderate average effect of alliance on dropout (d = .55), regardless of the source of alliance ratings (patient, therapist or observer-rated; Sharf et al., 2010). The therapist's ability to establish and maintain a positive therapeutic alliance with patients may therefore be a further source of therapist variance for dropout rates, beyond sociodemographic traits therapists bring into therapy. Various therapist behaviors have been associated with building a positive therapeutic alliance, including empathy, positive regard, and collaboration (Crits-Christoph et al., 2006). Further, therapist emotional supportiveness (Roos & Werbart, 2013) has been directly associated with lower dropout rates, as has therapist emotional intelligence (Kaplowitz et al., 2011), and more frequent repair of alliance ruptures (Muran et al., 2009). Therefore, therapists' competence in accepting the patient, conveying emotional understanding and successfully addressing tensions in the alliance may vary, which is also associated with therapists' dropout rates. Therapist competence has been widely studied in association with treatment outcome and findings have been mixed (Webb et al., 2010). However, investigations of a competence-dropout relation are lacking.

Beyond patient, therapist, and relationship factors, treatment factors have also been investigated as predictors of dropout. In their meta-analysis, Swift and Greenberg (2012) found time-limitation, manualization, and treatment setting (e.g., inpatient, outpatient, training clinics) to be associated with dropout, but not treatment orientation (e.g., cognitive behavioral therapy, psychodynamic therapy) or format (e.g., individual or group therapy). In a meta-analysis summarizing 587 studies, the same authors investigated whether dropout rates differed by treatment for specific disorders (Swift & Greenberg, 2014). The authors found treatment effects on dropout for only three of the 12 investigated diagnoses: depression, posttraumatic stress disorder, and eating disorders. For the first two diagnoses, integrative treatments showed the lowest dropout rates, while dialectical-behavior therapy had the lowest dropout rate for eating disorders.

To date, the investigation of treatment factors on dropout has focused on broader treatment characteristics, comparing, for example, treatment format, setting or entire treatment protocols (Swift & Greenberg, 2012, 2014). However, it remains unclear whether the application of specific therapeutic interventions, techniques, or strategies is systematically associated with dropout rates. Such findings would be especially applicable to current developments in CBT that put more emphasis on therapeutic strategies and processes (Hayes & Hofmann, 2018), moving away from strict manualization. In these contexts, therapists have the flexibility to select and order interventions as they deem appropriate for the individual patient. Currently, these clinical decisions are made unsystematically, largely based on the individual therapist's clinical experience, rather than on empirical findings (Lutz et al., 2019). However, gaining a better understanding of which treatment components are associated with dropout risk could inform treatment planning, especially in combination with an increasing understanding of patient dropout factors (e.g., avoiding interventions with a higher dropout risk with patients, whose intake characteristics already point toward an increased dropout risk). Such developments could possibly lower dropout rates in naturalistic settings, which have been shown to suffer from higher dropout rates than clinical trials (Swift & Greenberg, 2012).

The aim of the current study was to apply machine learning techniques to video rating data to examine therapist interventions and skills in early sessions of psychotherapy as further predictors of dropout, alongside patient intake characteristics and clinician intake assessments. It was hypothesized that including these observer-rated variables would improve dropout prediction beyond patient and clinician intake variables alone, as the interaction of the specific patient-therapist dyad was observed and further potential factors influencing dropout were considered.

Methods

Treatment

Therapies were conducted at a university outpatient CBT clinic in southwest Germany between 2017 and 2019 within the context of a randomized controlled trial investigating the effects of personalized prediction and adaptation tools on treatment outcome in outpatient psychotherapy (Lutz et al., 2017). The ethics committees of the University of Trier and the German Research Foundation approved the study (DFG, Grant no. LU 660/10-1). The trial was also registered at Current Controlled Trials (registration no. NCT03107845, registered on 30 March, 2017). Within the trial, process and outcome data were routinely collected and therapy sessions consistently videotaped. All patients and therapists included in the study consented to the use of their data (psychometric data and therapy videos) for research. Treatments had a mean length of 24.9 sessions (*SD* = 16.3; range: 1-72 sessions), which normally took place weekly. 69.1 % of patients completed treatment, while 30.9 % dropped out of therapy prematurely. Treatment length was significantly shorter when patients dropped out than when they completed therapy (M_{dropout} = 13.0 sessions; $M_{\text{completion}}$ = 30.1 sessions; $t(198.645) = 10.01$; $p < .001$).

Patients and Therapists

The sample consisted of a total of 259 patients. Patients were on average 36.3 years of age (*SD* = 13.3; range: 16–77 years), the majority were female (61.8 %). Patients were diagnosed at intake based on the German version of the Structured Clinical Interview for Axis I DSM-IV Disorders (SCID-I; Wittchen et al., 1997), which were conducted by intensively trained independent clinicians. SCID interviews were videotaped and discussed in expert consensus teams to enhance the validity of the intake diagnosis. At least four senior clinicians were part of each team and final diagnoses were determined by consensual agreement of at least 75% of the team members. Subsequently, patients were randomly assigned to their treating therapist. Table 1 provides an overview of patient characteristics and diagnoses.

Treatments were conducted by 65 therapists (83.1 % female). Therapists treated 1-13 patients each (*M* = 4.0 patients per therapist). All therapists had a Master’s degree in clinical psychology and were either currently participating in a 3-5 year post-graduate psychotherapy training program with a cognitive-behavioral therapy (CBT) focus or were already licensed CBT psychotherapists. Trainee therapists had at least 1.5 years of clinical experience prior to study participation. Therapists used individual case conceptualizations discussed in supervision as well as psychometric feedback and empirically-based treatment recommendations to personalize treatment to the individual patient.

Table 1
Patient characteristics

Characteristics	**Total sample (*N* = 259)**
Age (*M* years (*SD*))	36.3 (13.3)
Gender female (*n* (%))	160 (61.8 %)
German nationality (*n* (%))	242 (93.4 %)
Marital status (*n* married (%))	63 (24.0 %)
Education (*n* > 12 years (%))	129 (49.8 %)
Inability to work (*n* (%))	42 (16.2 %)
Primary diagnosis (*n* (%))	
Affective disorder	116 (44.8 %)
Anxiety disorder	34 (13.1 %)
Adjustment disorder / PTSD	57 (22.0 %)
Other	52 (20.1 %)
Comorbidity (*n* (%))	
Two diagnoses	96 (37.1 %)
Three or more diagnoses	82 (31.7 %)

Note: PTSD: Post-traumatic stress disorder; other diagnoses included e.g. substance use disorders, eating disorders, personality disorders, psychosis

Measures

Dropout

Dropout was assessed via therapist judgement at the end of treatment. If termination was planned and consensual, termination status was considered completed. In contrast, if the patient stopped coming to treatment, despite the therapist deeming continuation of treatment necessary, termination status was considered a dropout. Examples of dropout according to this definition include cases when a patient no longer showed up to appointments or was unable to be reached. Another example might be a patient telling the therapist that they are no longer interested in therapy, despite the therapist advising continuation.

Intake variables as dropout predictors

A total of 95 variables measured at intake or after the first session were examined as potential predictors of dropout. Variables stemmed from a questionnaire battery implemented in a comprehensive feedback and clinical support system (Lutz et al., 2019). Variables assessed demographics such as age, sex, and employment status as well as a variety of clinical factors including severity and chronicity of mental illness, interpersonal problems, functional impairment, dysfunctional attitudes, and treatment expectations. Furthermore, clinician-rated variables such as the alliance and expected improvement were assessed. For the complete list of intake variables, see Appendix 1.

Inventory of Therapeutic Interventions and Skills (ITIS)

The ITIS (Boyle et al., 2020) is a therapy video rating instrument that was developed to adequately assess the range of interventions and skills observable in modern, personalized CBT. The ITIS covers interventions from all three waves of CBT, including mindfulness-based, emotion-focused, interpersonal and resource-oriented interventions as well as therapeutic strategies corresponding to Grawe’s (2004) general mechanisms of change. As the inventory is conceptualized to assess treatment integrity in naturalistic and empirically personalized settings, adherence is defined as lege artis (i.e., state of the art) application of empirically-based interventions, rather than manual adherence.

The inventory comprises 20 intervention items, which are coded “0” if not observable and “1” to “3” if observable, whereby “1” reflects a low degree of lege artis application and “3” a high degree of lege artis application (i.e., adherence). Further, the inventory comprises 11 skills items, which are coded on a 7-point Likert scale ranging from “0” (poor) to “6” (excellent). The inventory also includes overall adherence and competence ratings as well as ratings of treatment difficulty and patient motivation, each also responded to on 7-point Likert scales. Therefore, the inventory comprises a total of 35 observer-rated variables. Average inter-rater reliability has been reported to be excellent for intervention ratings and good for skills ratings, independent of raters’ clinical experience (Boyle et al., 2020). For the complete list of ITIS variables, see Appendix 1.

Video Selection and Rating Procedure

In the current study, one early session in therapy was rated per case using the ITIS (*N* = 263). Routinely, the third session with the treating therapist was selected. If the video of the third session was a) unavailable (26 (9.9 %) videos), b) of too poor video quality (e.g., indecipherable voices; 22 (8.4 %) videos) or c) too long (> 70 minutes, 13 (4.9 %) videos), the next available subsequent videotaped session was rated. If patients dropped out of treatment before the third session (30 (11.6%)), the last available session was rated. Across the entire sample, on average, session 3.2 was rated (*SD* = 1.0; range: sessions 1-9).

The ITIS was applied by eight extensively trained raters: four graduate students of clinical psychology and four post-graduate clinicians who all participated in a 30-hour training program to learn to rate therapy videos using the inventory. During the independent rating phase, all raters received regular supervision and ratings were periodically compared to counter rater drift. Inter-rater reliability was excellent for the Interventions items (average Kendall's *W* across all raters: *M* = .821; *N* = 59 videos) and good for the Skills items (average Kendall's *W* across all raters: *M* = .738; *N* = 59 videos). All raters were blind to diagnoses, termination status (completion vs. dropout), and outcome.

Data Analytic Strategy

In order to generate the optimal dropout model, a three-step procedure was applied. All analyses were conducted using the free software environment R version 3.6.1 (R Core Team, 2013). In a first step, missing values were imputed using the R package missForest v4.6-14 (Stekhoven, 2015). All variables with less than 10% missingness were imputed; otherwise they were excluded from the analyses (for the exact number of missing data per variable, see Appendix 1). After all categorical variables had been dichotomized (i.e., dummy-coded if they had more than two categories for ease of interpretation (Asteriou & Hall, 2016)) and all continuous predictors had been centered, bivariate correlations were calculated to screen for possible predictors. Variables that significantly ($p < .05$) correlated with dropout were included in the subsequent models. This procedure was applied in order to exclude variables that have no predictive power concerning dropout and therefore would weaken the later model (see e.g., Lutz et al., 2019).

Second, a machine learning approach (i.e., elastic net) was used to generate dropout models. The application of machine learning algorithms to select predictive variables and generate prediction models has been considered an important analytical advancement on the road to precision mental health care (Maj et al., 2020). The R Package caret v6.0-84 (Kuhn, 2019) was applied, which adjusts algorithm parameters to their optimal settings depending on which algorithm is being used. For elastic net regularization, the R package glmnet v2.0-18 (Friedman et al., 2010) was used. An elastic net algorithm was chosen because it seems to handle clinical data well and therefore seemed very suitable for the analyses (Fisher & Bosley, 2020; Lutz et al., 2019; Webb et al., 2020). Furthermore, elastic net seems to protect well against overfitting (Pavlou et al., 2016) leading to better generalizability of the model. Elastic net regularization combines LASSO (least absolute shrinkage and selection operator) and ridge procedures. Both procedures shrink regression coefficients due to their predictive power, but only LASSO can shrink variable coefficients to zero.

For elastic net regularization we defined a range of values to be tested and allowed caret to identify the best fitting settings in order to achieve an optimal ratio of LASSO and ridge regression for our models. We set alpha to 0.1 for the first analysis and then altered alpha in increments of 0.1 until 1 was reached. An alpha of 0 is equal to a ridge regression, while an alpha of 1 equals a LASSO regression. We also defined lambda’s range analogue to the alpha parameter. Lambda defines the magnitude of the regression penalty. This resulted in 100 different possible combinations of these two parameters (10 values for alpha x 10 values for lambda) to identify the best fitting model. Identification of the best model was always based on the receiver operating characteristic (ROC) curve. The model with the highest value was considered the best model.

In this step, the dataset was split into a training (70% of the dataset) and a test set (remaining 30% of the dataset). The models were generated based on the training data and then evaluated in the test set to minimize overfitting (Rudin & Carlson, 2019). Each model was generated via a *k*-fold repeated cross-validation with 10 folds and three repetitions to select the optimal model (Collins et al., 2015). In addition, a sampling method (“up-sampling”) was used to minimize the impact of class imbalance (i.e., more completion cases than dropout cases). Class imbalance can lead to major consequences, such as very low predictive accuracy for the infrequent class (Hand & Vinciotti, 2003; Japkowicz & Stephen, 2002), which is why we applied an up-sampling procedure. Up-sampling equalizes class imbalance by randomly duplicating cases in the minority class and was preferred to down-sampling in order to prevent a decrease of the training sample size (McCarthy et al., 2005).

In order to minimize the influence of the training set’s specific sample characteristics, the split and model generation were repeated 100 times. Thereby, the training and test sets varied in each repetition. To assess the impact of the ITIS variables, this procedure was conducted twice, once under inclusion and once under exclusion of the ITIS variables. This approach was chosen to avoid multicollinearity, which would occur in multiple regression with a large number of variables (e.g., Slinker & Glantz, 1985).

The means of the following parameters across all 100 models generated for each of the respective test sets were examined for the set with the ITIS variables as well as for the set without them to identify the best overall model.

Brier score. The Brier score is a parameter that measures probabilistic predictions ranging from 0 (best prediction) to 1 (worst prediction) and was first proposed by Brier (1950). In contrast to accuracy, it takes the certainty of the prediction into account. In effect, it is the mean squared error of the forecast:

$$\frac{1}{N}\sum_{t=1}^{N}(f_t - o_t)^2$$

Hereby, *N* is the total number of observations, *f* is the probability of the event (i.e., dropout) and *o* is the actual outcome (i.e., 0 or 1) of the event at instance *t*.

Accuracy. Accuracy is the percentage of correctly identified dropouts ranging from 0 (worst prediction) to 1 (best prediction).

Area under the ROC curve (AUC). The AUC is a parameter that is calculated using the sensitivity and specificity of a prediction and ranges from 0 (worst prediction) to 1 (best prediction). Based on signal detection theory (Green & Swets, 1966), the AUC takes the base rate of the dependent variable into account.

Finally, the distributions of each parameter (i.e., Brier score, accuracy, AUC) were compared via three one-sided pairwise t-tests, one for each parameter. In order to assess the impact of the ITIS variables, Cohen's *d* was calculated for each of the three scores comparing the best score based on the data including the ITIS variables with the best score based on the data excluding the ITIS variables.

Results

Screening

Initial screening identified 56 of 130 variables significantly correlated with dropout (see Appendix 1), including three ITIS items (cognitive techniques (I6), use of feedback/ summaries (S3), and treatment difficulty). The application of cognitive techniques was associated with a significantly lower dropout rate, while the use of feedback/ summaries and treatment difficulty were related to higher dropout rates. The average inter-rater reliability of these three items across all raters was Kendall's *W* = .796, .681 and .768, respectively. This represents good to excellent reliability (Cicchetti, 1994).

Model Generation and Descriptives

The means of the relative importance of the pre-selected predictors across all 100 elastic net model generations are displayed with their respective error bars in Figure 1. Relative importance is a caret function that indicates the importance of predictors for model generation compared to the most important variable. The most important variable always has a score of 100, while all variables that have a value of 0 are excluded from the model. The elastic net algorithm included the ITIS variables when building models for most of the 100 different training sets (73% for cognitive techniques (I6), 96% for use of feedback/ summaries (S3), and 97% for treatment difficulty), indicating the importance of these items for dropout model generation.

Examining the mean prediction scores revealed that the best models were generated by the elastic net algorithm using the dataset including ITIS variables. This was the case for all three parameters (i.e., Brier score, accuracy, AUC; see Table 2). For the confusion matrices, see Table 3.

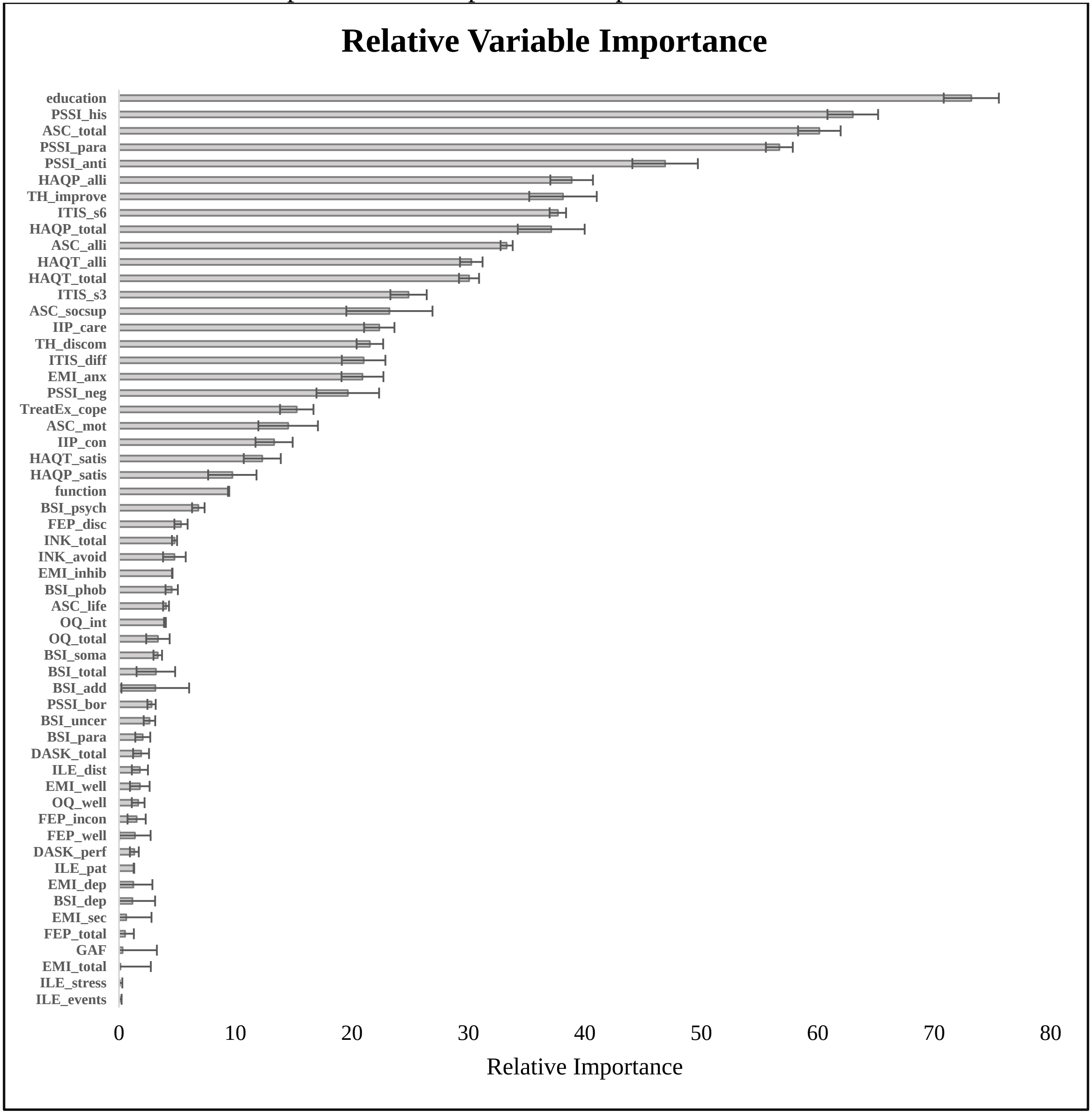
Table 2
Mean scores of the 100 models for each dataset and parameter

Dataset	**Brier score**	**Accuracy**	**AUC**
Dataset without ITIS variables	0.2138 (0.0140)	0.6687 (0.0474)	0.7130 (0.0588)
Dataset with ITIS variables	0.2096 (0.0152)	0.6758 (0.0533)	0.7226 (0.0592)

Note: AUC: area under the curve.

Figure 1

Means of the relative importance of the pre-selected predictor variables across all 100 models.



Note: PSSI: Personality Style and Disorder Inventory; ASC: Assessment for Signal Clients; HAQP: Helping Alliance Questionnaire – Patient version; TH: Therapist Expectations; ITIS: Inventory of Therapeutic Interventions and Skills; HAQT: Helping Alliance Questionnaire – Therapist version; IIP: Interpersonal Problems; EMI: Emotionality Inventory; TreatEx: Treatment Expectations; BSI: Brief Symptom Inventory; FEP: Questionnaire for the Evaluation of Psychotherapeutic Progress; INK: Incongruence Questionnaire; OQ: Outcome Questionnaire; DASK: Dysfunctional Attitudes Scale – Short form; ILE: Inventory of Stressful Life Events; GAF: Global Assessment of Functioning.

Table 3
Mean scores across all 100 confusion matrices for the models excluding and including ITIS variables.

Model including intake variables only	**Observed**		
Predicted	Regular	Dropout	Total
Regular	37.63	9.13	46.76
Dropout	16.71	14.53	31.24
Total	54.34	23.66	78
Model including intake and ITIS variables	**Observed**		
Predicted	Regular	Dropout	Total
Regular	38.00	8.95	49.58
Dropout	16.34	14.71	28.42
Total	54.34	23.66	78

Note: ITIS: Intervention of Therapeutic Interventions and Skills

Model Comparisons

Confirming our hypothesis, the *t*-tests revealed a significant effect between the dropout model including the ITIS variables as predictors and the model based on intake variables alone for the mean of the Brier score ($t(99) = 6.49$; $p < .001$; Cohen's $d = 0.65$), accuracy ($t(99) = 2.10$; $p < .05$; Cohen's $d = 0.21$) and the AUC ($t(99) = 3.93$; $p < .001$; Cohen's $d = 0.39$). According to Cohen's (1988) conventions, a small to medium effect of including the ITIS variables was found. Table 4 displays the predictors that were included in 95 of the dropout prediction models generated with elastic net. Although these six variables did not always have the highest mean importance, they remain the only ones to be included in a significant number of model generations. Therefore, it is unclear whether the other variables have an impact on model generation independent of the specific training set.

Table 4
Mean effects (n=100) of predictor variables included in at least 95 of the models generated by elastic net.

Variables	**Estimate**	**SE**	**Relative importance**	**% Model inclusion**
Intercept	-0.029	0.007		
ITIS treatment difficulty	0.084	0.005	21.008	97%
ITIS-S3 use of feedback /summaries	0.101	0.006	24.868	96%
PSSI-K subscale paranoid personality style	0.229	0.012	56.709	98%
ASC total score	-0.244	0.014	60.145	96%
PSSI-K subscale histrionic personality style	0.263	0.014	63.018	97%
High education (yes/no)	-0.310	0.017	73.195	96%

Note: PSSI-K: Personality style and disorder inventory – short form; ASC: Assessment for signal clients; High education: Does the patient have a university entrance qualification?; SE is the error of the mean for all 100 models; Relative importance: Importance of the variables compared to all other variables, a value of 0 means the variable was excluded during model generation, the maximum value is 100; % Model inclusion: Percentage of the 100 models in which the variable was included.

Discussion

The aim of the current study was to examine therapist interventions and skills in early sessions of psychotherapy as further predictors of dropout, alongside intake variables. Prediction models with and without the inclusion of video ratings of interventions and skills in the third session were built and compared using three different indices to gain a comprehensive picture of the supplemental variables' contribution to dropout prediction.

During predictor pre-selection, three video rating variables were found to correlate significantly with dropout: therapist application of cognitive techniques, therapist use of feedback and summaries, and treatment difficulty. The model that also included ITIS-variables outperformed the model that only included variables at intake and after the screening session. The added predictive power of therapist interventions and skills ratings corresponded to a small to medium effect.

In line with previous findings (Lutz et al., 2019; Swift & Greenberg, 2012; Zimmermann et al., 2017), several intake variables, including histrionic personality style and lower education level significantly predicted dropout. Furthermore, examining observer-rated therapist and treatment variables early in therapy seems to provide information that is relevant to later premature termination of therapy. Despite a much larger number of intake variables being correlated with dropout than video rating variables, these observations of the dyad interacting in an early session seemed to, at least to some degree, measure something that was not already captured by intake assessments. These results are in line with previous findings that not only patient characteristics, but also therapist and treatment characteristics affect dropout probability (Saxon et al., 2017; Swift & Greenberg, 2012, 2014).

The finding that the application of *cognitive techniques* seems to be protective of dropout is also interesting. Cognitive techniques are a core therapeutic strategy in CBT, which work to identify and modify biased and dysfunctional thoughts and beliefs that maintain problematic behavior and psychological symptoms (e.g., depressive mood; Beck et al., 1979). Cognitive procedures have been shown to effectively reduce symptoms and cognitive change has been demonstrated to contribute to symptom change (Lorenzo-Luaces et al., 2015). When cognitive techniques are applied effectively early in treatment, they may facilitate symptom relief, in turn strengthening the patient’s confidence in the treatment’s effectiveness and increasing commitment to continue treatment. However, it is also possible that therapists of patients at risk of dropping out were more likely to delay the implementation of cognitive strategies, as they may have felt that they first had to work on trying to keep the patient in treatment (i.e., facilitate basic patient behavior, see Schulte & Eifert, 2002).

The result that therapists’ competent application of *feedback and summaries* was associated with a higher dropout risk seems surprising. Regularly asking patients for feedback and providing summaries of session content are considered important therapeutic skills that are applied to structure the session, provide patients with a sense of being heard and taken seriously, and facilitate patient processing and recall of therapy content (Beck et al., 1979). It is possible that the intensified use of these strategies are not necessarily a cause of dropout, but rather a therapeutic reaction to an at-risk patient. Should therapists perceive the patient as difficult or at risk of dropping out of treatment, they may react by making a stronger effort to gain feedback on the patient’s perspective and provide summaries to give the patient with a sense of orientation in therapy. In fact, Cooper and colleagues (2016) also found more observer-rated negotiating and structuring of early CBT sessions to be predictive of dropout and came to a similar interpretation of their findings. Therefore, the therapist’s increased use of these techniques in early sessions may signal at-risk patients rather than cause dropout.

Higher observer-rated *treatment difficulty* was also predictive of dropout. This result is consistent with findings showing patient intake symptom severity, interpersonal impairment, and personality disorders to be associated with a higher dropout probability (Lutz et al., 2018; Lutz et al., 2019; Swift & Greenberg, 2012). Corresponding intake variables were also associated with dropout in this study (e.g., interpersonal impairment, histrionic personality style; see Appendix 1). However the observer perspective on the difficulty of treating the specific patient improved prediction accuracy beyond these variables. Therefore, an independent observer may be able to assess aspects of patient difficulty relevant to premature termination that are not fully captured by patient self-reports or therapist intake assessments of patient impairment.

Finally, when considering previous findings in the literature (Crits-Christoph et al., 2006; Roos & Werbart, 2013; Sharf et al., 2010), the lack of an association between the video rating variables *therapeutic relationship/collaboration* and *empathic understanding* and dropout seems somewhat surprising. However, the alliance is a dynamic construct (Zilcha-Mano, 2019) and it is possible that the observer-rated alliance level in a single early session is not predictive of later dropout, while negative developments in a dyad's alliance over time (i.e., alliance ruptures, Eubanks-Carter et al., 2010) may very well be. In line with the lack of a consistent competence-outcome relation in the literature (Webb et al., 2010), overall competence, as measured with the ITIS, was not significantly associated with dropout.

Strengths and Limitations

To our knowledge, this study was the first to examine a wide range of observer-rated therapist interventions and skills in early sessions of psychotherapy as predictors of dropout beyond questionnaire data. A sophisticated machine learning approach (i.e., elastic net) was applied to fit prediction models to the data and the Brier score, accuracy and the AUC were all reported to uncover possible inconsistencies in the results. Such studies are rare, because having therapy video content rated by independent observers for a sample large enough for the application of machine learning approaches is costly and time-consuming. However, this approach seems fruitful to better understand and predict dropout, as it provides a further, valuable perspective on therapy process beyond commonly applied questionnaire assessments. Although the added predictive benefit of the observer-rated variables may seem descriptively small, it is important to remember that these variables made a significant contribution to dropout prediction beyond intake variables that already covered a very wide range of potentially influencing factors. Therefore, this study speaks to the feasibility of combining routine data collection, video coding, and machine learning to further our understanding of dropout.

Several limitations must also be considered when interpreting the results of this study. Firstly, the dropout literature has emphasized the importance of the dropout definition for both the prevalence of dropout and its interpretation (Zimmermann et al., 2017). In this study, dropout status was determined by the therapist at the end of treatment, whereby dropout was defined as non-consensual termination of treatment (the patient discontinued treatment despite the therapist’s indication for continuation). This definition does not consider therapy phase and dropout within the first few sessions may be qualitatively different from dropout later in treatment, especially as treatment length was comparatively long in this study.

Further, as we conducted a standardized assessment of the third session, the number of sessions between the video-rated session and the dropout event, and therefore the temporal association between therapist interventions and skills and dropout, varied largely. We may have found stronger associations if we had consistently assessed the pre-dropout session, regardless of therapy phase. However, with the aim of developing dropout risk predictions that can be provided to therapists early in treatment, the standardized assessment of an early session seemed more advantageous.

In addition, we did not assess the dynamic nature of constructs such as aspects of the therapeutic alliance via repeated measures. However, recent psychotherapy process literature has emphasized the importance of doing so (Zilcha-Mano, 2019). A general limitation of machine learning approaches is further that no causal conclusions can be drawn (Wilkinson et al., 2020). Therefore, our results must be interpreted with a degree of caution. Also, treatments took place in the context of a university outpatient clinic with a comprehensive feedback and clinical support system. Future research must evaluate whether these findings are generalizable to other populations and settings. Finally, for many routine care settings, it is likely not feasible to implement resource-intensive observer ratings in a timely and routine manner, limiting the general applicability of this approach. However, this approach could be especially interesting in psychotherapy training contexts, particularly where observer ratings of treatment integrity are already an integral part of training and supervision. Ratings may help to improve dropout predictions and support novice therapists to recognize and appropriately react to at-risk patients.

Summary, Implications, and Future Directions

The current study demonstrated that predictions of dropout from outpatient psychotherapy can be improved when therapist interventions and skills in an early session with the patient are considered. These findings highlight that there is knowledge to be gained by actually watching therapists at work and that investments in video ratings are valuable to psychotherapy process research.

Future research should further explore treatment data in the context of dropout to improve existing prediction models and incorporate these advancements into clinical support systems. Warning therapists of their patients’ increased dropout risk early on may help them to change their strategy or seek supervision, therefore reducing dropout rates and the associated burdens on patients, service providers, and society.

References

Asteriou, D., & Hall, S. G. (2016). *Applied econometrics* (3rd edition). *Macmillan education*. Palgrave. https://doi.org/10.1057/978-1-137-41547-9_1

Barrett, M. S., Chua, W.-J., Crits-Christoph, P., Gibbons, M. B. C., & Thompson, D. (2008). Early withdrawal from mental health treatment: Implications for psychotherapy practice. *Psychotherapy: Theory, Research, Practice, Training*, *45*(2), 247. https://doi.org/10.1037/0033-3204.45.2.247

Beck, A. T., Rush, A. J., Shaw, B. F., & Emery, G. (1979). *Cognitive therapy of depression*. Guilford Press.

Björk, T., Björck, C., Clinton, D., Sohlberg, S., & Norring, C. (2009). What happened to the ones who dropped out? Outcome in eating disorder patients who complete or prematurely terminate treatment. *European Eating Disorders Review : The Journal of the Eating Disorders Association*, *17*(2), 109–119. https://doi.org/10.1002/erv.911

Boyle, K., Deisenhofer, A.-K., Rubel, J. A., Bennemann, B., Weinmann-Lutz, B., & Lutz, W. (2020). Assessing treatment integrity in personalized CBT: the inventory of therapeutic interventions and skills. *Cognitive Behaviour Therapy*, *49*(3), 210–227. https://doi.org/10.1080/16506073.2019.1625945

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3.

Cahill, J., Barkham, M., Hardy, G., Rees, A., Shapiro, D. A., Stiles, W. B., & Macaskill, N. (2003). Outcomes of patients completing and not completing cognitive therapy for depression. *British Journal of Clinical Psychology*, *42*(2), 133–143. https://doi.org/10.1348/014466503321903553

Carpenter, P. J., Del Gaudio, A. C., & Morrow, G. R. (1979). Dropouts and terminators from a community mental health center: Their use of other psyciatric services. *Psychiatric Quarterly*, *51*(4), 271–279.

Cicchetti, D. V. (1994). Interreliability Standards in Psychological Evaluations. *Psychol Assess*, 284–290.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge.

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. M. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMC Medicine*, *13*(1), 1. https://doi.org/10.1186/s12916-014-0241-z

Cooper, A. A., Strunk, D. R., Ryan, E. T., DeRubeis, R. J., Hollon, S. D., & Gallop, R. (2016). The therapeutic alliance and therapist adherence as predictors of dropout from cognitive therapy for depression when combined with antidepressant medication. *Journal of Behavior Therapy and Experimental Psychiatry*, *50*, 113–119. https://doi.org/10.1016/j.jbtep.2015.06.005

Crits-Christoph, P., Gibbons, M. B. C., Crits-Christoph, K., Narducci, J., Schamberger, M., & Gallop, R. (2006). Can therapists be trained to improve their alliances? A preliminary study of alliance-fostering psychotherapy. *Psychotherapy Research*, *16*(03), 268–281. https://doi.org/10.1080/10503300500268557

Eubanks-Carter, C., Muran, J. C., & Safran, J. D. (2010). Alliance ruptures and resolution. *The Therapeutic Alliance: An Evidence-Based Guide to Practice*, 74–94.

Fisher, A. J., & Bosley, H. G. (2020). Identifying the presence and timing of discrete mood states prior to therapy. *Behaviour Research and Therapy*, *128*, 103596. https://doi.org/10.1016/j.brat.2020.103596

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1. https://doi.org/10.18637/jss.v033.i01

Garfield, S. L. (1994). Research on client variables in psychotherapy. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change* (4th ed., pp. 190–228). Wiley.

Grawe, K. (2004). *Psychological therapy*. Hogrefe Publishing.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Wiley New York.

Hamilton, S., Moore, A. M., Crane, D. R., & Payne, S. H. (2011). Psychotherapy dropouts: Differences by modality, license, and DSM-IV diagnosis. *Journal of Marital and Family Therapy*, *37*(3), 333–343. https://doi.org/10.1111/j.1752-0606.2010.00204.x

Hand, D. J., & Vinciotti, V. (2003). Choosing k for two-class nearest neighbour classifiers with unbalanced classes. *Pattern Recognition Letters*, *24*(9-10), 1555–1562. https://doi.org/10.1016/S0167-8655(02)00394-X

Hayes, S. C., & Hofmann, S. G. (2018). *Process-based CBT: The science and core clinical competencies of cognitive behavioral therapy*. New Harbinger Publications.

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis*, *6*(5), 429–449. https://doi.org/10.3233/IDA-2002-6504

Kaplowitz, M. J., Safran, J. D., & Muran, C. J. (2011). Impact of therapist emotional intelligence on psychotherapy. *The Journal of Nervous and Mental Disease*, *199*(2), 74–84. https://doi.org/10.1097/NMD.0b013e3182083efb

Kuhn, M. (2019). *caret: Classification and Regression Training*. R package version 6.0-84.

Lorenzo-Luaces, L., German, R. E., & DeRubeis, R. J. (2015). It's complicated: The relation between cognitive change procedures, cognitive change, and symptom change in cognitive therapy for depression. *Clinical Psychology Review*, *41*, 3–15. https://doi.org/10.1016/j.cpr.2014.12.003

Lutz, W., Rubel, J. A., Schwartz, B., Schilling, V., & Deisenhofer, A.-K. (2019). Towards integrating personalized feedback research into clinical practice: Development of the Trier Treatment Navigator (TTN). *Behaviour Research and Therapy*, *120*, 103438. https://doi.org/10.1016/j.brat.2019.103438

Lutz, W., Schwartz, B., Hofmann, S. G., Fisher, A. J., Husen, K., & Rubel, J. A. (2018). Using network analysis for the prediction of treatment dropout in patients with mood and anxiety disorders: A methodological proof-of-concept study. *Scientific Reports*, *8*(1), 7819. https://doi.org/10.1038/s41598-018-25953-0

Lutz, W., Zimmermann, D., Müller, V. N. L. S., Deisenhofer, A.-K., & Rubel, J. A. (2017). Randomized controlled trial to evaluate the effects of personalized prediction and adaptation tools on treatment outcome in outpatient psychotherapy: Study protocol. *BMC Psychiatry*, *17*(1), 306. https://doi.org/10.1186/s12888-017-1464-2

Maj, M., Stein, D. J., Parker, G., Zimmerman, M., Fava, G. A., Hert, M. de, Demyttenaere, K., McIntyre, R. S., Widiger, T., & Wittchen, H.-U. (2020). The clinical characterization of the adult patient with depression aimed at personalization of management. *World Psychiatry*, *19*(3), 269–293. https://doi.org/10.1002/wps.20771

McCarthy, K., Zabar, B., & Weiss, G. (2005). Does cost-sensitive learning beat sampling for classifying rare classes? *Proceedings of the 1st International Workshop on Utility-Based Data Mining*, 69–77. https://doi.org/10.1145/1089827.1089836

Muran, J. C., Safran, J. D., Gorman, B. S., Samstag, L. W., Eubanks-Carter, C., & Winston, A. (2009). The relationship of early alliance ruptures and their resolution to process and outcome in three time-limited psychotherapies for personality disorders. *Psychotherapy (Chicago, Ill.)*, *46*(2), 233–248. https://doi.org/10.1037/a0016085

Owen, J., Imel, Z., Adelson, J., & Rodolfa, E. (2012). 'no-show': Therapist racial/ethnic disparities in client unilateral termination. *Journal of Counseling Psychology*, *59*(2), 314–320. https://doi.org/10.1037/a0027091

Pavlou, M., Ambler, G., Seaman, S., Iorio, M. de, & Omar, R. Z. (2016). Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. *Statistics in Medicine, 35*(7), 1159–1177. https://doi.org/10.1002/sim.6782

R Core Team. (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/

Reis, B. F., & Brown, L. G. (2006). Preventing therapy dropout in the real world: The clinical utility of videotape preparation and client estimate of treatment duration. *Professional Psychology: Research and Practice*, *37*(3), 311. https://doi.org/10.1037/0735-7028.37.3.311

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

Rudin, C., & Carlson, D. (2019). The Secrets of Machine Learning: Ten Things You Wish You Had Known Earlier to be More Effective at Data Analysis. *ArXiv Preprint ArXiv:1906.01998*. https://doi.org/10.1287/educ.2019.0200

Saxon, D., Barkham, M., Foster, A., & Parry, G. (2017). The Contribution of Therapist Effects to Patient Dropout and Deterioration in the Psychological Therapies. *Clinical Psychology & Psychotherapy*, *24*(3), 575–588. https://doi.org/10.1002/cpp.2028

Schulte, D., & Eifert, G. H. (2002). What to do when manuals fail? The dual model of psychotherapy. *Clinical Psychology: Science and Practice, 9*(3), 312–328. https://doi.org/10.1093/clipsy.9.3.312

Sharf, J., Primavera, L. H., & Diener, M. J. (2010). Dropout and therapeutic alliance: A meta-analysis of adult individual psychotherapy. *Psychotherapy: Theory, Research, Practice, Training*, *47*(4), 637. https://doi.org/10.1037/a0021175

Slinker, B. K., & Glantz, S. A. (1985). Multiple regression for physiological data analysis: the problem of multicollinearity. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, *249*(1), R1-R12. https://doi.org/10.1152/ajpregu.1985.249.1.R1

Stekhoven, D. J. (2015). missForest: Nonparametric missing value imputation using random forest. *Astrophysics Source Code Library*.

Sue, S. (1998). In Search of Cultural Competence in Psychotherapy and Counseling. *American Psychologist*, *53*(4), 440–448. https://doi.org/10.1037/0003-066X.53.4.440

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. (2014). A treatment by disorder meta-analysis of dropout from psychotherapy. *Journal of Psychotherapy Integration*, *24*(3), 193–207. https://doi.org/10.1037/a0037512

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 (Chicago, Ill.)*, *54*(1), 47–57. https://doi.org/10.1037/pst0000104

Webb, C. A., Cohen, Z. D., Beard, C., Forgeard, M., Peckham, A. D., & Björgvinsson, T. (2020). Personalized prognostic prediction of treatment outcome for depressed patients in a naturalistic psychiatric hospital setting: A comparison of machine learning approaches. *Journal of Consulting and Clinical Psychology*, *88*(1), 25–38. https://doi.org/10.1037/ccp0000451

Webb, C. A., DeRubeis, R. J., & Barber, J. P. (2010). Therapist adherence/competence and treatment outcome: A meta-analytic review. *Journal of Consulting and Clinical Psychology*, *78*(2), 200–211. https://doi.org/10.1037/a0018912

Wierzbicki, M., & Pekarik, G. (1993). A meta-analysis of psychotherapy dropout. *Professional Psychology: Research and Practice*, *24*(2), 190. https://doi.org/10.1037/0735-7028.24.2.190

Wilkinson, J., Arnold, K. F., Murray, E. J., van Smeden, M., Carr, K., Sippy, R., Kamps, M. de, Beam, A., Konigorski, S., & Lippert, C. (2020). Time to reality check the promises of machine learning-powered precision medicine. *The Lancet Digital Health*. https://doi.org/10.1016/S2589-7500(20)30200-4

Wittchen, H.-U., Wunderlich, U., Gruschwitz, S., & Zaudig, M. (1997). SKID I. Strukturiertes Klinisches Interview für DSM-IV. Achse I: Psychische Störungen. Interviewheft und Beurteilungsheft. Eine deutschsprachige, erweiterte Bearb. d. amerikanischen Originalversion des SKID I.

Zilcha-Mano, S. (2019). Major developments in methods addressing for whom psychotherapy may work and why. *Psychotherapy Research : Journal of the Society for Psychotherapy Research*, *29*(6), 693–708. https://doi.org/10.1080/10503307.2018.1429691

Zimmermann, D., Rubel, J., Page, A. C., & Lutz, W. (2017). Therapist Effects on and Predictors of Non-Consensual Dropout in Psychotherapy. *Clinical Psychology & Psychotherapy*, *24*(2), 312–321. https://doi.org/10.1002/cpp.2022

Supplementary Material

Therapist Interventions and Skills as Predictors of Dropout in Outpatient Psychotherapy

Kaitlyn Poster[1], Björn Bennemann[1], Stefan G. Hofmann[2], Wolfgang Lutz[1]

[1]University of Trier

[2]Boston University

Appendix 1
All predictor variables

Categorical
Sex (0)
Education* (1)
Employment status (4)
Children (0)
Marital status (1)
Medication (4)
Inventory of Therapeutic Interventions and Skills (ITIS)[1]
ITIS-I1 Therapy goals (0)
ITIS-I2 Functional analysis (0)
ITIS-I3 Psychoeducation (0)
ITIS-I4 Suicidality / Crisis intervention (0)
ITIS-I5 Behavior modification (0)
ITIS-I6 Cognitive techniques (0)*
ITIS-I7 Skills training (0)
ITIS-I8 Exposure / Behavioral experiments (0)
ITIS-I9 Emotion regulation (0)
ITIS-I10 Mindfulness / Acceptance (0)
ITIS-I11 Clarification of schemas/ needs/ motives/ values/ goals (0)
ITIS-I12 Emotion-focused techniques (0)
ITIS-I13 Interpersonal techniques (0)
ITIS-I14 Resource/ Solution-oriented techniques (0)
ITIS-I15 Homework (0)
ITIS-I16 Motivational clarification (0)
ITIS-I17 Problem actuation (0)
ITIS-I18 Mastery (0)
ITIS-I19 Resource activation (0)
ITIS-I20 Use of psychometric feedback (0)

Continuous
Age (0)
OQ total score (1)*
OQ Symptom Distress (SD) subscale (1)*
OQ Social Role functioning (SR) subscale (12)
OQ Interpersonal Relationship (IP) subscale (1)*
Questionnaire for the evaluation of psychotherapy (FEP2) total score (1)*
FEP2 Well-being subscale (2)*
FEP2 Discomfort subscale (1)*
FEP2 Incongruence subscale (1)*
FEP2 Interpersonal Problem subscale (2)
Emotionality Inventory (EMI) total score (5)*
EMI Anxiety subscale (8)*
EMI Depression subscale (8)*
EMI Inhibition subscale (7)*
EMI Security subscale (8)*
EMI Wellbeing subscale (6)*
Brief Symptom Inventory (BSI) total score (2)*
BSI Somatic problem subscale (3)*
BSI Obsessive Compulsive Subscale (2)
BSI Uncertainty subscale (3)*
BSI Depression subscale (2)*
BSI Anxiety subscale (2)*
BSI Hostility subscale (2)
BSI Phobia subscale (2)
BSI Paranoid subscale (2)*
BSI Psychoticism (2)*
BSI Additional (4)*
Interpersonal Problems (IIP-32) total score (1)
IIP-32 Autocratic/dominant subscale (4)
IIP-32 Confrontational subscale (4)*
IIP-32 Unapproachable subscale (5)
IIP-32 Introverted subscale (3)
IIP-32 Submissive subscale (6)
IIP-32 Exploitable (2)
IIP-32 Caring subscale (3)*
IIP-32 Expressive subscale (2)
Incongruence questionnaire (INK-23) total score (2)*
INK-23 Approach subscale (2)
INK-23 Avoidance subscale (4)*
Dysfunctional attitudes scale – short form (DAS-K) total score (2)*
DAS-K Recognition subscale (2)
DAS-K Performance subscale (3)*
Inventory of Stressful Life-Events (ILE) – Score for number of events (2)*
ILE – Score for stress (2)*
ILE – Number of events in patient's life subscale (2)*
ILE – Number of events in life of close relationships subscale (2)
ILE – Number of events in life of distant relationships subscale (2)*
General Perceived Self-Efficacy Scale (GSE) total score (4)
Personality style and disorder inventory – short form (PSSI-K)[1]
PSSI-K subscale – Antisocial personality style (4)*
PSSI-K subscale – Paranoid personality style (7)*
PSSI-K subscale – Schizoid personality style (9)
PSSI-K subscale – Avoidant personality style (7)
PSSI-K subscale – Compulsive personality style (7)
PSSI-K subscale – Schizotypal personality style (6)
PSSI-K subscale – Rhapsodic personality style (9)
PSSI-K subscale – Narcissistic personality style (6)
PSSI-K subscale – Negativistic personality style (7)*
PSSI-K subscale – Dependent personality style (13)
PSSI-K subscale – Borderline personality style (10)*
PSSI-K subscale – Histrionic personality style (7)*
PSSI-K subscale – Depressive personality style (9)
PSSI-K subscale – Altruistic personality style (5)
Patient rated well-being (4)*
Current emotional and psychological functioning (6)
Therapy expectations – Importance of psychotherapy (5)
Therapy expectations – Difficulties attending psychotherapy (4)
Therapy expectations – Confidence in the helpfulness of psychotherapy in dealing with problems (3)
Therapy expectations – Amount of previous psychotherapy (5)
Therapy expectations – Chronicity of the problem (4)
Therapy expectations – Estimated future coping (7)*
Therapist rated wellbeing – Patient's recent discomfort (10)*
Therapist rated wellbeing – Current effect of psychotherapy on the patient (10)
Therapist rated wellbeing – Expected patient improvement with further psychotherapy (10)*
Affective Style questionnaire (ASQ) total score (1)
ASQ Concealing subscale (1)
ASQ Adjusting subscale (1)
ASQ Tolerating subscale (2)
Assessment for Signal Clients (ASC) total score (10)*[2]
ASC Alliance subscale (10)*
ASC Social Support subscale (10)*
ASC Motivation subscale (11)*
ASC Life Events subscale (11)*
GAF last week before the start of therapy (10)*
Help Alliance Questionnaire - Patient Version (HAQ–P) total score (21)*
HAQ-P subscale Alliance (20)*
HAQ-P subscale Satisfaction with Therapy (23)*
Help Alliance Questionnaire - Therapist Version (HAQ–T) total score (24)*
HAQ-T subscale Alliance (24)*
HAQ-T subscale Satisfaction with Therapy (24*
Inventory of Therapeutic Interventions and Skills (ITIS)[1]
ITIS-S1 Pacing and efficient use of time (0)
ITIS-S2 Clarity of communication (0)
ITIS-S3 Use of feedback/ summaries (0)*
ITIS-S4 Rationale (0)
ITIS-S5 Guided discovery (0)
ITIS-S6 Therapeutic relationship/ collaboration (0)
ITIS-S7 Handling problems/ questions/ objections (0)
ITIS-S8 Empathic understanding (0)
ITIS-S9 Focusing on key cognitions and behaviors (0)
ITIS-S10 Strategy for change (0)
ITIS-S11 Application of techniques (0)
ITIS Overall adherence (0)
ITIS Overall competence (0)
ITIS Treatment difficulty (0)*
ITIS Patient motivation (0)

Note:
Missings for each variable are provided in brackets (total number of patients *N* = 259).
*Items correlate significantly with dropout (p < .05).

[1] There is no total score for this instrument.

[2] The ASC subscales were poled so higher values corresponded to higher functioning. The ASC total score is the mean of all 40 ASC items.

Study III

Bennemann, B., Schwartz, B., & Lutz, W. (2023). Fostering the upward spiral after a sudden gain in routine care cognitive behavioral therapy. [Manuscript submitted for publication].

Author Contributions

Conceptualization: B. Bennemann and W. Lutz; Methodology: B. Bennemann and W. Lutz; Formal Analysis: B. Bennemann, B. Schwartz, and W. Lutz; Investigation: W. Lutz; Writing – Original Draft: B. Bennemann; Writing – Review & Editing: B. Schwartz and W. Lutz; Funding Acquisition: W. Lutz.

Fostering the Upward Spiral After a Sudden Gain in Routine Care Cognitive Behavioral Therapy.

Björn Bennemann [a], Brian Schwartz [a], Wolfgang Lutz [a],

[a] University of Trier, Universitätsring 15, 54296 Trier, Germany

Please address correspondence concerning this article to:

Björn Bennemann
Clinical Psychology and Psychotherapy, Department of Psychology
University of Trier
D-54296 Trier, Germany
Phone: +49-651-201-3108
Fax: +49-651-201-2886
E-mail: bennemann@uni-trier.de

E-mail Brian Schwartz: schwartzb@uni-trier.de
E-mail Wolfgang Lutz: lutzw@uni-trier.de

Abstract

Objective: Sudden gains (SGs) are sudden symptom improvements. This study identifies variables that facilitate treatment outcome after a SG and quantifies them.
Method: The sample consisted of 1588 patients. Ratings of general change factors after a SG were investigated as predictors of outcome. Propensity score matching was used to compare SG patients with similar non-gain patients.
Results: A consistent use of problem-solving strategies and focusing on relevant goals had a larger impact on outcome after a SG, whereas it was more important for non-gainers to establish a fitting approach. Alliance ruptures had a larger negative effect on outcome in non-gainers than sudden gainers. Further, patients' coping improved outcome for non-gainers only.
Conclusion: Results indicate that a stable problem-solving approach is important after a SG. While for non-gainers, a fitting approach in therapy is more important than focusing on relevant goals, this effect was reversed for sudden gainers.

Keywords: sudden gain; upward spiral; outcome prediction; general change mechanisms; routine care

Abbrevations: Bern postsession reports (BPSR); Bern postsession report - patient (PSTB); Bern postsession report - therapist (TSTB); Brief symptom inventory short form (BSI); Coefficient of variation (CV); Elastic net (EN); Global assessment of functioning scale (GAF); Global severity index (GSI); Hopkins symptom checklist short form (HSCL-11); Propensity score matching (PSM); Pseudo gain (PG); Reliable change index (RCI); root-mean-square error (RMSE); Session rating scale (SRS); Sudden gain (SG); Symptom checklist-90-r (SCL-90-R); the least absolute shrinkage and selection operator (LASSO).

Fostering the Upward Spiral After a Sudden Gain in Routine Care Cognitive Behavioral Therapy.

Change and its measurement has been of great interest since the beginning of psychotherapy research (Barkham et al., 2021). For a long time, it was assumed that the courses of change were linear or log-linear in the sense of a *dose-response* pattern, with more sessions leading to more improvement (Howard et al., 1986; Kadera et al., 1996). More recent findings have shown that the number of sessions is not always the decisive factor and that patients show different levels of responsiveness (Barkham et al., 2006). According to the *good-enough level model* (GEL, Barkham et al., 2006; Stiles et al., 2008), patients stay in treatment until they have reached a level of change that they perceive to be *good enough* showing a linear course of change individually. In contrast, trajectories of individual patients often appear non-linear (Lutz et al., 2013), making it difficult to predict change on an individual level. A common phenomenon that reflects this varying responsiveness is the sudden gain (SG). A sudden gain is a disproportionate improvement in symptomatology between two sessions that is (a) relevant in absolute terms, (b) relevant in relative terms, and (c) relevant in terms of the stability of change. Tang and DeRubeis (1999) were the first to provide a quantitative description. A recent meta-analysis by Shalom and Aderka (2020) reported a moderate effect of the presence of SGs during treatment on primary outcome measures at post-treatment (Hedges's g = 0.68) and follow-up (Hedges's g = 0.51), suggesting that patients who experience SGs achieve superior treatment outcomes compared to patients without SGs. This effect persists even when patients with a SG are compared to patients who improved gradually (Aderka, Appelbaum-Namdar et al., 2011; Greenfield et al., 2011; Hedman et al., 2014; Lemmens et al., 2016).

Although SGs were first observed during the treatment of depressed patients with cognitive behavioral therapy, this phenomenon has also been found in interpersonal psychotherapy (Kelly et al., 2007; Lemmens et al., 2016), family therapy (Gaynor et al., 2003), group therapy (Kelly et al., 2005), and pharmacotherapy (Vittengl et al., 2005). Moreover, SGs were identified in treatments for posttraumatic stress disorder (Aderka, Foa et al., 2011; Wiedemann et al., 2020), panic disorder (Clerkin et al., 2008), social anxiety disorder (Hofmann et al., 2006), and obsessive-compulsive disorder (Aderka, Anholt et al., 2012). Even among patients receiving pill placebo (Vittengl et al., 2005) and patients receiving no treatment at all (Krüger et al., 2014), SGs have been found. Finally, SGs are found in both randomized controlled trials and naturalistic samples, suggesting that this phenomenon also occurs under routine conditions (Wucherpfennig, Rubel, Hollon et al., 2017). Interestingly, there also seems to be a therapist effect on SGs, indicating that some therapists are better at facilitating and initiating SGs (Deisenhofer et al., 2021). Tang and DeRubeis (1999) developed the theory that cognitive changes are elevated in the pre-gain sessions, eventually leading to a SG. Furthermore, Tang and DeRubeis (1999) hypothesized that a SG sparks a positive *upward spiral* of changes in cognitions and therapeutic alliance, which leads to further improvement in symptoms and thus to superior treatment outcomes. This theory led to two branches of research, one investigating possible predictors of the emergence of a SG (e.g., cognitive change) and the other investigating moderators of the upward spiral (i.e., outcome predictors after a SG).

Findings are inconsistent regarding predictors of a SG in the pre-gain session. Although some studies have found that elevated cognitive changes in the pre-gain session are responsible for the SG (e.g., Cavallini & Spangler, 2013; Norton et al., 2010; Tang et al., 2005), other studies could not replicate this finding (e.g., Aderka et al., 2021; Andrusyna et al., 2006; Bohn et al., 2013; Hunnicutt-Ferguson et al., 2012). Aderka and Shalom (2021) describe how difficult it is to find universally valid predictors. This is further confirmed by a recent study by Zilcha-Mano et al. (2019) that could not identify any predictors despite the use of sophisticated methodology. These heterogeneous findings show that the emergence of SGs seems to be more complex than initially assumed and requires more research.

The second branch of research primarily examines the upward spiral after a SG. Although understanding this process is of high clinical relevance, it has been comparatively rarely investigated (e.g., Lutz et al., 2013; Zilcha-Mano et al., 2019). Since not all patients benefit equally from SGs (Aderka, Nickerson et al., 2012; Hardy et al., 2005; Stiles et al., 2003; Tang et al., 2002), from a clinical point of view, it is important to understand which factors favor the upward spiral. There is evidence that some patients experience long persistent improvements after a SG, whereas others only temporarily improve with no effect on final treatment outcome (see Shalom & Aderka, 2020). However, studies that have investigated the processes after a SG show findings mostly consistent with Tang and DeRubeis' (1999) theory. For example, Lutz et al. (2013) examined changes in the therapeutic alliance before and after a SG. Consistent with Tang and DeRubeis' (1999) theory, a significant improvement in the therapeutic alliance was found after a SG, even though this was not the case in the pre-gain session. Wucherpfennig, Rubel, Hofmann et al. (2017) also found that there is an improvement in the therapeutic alliance after a SG that leads to a better outcome. Similarly, Zilcha-Mano et al. (2019) were able to show that improvements in the therapeutic alliance after a SG mediate the SG–outcome relationship. Finally, Bohn et al. (2013) found significant changes in cognitions only after a SG, but not before. These findings provide support for upward spirals in alliance and cognitions following a SG.

Even if these studies have already identified important factors for the upward spiral, the question remains whether other factors also influence the upward spiral. For example, Grawe (1997) was able to show that, in addition to the therapeutic alliance, four other common factors facilitate improvement in therapy. These include resource activation, clarification of meaning, problem actuation, and coping skills. It remains to be clarified whether these change mechanisms can also predict outcome after a SG, because, to our knowledge, these factors have not previously been linked to it.

In addition, new statistical approaches recently introduced to psychotherapy research might facilitate the further investigation of change mechanisms associated with SGs. Elastic net (EN) regressions, for example, avoid overfitting (Bennemann et al., 2022) and are thus well-suited to ensure the generalizability of results. EN removes predictors from the model that have no influence on the criterion (for a further explanation, see Brownlee (2019a)) and might therefore be well-suited to identify variables relevant to outcome after a SG.

Furthermore, the effect sizes of outcome predictors after a SG remain unclear. Since the therapeutic alliance has a general influence on therapy outcome (Flückiger et al., 2018), it is difficult to quantify *how much* the alliance influences outcome after a SG, independent of the general influence that the therapeutic alliance has on outcome. Wucherpfennig, Rubel, Hofmann et al. (2017) were the only ones to use a control group to quantify effects, however, they limited their investigation to alliance and coping skills. It therefore remains unclear, whether other change mechanisms also have an impact on outcome after a SG. Therefore, the purpose of this study is to investigate which predictors have an impact on outcome after a SG besides alliance and coping skills, independent of general outcome predictors. Here, a comparable control group that has not experienced Sudden Gains will be used. The effects of general outcome predictors and those after a SG can thereby be disentangled and quantified, revealing the impact of important outcome predictors after a SG. This study therefore has two objectives:

1.) The replication of findings from previous studies on outcome predictors after a sudden gain using a large sample and an elastic net analysis to ensure better generalizability.

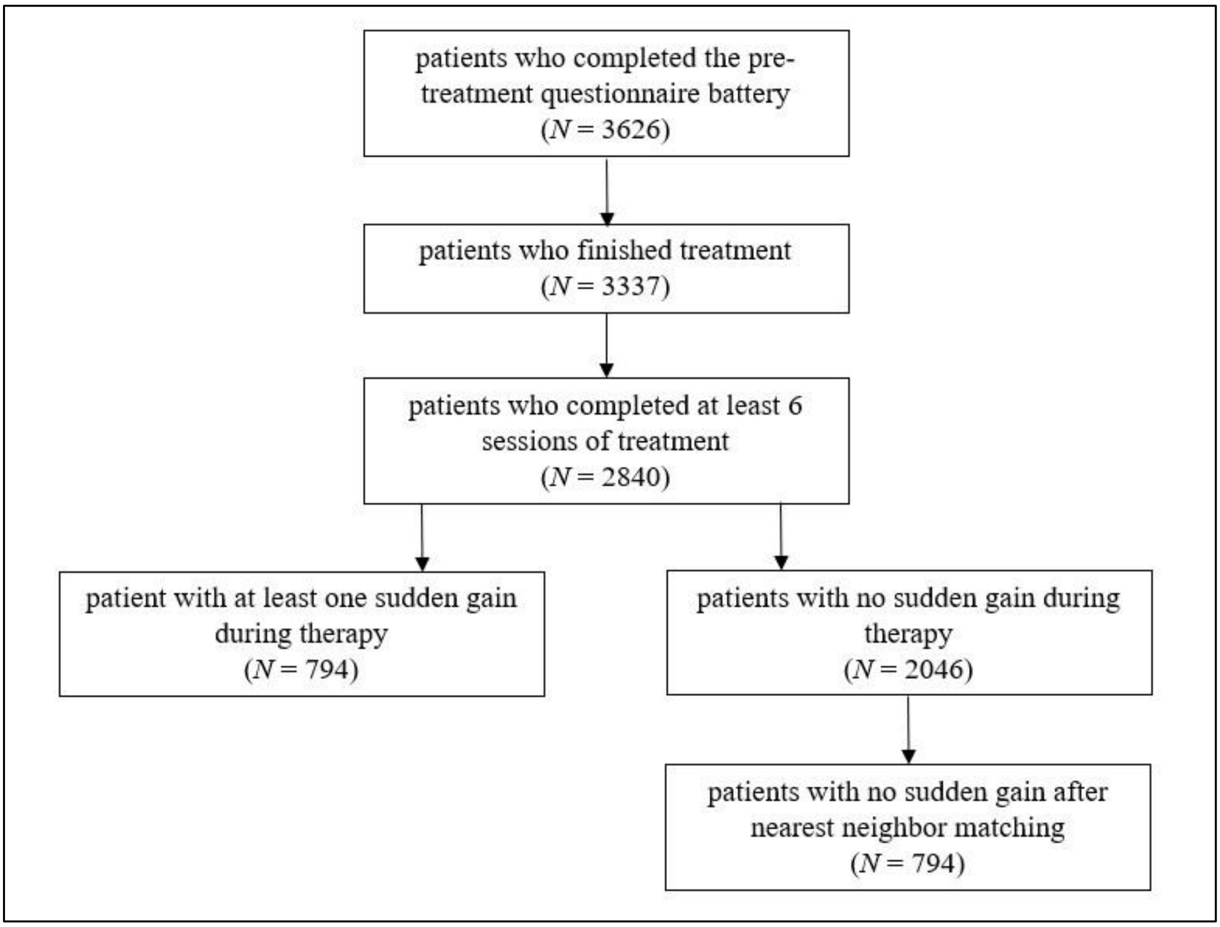
2.) The identification of further relevant outcome predictors after a SG, as well as their quantification in the form of effect sizes with the help of a control group.

Method

Patients and Treatment

The routine care sample included a total of 3626 patients treated at the University of Trier's outpatient clinic between 2007 and 2022. Patients were included in the analyses if they completed a questionnaire battery before therapy, completed therapy (i.e., regular termination or dropout), and their therapy lasted at least 6 sessions. Of the remaining patients, those with at least one sudden gain were included in the analyses. A corresponding patient who had not experienced a SG in therapy was identified and included in the analyses for each SG patient using propensity score matching via the nearest neighbor method. This resulted in a total sample of 1588 patients, half of whom experienced a SG during therapy and half of whom did not (see Figure 1 for the detailed procedure).

Figure 1
Patient selection flowchart


patients who completed the pre-treatment questionnaire battery
(*N* = 3626)
patients who finished treatment
(*N* = 3337)
patients who completed at least 6 sessions of treatment
(*N* = 2840)
patient with at least one sudden gain during therapy
(*N* = 794)
patients with no sudden gain during therapy
(*N* = 2046)
patients with no sudden gain after nearest neighbor matching
(*N* = 794)

Diagnoses were based on the German version of the Structured Clinical Interview for Axis I DSM–IV Disorders - Patient Edition (SCID-I; Wittchen et al., 1997) and DSM-5 disorders (SCID-5-CV; Beesdo-Baum et al., 2019) as well as the International Diagnostic Checklist for Personality Disorders (IDCL-P; Bronisch et al., 1996). Each session was videotaped. All patients who participated in the study gave written consent for their data to be used for research purposes. The interviews were conducted by well-trained therapists before treatment began. The interview and diagnoses were discussed in expert teams consisting of at least four senior clinicians. Final diagnoses were determined by consensual agreement of at least 75% of the team members. Table 1 provides an overview of patient characteristics.

The mean score of the Brief Symptom Inventory short form (BSI; Franke, 2000; German translation of Derogatis, 1975) was 1.48 (*SD* = 0.70) and did not differ significantly between groups, indicating a moderate to severe general level of distress.

The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2013.

Therapists

Treatment was provided by 218 therapists who took part in a 3- (full-time) or 5-year (part-time) postgraduate training program with a CBT focus. All therapists initially received one year of clinical training before beginning to treat outpatients. On average, each therapist treated 7.3 patients (*SD* = 3.2, range 1−14). Each therapist received one hour of supervision per month and patient, supported by session videos. Supervisors were clinicians with at least five years of clinical experience after completing training.

Measures

Hopkins Symptom Checklist short form. The Hopkins Symptom Checklist short form (HSCL-11; Lutz et al., 2006) is a short version of the Brief Symptom Inventory (Franke, 2000). This short form consists of 11 items and measures symptomatic distress with a focus on depressive and anxiety symptomatology. Items are based on a 4-point Likert scale ranging from 1 (*not at all*) to 4 (*extremely*). The HSCL-11 highly correlates with the BSI (r = .91) and has high internal consistency (α =.92; Lutz et al., 2006). The mean of the 11 items was used to identify sudden gains. It was assessed at the beginning of each session.

Table 1
Patient characteristics and matched group comparisons

Variable	Sudden gainers (*n* = 794)	Non-gainers (*n* =794)	*t*-value / *chi*²-value	*p*-value
Treatment length (*M* session (*SD*))	49.40 (20.25)	49.62 (21.02)	0.22	0.828
Session number of the sudden (pseudo) gain (*M* session (*SD*))	19.59 (16.03)	19.92 (16.28)	0.42	0.677
Brief Symptom Inventory (BSI) – pre total score (*M* (*SD*))	1.49 (0.68)	1.47 (0.73)	0.50	0.615
BSI – post total score (*M* (*SD*))	0.72 (0.61)	0.84 (0.70)	3.59**	0.000
Age (*M* years (*SD*))	35.57 (12.62)	35.99 (13.03)	0.66	0.508
Therapy expectation (*M* expectation (*SD*))	3.10 (0.71)	3.07 (0.69)	0.89	0.373
Sick leave (*n* left (%))	149 (18.77%)	152 (19.14%)	0.02	0.898
Marital status (*n* married (%))	216 (27.20%)	223 (28.09%)	0.11	0.736
Education (*n* > 12 years (%))	404 (50.88%)	404 (50.88%)	0	1
Gender female (*n* (%))	552 (69.52%)	546 (68.77%)	0.07	0.786
Medication intake (*n* (%))	567 (71.41%)	582 (73.30%)	0.62	0.432
Primary diagnosis (*n* (%))				
Affective disorder	411 (51.76 %)	387 (48.74 %)	5.21*	0.022
Anxiety disorder	85 (10.71 %)	126 (15.87 %)	6.57*	0.010
Adjustment disorder / PTSD	110 (13.85 %)	114 (14.36 %)	0.00	0.950
Other	188 (23.68 %)	167 (21.03 %)	1.45	0.228
Comorbidity (*n* (%))				
Two diagnoses	238 (29.97 %)	250 (31.49 %)	0.36	0.550
Three or more diagnoses	325 (40.93 %)	327 (41.18 %)	0.00	0.960

Note: For continuous variables (first 6 variables) a *t*-test was used, for dichotomous variables a chi²-test was used; * = p < .05; ** = p < .001. Due to the hardly existing variance, we have excluded the variable nationality (94.8% were German).

Brief Symptom Inventory The Brief Symptom Inventory (BSI; Franke, 2000; German translation of Derogatis, 1975) is a 53-item self-report symptom inventory assessing symptomatology during the last week. Nine subscales are included in the BSI: *somatization, obsessive–compulsive, interpersonal sensitivity, depression, anxiety, hostility, phobic anxiety, paranoid ideation,* and *psychoticism*. The BSI is a brief form of the Symptom Checklist-90-R (SCL-90-R; Derogatis, 1977). The items score on a 5-point Likert scale ranging from 0 (*not at all*) to 4 (*extremely*). In this study, the Global Severity Index (GSI) was calculated to capture symptomatic distress at the beginning and end of therapy.

Bern Post Session Reports. The Bern Post Session Reports (BPSR; Flückiger et al., 2010) capture therapeutic processes rated by patients (PSTB) and therapists (TSTB) on a session-to-session basis. The assessed processes are based on the model of general mechanisms of change (Grawe, 1997). We assessed the scales *therapeutic relationship, problem solving, problem actualization, motivational clarification,* and *resource activation* from both the patient and therapist perspectives after each session. All items are based on a 7-point Likert scale ranging from −3 (*not at all*) to +3 (*yes, exactly*). Each scale had good to excellent psychometric properties in our sample (α = .85 − .93 for the patient version; α = .87 − .94 for the therapist version). The scales *motivational clarification* and *problem actualization* consist of only one item each, therefore no internal consistency could be calculated for these scales. Nevertheless, confirmatory factor analysis (Flückiger et al., 2010) has shown them to be reliable and valid.

Session Rating Scale - Version 3. The Session Rating Scale - Version 3 (SRS; Duncan et al., 2003) is an ultra-brief alliance measure consisting of 4 items. Item 1 captures the *relationship* ranging from 0 (*I did not feel heard, understood, and respected*) to 100 (*I felt heard, understood, and respected*), item 2 captures the *goals and topics* ranging from 0 (*We did not work on or talk about what I wanted to work on and talk about*) to 100 (*We worked on and talked about what I wanted to work on and talk about*). Next, item 3 assesses *approach or method* ranging from 0 (*The therapist's approach is not a good fit for me*) to 100 (*The therapist's approach is a good fit for me*) and item 4 assesses *overall* ranging from 0 (*There was something missing in the session today*) to 100 (*Overall, today's session was right for me*). Each item is assessed by a visual analogue scale and has good psychometric properties that are comparable to the Health Assessment Questionnaire-II (α =.88, rtt =.64; Duncan et al., 2003). Each item was assessed after each session and used separately for the analyses.

Additional Measurements. Additionally, 4 more items were assessed on a session-to session basis. First the Global Assessment of Functioning Scale (GAF; American Psychiatric

Association, 2005) was rated by the therapist after each session. Next, the therapist rated patient coping with the item *How well is your patient coping emotionally and psychologically?* on a visual analog scale ranging from 0 (*very poorly*) to 100 (*very well*). Last, alliance ruptures were assessed via the item *During today's session, did you perceive any tension, misunderstandings, or inconsistencies in the relationship with your patient/therapist?* These were rated on a 5-point Likert scale ranging from 0 (*not at all*) to 5 (*constantly*) by the patient as well as the therapist.

Missing Data

Following recommendations for the imputation of nested data (van Buuren, 2018), we first imputed the level 1 data only. In the next step, we used the mean values of each level 1 variable for each patient to impute the level 2 variables. To do so, we used the randomForest method of the mice package (van Buuren, 2021) in R. The randomForest algorithm is recommended for imputation of mixed-type data with both categorical and continuous variables (Waljee et al., 2013). No variable had more than 30% missings, for which good algorithm performance has previously been shown in simulation analyses (Stekhoven & Bühlmann, 2012).

Identification of Sudden Gains

The criteria developed by Tang and DeRubeis (1999) were used to identify SGs. However, the first criterion had to be adapted, as we used the HSCL-11 instead of the BDI-II to identify SGs. Following Stiles et al.'s (2003) suggestions, we modified the first criterion (improvement of at least 7 BDI points) by using the reliable change index (RCI) to identify meaningful improvement between two sessions. The RCI is defined as the difference between the pre-treatment and post-treatment scores, divided by the standard error of the difference (Jacobson & Truax, 1991). Based on the data from the naturalistic sample, the RCI for the HSCL-11 was 0.61. Thus, a SG was identified when the following criteria were fulfilled between the pre-gain session (*N*) and the post-gain session (*N* + 1):

(a) The HSCL-11 decreased by at least 0.61 between two subsequent sessions (HSCL-11_{N} – HSCL-$11_{N+1} \geq 0.61$).

(b) This decrease was at least 25% of the HSCL-11 score in the pre-gain session (HSCL-11_{N} – HSCL-$11_{N+1} \geq 0.25$ * HSCL-11_{N}).

(c) The mean score of the three sessions before (sessions $N - 2$, $N - 1$, and N) and after (sessions $N + 1$, $N + 2$, $N + 3$) the gain were significantly different, based on a two sample *t*-test with a 5% significance level ($t_{(4;95\%)} > 2.78$).

Generating a Control Group

To generate a comparable control group, propensity score matching (PSM; see Stuart et al., 2004 for an overview) was used. This method has been shown to reduce bias by balancing two samples for comparability (Lutz et al., 2016; Rosenbaum & Rubin, 1983; Wucherpfennig, Rubel, Hollon et al., 2017). We applied the nearest neighbor method according to Ho et al.'s (2007) suggestions, which has shown good performance in a sophisticated simulation study (Geldof et al., 2020). Nearest neighbor matching utilizes propensity scores to adjust for confounding baseline variables. As West et al. (2014) recommended, we included baseline variables that could potentially confound the comparison between gainers and non-gainers. Adapted from the work of Delgadillo et al. (2016) and Wucherpfennig, Rubel, Hofmann et al. (2017), the following variables were considered for matching: intake symptom severity (pre-treatment BSI), age, therapy expectations, medication at intake, gender, education status, marital status, and employment status (see Table 1 for an overview). Based on 1:1 nearest neighbor matching, for each sudden gainer, the most similar patient was selected from the group of non-gainers using these baseline variables. After matching, we calculated several *t*-tests and χ^2- tests to examine the differences between these groups and ensure comparability (see Figure 1).

Pseudo Gains

After the matching process, a so-called *pseudo gain* (PG) session was selected for each patient in the non-gainer (i.e., control) group (a non-gain session comparable to a sudden gain session with regard to the time point at which it occurred during treatment). Further, this time point was set in relation to the total number of sessions. When, for example, a SG occurred in session 10 and the SG patient had 30 sessions in total, we first reviewed the matched counterpart's total number of sessions. For example, if they had a total of 60 sessions, the 20th session was then selected as a PG session. Using this procedure, we obtained a control group with corresponding pseudo gain sessions at which the outcome predictors after a SG could be relativized and quantified. In this way, effect sizes could be calculated that provide more precise information about the importance of the processes after a SG.

Data Analytic Strategy

All analyses were conducted using the free software environment R version 4.1.1 (R Core Team, 2021). To quantify the outcome predictors after a SG independently of the influence that these variables generally have on outcome, two comparable models were calculated, one with the sudden gainers and the other with the non-gainers. The first three sessions after a sudden (or pseudo) gain are relevant for these calculations, which are also crucial for identifying a sudden gain according to Tang and DeRubeis' (1999) criteria. We used the mean scores and the coefficient of variation (CoV) of each variable over all three sessions after a SG/PG as potential predictors. The CoV is defined as the standard deviation standardized by the mean. Small scores indicate high consistency, that is, low fluctuation over time. The criterion in these models was the post-treatment BSI value.

Model generation process. For model generation, we conducted a nested cross-validation with ten outer and five inner loops, repeating the inner cross-validations three times (Brownlee, 2019b) to further avoid overfitting (Cawley & Talbot, 2010). This procedure was performed separately for the sudden gainer and non-gainer groups using 19 predictors (i.e., total HSCL value, the five PSTB and TSTB scales, respectively, four SRS items, the GAF value, estimated coping, degree of alliance rupture rated by therapist and patient). Here, the mean and CoV of each predictor were used, resulting in a total of 38 predictors with the post-treatment BSI value as criterion.

We used elastic net (EN) regularization (Friedman et al., 2010) as the model generation algorithm. EN protects well against overfitting (Pavlou et al., 2016) and performs very well in clinical and naturalistic contexts (Bennemann et al., 2022; Lutz et al., 2019). EN combines two regularized regression procedures, the least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996) and ridge regression. Both penalizations shrink regression coefficients to reduce overfitting, but only LASSO can shrink variable coefficients to zero. For EN regularization, we defined a range of values to be tested to find an optimal ratio of LASSO and ridge regression for our models as well as the optimal value for the penalization parameter. We set alpha (i.e., the parameter for the ratio of LASSO to ridge regression) to 0.1 for the first analysis and then altered alpha in increments of 0.1 until 1 was reached. An alpha of 0 is equal to a ridge regression, while an alpha of 1 equals a LASSO regression. We also defined lambda’s range analogue to the alpha parameter. Lambda defines the magnitude of the regression penalty. This resulted in 100 different possible combinations of these two parameters (10 values for alpha x 10 values for lambda) to identify the best fitting model. To run each of the ten outer cross-validation loops, this procedure was performed with the help of the R package caret v6.0-90 (Kuhn, 2021). Identification of the best model was based on the root-mean-square error (RMSE; Hyndman & Koehler, 2006). The model with the lowest value was considered the best model.

Following Zou and Hastie's (2005) recommendations, before model generation all continuous variables were centered separately for each outer cross-validation loop of the training and test sets. For the centering process, only the mean value from the training data set was subtracted for each corresponding variable to avoid data leakage.

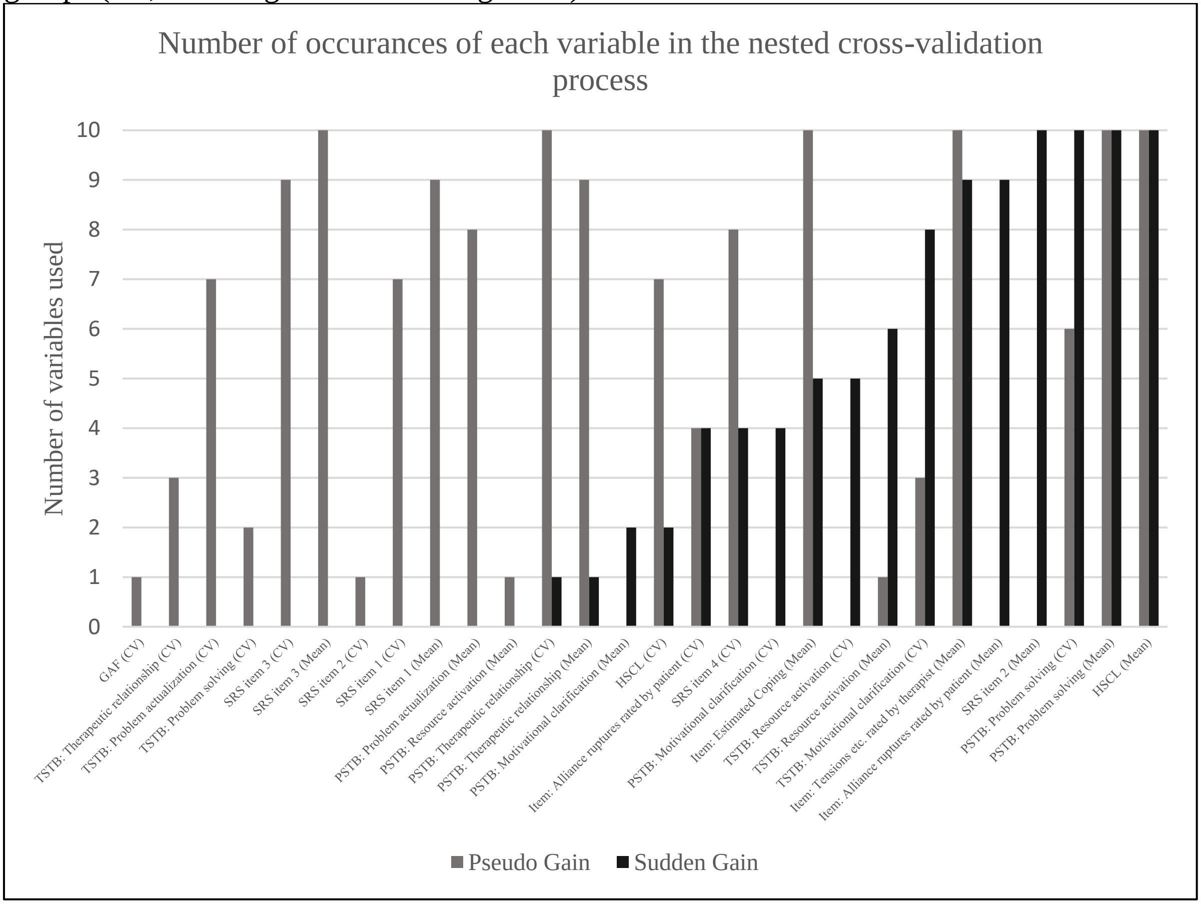
Predictor comparisons. For each outer cross-validation run, we obtained an RMSE value and a regression weight for all predictors unless they were set to 0 by the penalization procedure (i.e., maximum of 10 values for a predictor per group). Only predictors that were included in all 10 outer runs of the EN penalization, either in the SG group, the PG group, or both groups were used for further analysis. When the predictor is included in all 10 outer runs, the inclusion rate lies above 95% and is thus significantly above chance. Due to the maximum of 10 values per group, distributions were available for each predictor in each group. These were compared between the two groups using *t*-tests to determine how large the differences in each predictor's effects on outcome were after a SG vs. after a PG. Using this procedure, the outcome predictors' effects after a SG could be quantified while considering the general influence on outcome that was assessed via the control group (i.e., the non-gainer group). Further, we applied the same procedure to the model evaluation parameter used (i.e., RMSE) to compare the two groups regarding model performance.

Results

After the matching process, two groups of equal size were generated, which did not differ significantly with respect to the input variables. At the end of therapy, the sudden gainer group (*M* = 0.72, *SD* = 0.61) was significantly less distressed than the control group (*M* = 0.84, *SD* = 0.70) according to the BSI (*t*(1586) = 3.59, *p* < .001, Cohen's *d* = 0.18), indicating that sudden gainers had a significantly better therapy outcome, despite being as distressed as the control group at the beginning of therapy. In addition, the two groups differed with respect to the distribution of diagnoses. While there were significantly more affective disorders in the sudden gain group, there were significantly more anxiety disorders in the control group (for an overview of the group comparisons, see Table 1).

After the model building process with EN, some variables were excluded that did not have predictive power for treatment outcome. Figure 2 shows an overview of all variables that were included at least once in at least one group (i.e., SG group and/or PG group). All variables that were not included in any model in either group were excluded completely and are not shown in the figure. As described above, for the final comparisons, only variables that were included in at least one of the two groups in all 10 outer runs of the nested cross-validation were included.

Figure 2: Number of occurrences of each variable in each outer cross-validation run for both groups (i.e., sudden gainers and non-gainers).



Note: HSCL = Hopkins Symptom Checklist short form; PSTB = Bern Post-Session Reports for patients; TSTB = Bern Post-Session Reports for therapists; SRS = Short Rating Scale; GAF = Global Assessment Functioning; CoV = Coefficient of variation; All variables not included in the figure were excluded in each run in both groups

Table 2 provides an overview of the included variables and their comparisons via a two-tailed independent *t*-test. Concerning mean values, the HSCL, the PSTB *problem solving* scale, the SRS scales *goals and topics* and *approach or method*, the single coping item and the single alliance rupture item rated by the therapist were considered. In addition, the CoV of the PSTB scales *problem solving* and *therapeutic relationship* were included in the comparisons.

Table 2
Mean regression weights for all relevant variables across the three sessions after a sudden (pseudo) gain and the mean differences between these two groups.

	Sudden gainers (n = 794)	Non-gainers (n = 794)			
Variable	*M* (*SD*)	*M* (*SD*)	*t*-value	*p*-value	Cohen's *d*
Intercept	.000 (.000)†	.000 (.000)†	1.54	0.140	0.69
HSCL (mean)	.213 (.007)	.348 (.013)	27.91***	0.000	12.48
PSTB: Problem solving (mean)	-.013 (.010)	-.011 (.008)	0.55	0.589	0.25
PSTB: Problem solving (CoV)	.029 (.011)	.005 (.009)†	5.24***	0.000	2.34
PSTB: Therapeutic relationship (CoV)	.000 (.001)†	.025 (.001)	10.90***	0.000	4.88
SRS: Item 2 (mean)	-.017 (.010)	.000 (.000)†	5.37***	0.000	2.40
SRS: Item 3 (mean)	.000 (.000)†	-.018 (.005)	11.11***	0.000	4.97
Item: Alliance ruptures (mean)	.009 (.006)	.017 (.009)	2.42*	0.026	1.08
Item: Estimated coping (mean)	.003 (.005)†	-.013 (.008)	5.37***	0.000	2.40
Model parameter (RMSE)	.546 (.050)	.539 (.056)	0.33	0.748	0.15

Note: HSCL = Hopkins Symptom Checklist short form; PSTB = Bern Post Session Reports for patients; SRS = Short Rating Scale; RMSE = root mean square error; CoV = Coefficient of variation; * p > .05; ** p > .01; *** p > .001; † mean value is not significantly different from 0. To determine significance, the Benjamini-Hochberg procedure was applied to prevent false discovery rates through multiple testing.

Except for the mean score of the *problem solving* scale (t(18) = 0.55; p = .140; Cohen's d = 0.25), all differences between these items in the two groups were significant. A high mean HSCL value predicted a worse outcome, whereas this effect was significantly greater in the PG group (t(18) = 27.91; p < .001; Cohen's d = 12.48). A high variation in the application of problem-solving strategies seemed to lead to a significantly worse outcome for sudden gainers, whereas this was not found for non-gainers. The difference was also significant (*t*(18) = 5.24; *p* < .001; Cohen's *d* = 2.34). For the CoV of the therapeutic relationship, it was the other way around; here a large amount of variation seemed to lead to worse outcomes in the non-gainer group, whereas the sudden gain group was not affected by it. Here, the difference between the groups was also significant (*t*(18) = 10.90; *p* < .001; Cohen's *d* = 4.88). In addition, the *goals and topics* (SRS item 2) in therapy led to a better outcome only for sudden gainers, whereas the therapists' *approaches or methods* (SRS item 3) did not have such an effect. This result was reversed for non-gainers. The group difference was significant regarding *goals and topics* (*t*(18) = 5.37; *p* < .001; Cohen's *d* = 2.40), as well as *approaches and methods* (*t*(18) = 11.11; *p* < .001; Cohen's *d* = 4.97). The therapist assessment of alliance ruptures predicted a poorer outcome, but this effect was significantly smaller for sudden gainers (*t*(18) = 2.42; *p* = .026; Cohen's *d* = 1.08). Finally, the patient's coping skills assessed by the therapist significantly predicted better therapy outcome, but only for non-gainers, leading to a significant difference between groups (*t*(18) = 5.37; *p* < .001; Cohen's *d* = 2.40). Model quality as measured by the RMSE did not differ significantly between the two groups (*t*(18) = 0.33; *p* > .05; Cohen's *d* = 0.15).

Discussion

The aim of this study was to identify the most important predictors of treatment outcome after a SG that foster an upward spiral and to compare them to outcome predictors in a control group without SGs that did not differ significantly regarding relevant pre-treatment variables. For this purpose, an elastic net analysis with nested cross-validation was used that allows generalizability of the results and protects against overfitting. Furthermore, this procedure allows conclusions to be drawn about processes that are specifically decisive after a SG. It avoids overlap with generally important therapy factors by relativizing the effects via comparison to a control group (i.e., the non-gainer group). One model was generated per group (i.e., SG and PG), which did not differ significantly regarding their overall prediction performance. Through this process, only variables that had a significant effect on the outcome after a SG or PG were included in the analysis and further investigated. It was found that the predictive power of some variables differed significantly, depending on whether they predicted outcome after a SG or PG, with the differences showing large effect sizes (see Table 2).

First, the sudden gainers had a lower HSCL value than the non-gainers, which replicates previous research results (see Shalom & Aderka, 2020). Since the non-gainers did not experience a substantial improvement in therapy in form of a sudden gain, this result is consistent with the literature. In addition, a high average use of problem-solving strategies (patient rating) shows an equally positive effect on outcome for both groups. Interestingly, a high variation in the use of problem-solving strategies (patient rating) lead to a significantly worse outcome in the sudden gainer group, but not so in the non-gainer group. It seems that the constant use of problem-solving strategies is important for the maintenance of positive effects, even more so than other general change factors. One reason for this could be that the therapists in this study primarily had a CBT focus, which typically includes problem-solving strategies. A replication in a sample with another therapy focus could lead to different results. Nevertheless, SGs have higher effect sizes in CBT settings (Aderka, Nickerson et al., 2012), which further strengthens our findings.

Regarding the therapeutic relationship (patient rating), this effect seems to be reversed. This can be well explained by the fact that sudden gainers have also experienced significant symptom improvement on an interpersonal level, so that they have a more functional way of dealing with disagreements and interactional problems in therapy. This assumption is supported by the fact that alliance ruptures perceived by the therapist have a significantly larger negative effect on outcome for the group of non-gainers than for the sudden gainer group. Further, therapist-rated estimated patient coping lead to a better outcome for non-gainers only. This finding is strengthened by the fact that for sudden gainers, goals and topics are crucial (SRS item 2), while for non-gainers, the approach and method (SRS item 3) are vital. Therefore, the *what* seems to be more important for sudden gainers than the *how*, which is reversed in the non-gainer group. This also supports the finding that consistent use of problem-solving strategies is important for sudden gainers, as they have enough resources to apply concrete strategies and are better able to handle a directive approach that has the patient’s goals in mind.

Except for therapist-rated coping (“*How well is your patient coping emotionally and psychologically?*”) and alliance ruptures (“*During today’s session, did you perceive any tension, misunderstandings, or inconsistencies in the relationship with your patient?*”), no therapist variable was included in the model. Therefore, patient ratings seem to be better predictors of therapy outcome. Nevertheless, there is evidence for the predictive power of therapist ratings (Laws et al., 2017). One reason for the superiority of patient ratings could be that outcome was only assessed from the patient’s perspective. Although this is a common method of measuring outcome, the effects may change if other instruments were used.

Limitations

Although this study has many strengths, several limitations must be mentioned. For example, the two samples of sudden and non-gainers are not identical with respect to the primary diagnoses (see Table 1). More depressed patients were in the sudden gain group, while more anxiety patients were in the control group. Although we do not believe this had a causal impact on our results, as there is efficacy evidence for both diagnoses (Cuijpers et al., 2013), we cannot completely rule it out. However, Aderka and Shalom (2021) found evidence that the frequency of sudden gains is higher in depressed patients than in patients who suffer from an anxiety disorder. This finding fits to our data and therefore seems not to interfere with our findings.

In addition, we cannot rule out that crucial patient characteristics were overlooked in our study. Although we collected important variables associated with treatment outcome (Delgadillo et al., 2016), there is a possibility that others are missing. Furthermore, it should be noted that only three sessions were selected for the evaluation of relevant processes after a sudden or pseudo gain. Although this number of sessions was based on Tang and DeRubeis' (1999) original criteria and has already been used in other studies (Wucherpfennig, Rubel, Hofmann et al., 2017), it cannot be excluded that some longer-term processes were overlooked due to the limited number of sessions. Here, a replication with more sessions after a SG could provide a clearer picture and strengthen our results.

Another limitation could be that we did not consider the timing of the SG. Although our groups were comparable in terms of session number and length of therapy, it cannot be excluded that varying processes are decisive for an earlier SG versus later SG. Future studies should include the point in time of a sudden gain in the analyses to uncover possible influences.

Finally, it must be mentioned that our sample consisted almost exclusively of German patients (94.8%), which is why this variable was not included in the analysis. This limits the generalizability to other cultural groups and ethnicities, even if the data come from a naturalistic setting.

Conclusion

The present study shows that to maintain the upward spiral after a SG, processes other than those that have a general influence on outcome are prominent. This information is of great value for therapists, as it may help to maintain improvement after a SG and subsequently achieve better improvement rates in the context of data-informed psychological therapy (Lutz, Deisenhofer et al., 2022; Lutz, Schwartz et al., 2022). In addition, a sudden improvement in symptom severity may make it necessary to adjust the treatment strategy. This study provides important insights into the direction in which the therapy strategy must be adapted and reveals that a more directive problem-solving approach is more promising to maintain an upward spiral. The results again show that SGs are a highly complex therapeutic phenomenon that require a high level of flexibility and intuition on the therapist’s part. However, further research is needed to better understand this complex phenomenon and to derive more benefits for patients.

CRediT authorship contribution statement:

Björn Bennemann: Conceptualization, Methodology, Software, Formal analysis, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization, Project administration. **Brian Schwartz:** Methodology, Writing – Review & Editing. **Wolfgang Lutz:** Investigation, Supervision, Funding acquisition.

Additional Information

Conflicts of Interest: The authors declare that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding: This work was partly supported by the German Research Foundation (DFG) under Grant Nr. LU 660/10-1.

Acknowledgments: The authors thank Kaitlyn Poster for proofreading the manuscript.

References

Aderka, I. M., Anholt, G. E., van Balkom, A. J. L. M., Smit, J. H., Hermesh, H., & van Oppen, P. (2012). Sudden gains in the treatment of obsessive-compulsive disorder. *Psychotherapy and Psychosomatics*, *81*(1), 44–51. https://doi.org/10.1159/000329995

Aderka, I. M., Appelbaum-Namdar, E., Shafran, N., & Gilboa-Schechtman, E. (2011). Sudden gains in prolonged exposure for children and adolescents with posttraumatic stress disorder. *Journal of Consulting and Clinical Psychology*, *79*(4), 441–446. https://doi.org/10.1037/a0024112

Aderka, I. M., Foa, E. B., Applebaum, E., Shafran, N., & Gilboa-Schechtman, E. (2011). Direction of influence between posttraumatic and depressive symptoms during prolonged exposure therapy among children and adolescents. *Journal of Consulting and Clinical Psychology*, *79*(3), 421–425. https://doi.org/10.1037/a0023318

Aderka, I. M., Kauffmann, A., Shalom, J. G., Beard, C., & Björgvinsson, T. (2021). Using machine-learning to predict sudden gains in treatment for major depressive disorder. *Behaviour Research and Therapy*, *144*, 103929. https://doi.org/10.1016/j.brat.2021.103929

Aderka, I. M., Nickerson, A., Bøe, H. J., & Hofmann, S. G. (2012). Sudden gains during psychological treatments of anxiety and depression: A meta-analysis. *Journal of Consulting and Clinical Psychology*, *80*(1), 93–101. https://doi.org/10.1037/a0026455

Aderka, I. M., & Shalom, J. G. (2021). A Revised Theory of Sudden Gains in Psychological Treatments. *Behaviour Research and Therapy*, *139*, 103830. https://doi.org/10.1016/j.brat.2021.103830

American Psychiatric Association. (2005). *Diagnostic and Statistical Manual of Mental Disorders*. Arlinton, VA: American Psychiatric Association.

Andrusyna, T. P., Luborsky, L., Pham, T., & Tang, T. Z. (2006). The Mechanisms of Sudden Gains in Supportive–Expressive Therapy for Depression. *Psychotherapy Research*, *16*(5), 526–536. https://doi.org/10.1080/10503300600591379

Barkham, M., Lutz, W., & Castonguay, L. G. (Eds.). (2021). *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change* (7th ed.). Wiley.

Barkham, M., Connell, J., Stiles, W. B., Miles, J. N. V., Margison, F., Evans, C., & Mellor-Clark, J. (2006). Dose-effect relations and responsive regulation of treatment duration: The good enough level. *Journal of Consulting and Clinical Psychology*, *74*(1), 160–167. https://doi.org/10.1037/0022-006X.74.1.160

Beesdo-Baum, K., Zaudig, M., & Wittchen, H.-U. (2019). SCID-5-CV. Strukturiertes Klinisches Interview für DSM-5-Störungen - Klinische Version.

Bennemann, B., Schwartz, B., Giesemann, J., & Lutz, W. (2022). Predicting patients who will drop out of out-patient psychotherapy using machine learning algorithms. *The British Journal of Psychiatry : The Journal of Mental Science*, 1–10. https://doi.org/10.1192/bjp.2022.17

Bohn, C., Aderka, I. M., Schreiber, F., Stangier, U., & Hofmann, S. G. (2013). Sudden gains in cognitive therapy and interpersonal therapy for social anxiety disorder. *Journal of Consulting and Clinical Psychology*, *81*(1), 177–182. https://doi.org/10.1037/a0031198

Bronisch, T., Hiller, W., Mombour, W., & Zaudig, M. (1996). *International diagnostic checklists for personality disorders according to ICD-10 and DSM-IV—IDCL-P*. Seattle, WA: Hogrefe and Huber Publishers.

Brownlee, J. (2019a). *Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch*.

Brownlee, J. (2019b, November 3). *Machine Learning Mastery: Nested Cross-Validation for Machine Learning with Python*. https://machinelearningmastery.com/nested-cross-validation-for-machine-learning-with-python/

Cavallini, A. Q., & Spangler, D. L. (2013). Sudden Gains in Cognitive-Behavioral Therapy for Eating Disorders. *International Journal of Cognitive Therapy*, *6*(3), 292–310. https://doi.org/10.1521/ijct.2013.6.3.292

Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, *11*, 2079–2107.

Clerkin, E. M., Teachman, B. A., & Smith-Janik, S. B. (2008). Sudden gains in group cognitive-behavioral therapy for panic disorder. *Behaviour Research and Therapy*, *46*(11), 1244–1250. https://doi.org/10.1016/j.brat.2008.08.002

Cuijpers, P., Sijbrandij, M., Koole, S. L., Andersson, G., Beekman, A. T., & Reynolds, C. F. (2013). The efficacy of psychotherapy and pharmacotherapy in treating depressive and anxiety disorders: A meta-analysis of direct comparisons. *World Psychiatry: Official Journal of the World Psychiatric Association (WPA)*, *12*(2), 137–148. https://doi.org/10.1002/wps.20038

Delgadillo, J., Moreea, O., & Lutz, W. (2016). Different people respond differently to therapy: A demonstration using patient profiling and risk stratification. *Behaviour Research and Therapy*, *79*, 15–22. https://doi.org/10.1016/j.brat.2016.02.003

Derogatis, L. R. (1975). *SCL-90-R: Symptom Checklist-90-R: Administration, Scoring, and Procedures Manual*. NCS Pearson. https://books.google.de/books?id=fcxxtwAACAAJ

Derogatis, L. R. (1977). SCL-90-R: Administration, scoring and procedures: Manual 1. *Baltimore, MD: Clinical Psychometric Research*.

Duncan, B. L., Miller, S. D., Sparks, J. A., Claud, D. A., Reynolds, L. R., Brown, J., & Johnson, L. D. (2003). The Session Rating Scale: Preliminary psychometric properties of a "working" alliance measure. *Journal of Brief Therapy*, *3*(1), 3–12.

Flückiger, C., Del Re, A. C., Wampold, B. E., & Horvath, A. O. (2018). The alliance in adult psychotherapy: A meta-analytic synthesis. *Psychotherapy (Chicago, Ill.)*, *55*(4), 316–340. https://doi.org/10.1037/pst0000172

Flückiger, C., Regli, D., Zwahlen, D., Hostettler, S., & Caspar, F. (2010). Der Berner Patienten- und Therapeutenstundenbogen 2000. *Zeitschrift Für Klinische Psychologie Und Psychotherapie*, *39*(2), 71–79. https://doi.org/10.1026/1616-3443/a000015

Franke, G. H. (2000). *Brief symptom inventory (BSI) von LR Derogatis:(Kurzform der SCL-90-R)*. Beltz Test.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, *33*(1), 1–22.

Gaynor, S. T., Weersing, V. R., Kolko, D. J., Birmaher, B., Heo, J., & Brent, D. A. (2003). The prevalence and impact of large sudden improvements during adolescent therapy for depression: A comparison across cognitive-behavioral, family, and supportive therapy. *Journal of Consulting and Clinical Psychology*, *71*(2), 386–393. https://doi.org/10.1037/0022-006X.71.2.386

Geldof, T., Popovic, D., van Damme, N., Huys, I., & van Dyck, W. (2020). Nearest Neighbour Propensity Score Matching and Bootstrapping for Estimating Binary Patient Response in Oncology: A Monte Carlo Simulation. *Scientific Reports*, *10*(1), 964. https://doi.org/10.1038/s41598-020-57799-w

Grawe, K. (1997). Research-Informed Psychotherapy. *Psychotherapy Research*, *7*(1), 1–19. https://doi.org/10.1080/10503309712331331843

Greenfield, M. F., Gunthert, K. C., & Haaga, D. A. F. (2011). Sudden gains versus gradual gains in a psychotherapy training clinic. *Journal of Clinical Psychology*, *67*(1), 17–30. https://doi.org/10.1002/jclp.20748

Hardy, G. E., Cahill, J., Stiles, W. B., Ispan, C., Macaskill, N., & Barkham, M. (2005). Sudden gains in cognitive therapy for depression: A replication and extension. *Journal of Consulting and Clinical Psychology*, *73*(1), 59–67. https://doi.org/10.1037/0022-006X.73.1.59

Hedman, E., Lekander, M., Ljótsson, B., Lindefors, N., Rück, C., Hofmann, S. G., Andersson, E., Andersson, G., & Schulz, S. M. (2014). Sudden gains in internet-based cognitive behaviour therapy for severe health anxiety. *Behaviour Research and Therapy*, *54*, 22–29. https://doi.org/10.1016/j.brat.2013.12.007

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis*, *15*(3), 199–236. https://doi.org/10.1093/pan/mpl013

Hofmann, S. G., Schulz, S. M., Meuret, A. E., Moscovitch, D. A., & Suvak, M. (2006). Sudden gains during therapy of social phobia. *Journal of Consulting and Clinical Psychology*, *74*(4), 687–697. https://doi.org/10.1037/0022-006X.74.4.687

Howard, K. I., Kopta, S. M., Krause, M. S., & Orlinsky, D. E. (1986). The dose–effect relationship in psychotherapy. *American Psychologist*, *41*(2), 159–164. https://doi.org/10.1037/0003-066X.41.2.159

Hunnicutt-Ferguson, K., Hoxha, D., & Gollan, J. (2012). Exploring sudden gains in behavioral activation therapy for Major Depressive Disorder. *Behaviour Research and Therapy*, *50*(3), 223–230. https://doi.org/10.1016/j.brat.2012.01.005

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*(1), 12–19. https://doi.org/10.1037/0022-006X.59.1.12

Kadera, S. W., Lambert, M. J., & Andrews, A. A. (1996). How Much Therapy Is Really Enough? : A Session-by-Session Analysis of the Psychotherapy Dose-Effect Relationship. *The Journal of Psychotherapy Practice and Research*, *5*(2), 132–151.

Kelly, M. A. R., Cyranowski, J. M., & Frank, E. (2007). Sudden gains in interpersonal psychotherapy for depression. *Behaviour Research and Therapy*, *45*(11), 2563–2572. https://doi.org/10.1016/j.brat.2007.07.007

Kelly, M. A. R., Roberts, J. E., & Ciesla, J. A. (2005). Sudden gains in cognitive behavioral treatment for depression: When do they occur and do they matter? *Behaviour Research and Therapy*, *43*(6), 703–714. https://doi.org/10.1016/j.brat.2004.06.002

Krüger, A., Ehring, T., Priebe, K., Dyer, A. S., Steil, R., & Bohus, M. (2014). Sudden losses and sudden gains during a DBT-PTSD treatment for posttraumatic stress disorder following childhood sexual abuse. *European Journal of Psychotraumatology*, *5*. https://doi.org/10.3402/ejpt.v5.24470

Kuhn, M. (2021, November 3). *The caret Package: 6 Available Models*. https://topepo.github.io/caret/available-models.html

Laws, H. B., Constantino, M. J., Sayer, A. G., Klein, D. N., Kocsis, J. H., Manber, R., Markowitz, J. C., Rothbaum, B. O., Steidtmann, D., Thase, M. E., & Arnow, B. A. (2017). Convergence in patient-therapist therapeutic alliance ratings and its relation to outcome in chronic depression treatment. *Psychotherapy Research*, *27*(4), 410–424. https://doi.org/10.1080/10503307.2015.1114687

Lemmens, L. H. J. M., DeRubeis, R. J., Arntz, A., Peeters, F. P. M. L., & Huibers, M. J. H. (2016). Sudden gains in Cognitive Therapy and Interpersonal Psychotherapy for adult depression. *Behaviour Research and Therapy*, *77*, 170–176. https://doi.org/10.1016/j.brat.2015.12.014

Lutz, W., Deisenhofer, A.-K., Rubel, J., Bennemann, B., Giesemann, J., Poster, K., & Schwartz, B. (2022). Prospective evaluation of a clinical decision support system in psychological therapy. *Journal of Consulting and Clinical Psychology*, *90*(1), 90–106. https://doi.org/10.1037/ccp0000642

Lutz, W., Ehrlich, T., Rubel, J., Hallwachs, N., Röttger, M.-A., Jorasz, C., Mocanu, S., Vocks, S., Schulte, D., & Tschitsaz-Stucki, A. (2013). The ups and downs of psychotherapy: Sudden gains and sudden losses identified with session reports. *Psychotherapy Research: Journal of the Society for Psychotherapy Research*, *23*(1), 14–24. https://doi.org/10.1080/10503307.2012.693837

Lutz, W., Rubel, J. A., Schwartz, B., Schilling, V., & Deisenhofer, A.-K. (2019). Towards integrating personalized feedback research into clinical practice: Development of the Trier Treatment Navigator (TTN). *Behaviour Research and Therapy*, *120*, 103438. https://doi.org/10.1016/j.brat.2019.103438

Lutz, W., Schiefele, A.-K., Wucherpfennig, F., Rubel, J., & Stulz, N. (2016). Clinical effectiveness of cognitive behavioral therapy for depression in routine care: A propensity score based comparison between randomized controlled trials and clinical practice. *Journal of Affective Disorders*, *189*, 150–158. https://doi.org/10.1016/j.jad.2015.08.072

Lutz, W., Schwartz, B., & Delgadillo, J. (2022). Measurement-Based and Data-Informed Psychological Therapy. *Annual Review of Clinical Psychology*, *18*, 71–98. https://doi.org/10.1146/annurev-clinpsy-071720-014821

Lutz, W., Tholen, S., Schürch, E., & Berking, M. (2006). Reliabilität von Kurzformen gängiger psychometrischer Instrumente zur Evaluation des therapeutischen Fortschritts in Psychotherapie und Psychiatrie. *Diagnostica*, *52*(1), 11–25. https://doi.org/10.1026/0012-1924.52.1.11

Norton, P. J., Klenck, S. C., & Barrera, T. L. (2010). Sudden gains during cognitive-behavioral group therapy for anxiety disorders. *Journal of Anxiety Disorders*, *24*(8), 887–892. https://doi.org/10.1016/j.janxdis.2010.06.012.

Pavlou, M., Ambler, G., Seaman, S., Iorio, M. de, & Omar, R. Z. (2016). Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. *Statistics in Medicine*, *35*(7), 1159–1177. https://doi.org/10.1002/sim.6782

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Shalom, J. G., & Aderka, I. M. (2020). A meta-analysis of sudden gains in psychotherapy: Outcome and moderators. *Clinical Psychology Review*, *76*, 101827. https://doi.org/10.1016/j.cpr.2020.101827

Stekhoven, D. J., & Bühlmann, P. (2012). Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics (Oxford, England)*, *28*(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

Stiles, W. B., Barkham, M., Connell, J., & Mellor-Clark, J. (2008). Responsive regulation of treatment duration in routine practice in United Kingdom primary care settings: Replication in a larger sample. *Journal of Consulting and Clinical Psychology*, *76*(2), 298–305. https://doi.org/10.1037/0022-006X.76.2.298

Stiles, W. B., Leach, C., Barkham, M., Lucock, M., Iveson, S., Shapiro, D. A., Iveson, M., & Hardy, G. E. (2003). Early sudden gains in psychotherapy under routine clinic conditions: Practice-based evidence. *Journal of Consulting and Clinical Psychology*, *71*(1), 14–21. https://doi.org/10.1037/0022-006X.71.1.14

Stuart, E. A., Rubin, D. B., & Osborne, J. (2004). Matching methods for causal inference: Designing observational studies. *Harvard University Department of Statistics Mimeo*.

Tang, T. Z., & DeRubeis, R. J. (1999). Sudden gains and critical sessions in cognitive-behavioral therapy for depression. *Journal of Consulting and Clinical Psychology*, *67*(6), 894–904. https://doi.org/10.1037/0022-006X.67.6.894

Tang, T. Z., DeRubeis, R. J., Beberman, R., & Pham, T. (2005). Cognitive changes, critical sessions, and sudden gains in cognitive-behavioral therapy for depression. *Journal of Consulting and Clinical Psychology*, *73*(1), 168–172. https://doi.org/10.1037/0022-006X.73.1.168

Tang, T. Z., Luborsky, L., & Andrusyna, T. (2002). Sudden gains in recovering from depression: Are they also found in psychotherapies other than cognitive-behavioral therapy? *Journal of Consulting and Clinical Psychology*, *70*(2), 444–447. https://doi.org/10.1037//0022-006X.70.2.444

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.

van Buuren, S. (2021). *Multivariate Imputation by Chained Equations*. R package version 3.14.0.

Vittengl, J. R., Clark, L. A., & Jarrett, R. B. (2005). Validity of sudden gains in acute phase treatment of depression. *Journal of Consulting and Clinical Psychology*, *73*(1), 173–182. https://doi.org/10.1037/0022-006X.73.1.173

Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J., Balis, U., Marrero, J., Zhu, J., & Higgins, P. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open*, *3*(8). https://doi.org/10.1136/bmjopen-2013-002847

West, S. G., Cham, H., Thoemmes, F., Renneberg, B., Schulze, J., & Weiler, M. (2014). Propensity scores as a basis for equating groups: Basic principles and application in clinical treatment outcome research. *Journal of Consulting and Clinical Psychology*, *82*(5), 906–919. https://doi.org/10.1037/a0036387

Wiedemann, M., Stott, R., Nickless, A., Beierl, E. T., Wild, J., Warnock-Parkes, E., Grey, N., Clark, D. M., & Ehlers, A. (2020). Cognitive processes associated with sudden gains in cognitive therapy for posttraumatic stress disorder in routine care. *Journal of Consulting and Clinical Psychology*, *88*(5), 455.

Wittchen, H.-U., Wunderlich, U., Gruschwitz, S., & Zaudig, M. (1997). SKID I. Strukturiertes Klinisches Interview für DSM-IV. Achse I: Psychische Störungen. Interviewheft und Beurteilungsheft. Eine deutschsprachige, erweiterte Bearb. d. amerikanischen Originalversion des SKID I.

Wucherpfennig, F., Rubel, J. A., Hofmann, S. G., & Lutz, W. (2017). Processes of change after a sudden gain and relation to treatment outcome-Evidence for an upward spiral. *Journal of Consulting and Clinical Psychology*, *85*(12), 1199–1210. https://doi.org/10.1037/ccp0000263

Wucherpfennig, F., Rubel, J. A., Hollon, S. D., & Lutz, W. (2017). Sudden gains in routine care cognitive behavioral therapy for depression: A replication with extensions. *Behaviour Research and Therapy*, *89*, 24–32. https://doi.org/10.1016/j.brat.2016.11.003

Zilcha-Mano, S., Errázuriz, P., Yaffe-Herbst, L., German, R. E., & DeRubeis, R. J. (2019). Are there any robust predictors of “sudden gainers,” and how is sustained improvement in treatment outcome achieved following a gain? *Journal of Consulting and Clinical Psychology*, *87*(6), 491–500. https://doi.org/10.1037/ccp0000401

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320.

Eidesstattliche Erklärung

Ich versichere, dass ich meine Dissertation ohne Hilfe Dritter und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Trier, den 21 April 2023

Nachname: Bennemann Vorname: Björn

Unterschrift: ____________________________