Do we know what we think we know? – Transferability of findings from randomized controlled trials to routine care treatment settings

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

Vorgelegt im Mai 2017

von

Felix Wucherpfennig

Gutachter:

Prof. Dr. Wolfgang Lutz

Dr. Julian Rubel

Dissertationsort: Trier

Danksagung

Ich danke meinem Doktorvater und Chef Prof. Dr. Wolfgang Lutz für die Unterstützung und das in mich gesetzte Vertrauen. Ich glaube, dass ich sehr von den vielseitigen Lernerfahrungen in der Forschung und Lehre profitiert habe, die ich im Rahmen meiner Promotion sammeln durfte. Dabei weiß ich es zu schätzen, dass mir die Freiheiten und Möglichkeiten eingeräumt wurden, Antworten auf jene Fragen zu finden, die mein persönliches Interesse geweckt haben.

Die Befunde der vorliegenden Dissertation basieren auf Daten, die routinemäßig in der Poliklinischen Psychotherapieambulanz der Universität Trier erhoben wurden. Mein Dank gilt dem gesamten Leitungsteam der Psychotherapieambulanz, durch deren Arbeit ein besonders gut organisiertes Forschungsumfeld entstanden ist, von dem ich persönlich sehr profitiert habe.

Ganz besonders möchte ich mich bei Dr. Julian Rubel bedanken. Seine Anregungen und kritischen Rückmeldungen haben meine Forschung enorm bereichert. Bedanken möchte ich mich für seine Geduld und seine Zeit, die er über die Jahre immer wieder für mich aufgebracht hat. Durch seinen Sachverstand und Humor hat er mich aus einigen Sackgassen herausmanövriert und meinen Blick auf die Forschung erweitert.

Mein Dank gilt auch den Patienten der Psychotherapieambulanz. Deren Bereitschaft mit uns Forschern zusammenzuarbeiten und Informationen zu teilen, ist das Fundament auf dem diese Arbeit steht. Ich möchte mich bei meinen Kollegen für deren Inspiration, Anregung und Kritik bedanken. Vielen Dank an Kaitlyn Boyle, die diese Arbeit Korrektur gelesen hat. Im Rahmen meiner Promotion haben sich einige Freundschaften entwickelt, die vielleicht keine „sudden gains“ sind, aber doch „gains“, die mir besonders viel bedeuten.

Table of content

Abstract

Numerous RCTs demonstrate that cognitive behavioral therapy (CBT) for depression is effective. However, these findings are not necessarily representative of CBT under routine care conditions. Routine care studies are not usually subjected to comparable standardizations, e.g. often therapists may not follow treatment manuals and patients are less homogeneous with regard to their diagnoses and sociodemographic variables. Results on the transferability of findings from clinical trials to routine care are sparse and point in different directions. As RCT samples are selective due to a stringent application of inclusion/exclusion criteria, comparisons between routine care and clinical trials must be based on a consistent analytic strategy. The present work demonstrates the merits of propensity score matching (PSM), which offers solutions to reduce bias by balancing two samples based on a range of pretreatment differences.

The objective of this dissertation is the investigation of the transferability of findings from RCTs to routine care settings. In Study I, *n* = 574 CBT patients with major depression from a routine care outpatient clinic were matched stepwise to patients undergoing CBT in a high-quality RCT. First, the RCT's inclusion/exclusion criteria were applied to the routine care sample, subsequently PSM was implemented to adjust for confounding baseline variables and to match the distribution of covariates. Results suggest that CBT for depression in routine care is equally as effective as in the RCT, when applied to comparable patients.

The average efficacy of a treatment does not imply that a specific treatment is beneficial for a specific patient. Accordingly, the investigation of CBT for depression was expanded to include the analysis of individual change patterns (i.e. sudden gains). Previously, Tang and DeRubeis (1999) found that some patients experience sudden symptom improvements between session intervals and that these patients reveal treatment outcomes superior to patients without sudden gains. Study II investigated sudden gains in a routine care sample (*n* = 462) that was matched stepwise to patients from the RCT by Tang and DeRubeis (1999). Matching was performed by two different applications of PSM. Results suggest that similar rates and effects of sudden gains can be expected under routine care conditions, when patients are comparable to those examined in the original study. The closer the match between the

naturalistic sample and the RCT, the more similar the association between sudden gains and treatment success.

Study III assessed the processes that may facilitate treatment outcome after a sudden gain has occurred. A routine care sample of *n* = 211 depressed patients who underwent CBT was analyzed. Patient ratings of general change factors were investigated in the sessions before and after a sudden gain occured. General change factors increased in the sessions following sudden gains. This increase predicted symptom distress at termination. Results provide supporting evidence for the “upward spiral”, a concept previously proposed by Tang and DeRubeis (1999).

The findings of the three studies are discussed along with suggestions to improve replications in psychotherapy research by means of appropriate statistical methods and the utilization of data and protocols from original studies.

1 Introduction

Over 60 years ago, Hans Eysenck criticized that psychotherapy had not been shown to be effective. He stated that the majority of studies failed to prove that treatments facilitate recovery from mental distress (Eysenck, 1952). Eysenck's criticism initiated a controversial debate and stimulated efforts to provide valid empirical support that psychotherapy works. In the years to follow, numerous randomized controlled trials (RCTs) and meta-analyses of these trials demonstrated the efficacy of psychological treatments (Grawe, 1992; Smith & Glass, 1977). Today psychotherapy is well established in health care systems around the globe and consequently regarded as an effective treatment for psychological disorders, helping the majority of patients to overcome their mental distress (Lambert, 2013).

Randomized controlled trials (RCTs) are considered the gold standard of clinical trials. In RCTs, patients are randomly assigned to one of at least two alternative interventions. The assignment strategy, along with a highly structured treatment setting, can help to minimize selection bias and the impact of confounding factors. RCTs are considered to be the most reliable form of scientific evidence in health care. Not surprisingly, psychotherapy research is based to a large extend on RCTs.

The Treatment of Depression Collaborative Research Program (TDCRP, Elkin et al., 1989) is a particularly influential RCT that demonstrated the efficacy of psychotherapy. This program, initiated by the National Institute of Mental Health (NIMH), was the first collaborative multisite study in the field of mental health. The TDCRP revealed strong effects of psychotherapy for depression, comparable to the effects of pharmacotherapy (Elkin et al., 1989). Subsequent results of numerous meta-analyses point in a similar direction (e.g. Cuijpers et al., 2014; Cuijpers, van Straten, Andersson, & van Oppen, 2008; Cuijpers, van Straten, & Warmerdam, 2007).

The effects of psychotherapy observed in RCTs are encouraging, however not necessarily representative of treatments under routine care conditions. There are several characteristics of RCTs, which aim to strengthen the internal validity of study findings. In RCTs, treatments are usually carried out by intensively trained therapists using highly structured treatment manuals. Patients must meet a series of highly specific inclusion criteria and treatment duration is restricted to standardization. These characteristics of clinical trials may help to minimize systematic error, however at the expense of external validity, that is, the transfer of study findings to clinical practice. Treatments under routine care conditions are not usually subjected to comparable standardizations: often therapist may not follow treatment manuals and patients are less homogeneous with regard to their diagnoses and sociodemographic variables (Castonguay, Barkham, Lutz, & McAleavey, 2013; Shadish, Navarro, Matt, & Phillips, 2000). Studies conducted in routine care allow conclusions to be drawn about the generalizability of clinical findings to "real world conditions" (Seligman, 1995).

Both efficacy studies (highly controlled RCTs) and effectiveness studies (routine care conditions) are important aspects of psychotherapy research. However, there is a gap between the two research paradigms. While an abundance of RCTs can be found promoting the efficacy of psychotherapy for depression, there are significantly fewer naturalistic studies with high clinical representativeness (Shadish et al., 2000). Results on the transferability of findings from RCTs to naturalistic studies are sparse and point in different directions. Some studies have found similar treatment effects (Merrill, Tolbert, & Wade, 2003; Minami et al., 2008), whereas others report that RCTs tend to find larger effects than naturalistic studies (Gibbons et al., 2010; Hansen, Lambert, & Forman, 2002; Weisz, Weiss, & Donenberg, 1992). This points to the necessity of a consistent analytic strategy that allows for a sound comparison between treatments in RCTs and routine care.

This dissertation's analytic strategy is based on the application of propensity score matching (PSM). PSM offers solutions to reduce bias by balancing two samples based on a range of pretreatment differences (Rosensbaum & Rubin, 1983). This strategy seems particularly important as the comparison of treatment effects across different populations and study designs is potentially biased due to a range of confounding covariates. Baseline variables such as intake symptom severity, number of comorbid disorders, age, employment status and marital status have been repeatedly found to predict treatment response (Kessler, van Loo, Wardenaar, Bossarte, Brenner, Cai et al., 2016; Kessler, van Loo, Wardenaar, Bossarte, Brenner, Ebert et al., 2016). RCT samples are selective as many patients frequently encountered in clinical practice are excluded due to a stringent application of inclusion/exclusion criteria. Accordingly, we can expect an uneven allocation of baseline covariates in samples that stemmed from RCTs and naturalistic studies respectively.

Study I aims to bridge the gap between efficacy and effectiveness studies by comparing CBT for depression in the TDCRP (Elkin et al., 1989) with a routine care treatment setting. Two consecutive steps were performed to enhance the comparability of RCT and naturalistic study patients and to reduce the impact of potential confounders. First, the RCT's inclusion/exclusion criteria were applied to the routine care sample. In a second step, PSM was implemented to match the distribution of covariates between samples.

The average efficacy of a treatment does not imply that a specific treatment is beneficial for a specific patient. Thus, Study II and Study III expanded the comparison between RCTs and routine care studies to the field of process-outcome research. Process-outcome research refers to therapeutic processes (therapist behaviors, client behaviors and the interaction between therapist and client) and changes (e.g. symptomatic distress) that happen as a result of these processes. One of the objectives is to gain insight into individual change patterns and to utilize this knowledge for personalized, adaptive treatment strategies (Grawe, 2006; Orlinsky, Grawe, & Parks, 1994). Tang and DeRubeis (1999) were the first to define and empirically test a concept known as sudden gains, which is widely used in process-outcome research. Sudden gains are large symptom improvements that occur suddenly from one psychotherapy session to the next.

In a study of CBT for depression, Tang and DeRubeis (1999) found that some patients experienced a sudden gain in a single, between-session interval. At post-treatment, patients with sudden gains were less depressed than other patients in the sample. However, recent replications point in different directions. Some studies have found a significant association between sudden gains and treatment outcome (Abel, Hayes, Henley, & Kuyken, 2016; Doane, Feeny, & Zoellner, 2010; Hardy et al., 2005), whereas other studies did not find a considerable association (Kelly, Cyranowski, & Frank, 2007; Present et al., 2008; Stiles et al., 1996; Stiles et al., 2003). The original findings and the vast majority of subsequent findings are based on RCTs. For instance, Hardy et al. (2005) found a strong association between sudden gains and treatment outcome and their treatment context was subjected to standardizations comparable to RCTs. In contrast, the treatment context of the study by Stiles et al. (2003) can be characterized as routine care. However, they did not find a meaningful association between sudden gains and outcome. Inter-individual differences of how patients leverage a sudden gain may contribute to the explanation of divergent study findings. Some patients with sudden gains are perhaps more likely to sustain improvements and to end up recovered than other sudden gainers. Patients differential ability to leverage a sudden gain may be associated with specific baseline variables (cf. Kessler, van Loo, Wardenaar, Bossarte, Brenner, Cai et al., 2016; Kessler, van Loo, Wardenaar, Bossarte, Brenner, Ebert et al., 2016).

Study II investigates whether a similar association between sudden gains and treatment outcome can be expected under routine care CBT for depression, when the patients are comparable to those examined in the original study by Tang and DeRubeis (1999). Similar to the first study, PSM was used to enhance comparability between samples. The effects of sudden gains among routine care patients were analyzed before and after the application of PSM respectively.

The third study builds on the results of the second study. It examines therapeutic processes that may facilitate recovery after a sudden gain has occurred. Tang and DeRubeis (1999) assumed that sudden symptom improvements spark an upward spiral by improving the therapeutic alliance and cognitive changes in the following therapy sessions. Furthermore, they hypothesized that improved alliance quality and decreased cognitive bias sustain symptom relief and eventually lead to recovery. To date, only a few studies (Bohn, Aderka, Schreiber, Stangier, & Hofmann, 2013; Lutz et al., 2012), have focused on processes during the sessions following a sudden gain. Tang and DeRubeis (1999) concept of an upward spiral has yet to be replicated in a routine care treatment setting.

Accordingly, Study III investigates patients’ ratings of general change factors (therapeutic alliance; coping skills) in the sessions before and after a sudden gain as predictors of ultimate treatment outcome. PSM was used to compare sudden gain patients with similar patients who did not experience a sudden gain.

The three studies are described in detail in chapters five to seven. All studies presented in this dissertation stress the need for accurate replications under routine care conditions to test the generalizability of previous findings. Accordingly, chapter two highlights suggestions to improve replications in psychotherapy research by means of appropriate statistical methods. The specific research questions that motivated the three studies are described in chapter three. Chapter four describes the utility of different PSM applications as a method to reduce bias in comparative studies. Finally, the implications, limitations and future directions deduced from the three studies are discussed in chapter nine.

2 Theoretical Background

2.1 Replication in Psychological Science

This dissertation’s underlying question is: Do we know what we think we know? Of course, there is no final answer to this question. Most of the time, we simply do not know whether our knowledge is true or false. In line with Poppers (1983) epistemology, a hypothesis can be accepted as scientifically true until it is falsified by counterevidence. Thus, we must acknowledge a degree of uncertainty with regard to what we believe we already know. A single scientific study almost never provides a definitive resolution for or against an effect and its explanation. There are often alternative explanations that may account for the observed effect. Accurate replications are important to tackle the uncertainty of scientific evidence. In psychotherapy research, replications can help to control for sampling errors and treatment artifacts. Moreover, replications can help to assess the generalizability of the original findings across different treatment settings and populations. This dissertation investigates the reproducibility of the effects of CBT for depression found in RCTs under routine care conditions. The central goal is to assess the transferability of RCT findings to treatment conditions with high clinical representativeness based on samples with a comparable distribution of baseline covariates.

Recently, the Open Science Collaboration (2015) conducted replications of 100 studies published in major psychological journals. They found a mean effect size of only half the magnitude of the original effects. The average replication effect size (M = 0.2) comprised only 36 % significant effects, whereas the average original effect size (M = 0.4) comprised 97 % significant results. One possible explanation for this decline is that the replication methodology differs from the original in ways that interfere with observing the effect. Another explanation is simply that the original findings are exaggerated or even false positive (incorrect rejection of a null hypothesis). Exaggerated findings may be promoted by a biased research and publication practice towards significant positive results (cf. Easterbrook, Gopalan, Berlin, & Matthews, 1991).

Simmons, Nelson, and Simonsohn (2011) raised concerns that flexibility of data collection, analysis and reporting increases false positive findings. Accordingly, it is common practice to explore different analytic strategies “post hoc” and to exclusively report those strategies that yield significant results. This behavior is may be driven both by the ambiguity of analytic decisions and researchers' desire to find significant results. Researchers are likely to be self-serving in their interpretation of results. When faced with ambiguous analytic decisions, they may tend to conclude with self-justification that the appropriate decision is the one that results in statistical significance (Babcock & Loewenstein, 1997; Dawson, Gilovich, & Regan, 2002). On a related note, simulations demonstrate that treatment effects are often inflated or even false positive given the prevailing bias and statistical power that is to be observed in most clinical trials (Ioannidis, 2005, 2008).

Scientific credibility must be based on both the quality of the original findings and their replication success. Accurate replications are regarded as the final arbitrator when determining whether effects are true or false (Cohen, 1995). However, only a small number of published articles and book chapters address this topic. One possible explanation is that journal editors and reviewers are inclined to dismiss replications as unoriginal. Reproducibility is not well promoted in the scientific community and novelty is often prioritized over replication (Open Science Collaboration, 2015; Schmidt, 2009).

Ioannidis (2014) postulates a replication culture by means of collaborative research and sharing data. In line with Ioannidis recommendations, this dissertation's analytic strategy is based on a collaborative approach. The data and protocols from the original RCTs (Elkin et al., 1989; Hollon et al., 1992; Tang & DeRubeis, 1999) were made available to Prof. Lutz's research group at the University of Trier. This data was utilized for the application of PSM to match patients from a routine care university outpatient clinic to those from the original RCTs.

There are two basic types of replications: A direct replication is defined as the attempt to reproduce a finding based on the exact repetition of an experiment. In contrast, a conceptual replication attempts to replicate a research finding with a different experimental set-up (Schmidt, 2009). The present studies can be defined as conceptual replications. In the context of this dissertation, the attempt is made to replicate findings from RCTs in a routine care setting that differs with respect to the degree of standardization of pretreatment training of therapists, treatment compliance checks, treatment duration and the use of treatment manuals. Apart from these differences, comparisons were based on samples with a similar distribution of baseline covariates. Replication is the attempt to recreate the conditions that are believed to be sufficient to obtain a previously observed effect. The adjustment for sample differences can be regarded as the replication condition.

Shadish et al. (2000) suggest that efficacy and effectiveness studies should be seen on a continuum of clinical representativeness. At the one end of the continuum, there are highly structured and standardized RCTs. At the other end, there are routine care treatment settings. The closer the samples from RCTs and naturalistic studies are located on this continuum, the more similar the outcome effects. The concept of clinical representativeness points to the need for a consistent analytic strategy that allows a sound comparison between different study designs. As it has been repeatedly shown that a substantial proportion of variance in outcome is explained by patient characteristics, I believe that one important aspect of replication success is the adjustment for sample differences at baseline (Barber, 2007; Delgadillo, Moreea, & Lutz, 2016; DeRubeis, Gelfand, German, Fournier, & Forand, 2014; Kessler, van Loo, Wardenaar, Bossarte, Brenner, Ebert et al., 2016).

3 Research Questions

The first study compares the efficacy with the effectiveness of CBT for depression. The comparison is based on effect sizes that account for pre-existing differences between the samples. The second study investigates the association between sudden gains and treatment outcome. The study aims to replicate findings from the original study by Tang and DeRubeis (1999) under routine care conditions. The objective of the third study is to provide supporting evidence for the upward spiral, which was proposed by Tang and DeRubeis (1999). According to this concept, sudden gains may improve the therapeutic alliance and cognitive changes in the following therapy sessions. In turn, these improvements may sustain symptom relief and eventually lead to recovery.

3.1 Study I

1. Can we find similar effects for CBT under routine care conditions when the patients are comparable to those examined in RCTs?

3.2 Study II

1. Can we find similar rates of sudden gains (percentage of patients with a sudden gain) under routine care conditions when patients are comparable to those examined in the original study by Tang and DeRubeis (1999)?
2. It the association between sudden gains and treatment outcome comparable to the association found in the original study?

3.3 Study III

1. Can we find significant improvements in patients' perceived therapeutic alliance and coping skills in the sessions following a sudden gain?
2. Can we find comparable improvements of general change factors for patients who did not experience a sudden gain?
3. Do general change factors in the sessions following a sudden gain predict treatment success at termination?

4 Methodological Aspects

All three studies share a common procedure for the adjustment of pre-existing sample differences. The following section briefly describes the application of PSM in general and with an emphasis on the specific methods realized in the present studies.

4.1 Propensity Score Matching

PSM was originally applied in observational studies to aid in the evaluation of cause-effect hypotheses (Rosensbaum & Rubin, 1983). In observational or non-experimental studies, a random assignment is often not feasible or even unethical. Non-experimental studies are common in epidemiology and clinical effectiveness research. These designs are a suitable option in psychological research. However, a non-random allocation increases the influence of confounders. Potential confounders are pre-treatment differences (e.g. age, symptom severity, comorbid diagnosis) that introduce bias by influencing both the assignment (e.g. treatment group and control group) and the outcome.

Propensity score based techniques can be separated in two consecutive steps: the estimation of the propensity score and its application (cf. Harder, Stuart, & Anthony, 2010). The propensity score is an estimate of the probability (P) of receiving one of two treatments or interventions (T_i), given a vector of observed covariates (X_i): $e(X_i) = P(T_i = 1|X_i)$. Propensity scores range from 0 to 1 and are typically estimated by logistic regressions with a binary dependent variable (e.g. receiving treatment A[1] or B) and several covariates (e.g. pre-treatment differences) as independent variables (Guo & Fraser, 2014). In a randomized treatment assignment, the propensity score for each patient would be 0.5 (probability of 50 %) if the randomization procedure successfully produced comparable distributions of the relevant covariates. Accordingly, if two patients have the same propensity score they have comparable scores in observed covariates and therefore the same probability of receiving treatment A (B).

The ultimate goal of PSM is to generate a strong ignorability, which is a key assumption for estimating true causal effects (Rosensbaum & Rubin, 1983). Strong ignorability holds if the treatment

[1] In the context of this dissertation treatment A equals the RCT and treatment B the naturalistic study

assignment is independent of the outcome. In other words, there should be no unmeasured confounders of the association between the treatment and the outcome. Covariates should be chosen on the basis of their empirical and/or theorized potential to confound the relationship between the treatment and the outcome (Harder et al., 2010). Confounders are covariates that cause both treatment assignment and the outcome. Accordingly, it is key to identify all potential confounders and to include these covariates in the PSM model (West et al., 2014).

The application of PSM refers to the use of the estimated propensity scores to reduce bias by balancing two samples (e.g. receiving treatment A or B). Three different applications were performed in the present thesis: nearest neighbor (NN) matching, caliper matching and full matching. All three procedures have advantages and disadvantages.

The NN method (1:1) matches a patient to its counterpart based on the most similar or, in other words, nearest propensity score. The NN for each patient in treatment A is identified, starting with the highest propensity score in the sample. Consequently, each match drawn from treatment A and B is removed from the data set. The process is iterated until every patient in sample A has a match in sample B (Guo & Fraser, 2014).

The caliper method matches every patient within a pre-specified range or caliper to its counterpart. The caliper is the maximum tolerated difference between matched individuals. It is defined as the standard deviation of the sample estimated propensity score. In contrast to the NN method (1:1), caliper matching allows more matching patients per target patient. This strategy may help to improve the generalizability of a finding, however at the expense of a potential worse fit between samples (Guo & Fraser, 2014).

The NN and caliper matching approaches are sometimes criticized for discarding data, as the unmatched comparison individuals are excluded from subsequent analysis. Full matching is a method that makes use of all individuals in the data by providing a series of matched sets, i.e. subclasses. The subclasses are organized as follows: A patient from treatment A with many comparison individuals is grouped with several patients from treatment B, whereas a patient of treatment A with few comparison individuals is grouped with relatively fewer patients from treatment B. Hence, each subclass contains a match of a patient from treatment A with one or more patients from treatment B. There are weights attached to each subclass, which are scaled to equal the number of matched comparison individuals (Stuart & Green, 2008).

There are different tests to scrutinize the goodness of the match. These assessments are intended to check whether a sufficient balance of each of the baseline covariates was achieved by the application of PSM. A prvailing strategy is the comparison of standardized mean difference (smd) scores for each covariate before and after the matching procedure. The smd score is defined as the weighted difference of means of each covariate between sample A and sample B standardized by the standard deviation of sample A before adjustment. A covariate with an smd score < 0.25 indicates an acceptable match between samples (Rubin, 2001).

An estimate for the causal effect after the application of PSM is the average treatment effect (ATE). The ATE represents the estimate of the gain from receiving treatment A rather than treatment B for an individual randomly selected from the population. If the assumption of full treatment adherence is violated, an alternative estimate is proposed, that is, the average treatment effect of the treated (ATT). The ATT represents the average outcome for a patient from treatment A who actually received the treatment with those in treatment B, who would have accepted the treatment if offered (West et al., 2014). Different applications of PSM yield different estimates of treatment effects depending on the weighing that is done to groups A and B. The PS methods illustrated in this dissertation estimate the ATT (Harder et al., 2010). The ATT is typically calculated by outcome regression models that include the covariates used in PSM. These models usually account for the weights created by PSM to ensure that the matched individuals of treatment B resemble the units of treatment A (Ho, Imai, King, & Stuart, 2011; Kurth et al., 2006).

5 Study I: Clinical effectiveness of cognitive behavioral therapy for depression in routine care: A propensity score based comparison between randomized controlled trials and clinical practice

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.

5.1 Abstract

Background: The efficacy of cognitive behavioral therapy (CBT) for the treatment of depressive disorders has been demonstrated in many randomized controlled trials (RCTs). This study investigated whether for CBT similar effects can be expected under routine care conditions when the patients are comparable to those examined in RCTs.

Method: *N* = 574 CBT patients from an outpatient clinic were stepwise matched to the patients undergoing CBT in the National Institute of Mental Health Treatment of Depression Collaborative Research Program (TDCRP). First, the exclusion criteria of the RCT were applied to the naturalistic sample of the outpatient clinic. Second, propensity score matching (PSM) was used to adjust the remaining naturalistic sample on the basis of baseline covariate distributions. Matched samples were then compared regarding treatment effects using effect sizes, average treatment effect on the treated (ATT) and recovery rates.

Results: CBT in the adjusted naturalistic subsample was as effective as in the RCT. However, treatments lasted significantly longer under routine care conditions.

Limitations: The samples included only a limited amount of common predictor variables and stemmed from different countries. There might be additional covariates, which could potentially further improve the matching between the samples.

Conclusions: CBT for depression in clinical practice might be equally effective as manual-based treatments in RCTs when they are applied to comparable patients. The fact that similar effects under routine conditions were reached with more sessions, however, points to the potential to optimize treatments in clinical practice with respect to their efficiency.

Keywords: randomized controlled trial, clinical practice, cognitive behavioral therapy, depressive disorders, propensity score matching, treatment effects

5.2 Introduction

With a lifetime prevalence of 9.5% depressive disorders are the second most common mental disorder after anxiety disorders (Kessler, Chiu, Demler, & Walters, 2005). According to the World Health Organization (WHO) depression is even the leading disorder concerning the overall burden of diseases and it might be the second-leading cause of disability worldwide by 2020 (Murray & Lopez, 1996). Not surprisingly, depression therefore is one of the most intensively studied mental disorders (Cuijpers et al., 2014; Cuijpers et al., 2008). Actually, more than 350 randomized controlled trails (RCT) on the efficacy of depression treatment have been published. The effects of well-standardized depression treatments found in highly controlled RCTs have to be compared to the effects of depression treatment when delivered under routine care conditions, however. There are several peculiarities of RCTs which aim to strengthen the internal validity of study findings but which may hamper the external validity, that is, transfer of the study's findings to clinical practice:

RCTs usually use highly structured treatment manuals for psychosocial interventions and therapists are intensively trained to ensure that all patients receive a comparable treatment. Therapists in clinical practice may often not follow treatment manuals that strictly. Strict standardization of psychotherapeutic procedures and their one-to-one transfer from RCTs to clinical practice is therefore much more difficult in psychotherapy research than for other medical interventions (e.g. pharmacotherapy). Moreover, RCTs usually only include patients who meet a series of highly specific inclusion criteria in order to generate homogenous samples and hence to strengthen the validity of the causal inferences. Combined with the restriction on voluntary patients who accept to be randomly assigned to a treatment condition, these inclusion/exclusion criteria may lead to highly selective samples in RCTs that omit many patients encountered in clinical practice. For instance, studies on antidepressant medications often exclude more than 80% of the patients with a major depression disorder (MDD) due to any non-conformity with the inclusion criteria (Keitner, Posternak, & Ryan, 2003; Zetin & Hoepner, 2007). While comorbid disorders commonly represent an exclusion criterion in RCTs, patients with more than one mental disorder are frequently seen in clinical practice. Consequently, well-conducted efficacy studies increasingly became criticized in terms of their external validity (Rothwell, 2005), and several efforts have been made to improve the external validity in RCTs. The STAR*D research program, for example, used an equipoised stratified randomized design and gave each patient the possibility to accept the assignment to a particular treatment strategy (e.g., pharmacotherapy and CBT) or decline it and to move to another study arm. This procedure was intended to be more close to what happens in routine care and to reduce the number of non-consenters, resulting in a higher external validity of the study’s findings (Warden, Rush, Trivedi, Fava, & Wisniewski, 2007).

To date, it is generally accepted that both, efficacy (strictly controlled RCTs) and effectiveness studies (studies in naturalistic clinical settings that strengthen external validity at the cost of internal validity) are necessary to evaluate the usefulness of a treatment protocol (Castonguay et al., 2013; Finger & Rand, 2003; Green & Glasgow, 2006; Rothwell, 2005; Taylor & Asmundson, 2008). Results on the transferability of findings from RCTs to naturalistic studies are mixed: while some studies found similar effects (Merrill et al., 2003; Minami et al., 2008), others report that efficacy studies tend to find larger effect sizes than naturalistic studies (Gibbons et al., 2010; Hansen et al., 2002; Weisz et al., 1992). Furthermore, the outcome variance in naturalistic samples tends to be larger than in RCTs (McEvoy & Nathan, 2007). These findings point to the need for a further investigation of the comparability between treatment effects in RCTs and in naturalistic settings.

We therefore aimed to compare the effects of CBT for patients with MDD in (a) a high-quality RCT (Elkin et al., 1989) and (b) a naturalistic study performed under routine care conditions. As in previous research (Schindler, Hiller, & Witthöft, 2011; Shadish et al., 1997; Shadish et al., 2000), we first applied the inclusion/exclusion criteria of the RCT to the sample from routine care to enhance the comparability of the patients examined in both study designs. In addition, we subsequently implemented *propensity score matching* (PSM) to adjust for confounding baseline variables between samples and to match the variable distributions (Rosensbaum & Rubin, 1983; West, Cham, & Thoemmes, 2015).

5.3 Methods

The current study was based on data from the National Institute of Mental Health Treatment of Depression Collaborative Research Program (Elkin et al., 1989), which was a large multicenter RCT in the US, as well as on naturalistic outcome data, which was routinely assessed at the University Outpatient Clinic Trier in the Southwest of Germany.

Patients, instruments and data collection in the TDCRP

Details on the design and procedures of the TDCRP trial have already been published elsewhere (Elkin et al., 1989; Elkin, Parloff, Hadley, & Autry, 1985). Therefore, we will provide only a brief overview here. The TDCRP was a collaborative randomized controlled clinical trial comparing four treatments for MDD at three research sites (George Washington University, University of Pittsburgh, and University of Oklahoma). Eligible patients had to meet Research Diagnostic Criteria (Spitzer, Endicott, & Robins, 1978) for a current episode of a MDD and they had to have a score of 14 or greater on a modified version of the 17-item Hamilton Rating Scale for Depression (Hamilton, 1967) at intake. Further inclusion and exclusion criteria were (Figure 1): (a) presence of a MDD (with required symptomatology present for at least the previous two weeks); (b) male and female outpatients between the ages of 21 and 60; (c) minimum education of eighth grade; (d) sufficient reading and comprehension capabilities to complete self-report forms; (e) no specific additional psychiatric disorders (bipolar I or II, psychotic disorder, panic disorder, alcoholism, drug use disorder, antisocial personality, Briquet's syndrome, MDD with psychotic subtype); (f) no more than one schizotypal features; (g) no history of schizophrenia; (h) no organic brain syndrome; (i) no mental retardation; (j) no concurrent treatment; (k) no presence of specific physical illness or other medical contraindication for the use of imipramine; (l) no clinical state inconsistent with participation in the research protocol (e.g., current active suicide potential) (Elkin et al., 1985).

Of the 560 patients screened for the TDCRP study, 250 patients fulfilled the inclusion and exclusion criteria of the trial. In each of the three research sites these were randomly allocated to one of four treatment conditions: Cognitive Behavioral Therapy (Beck, Rush, Shaw, & Emery, 1979), Interpersonal Therapy (Klerman, Weissman, Rounsaville, & Chevron, 1984), imipramine plus clinical management (IMI-CM) or placebo plus clinical management (PLA-CM). Of the 250 patients who had been randomly assigned to one of the four treatment conditions, 239 actually began treatment (59 in CBT, 61 in IPT, 57 in IMI-CM, and 62 in PLA-CM) and 162 completed treatment (Figure 1). All treatments were planned to be 16 weeks in length, with a range of 16 to 20 sessions (Elkin et al., 1985). Of the 162 completers, 40 were treated with CBT, 47 with IPT, 38 with IMI-CM, and 37 with PLA-CM. However, the current study included only the 40 patients of the TDCRP who completed CBT.Figure 1 gives an overview of the selection of the actual sample. The descriptive statistics of the CBT completer sample are presented in Table 1.

Patients filled out several instruments before treatment, at several points during treatment (4, 8, and 12 weeks), at treatment termination (16 weeks), and at 6-, 12-, and 18-

month post-treatment. The following instruments were assessed in both the TDCRP and the University Outpatient Clinic Trier:

The *Hopkins Symptom Checklist-90* (Lipman, Covi, & Shapiro, 1979) includes 90 items and was assessed at all assessment points. It inquires physical and psychological symptoms within the last week and assesses 9 subscales with the following dimensions: somatization, obsessive-compulsive, interpersonal sensitivity, depression, anxiety, hostility, phobic anxiety, paranoid ideation, and psychoticism. Items are based on a 5-point Likert scale ranging from 0 ("not at all") to 4 ("extremely"). For the analyses in this paper we used only the 53 items of the original HSCL-90 version that can be matched with the *Brief Symptom Inventory* (Derogatis, 1977; Franke, 2000), shortened version of the HSCL-90 that was routinely assessed at the University Outpatient Clinic Trier. The Global Severity Index (GSI), which is computed by averaging all BSI items, served as the primary outcome measure in the current study. Hence the GSI at baseline was used as a covariate in the PSM. Psychometric properties of this index are good ($\alpha = .92$; $r_{tt} = .90$; Franke, 2000).

The *Beck Depression Inventory* (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) is a self-report instrument to capture depressed mood. The items relate to depressive symptoms such as mood, hopelessness, and cognitions like guilt or feelings of being punished as well as physical symptoms like the loss of appetite or the loss of libido. In the naturalistic sample of the University Outpatient Clinic Trier the BDI-II (Beck, Steer, Ball, & Ranieri, 1996), a revised version of the BDI, was assessed. Like the BDI, the BDI II contains 21 questions of which some have been changed due to adapted diagnostic criteria for MDD in the DSM IV (American Psychiatric Association, 2000). Each item can be answered with scores between 0 and 3 whereby higher total scores indicate more severe depressive symptoms. Psychometric properties of the BDI and BDI-II range between acceptable and excellent (Beck et al., 1996;

Beck, Steer, & Carbin, 1988). The sum scores of the BDI and the BDI-II, respectively, were used as covariates in the PSM.

The *Dysfunctional Attitude Scale* (Weissman, 1980) is an instrument to assess the intensity of dysfunctional attitudes that reflect a general cognitive vulnerability factor to depression. In the naturalistic sample of the University Outpatient Clinic Trier only the short version of the DAS was assessed (DAS-K). The DAS-K consists of 26-items (instead of 40 items in the DAS) and has high internal consistency (Floyd, Scogin, & Chaplin, 2004). Hence, the DAS-K was used as one of the covariates for PSM.

Patients, instruments and data collection in the naturalistic sample

The routine care sample compromised 574 patients treated by 94 therapists at the University Outpatient Clinic Trier between 2010 and 2014. All therapists took part in a three (full-time) or five year (part-time) postgraduate training program with a CBT focus and had at least one year of training before starting to see patients within this project. The data collection was part of the routine outcome monitoring at the clinic, which takes place before treatment, at each session during treatment and at termination. Instruments which were relevant for this study have been reported above: BSI, BDI-II and DAS-K.

Patients included in our analyses had received at least 3 sessions of individual treatment with a mean treatment length of 33.03 sessions (*SD* = 18.82, interquartile range 21-45). Diagnoses were based on the *Structured Clinical Interview for Axis I DSM-IV Disorders* (Spitzer, Gibbon, Skodol, Williams, & First, 2002) which was conducted before the actual therapy by intensively trained independent clinicians. Above this, SKID-I interviews were videotaped 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.

Data Analytic Strategy

First, all inclusion and exclusion criteria of the TDCRP were used to adjust the naturalistic sample to the experimental trial data (Figure 1). The HRSD was the only exception because this instrument was not available in the naturalistic sample. Second, we used *propensity score matching* (PSM; Rosenbaum and Rubin, 1983) to adjust for confounding baseline variables as well as to match the variable distributions between the two samples deriving from clinical practice and a RCT, respectively. PSM has become increasingly popular in other disciplines such as epidemiology, economics, and political and social sciences (Barabas, 2004; Boyd, Epstein, & Martin, 2010; Imbens, 2004) in psychology the application of PSM is still rare (West et al., 2014; West et al., 2015). In RCTs randomization is used to generate intervention and control groups that have comparable distributions of observed and unobserved baseline covariates. The rationale behind random allocation of patients to study arms in RCTs is to ensure that potentially observed differences in outcome variables are due to differences in the treatment and not due to confounding variables (pre-existing differences between treatment and control groups in other relevant variables). In contrast, in naturalistic studies randomization and control for confounders are not part of the study design. Traditionally, analysis of covariance (ANCOVA) or structural equation modeling (SEM) were used to equate the study groups undergoing various interventions and to control for confounders post-hoc. Rubin (2001) however pointed out that ANCOVA can lead to biased estimations if there are large differences between the covariate distributions of the compared groups or if regression assumptions are not fully met. PSM allows to match or to equate groups based on comparable propensity scores for patients in both groups (region of common support, West et al., 2014).

Covariates. The selection of covariates is crucial in PSM since all potential variables influencing or predicting treatment outcome should be reliably measured and included to balance comparison groups. This optimally results in a strong ignorability as described by

Rosenbaum and Rubin (1983). There is an ongoing debate about whether PSM should include as many potential confounders as possible (e.g. Rubin, 2001) or whether it should focus only on covariates which have a significant impact on the selection of cases or on the outcome variable (Augurzky & Schmidt, 2001). For psychological interventions initial severity of symptoms is known to be one of the best predictors, that is, one of the case-mix variables with the largest impact on patient progress and outcome (Elkin et al., 1989; Garfield, 1994; Lambert, 2013; Lutz et al., 2014; Stulz, Lutz, Leach, Lucock, & Barkham, 2007). Therefore, the pre-scores of the following instruments, all capturing initial severity, were implemented as important covariates in the analysis: BSI as the primary outcome measure and BDI as covariate measuring the severity of depression. Additionally, we included the pre-score of the DAS-K, which measures cognitive vulnerability for depression and represents a core target variable in CBT. Furthermore, we included the following available socio-demographic variables as potential cofounders of outcome: sex, age, education and employment status. Table 1 gives an overview of the seven baseline covariates used in this study[2].

Propensity scores. The propensity score $e(X_i)$ for subject *i*, defined by Rosenbaum and Rubin (1983), is the probability (P) of receiving one of the two treatments (T_i), given a vector of observed covariates (X_i):

$$e(X_i) = P(T_i = 1|X_i)$$

The propensity score is a probability score which ranges from 0 to 1. If the concept of propensity scores would be implemented in an RCT where an experimental condition is compared to a control group, the propensity score for each patient would be .5 (probability of 50%), when the randomization procedure did successfully produce comparable distributions in

[2] As suggested by West et al. (2014), multiple imputations were used to handle missing values in the baseline covariates. Multiple imputations were generated with the Amelia II package in R (Honaker, King, and Blackwell (2011). Each missing value was replaced by the average score derived from five iteratively simulated datasets each taking all covariates into account (Honaker et al., 2011). No patient had more than two missing values on baseline variables.

the relevant covariates. Transferred to the context of this study: If two patients have the same propensity score then they have comparable scores in observed covariates and therefore the same probability of receiving the treatment in the RCT (TDCRP sample) or, in our case, the treatment in routine care (University Outpatient Clinic Trier sample). Propensity scores were calculated using logistic regression with the binary dependent variable $T_i = 1$ for TDCRP sample and $T_i = 0$ for University Outpatient Clinic Trier sample and the seven covariates (BSI pre, DAS pre, BDI pre, sex, age, education and employment status) as independent variables (Table 1). Furthermore, interaction and quadratic terms of baseline covariates were added through an iterative process as described by Dehejia and Wahba (1999) to improve the balance of the PSM. Propensity scores were calculated twice for every patient in the naturalistic sample to match cases to the CBT condition of the RCT.

Subsequently, patients of the naturalistic sample were matched to patients of the TDCRP samples applying either the nearest neighbor (NN) or the caliper matching procedure (e.g. West et al., 2014).

The NN method enables matching each patient of the RCT to his or her NN in the naturalistic sample based on the most similar (nearest) propensity score. Whenever the NN for a patient of the RCT was identified, this matched pair was removed from the data sets. The process was iterated until every patient in the RCT sample had a matched partner in the naturalistic sample.

Similar to the NN method, the caliper procedure enables matching patients of the RCT to similar cases in the naturalistic sample based on covariates that are potentially relevant for treatment outcomes. However, with the caliper method, a matching counterpart in the naturalistic sample is selected only if the absolute distance of the propensity scores between the two patients is within a prespecified caliper, i.e., a predetermined maximal tolerance for proximity. As suggested by Rosenbaum and Rubin (1985) we used a caliper size of a quarter of a standard deviation of the sample estimated propensity score. Thus, a patient in the naturalistic sample is regarded as a match only if its propensity score falls into the caliper of 0.25 *SD*. The caliper method enables to select more than one matching patient per target patient in the RCT sample, if multiple patients in the naturalistic sample are within the prespecified range of minimal closeness. This allowed us to incorporate more information, resulting in a potentially better fit between the two samples, whereas it had the possible disadvantage of eventually losing patients without any sufficiently fitting counterpart in the naturalistic sample. In other words, some patients in the RCT may have more than one matching patient in the naturalistic sample within the prespecified caliper whereas some other patients might have no fitting case in the prespecified caliper at all.

The goodness of the propensity-score model and of the matching procedure is indicated by the degree to which they result in similar distributions of covariates in the matched samples. For this study, we implemented the *standardized mean difference* (smd) technique, a widely used method to check covariate balance between samples (Guo & Fraser, 2014). The smd method is similar to Cohen`s *d* and allows to compare differences in matched and unmatched conditions for each covariate (with the *SD* of the unmatched condition being used as the denominator). A smd < .25 indicates acceptable match between samples on the respective covariate (Rubin, 2001).

Outcome comparisons. Finally, the matched sample from the RCT and the naturalistic setting were compared concerning treatment effects based on effect sizes, average treatment effects of the treated (ATT) and clinical significance. Pre-post effects sizes were calculated by dividing the mean BSI pre-post difference by the pooled pre- and post-standard deviation of each sample. Confidence intervals around effect sizes were bootstrapped (Wilcox, 2011).

To calculate the differential effects between the naturalistic sample and the RCT, the average treatment effect on the treated (ATT) was calculated. The ATT allowed an effect size comparison between the samples, which controlled for differences in potential confounder variables (Guo & Fraser, 2014). Finally, patients' outcomes were classified according to the concept of clinical significant change[3] (Jacobson & Truax, 1991; Lutz, Stulz, & Köck, 2009). All analyses were performed in R using the MatchIT package (Ho et al., 2011).

5.4 Results

Independent *t* tests and χ^2 tests were calculated to compare the baseline variables (BSI, BDI, DAS-K, sex, age, education and employment status; Table 1) and treatment length between the full naturalistic sample (N = 574) and the CBT subsample of the TDCRP used in this study (n = 40). Pretreatment scores in the BSI ($t(612) = 2.30$, $p = .02$) and the BDI ($t(53.28) = 2.48$, $p = .02$) both were significantly lower in the full naturalistic sample than in the RCT sample whereas education status was significantly higher in the RCT sample ($\chi^2(1, N = 614) = 14.96$, $p < .001$). All other baseline variables did not differ significantly between the two samples.

Patients in the full naturalistic sample were treated for significantly more sessions than those in the RCT ($t(612) = -6.65$, $p < .001$).

In a first step to adjust the naturalistic sample, we applied the inclusion/exclusion criteria of the RCT to the naturalistic sample, which resulted in an adjusted sample of 161 eligible cases (Figure 1 and Table 1). Subsequently, we used these 161 cases to select (a) via caliper matching a sample of patients who fell into the predetermined caliper size when compared to the patients in the RCT (n = 83)[4] and (b) via NN matching another sample of patients who were most similar

[3] To classify patients according to the concept of clinical significant change the reliable change index (RCI) as well as the cut-off score is required. Reliable change is reached when the pre-post difference exceeds the measurement error of the instrument. The cut-off score separates between a dysfunctional and a functional population. For the classification we used the BSI cut-off score of X = .61 and a RCI of X = .27, calculated based on Derogatis and Melisaratos (1983). Recovered: pre score above cut-off, post score below cut-off and RCI fulfilled; improved: RCI fulfilled; no change: RCI is not fulfilled; deteriorated: RCI is fulfilled but in the negative direction.

[4] It should be noted that for three patients of the RCT no matching counterparts were found within the prespecified caliper size of .25 *SD*.

to the CBT subsample of the RCT (*n* = 40). As can be seen in Table 2, after the application of both PSM approaches all baseline variables under consideration were sufficiently well balanced. None of the smd scores exceeded .25. The smd scores of the caliper matching ranged from .02 for BDI pre scores to .08 for employment status and sex. The NN matching yielded smd scores ranging from .06 for education and sex to .18 for BSI pre scores. When compared to the RCT sample, treatment length was still significantly longer for the naturalistic samples after caliper matching (*t*(85.18) = 6.18 , *p* < .001) and after NN matching (*t*(39.71) = 6.10, *p* < .001), respectively.

In all samples under consideration, psychological distress decreased over the course of treatment (Table 1). The matching procedure affected both the initial scores (e.g. in the BSI) and the magnitude of the effect sizes in the selected samples (Table 1 and Figure 2). The smallest pre-post effect size was observed in the naturalistic sample (*d* = .94, 95% CI [.86, 1.03]). Adjusting the naturalistic sample for the inclusion/exclusion criteria of the RCT resulted in an increased effect size of *d* = 1.16, 95% CI [1.01, 1.32]. Caliper matching (*d* = 1.44, 95% CI [1.02, 1.87]) and NN matching (*d* = 1.72, 95% CI [1.31, 2.17]) both further increased the effect sizes of CBT in the resulting subsamples. The highest pr- post effect size on the BSI was observed for the RCT (*d* = 1.85, 95% CI [1.39, 2.40]).

Using the ATT to compare effect-sizes between the samples, the effects of CBT turned out to be significantly larger in the RCT than in the full naturalistic sample (ATT =.21, 95% CI [.05, .37]) and than in the naturalistic sample adjusted for the RCT inclusion/exclusion criteria (ATT = .18, 95% CI [.01, .38]). The effect sizes in the naturalistic subsamples resulting from caliper matching (ATT = .08, 95% CI [-.11, .26]) and from NN matching (ATT = .04, 95% CI [-.18, .24]), however, were not significantly lower than in the RCT sample (Figure 3).

Finally, we examined the recovery rates according to the concept of clinical significant change in all samples. As can be seen in Table 3, the full and the adjusted samples clearly differed from the RCT sample in terms of number of patients classified as recovered, improved, not changed or deteriorated. The recovery rate in the RCT sample was significantly higher than in the full naturalistic sample (χ^2(1, *N* = 614) = 3.97, *p* = .04). This difference in recovery rates disappeared after the application of caliper (χ^2(1, *N* = 123) = .83, *p* = .36) and NN (χ^2(1, *N* = 80) = .05, *p* = .98) matching.

5.5 Discussion

The present study examined whether the effects of CBT for depressive patients in routine care are similar to the effects in a high-quality RCT, if the naturalistic sample is adjusted for inclusion/exclusion criteria of the RCT and matched for further baseline covariates that might affect treatment outcome. PSM, which is a sample matching procedure that takes the distribution of confounding baseline variables into account, was used to select a subsample of patients treated with CBT at a university outpatient clinic who most closely matched the patients treated with CBT in the TDCRP trial.

Previous studies comparing the efficacy (observed in RCTs) and the effectiveness (observed under routine care conditions) of psychological interventions provided inconclusive findings. Some research found efficacy studies to produce higher effect sizes (Gibbons et al., 2010; Hansen et al., 2002; Schindler et al., 2011; Weisz et al., 1992). Some authors, however, argued that both study designs (efficacy and effectiveness studies) should be seen on a continuum of clinical representativeness (Shadish et al., 1997; Shadish et al., 2000). For example, Shadish et al. (2000) showed that the closer the samples from RCTs and naturalistic studies are located on the continuum of clinical representativeness the more similar their outcome effects are. However, so far research in this field has not used a consistent strategy, which allowed for a sound comparison between the outcomes of the two study designs by matching baseline variables to systematically reduce bias.

The present study used a stepwise procedure to adjust for baseline differences. Each step was intended to make the naturalistic subsample more similar to the samples seen in the RCT. The initial comparison of the treatment effects between the full naturalistic sample and the RCT samples revealed significantly smaller effect sizes, ATTs, and recovery rates in the unselected naturalistic sample. After the application of the RCT inclusion/exclusion criteria to the naturalistic sample, the treatment effects in the adjusted naturalistic sample became more similar to the treatment effects in the RCT but they still were significantly smaller. This finding is in line with the results of Schindler and colleagues (2011) who found that even after the application of RCT inclusion/exclusion criteria to naturalistic samples the effects of the naturalistic treatments were weaker than those in RCTs. In the current study we additionally used two methods of PSM (caliper and NN) based on baseline covariates to improve the match between the patients within both designs. After these adjustments for baseline covariates the naturalistic sample and the RCT samples did no more differ significantly with respect to treatment effects. Furthermore, our study replicates the finding from McEvoy and Nathan (2007) that the variance in naturalistic samples (pre- and post-treatment) tends to be larger than in RCTs. However, this study also shows that these differences disappear if a specific matching procedure like PSM (caliper and NN) is applied.

In summary these findings suggest that we may expect similar treatment effects for CBT under routine care or RCT conditions, if the patients seen in both settings are comparable not only regarding the inclusion/exclusion criteria, but also in other important baseline variable distributions. The simple application of RCT inclusion/exclusion criteria to naturalistic data doesn’t seem to be sufficient for a fair comparison, since samples are still to heterogeneous. This emphasizes the importance of matching procedures such as variants of PSM to control for confounders when comparing RCTs and naturalistic studies. For example, even after applying the inclusion/exclusion criteria of the TDCRP to our naturalistic sample, only 40.25% of those patients treated under routine care conditions had more than 12 years of education (compared to 80% of the patients in the RCT). After PSM, the resulting naturalistic subsamples comprised 72.70% (caliper machting) and 85.50% (NN matching) cases with more than 12 years of education which was much more comparable to the RCT.

However, some differences between the samples still existed even after PSM adjustments. With an average treatment length of *M* = 34.58 (*SD* = 18.56) sessions for the caliper matched sample and *M* = 34.70 (*SD* = 18.90) sessions for the NN method, the naturalistic treatment lasted about twice as long as in the RCT, which showed treatment durations of approximately *M* =16.38 sessions (*SD* = 1.08) (see Table 1). These findings raise the question whether treatments under routine conditions could be shortened or whether these differences in treatment length are the result of an uncompleted matching procedure and further covariates would result in even more homogeneity between both designs. This could be an area of further investigation related to concrete consequences for clinical practice.

The main strength of the study is also a limitation and relates to cultural differences and differences concerning the data actuality. This is especially important, since recent findings by Johnsen and Friborg (2015) suggest that the effects found in CBT trials for unipolar depression changed over the last decades and effect sizes diminished over the years. The TDCRP was a clinical trial conducted at three different sites in the US many years ago whereas the data of the naturalistic sample comprised recent data from an outpatient clinic in Germany. Furthermore, differences concerning the weaker diagnostic procedures, the lack of controlling for additional treatments (e.g. medication), the lack of adherence data in the routine care sample as well as dissimilarities in the translation of psychometric instruments and the exclusive use of self-report measures might have hampered the bias reduction or the possibility to add further covariates in the matching procedure. The goal of PSM is a strong ignorability, which means that there are no unmeasured confounders that influence the association between treatment and outcome

(Rosensbaum & Rubin, 1983; Shadish, 2013). Given the described limitations, the matching procedure in this study is unlikely to generate strong ignorability. Hence, the PSM analysis used in this case is by definition limited. There is still a large opportunity of unobserved variables that might have had an impact on treatments and outcomes in the present data. However, although cultural and time differences clearly exist the similarity, which actually resulted after the PSM procedure, is astonishing and indicates the potential of this method for future research.

Another limitation concerns the relatively small sample size. The matching process was based on a caliper and NN approach where each patient in the RCT was matched to the most similar counterpart or to the most similar counterparts in the adjusted naturalistic sample (*n* = 161) based on propensity scores. The smaller the samples in the original dataset are, the more difficult it is to find good matching partners in the target dataset. Even though we were able to achieve an adequate balance, it is nevertheless important to conduct similar analyses in larger samples in order to replicate the present findings within other RCTs and naturalistic samples.

Efficacy and naturalistic studies seem to fall on a continuum of clinical representativeness. The (lower) outcomes in naturalistic studies describe the expected outcomes of a treatment when delivered in that setting. The current study shows that raw effect sizes do not allow us to support conclusions about relative effectiveness. Our findings suggest that some form of matching procedure such as PSM should be considered when comparisons between efficacy and effectiveness studies are intended.

5.6 Tables and Figures

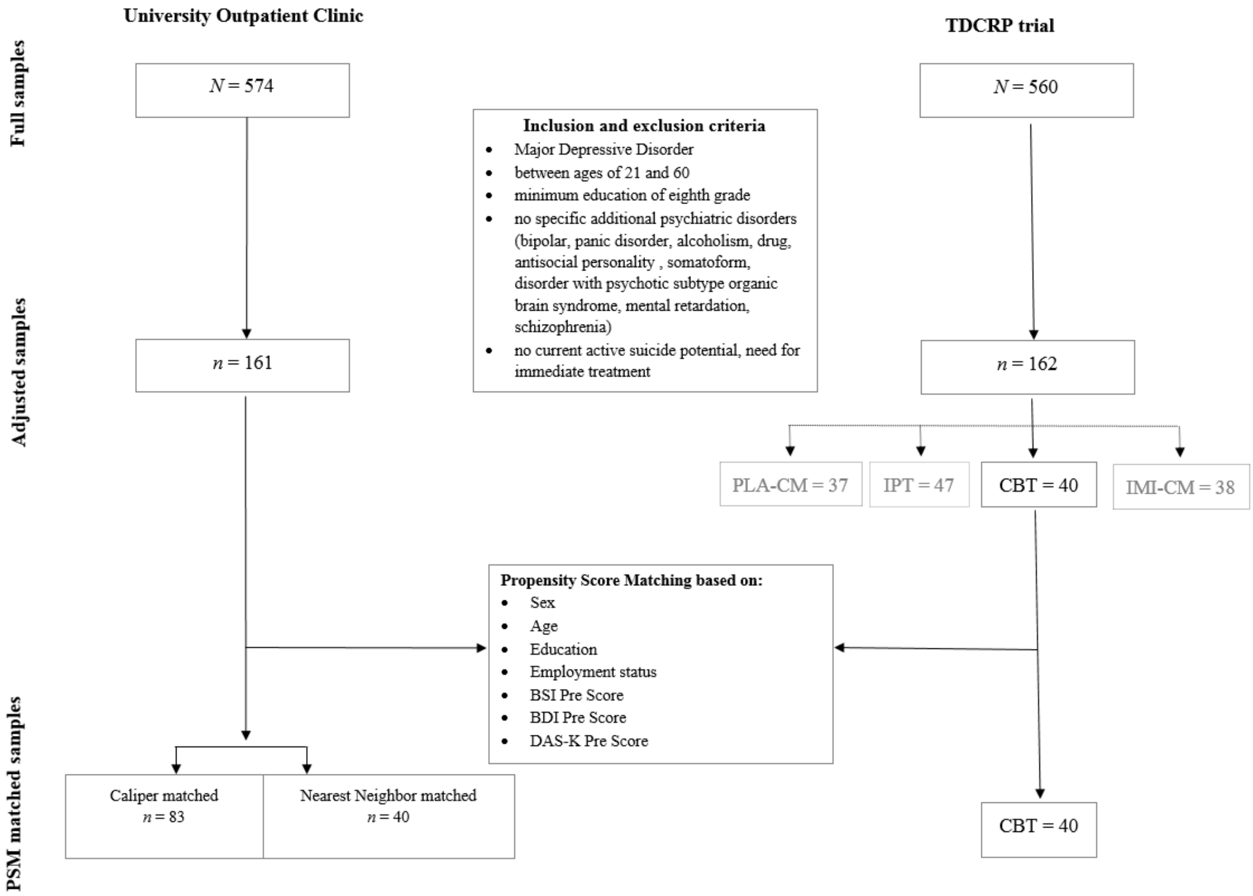
University Outpatient Clinic
TDCRP trial
Full samples
Adjusted samples
PSM matched samples
N = 574
N = 560
Inclusion and exclusion criteria
Major Depressive Disorder
between ages of 21 and 60
minimum education of eighth grade
no specific additional psychiatric disorders (bipolar, panic disorder, alcoholism, drug, antisocial personality , somatoform, disorder with psychotic subtype organic brain syndrome, mental retardation, schizophrenia)
no current active suicide potential, need for immediate treatment
n = 161
n = 162
PLA-CM = 37
IPT = 47
CBT = 40
IMI-CM = 38
Propensity Score Matching based on:
Sex
Age
Education
Employment status
BSI Pre Score
BDI Pre Score
DAS-K Pre Score
Caliper matched
n = 83
Nearest Neighbor matched
n = 40
CBT = 40

Figure 1. Flow chart of the full, adjusted and propensity score matched (PSM) samples of the University Outpatient Clinic and the TDCRP trial. TDCRP = Treatment of Depression Collaborative Research Program; CBT = Cognitive Behavioral Therapy; PLA-CM = Placebo plus clinical management; IMI-CM = Imipramine plus clinical management; BSI = Brief Symptom Inventory; BDI = Beck Depression Inventory; DAS-K = Dysfunctional Attitude Scale

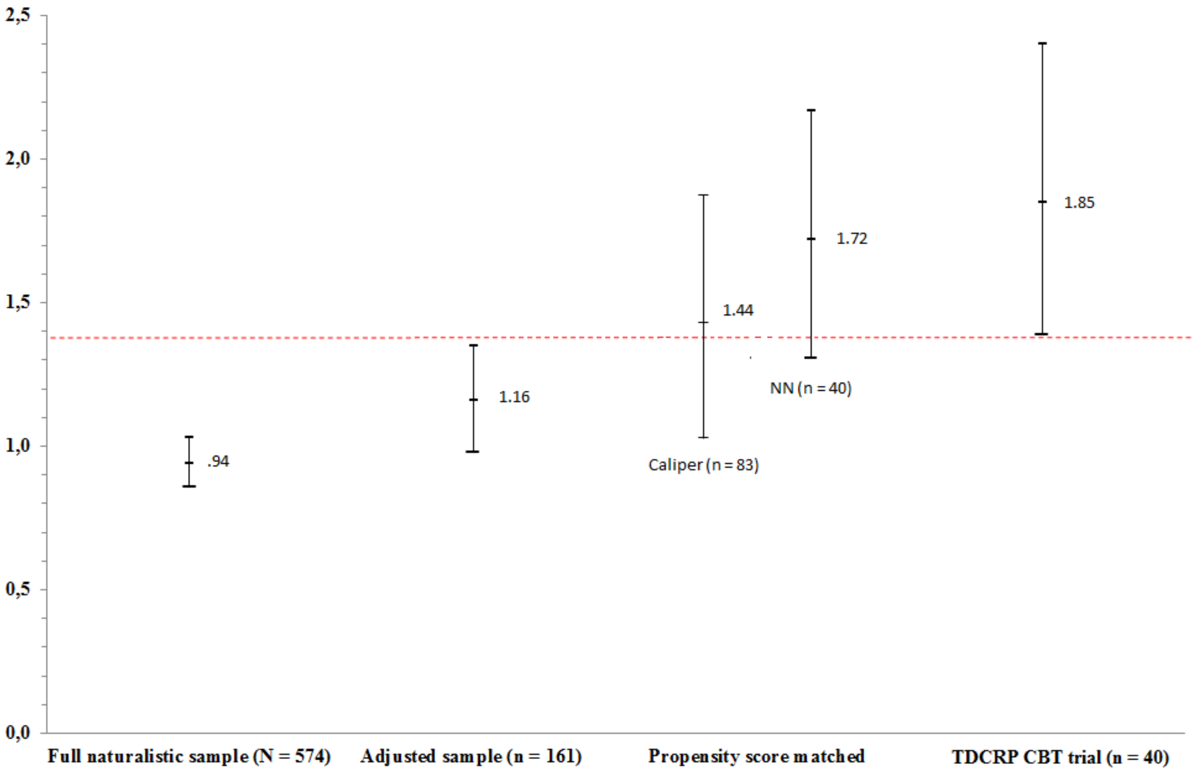

Figure 2. Effect size comparison between the RCT and the naturalistic sample following the matching procedure. Effect sizes for the stepwise adjustment of the full naturalistic sample to the TDCRP CBT trial. Application of exclusion criteria resulted in the adjusted dataset. Matching resulted in two propensity score matched naturalistic subsamples adjusted to the CBT subsample of the TDCRP. Error bars represent 95 % confidence intervals of the effect sizes. The dashed line represents the lower boundary of the 95 % confidence interval of the TDCRP trial.

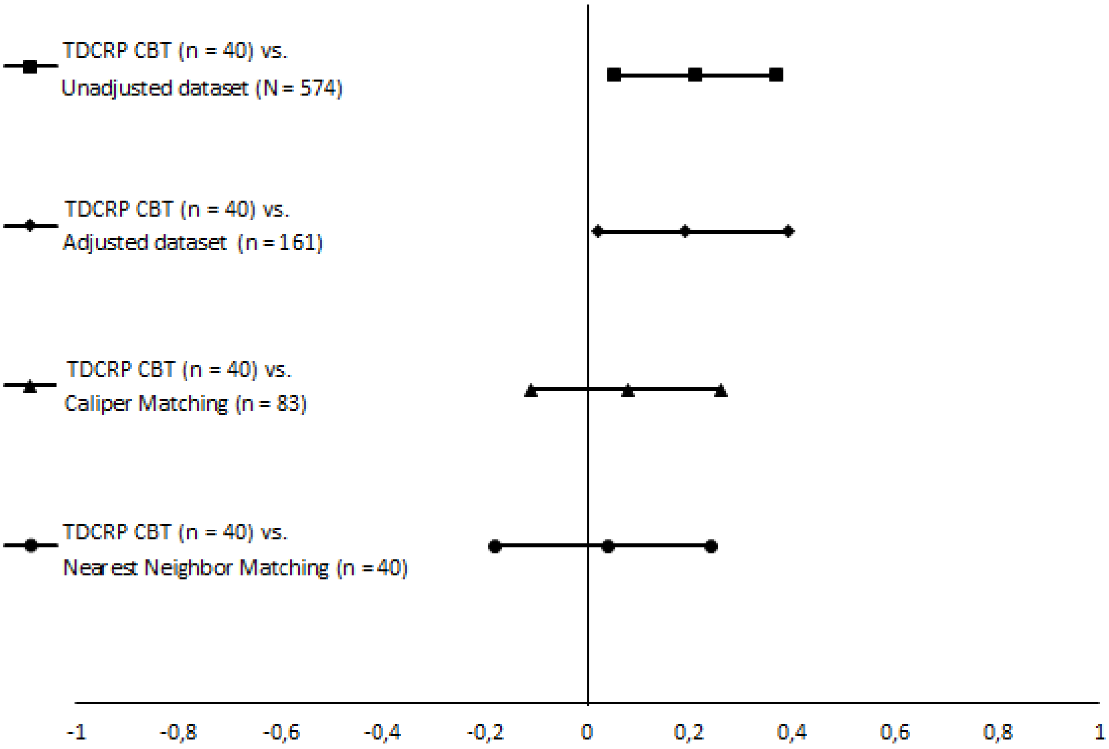

Figure 3. Dot plots of ATT for the naturalistic sample matched to the TDCRP trial. The lines represent 95% confidence intervals for the ATT. ATT = Average treatment effect on the treated

Table 1

Sample characteristics of the naturalistic sample and the TDCRP trial

	Naturalistic sample				TDCRP trial
	Full sample (*N* = 574)	Adjusted sample (*n* = 161)	Caliper matched (*n* = 83)	NN matched (*n* = 40)	CBT trial (*n* = 40)
	Mean (*SD*) or %	Mean (*SD*) or %	Mean (*SD*) or %	Mean (*SD*) or %	Mean (*SD*) or %
BSI_{pre}	1.24 (.68)	1.37 (.63)	1.36 (.58)	1.38 (.55)	1.49 (.63)
BSI_{post}	.64 (.59)	.68 (.58)	.61 (.48)	.55 (.44)	.54 (.55)
BDI_{pre}	23.77(11.2)	26.70 (9.82)	27.05 (8.81)	26.74 (8.57)	27.33 (8.39)
$\text{DAS-K}_{\text{pre}}$	3.48 (1.01)	3.59 (.99)	3.55 (.95)	3.55 (.88)	3.52 (.95)
Sex (% female)	67.42	70.00	74.40	74.80	75.00
Age	36.87 (12.72)	39.07 (11.54)	35.23 (10.90)	33.49 (10.33)	33.83 (9.16)
Education (% more than 12 years)	46.94	40.25	72.70	85.50	80.00
Employment status (% unemployed or unskilled employee)	11.83	12.58	10.87	11.00	5.00
Treatment length (sessions)	33.03 (18.82)	34.73 (18.14)	34.58 (18.56)	34.70 (18.90)	16.38 (1.08)

Note. TDCRP = Treatment of Depression Collaborative Research Program; NN matched= Nearest Neighbor matched; CBT = Cognitive Behavioral Therapy; BSI = Brief Symptom Inventory; BDI = Beck Depression Inventory (sum-score); DAS-K = Dysfunctional Attitude Scale.

Table 2

Standardized mean difference (smd) of the naturalistic sample matched to the TDCRP trial across baseline covariates

	Caliper matched (*n* = 83)		NN matched (*n* = 40)	
Covariate	smd pre-PSM	smd post-PSM	smd pre-PSM	smd post-PSM
BSI_{pre}	.19	.04	.19	.18
BDI_{pre}	.03	.02	.04	.07
$\text{DAS-K}_{\text{pre}}$	.07	.04	.07	.10
Sex	.12	.08	.12	.06
Age	.57	.03	.57	.08
Education	1.79	.05	1.80	.06
Employment status	.16	.08	.17	.11

Note. TDCRP = Treatment of Depression Collaborative Research Program; CBT = Cognitive Behavior Therapy; NN matched= Nearest Neighbor matched; smd = standardized mean difference; PSM = propensity score matching; BSI_{pre} = Brief Symptom Inventory initial patient score; BDI_{pre} = Beck Depression Inventory initial patient score; $\text{DAS-K}_{\text{pre}}$ = Dysfunctional Attitude Scale initial patient score.

Table 3

Recovery rates following the concept of clinical significant change of the naturalistic sample and the TDCRP trial

	Naturalistic Sample				TDCRP trial
	Full sample (*N* = 574)	Adjusted sample (*n* = 161)	Caliper matched (*n* = 83)	NN matched (*n* = 40)	CBT trial (*n* = 40)
Clinical significant change:	% (*n*)	% (*n*)	% (*n*)	% (*n*)	% (*n*)
recovered	42.51 (244)	45.96 (74)	50.00 (41)	57.50 (22)	60.00 (24)
improved	25.26 (145)	24.84 (40)	24.00 (20)	27.50 (11)	22.50 (9)
no change	26.66 (153)	24.84 (40)	23.10 (19)	15.00 (6)	15.00 (6)
deteriorated	5.57 (32)	4.35 (7)	2.90 (3)	.00 (0)	2.50 (1)

Note. TDCRP = Treatment of Depression Collaborative Research Program; CBT = Cognitive Behavioral Therapy; NN matched= Nearest Neighbor matched; Clinical significant change: Cut-off = .61, RCI = .27 (Derogatis & Melisaratos, 1983); Recovered: pre score above cut-off, post score below cut-off and RCI fulfilled; improved: RCI fulfilled; no change: RCI is not fulfilled; deteriorated: RCI is fulfilled but in the negative direction.

5.7 Study I: References

American Psychiatric Association. (2000). *Diagnostic criteria from dsm-iv-tr*: American Psychiatric Pub.

Augurzky, B., & Schmidt, C. M. (2001). The Propensity Score: A Means to An End. *IZA Discussion paper series*. (271).

Barabas, J. (2004). How Deliberation Affects Policy Opinions. *American Political Science Review, 98*(04), 687–701. https://doi.org/10.1017/S0003055404041425

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

Beck, A. T., Steer, R. A., Ball, R., & Ranieri, W. F. (1996). Comparison of Beck Depression Inventories-IA and-II in psychiatric outpatients. *Journal of personality assessment*, *67*(3), 588–597.

Beck, A. T., Steer, R. A., & Carbin, M. G. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical psychology review, 8*(1), 77–100.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & ERBAUGH, J. (1961). An inventory for measuring depression. *Archives of general psychiatry, 4*(6), 561–571.

Boyd, C. L., Epstein, L., & Martin, A. D. (2010). Untangling the causal effects of sex on judging. *American journal of political science, 54*(2), 389–411.

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

Cuijpers, P., Karyotaki, E., Weitz, E., Andersson, G., Hollon, S. D., & van Straten, A. (2014). The effects of psychotherapies for major depression in adults on remission, recovery and improvement: a meta-analysis. *Journal of affective disorders, 159,* 118–126.

Cuijpers, P., van Straten, A., Andersson, G., & van Oppen, P. (2008). *Psychotherapy for depression in adults: a meta-analysis of comparative outcome studies*: American Psychological Association.

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, *94*(448), 1053–1062.

Derogatis, L. R. (1977). Administration, scoring, and procedures manual for the SCL-90-R. *Baltimore: Clinical Psychometrics Research*.

Derogatis, L. R., & Melisaratos, N. (1983). The brief symptom inventory: an introductory report. *Psychological medicine*, *13*(03), 595–605.

Elkin, I., Parloff, M. B., Hadley, S. W., & Autry, J. H. (1985). NIMH treatment of Depression Collaborative Research Program: Background and research plan. *Archives of general psychiatry*, *42*(3), 305–316.

Elkin, I., Shea, M. T., Watkins, J. T., Imber, S. D., Sotsky, S. M., Collins, J. F.,. . . 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.

Finger, M. S., & Rand, K. L. (2003). Addressing validity concerns in clinical psychology research. *Handbook of research methods in clinical psychology*, 13–30.

Floyd, M., Scogin, F., & Chaplin, W. F. (2004). The Dysfunctional Attitudes Scale: factor structure, reliability, and validity with older adults. *Aging & mental health*, *8*(2), 153–160.

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

Garfield, S. L. (1994). Research on client variables in psychotherapy. In A. E. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change* (pp. 190–228). New York: John Wiley & Sons.

Gibbons, C. J., Fournier, J. C., Stirman, S. W., DeRubeis, R. J., Crits-Christoph, P., & Beck, A. T. (2010). The clinical effectiveness of cognitive therapy for depression in an outpatient clinic. *Journal of affective disorders*, *125*(1), 169–176.

Green, L. W., & Glasgow, R. E. (2006). Evaluating the relevance, generalization, and applicability of research: issues in external validation and translation methodology. *Evaluation & the health professions*, *29*(1), 126–153.

Guo, S., & Fraser, M. W. (2014). *Propensity score analysis: Statistical methods and applications*: Sage Publications.

Hamilton, M. A. (1967). Development of a rating scale for primary depressive illness. *British journal of social and clinical psychology*, *6*(4), 278–296.

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.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: nonparametric preprocessing for parametric causal inference. *J Stat Soft*, *42*, 1–28.

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of statistical software*, *45*(7), 1–47.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, *86*(1), 4–29.

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.

Johnsen, T. J., & Friborg, O. (2015). *The effects of cognitive behavioral therapy as an anti-depressive treatment is falling: A meta-analysis*: American Psychological Association.

Keitner, G. I., Posternak, M. A., & Ryan, C. E. (2003). How many subjects with major depressive disorder meet eligibility requirements of an antidepressant efficacy trial? *The Journal of clinical psychiatry*, *64*(9), 1091–1093.

Kessler, R. C., Chiu, W. T., Demler, O., & Walters, E. E. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of general psychiatry*, *62*(6), 617–627.

Klerman, G. L., Weissman, M. M., Rounsaville, B. J., & Chevron, E. S. (1984). *Interpersonal psychotherapy of depression*. New York: Basic Books.

Lambert, M. J. (2013). The efficacy and effectiveness of psychotherapy. In M. J. Lambert (Ed.), *Bergin and Garfield's handbook of psychotherapy and behavior change* (pp. 169–218). New York: John Wiley & Sons.

Lipman, R. S., Covi, L., & Shapiro, A. K. (1979). The Hopkins Symptom Checklist (HSCL): factors derived from the HSCL-90. *Journal of affective disorders*, *1*(1), 9–24.

Lutz, W., Hofmann, S. G., Rubel, J., Boswell, J. F., Shear, M. K., Gorman, J. M.,. . . Barlow, D. H. (2014). Patterns of early change and their relationship to outcome and early treatment termination in patients with panic disorder. *Journal of consulting and clinical psychology*, *82*(2), 287.

Lutz, W., Stulz, N., & Köck, K. (2009). Patterns of early change and their relationship to outcome and follow-up among patients with major depressive disorders. *Journal of affective disorders*, *118*(1), 60–68.

McEvoy, P. M., & Nathan, P. (2007). Effectiveness of cognitive behavior therapy for diagnostically heterogeneous groups: A benchmarking study. *Journal of consulting and clinical psychology*, *75*(2), 344.

Merrill, K. A., Tolbert, V. E., & Wade, W. A. (2003). Effectiveness of cognitive therapy for depression in a community mental health center: A benchmarking study. *Journal of consulting and clinical psychology*, *71*(2), 404.

Minami, T., Wampold, B. E., Serlin, R. C., Hamilton, E. G., Brown, G. S. J., & Kircher, J. C. (2008). Benchmarking the effectiveness of psychotherapy treatment for adult depression in a managed care environment: A preliminary study. *Journal of consulting and clinical psychology*, *76*(1), 116.

Murray, C. J. L., & Lopez, A. D. (1996). Evidence-based health policy—lessons from the Global Burden of Disease Study. *Science*, *274*(5288), 740.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 41–55.

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.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, *2*(3), 169–188.

Schindler, A. C., Hiller, W., & Witthöft, M. (2011). Benchmarking of cognitive-behavioral therapy for depression in efficacy and effectiveness studies—How do exclusion criteria affect treatment outcome? *Psychotherapy Research*, *21*(6), 644–657.

Shadish, W. R. (2013). Propensity score analysis: promise, reality and irrational exuberance. *Journal of Experimental Criminology*, *9*(2), 129–144.

Shadish, W. R., Matt, G. E., Navarro, A. M., Siegle, G., Crits-Christoph, P., Hazelrigg, M. D.,. . . Robinson, L. (1997). *Evidence that therapy works in clincally representative conditions*: American Psychological Association.

Shadish, W. R., Navarro, A. M., Matt, G. E., & Phillips, G. (2000). The effects of psychological therapies under clinically representative conditions: a meta-analysis. *Psychological bulletin*, *126*(4), 512.

Spitzer, R. L., Endicott, J., & Robins, E. (1978). Research diagnostic criteria: rationale and reliability. *Archives of general psychiatry*, *35*(6), 773–782.

Stulz, N., Lutz, W., Leach, C., Lucock, M., & Barkham, M. (2007). Shapes of early change in psychotherapy under routine outpatient conditions. *Journal of consulting and clinical psychology, 75*(6), 864.

Taylor, S., & Asmundson, G. J. G. (2008). Internal and external validity in clinical research. *Handbook of research methods in abnormal and clinical psychology. Sage Publications, Los Angeles*, 23–34.

Warden, D., Rush, A. J., Trivedi, M. H., Fava, M., & Wisniewski, S. R. (2007). The STAR*D project results: A comprehensive review of findings. *Current Psychiatry Reports*, *9*(6), 449–459. https://doi.org/10.1007/s11920-007-0061-3

Weissman, A. N. (Ed.) 1980. *Assessing depressogenic attitudes: A validation study.*

Weisz, J. R., Weiss, B., & Donenberg, G. R. (1992). The lab versus the clinic: Effects of child and adolescent psychotherapy. *American Psychologist*, *47*(12), 1578.

West, S. G., Cham, H., & Thoemmes, F. (2015). Propensity score analysis. In R. L. Cautin & S. O. Lilienfeld (Eds.), *The encyclopedia of clinical psychology* (pp. 1–10). New York: John Wiley & Sons.

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.

Wilcox, R. (2011). *Modern statistics for the social and behavioral sciences: A practical introduction*. Boca Raton, FL: CRC press.

Zetin, M., & Hoepner, C. T. (2007). Relevance of exclusion criteria in antidepressant clinical trials: a replication study. *Journal of clinical psychopharmacology*, *27*(3), 295–301.

6 Study II: Sudden gains in routine care cognitive behavioral therapy for depression: A replication with extensions

Wucherpfennig, F., Rubel, J.A., Hollon, S.D., & Lutz, W. (2016). Sudden gains in routine care cognitive behavioral therapy for depression: A replication with extensions. *Behaviour Research and Therapy, 89*, 24–32.

6.1 Abstract

Background: Over the last decade, a substantial amount of findings have been reported on the association between sudden gains (large symptom improvements in a between-session interval) and treatment outcome. Accurate replications of previous findings are needed to tackle inconsistencies and to shed light on the clinical implications of sudden gains. This study investigates whether similar effects of sudden gains can be expected under routine care conditions, when the patients are comparable to those examined in the original study by Tang and DeRubeis (1999).

Method: Using propensity score matching (PSM), 462 patients treated with cognitive behavioral therapy (CBT) under routine conditions were matched stepwise to patients of the original study on sudden gains, a randomized controlled CBT trial (RCT).

Results: After the application of PSM, the effects of sudden gains on treatment outcome were similar to those found by Tang and DeRubeis (1999). The closer the match between the RCT and the naturalistic sample, the more similar the association between sudden gains and treatment outcome.

Conclusion: Sudden gains seem to have a significant impact on recovery rates, even in treatments under routine care. Results suggest that one important aspect of replication success is to control for confounding baseline covariates.

Keywords: sudden gains; replication; routine care; propensity score matching

6.2 Introduction

Recently, the Open Science Collaboration (2015) conducted replications of 100 studies published in psychological journals and revealed a mean effect size of only half the magnitude of the original effects.

This substantial decline emphasizes the need to acknowledge a degree of uncertainty to what we believe we already know. Accordingly, concerns have been raised that publishing and analytic strategy are likely to be biased toward false positive findings (Ioannidis, 2005; Simmons et al., 2011). Collaborative research and accurate replications are needed to verify previous findings and to overcome such bias. Reproducibility is, however, not well promoted in the scientific community and novelty is often prioritized over replication (Ioannidis, 2014; Schmidt, 2009).

Following these considerations, we want to address the reproducibility of a framework known as *sudden gains*. This framework was developed by Tang and DeRubeis (1999) and can be utilized for a fine-grained analysis of individual change patterns. *Sudden gains* are defined as large between-session symptom improvements. Three criteria must be fulfilled to consider a rapid symptom shift a *sudden gain*: The improvement from one session to the next must be meaningful (a) in absolute terms, (b) in relation to symptom severity before the gain, and (c) relative to symptom fluctuations observed for that patient.

In recent years, a substantial amount of findings have been reported on sudden gains in a variety of treatments and psychopathologies. Initially, sudden gains were investigated in cognitive behavioral therapy for depression (Hardy et al., 2005; Lutz et al., 2012; Tang & DeRubeis, 1999; Tang, DeRubeis, Beberman, & Pham, 2005), subsequently in other treatments for depression such as interpersonal psychotherapy (Kelly, Cyranowski et al., 2007; Lemmens, DeRubeis, Arntz, Peeters, & Huibers, 2016), family therapy (Gaynor et al., 2003), group therapy (Kelly, Roberts, & Ciesla, 2005) and even pharmacotherapy (Vittengl, Clark, & Jarrett, 2005). Sudden gains have also been found in various treatments for anxiety disorders (Hofmann, Schulz, Meuret, Moscovitch, & Suvak, 2006; Norton, Klenck, & Barrera, 2010), obsessive-compulsive disorders (Aderka, Anholt et al., 2012), posttraumatic stress disorders (Keller, Feeny, & Zoellner, 2014; Kelly, Rizvi, Monson, & Resick, 2009), bulimia nervosa and alcohol abuse (Wilson, 1999). Moreover, the reverse phenomenon of sudden gains, known as sudden losses, has been discussed (Lutz et al., 2012).

Although sudden gains seem to be a widespread phenomenon prevalent in several different interventions, there are inconsistencies regarding the association between sudden gains and ultimate treatment outcome. Tang and DeRubeis (1999) found that patients who experienced sudden gains (39.34 % of the sample) revealed treatment outcomes superior to patients without sudden gains (Hedges'*g* = 0.98). Previous replications point in different directions. Hardy et al. (2005) were able to confirm Tang and DeRubeis' (1999) findings, whereas Stiles et al. (2003) revealed no considerable association between sudden gains and outcome. In their meta-analysis, Aderka, Nickerson, Bøe, and Hofmann (2012) found a mean effect size of sudden gains on outcome of Hedges' *g* = 0.62 (range: 0.03 - 1.19). The mean effect is composed of 19 studies ranging from large effects (Doane et al., 2010; Hardy et al., 2005; Tang & DeRubeis, 1999) to small or no effects (Kelly, Cyranowski et al., 2007; Present et al., 2008; Stiles et al., 1996; Stiles et al., 2003). Further, Aderka, Nickerson et al. (2012) showed that smaller effects of sudden gains can be expected for so-called secondary outcomes, that is, when treatment outcome and sudden gains are assessed with different measures. The mean effect size of sudden gains on secondary outcomes was Hedges' *g* = 0.34 (range: 0.01 – 1.01).

There may be different explanations of these inconsistent findings concerning the association between sudden gains and treatment outcome. Apparently, it is important to apply a procedure for the identification of sudden gains comparable to that of Tang and DeRubeis (1999) in order to investigate the very same construct (Stiles et al., 2003). Moreover, divergent findings may be due to variation in the time points when sudden gains occur. Sudden gains experienced early in treatment tend to yield stronger effects than sudden gains experienced in later treatment sessions (Busch, Kanter, Landes, & Kohlenberg, 2006; Kelly et al., 2005; Stiles et al., 2003).

Results of process outcome research have shown that a significant proportion of variance in outcome is explained by the variance attributable to patient characteristics (Barber, 2007; Delgadillo et al., 2016; DeRubeis et al., 2014). Similarly, there is a substantial variance across patients with regard to how they sustain a sudden gain. Some patients experience long lasting improvements, others only temporary improvements with a marginal effect on treatment outcome (Hardy et al., 2005; Stiles et al., 2003; Tang, Luborsky, & Andrusyna, 2002). Accordingly, we expect that even within the same treatment, the experience of a sudden gain may be more beneficial to some patients than to others. In a recent review, Kessler, van Loo, Wardenaar, Bossarte, Brenner, Ebert et al. (2016) showed that baseline variables such as intake symptom severity, number of comorbid disorders, age, employment status and marital status have been repeatedly found to predict treatment response for depressed patients. Consequently, these baseline variables may be associated with patients' differential ability to sustain a sudden gain and eventually to recover.

The analysis by Tang and DeRubeis (1999) is based on data drawn from two randomized controlled trials (RCT). Their sample is selective, as many patients encountered in clinical practice were excluded (see Figure 1).

Different study designs (RCT and naturalistic studies) may vary in their degree of clinical representativeness (Shadish et al., 2000). Treatments in RCTs are usually carried out by intensively trained therapists using highly structured treatment manuals. Patients have to meet a series of specific inclusion criteria and treatment duration is restricted by standardization. However, in clinical practice, treatments are not subjected to comparable standardizations and patients are less homogeneous with respect to their diagnosis and socio-demographic variables. Findings observed in RCTs are not necessarily representative for treatments under routine care conditions (Castonguay et al., 2013; Shadish et al., 2000). Currently, there is only little knowledge about the association of sudden gains and treatment outcome in clinical practice. The vast majority of findings are based on RCTs. For instance, Hardy et al. (2005) found effects, but their treatment context was subjected to standardizations comparable to RCTs. On the contrary, the treatment context of the study by Stiles et al. (2003) was less standardized, however, they were unable to show a meaningful association between sudden gains and outcome.

This points to the necessity of a further investigation of the generalizability of the original findings under routine care conditions. Ioannidis (2005; 2014) suggests improving practice by means of a culture of replication, which is based on appropriate statistical methods and on utilizing data and protocols from original studies. Following this recommendation, we based our replication on data adjusted by a statistical method known as propensity score matching (PSM). PSM offers a solution to reduce bias by balancing two samples based on a range of pretreatment differences (Rosensbaum & Rubin, 1983).

In a previous study, Lutz and colleagues demonstrated the merits of PSM adjustment for the comparison of treatments under routine care with those in RCTs. Their results suggest that cognitive behavioral therapy (CBT) for depression in clinical practice is equally effective as in RCTs when applied to comparable patients (Lutz, Schiefele, Wucherpfennig, Rubel, & Stulz, 2016). To our knowledge, Tang and DeRubeis' (1999) findings have yet to be replicated based on PSM adjustment.

This study aims at assessing the reproducibility of the original findings under routine care conditions with a high level of clinical representativeness according to the criteria by Shadish et al. (2000). In a first step, we applied the inclusion/exclusion criteria of the original sample to a routine care sample. Subsequently, we implemented PSM to enhance the comparability between samples and to adjust for confounding baseline variables. By doing so, we wanted to see if we could find a similar association between sudden gains and treatment outcome, when our patients are comparable to those examined by Tang and DeRubeis (1999).

6.3 Methods

Setting and patients

The routine care sample comprised a total of 462 patients treated at the University Outpatient Clinic Trier between 2010 and 2014. Applying the same inclusion and exclusion criteria used by Tang and DeRubeis (1999), 227 patients were eligible for this study (see section 2.3 sample selection).

All 227 patients included in our analysis had a primary diagnosis of major depression and received at least 8 sessions of individual treatment, with a mean treatment length of 36.67 sessions (*SD* = 17.32, interquartile range = 24 – 45). Treatment was provided by 89 therapists who took part in a three (full-time) or five year (part-time) postgraduate training program with a cognitive behavioral therapy (CBT) focus. All therapists had received at least one year of training before entering the study and were supervised by licensed CBT clinicians. According to German healthcare requirements, therapists were obligated to provide case formulations at the beginning of treatment. All case formulations were examined by independent surveyors (commissioned by health insurance companies), who endorsed the suggested treatment strategies as lege artis CBT interventions. Therapists were familiar with treatment manuals, though not constrained to follow a strict protocol. Data collection was part of the outpatient clinic's routine outcome monitoring and took place before treatment, at each session and at termination. Diagnoses were based on the Structured Clincial Interview for Axis I DSM-IV Disorders (Spitzer et al., 2002), which was conducted before treatment 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.

Measures

Hopkins Symptom Checklist short form (HSCL-11). The HSCL-11 (Lutz, Tholen, Schürch, & Berking, 2006) is a short version of the Brief Symptom Inventory (Franke, 2000). It is comprised of 11 items capturing self-reported symptomatic distress with a focus on depressive symptoms. The items are based on a four-point Likert scale ranging from 1 (not at all) to 4 (extremely). The mean score of global symptomatic distress assessed by the 11 items at the beginning of each session was used to identify sudden gains. The HSCL-11 correlates highly with the BSI (r = 0.91) and substantially with the BDI-II (r = 0.70). The instrument has a high internal consistency (α = 0.92; Lutz et al., 2006)

Brief Symptom Inventory (BSI). The BSI (Franke, 2000) is a brief form of the Derogatis Symptom Check-List-90 Revised (Derogatis, 1992). It is a self-report instrument based on 53 items that inquires nine subscales with the following dimensions: somatization, obsessive-compulsive, interpersonal sensitivity, depression, anxiety, hostility, phobic anxiety, paranoid ideation and psychoticism. The items are rated on a five-point Likert scale ranging from 0 (not at all) to 4 (extremely). The Global Severity Index (GSI; mean score) assessed before treatment and at termination, was used as the primary outcome measure. The psychometric properties of the BSI are excellent (α = 0.92, r_{tt} = 0.90; Franke, 2000).

Beck Depression Inventory II (BDI-II). The BDI-II is the revised version of the BDI and contains 21 items (Beck et al., 1996). It is a self-report instrument developed to assess depressed mood based on both mental symptoms (e.g. hopelessness, guilt, feelings of being punished) and physical symptoms (e.g. loss of libido or appetite). The items can be answered on a four-point Likert scale ranging from 0 to 3. Higher scores indicate higher symptom severity. The sum score of the BDI-II assessed before treatment and at termination, was used as the secondary outcome measure. That is, a measure not related to the HSCL-11, which was used to identify sudden gains. Moreover we used the BDI-II score before treatment as a covariate in the PSM. This instrument has good psychometric properties ($\alpha = 0.76$ - 0.95, $r_{tt} = 0.90$; Beck et al., 1996).

Sample selection and application of PSM

Details on the sample analyzed by Tang and DeRubeis (1999) have been published elsewhere (Elkin et al., 1989; Hollon et al., 1992), thus we will provide only a brief overview. Data were obtained from two different RCTs: $N = 239$ (Elkin et al., 1989) and $N = 107$ (Hollon et al., 1992), which tested the efficacy of CBT for major depression. In both studies, the treatment lasted up to 20 sessions. Eligible patients had to meet the following inclusion/exclusion criteria: a) presence of a current episode of major depression, b) male and female outpatients aged between 21 and 60, c) at least eight years of education and d) no specific additional psychiatric disorders (bipolar I or II, psychotic disorder, alcoholism or other drug use disorder, antisocial personality, schizophrenia, organic brain syndrome). Tang and DeRubeis (1999) selected $n = 61$ patients from the combined sample of the two RCTs based on the following additional criteria: e) intake BDI score of at least 15 points, f) at least 8 sessions of psychotherapy and g) receiving treatment of CBT.

In a first step, we applied all inclusion/exclusion criteria from the original sample to our routine care sample ($N = 462$). In a second step, we used PSM to match our adjusted sample ($n = 227$) to the RCT sample (Figure 1). This was done to enhance the comparability between the samples by controlling for confounding baseline variables. The ultimate goal of PSM is to generate a strong ignorability (Rosensbaum & Rubin, 1983). The assumption of strong ignorability holds if the treatment assignment is independent of the outcome. Therefore, it is crucial to consider all covariates that have a significant impact on both treatment assignment and outcome (West et al., 2014). We were able to control for the following covariates: BDI pre-score, sex, age, marital status and employment status. The selection of covariates was restricted to their availability in all samples. However, we are confident to control for meaningful baseline variables that potentially confound the comparison of samples. Initial symptom severity, age, marital status and employment status have been repeatedly found to predict outcome in treatments for depression (Kessler et al., 2016). In addition, a recent study by Zimmermann, Rubel, Page, and Lutz (2016) suggests that, among other characteristics, male sex predicts premature treatment termination (drop out).

Logistic regressions were performed to calculate propensity scores with the binary dependent variable RCT sample (1) or routine care sample (0) and the five covariates as independent variables. Patients with the same propensity scores have comparable scores in the observed covariates and therefore the same probability of receiving treatment in the RCT or under routine care.

PSM approaches are based on two consecutive steps, that is the estimation of propensity scores and their application. Initially, we estimated propensity scores by adding interaction and quadratic terms of baseline covariates by means of an iterative process, as described by Dehejia and Wahba (1999), to improve the balance of the PSM.[5] Subsequently, we applied two different matching procedures: caliper-matching and nearest neighbor (NN) matching. Both procedures have advantages and disadvantages.

The NN method matches a patient to its counterpart based on the most similar or, in other words, nearest propensity score. The NN for each patient in the RCT was identified, starting with the highest propensity scores in the sample. Consequently, each match drawn from the RCT and the naturalistic sample was removed from the data set. The process was iterated until every patient in the RCT sample had a match in the naturalistic sample. After the application of NN-matching, we received a routine care sample of n = 61 patients.

[5] As suggested by Waljee et al. (2013), the missForest method was used for missing value imputation. This method creates a random forest model for each variable and predicts missing values based on the rest of the variables in the data set. The imputation technique was implemented by the missForest package in R (Stekhoven and Bühlmann (2012). There was no patient with more than two missing values on baseline variables.

The caliper method matches every patient within a pre-specified range or caliper to its counterpart. The caliper is the maximum tolerated difference between matched individuals. It is defined as the standard deviation of the sample estimated propensity score. We used a caliper size of 1 *SD,* thus a patient in the naturalistic sample was considered a match if its propensity score fell into this caliper. We chose a relatively large caliper to have more matching patients per target patient in the RCT sample. We chose a relatively large caliper to have more matching patients per target patient in the RCT sample. In comparison to the NN method, more routine care patients were included in the analysis to improve the generalizability of our findings, however at the expense of a potentially worse fit between the two samples. After the application of caliper-matching, we received a naturalistic sample of $n = 180$ patients.

We scrutinized the goodness of our propensity score models by calculating standardized mean difference scores (smd). The smd method is recommended (Guo & Fraser, 2014; West et al., 2014) to check the balance of covariates between samples. In our context, smd scores indicate the difference in means of each covariate between the RCT and the naturalistic sample, standardized by the standard deviation of the naturalistic sample. Smd scores were calculated before and after the application of PSM. A covariate with an smd < 0.25 indicates an acceptable match between samples (Rubin, 2001).

In summary, the application of inclusion/exclusion criteria and two different forms of PSM resulted in three samples ($n_{Adjusted} = 227$, $n_{Caliper\text{-}matched} = 180$, $n_{NN\text{-}matched} = 61$) treated by routine care. For each sample, we identified sudden gains and analyzed the effects on treatment outcome.

Identification of sudden gains

For the identification of sudden gains, we used the criteria developed by Tang and DeRubeis (1999) in order to ensure comparability. Modifications as suggested by Stiles et al. (2003) were, however, necessary, as we used a different measure for the identification of sudden gains. Tang and DeRubeis (1999) regarded a change of 7 BDI points as a meaningful improvement. In accordance with the suggestions provided by Stiles et al. (2003), we reframed this criterion and considered a reliable improvement in the HSCL-11, indicated by the reliable change index (RCI), to be meaningful. 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.

Following Tang and DeRubeis (1999), a sudden gain between the pre-gain session (N) and the after-gain session (N+1) occurred if:

a) the gain represented a difference between two subsequent sessions of at least 0.61 scores in the HSCL-11 ($\text{HSCL-11}_{N} - \text{HSCL-11}_{N+1} \geq 0.61$).
b) the gain represented at least 25% of the HSCL-11 score in the pre-gain session ($\text{HSCL-11}_{N} - \text{HSCL-11}_{N+1} \geq 0.25 \text{ x } \text{HSCL-11}_{N}$)
c) the mean score of the two or 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 the following critical t-values (5 % significance level): $t_{(4;97.5\%)} > 2.78$; $t_{(3;97.5\%)} > 3.18$; $t_{(2;97.5\%)} > 4.30$.

Association of sudden gains and treatment outcome

We calculated Cohen’s d pre-post effect sizes separately for sudden gainers and non-gainers. Cohen’s d was assessed by dividing mean pre-post differences by the pre standard deviation. Hedges’ g was used to calculate the differential effect between patients with sudden gains and without sudden gains. Hedges’ g is a robust estimator for between-group differences based on the mean pre-to post-treatment change, corrected for bias due to divergent sample sizes (Hedges & Olkin, 1985). According to suggestions provided by Cohen (1988), effect sizes can be categorized into small (0.2), medium (0.5) and large (0.8). Additionally, at the end of treatment, we assessed the recovery rates of sudden gainers and non-gainers according to the concept of clinical significance (Jacobson & Truax, 1991). Recovery rates for the BSI were calculated based on a cut-off score of 0.61 and RCI score of 0.27 (Derogatis & Melisaratos, 1983). Patients with a pre-score > 0.61, a post-score < 0.61 and an RCI score ≥ 0.27 were termed as recovered. For the secondary outcome measure (BDI-II), we used a cut-off score of 11 and RCI score of 7.83 to define recovery (Beck et al., 1988). Additionally, according to the definition by

Tang and DeRubeis (1999), we examined recovery as a post-treatment score of less than 10 points in the BDI-II.

6.4 Results

Application of PSM

After the application of the inclusion/exclusion criteria, the covariates of the adjusted sample (*n* = 227) were still substantially different from the RCT sample. All baseline covariates other than sex revealed smd scores > 0.25, ranging from 0.2 to 0.58 (see Table 1). After the application of PSM, all baseline covariates were sufficiently well balanced. None of the smd scores exceeded 0.25. In the caliper-matched sample (*n* = 180), the smd scores ranged between 0.24 and 0.12. The NN-matched sample (*n* = 61) revealed the smallest smd scores, ranging between 0.01 and 0.24. Even after PSM, treatment length was, however, still significantly longer in the naturalistic samples than in the RCT (caliper-matched: $t(214.19) = 18.87$, $p <.001$; NN-matched: $t(64.36) = 10.79$, $p <.001$).

There were 76 patients (33.48%) in the adjusted sample, who experienced at least one sudden gain and a total of 99 sudden gains. In the caliper-matched sample, 62 (34.44%) patients experienced at least one sudden gain and in total there were 84 sudden gains. There were 26 (42.62%) patients with at least one sudden gain in the NN-matched sample and a total of 33 sudden gains.

Sample differences

We compared all three naturalistic samples on a number of patient characteristics (Table 2). We found more female patients in the NN-matched sample, than in the caliper-matched and adjusted sample. This difference was, however, not significant ($\chi^2(2, N = 468) = 2.59$, $p = .27$). Patients in the NN-matched sample were younger ($F(1, 466) = 4.19$, $p = .003$), revealed higher BDI-II intake scores ($F(1, 466) = 3.78$, $p = .053$) and showed a clear trend toward a lower rate of unemployment ($\chi^2(2, N = 468) = 4.73$, $p = .09$) than in the other naturalistic samples. All three naturalistic samples showed no significant differences in BSI intake scores ($F(1, 466) = 2.48$, $p = .12$), treatment length ($F(1, 466) = .001$, $p = .972$) or marital status ($\chi^2(2, N = 468) = 0.42$, $p = .81$). Sudden gains tended to occur slightly earlier in the NN-matched sample (Median = 11_{th} session) than in the caliper-matched (Median = 13.50_{th} session) and adjusted samples (Median = 13.00$_{th}$ session). The occurrence of sudden gains standardized by the individual treatment length did not differ significantly between routine care samples (NN-matched: *M* = 0.32, *SD* = 0.25; caliper-matched: *M* = 0.36; *SD* = 0.26; adjusted: *M* = 0.37, *SD* = 0.26; *F*(1,162) = 0.55, *p* = .46).

Association between sudden gains and primary outcome (BSI)

The association between sudden gains and treatment outcome differed substantially between the naturalistic samples (Table 3). In comparison to other sudden gainers, those in the adjusted sample revealed the smallest BSI pre-post effect sizes with *d* = 1.32, 95% CI [0.97,1.67]. They showed a medium differential effect, suggesting that their treatment outcome was superior to other patients in the adjusted sample, who did not experience a sudden gain (Hedges' *g* = 0.51, 95% CI [0.23,0.79]). In the caliper-matched sample, sudden gainers yielded a pre-post effect of *d* = 1.63, 95% CI [1.20,2.01] and a medium differential effect of Hedges' *g* = 0.60, 95% CI [0.28,0.92]. The highest effects of sudden gains on outcome were observed for the NN-matched sample with *d* = 2.06, 95% CI [1.37,2.71] and a large differential effect of Hedges' *g* = 0.79, 95% CI [0.25,1.34]. The 95% confidence interval of the NN-matched sample contains the original effect size of Hedges' *g* = 0.98 (Tang & DeRubeis, 1999) whereas this effect can not be found within the 95% CI of the adjusted and caliper-matched samples.

In the adjusted sample, only 46.05% (35 of 76) of sudden gainers and 41.06% (62 of 151) of non-sudden gainers were classified as recovered. This difference in recovery rates was not significant (χ^2(3, *N* = 227) = 6.25, *p* = .10). In the caliper-matched sample, 53.23% (33 of 62) of patients with sudden gains and 41.53% (49 of 118) of patients without sudden gains recovered. Sudden gainers in the NN-matched sample revealed the highest recovery rate of 69.23% (18 of 26), whereas only 45.71% (16 of 35) of the non-gainers recovered. The recovery rates in the two PSM adjusted samples differed significantly between gainers and non-gainers (caliper-matched: χ^2(3, *N* = 180) = 9.19, *p* = .03; NN-matched: χ^2(3, *N* = 61) = 8.26, *p* = .03).

Association between sudden gains and secondary outcome (BDI-II)

Sudden gainers in the adjusted sample yielded a BDI-II pre-post effect size of *d* = 1.91 (95% CI [1.52,2.29]) and Hedges' *g* = 0.43 (95% CI [0.15,0.72]). After PSM adjustment, the association between sudden gains and treatment outcome increased. The caliper-matched sample revealed effect sizes of sudden gainers of *d* = 2.28 (95% CI [1.81,2.72]) and Hedges' *g* = 0.46 (95% CI [0.15,0.78]). The highest effects among the routine care patients were observed for gainers in the NN-matched sample (*d* = 3,45, 95% CI [2.57,4.30]; Hedges' *g* = 0.57 95% CI [0.03,1.11]).

The BDI-II pre-post effect sizes of sudden gainers in the adjusted and the caliper-matched samples were significantly lower than observed by Tang and DeRubeis (1999, *d* = 3.76, 95% CI [2.80,4.70]), whereas sudden gainers in the NN-matched sample yielded comparable pre-post effect sizes, as indicated by the overlap in confidence intervals[6] (see Figure 2).

6.5 Discussion

We investigated whether similar effects of sudden gains and associated therapeutic processes can be expected under routine care conditions, when patients are comparable to those in the RCT examined by Tang and DeRubeis (1999). We adjusted for the inclusion/exclusion criteria of the RCT and subsequently matched for further baseline covariates. Matching was performed by the application of two different PSM approaches (caliper and NN matching). Each step was indended to enhance comparability and to improve the balance of baseline covariates between the RCT and the naturalistic samples.

In contrast to previous research, we used a short version of a global symptom severity measure for the identification of sudden gains. Routine care patients' sudden gains tended to occur later in therapy (Median-range: 11.00th -13.50th session) than in the RCT (Tang & DeRubeis, 1999; Median: 5th session). After adjustment for the inclusion/exclusion criteria, sudden gainers in the naturalistic sample still

[6] The BDI-II pre-post effect size was not significantly different between sudden gainers from the NN-matched sample and the original study, although a sensitivity analysis indicate that we had enough power to find a substantial between group effect size of r= 0.37 (given sample size of n=50 $\alpha = 0.05$ and $1-\beta = 0.8$)

revealed effect sizes significantly smaller than gainers in the RCT. This result resembles that of Stiles et al. (2003), who were unable to replicate the original findings under routine care conditions.

This might in part be due to the fact that the application of the same inclusion/exclusion criteria did not lead to a sufficient balance of covariates between samples (i.e. the samples were still very different with regard to several intake variables). However, the application of PSM resulted in an acceptable match as indicated by smd scores below 0.25. The NN-matched sample was the closest match to the RCT. This naturalistic sample revealed effects of sudden gains comparable to those found by Tang and DeRubeis (1999). The proportion of patients experiencing a sudden gain in the NN-matched sample (42.60%) was about the same as in the in the RCT (39.34%). The BDI-II pre-post effects of patients experiencing a sudden gain no longer differed significantly between samples. Sudden gainers in the RCT revealed high differential effects, indicating superior treatment outcome in comparison to non-gainers. The differential effect in the NN-matched sample was, however, only medium. This may be a result of longer treatment durations in the naturalistic sample than in the RCT. We found consistently higher treatment effects of non-gainers in the routine care sample than in the RCT. Accordingly, in the naturalistic sample, non-gainers may have a higher chance of accomplishing their individual good enough level of improvement in a later treatment phase. We did not find a significant difference in treatment length between sudden gainers and non-gainers (t(179.16) = -1.81, p = 0.07, Mgainer = 40.11). However, the non-gainers from the naturalistic sample (M = 35.75) received treatments of about twice the length as the non-gainers from the RCT (M = 16.59). Once the good enough level is accomplished, additional benefits due to sudden gains may only have a smaller effect on treatment outcome (Barkham et al., 1996; Stiles, Barkham, Connell, & Mellor-Clark, 2008; Stulz, Lutz, Kopta, Minami, & Saunders, 2013). More important, however, seems that the BDI-II was only a secondary outcome for the routine care patients. Based on the BSI – the primary outcome – sudden gainers in the NN-matched sample showed high differential effects comparable to those in the RCT. Accordingly, the recovery rates based on the BSI were significantly higher among the gainers. This finding is in line with Aderka, Nickerson et al. (2012), who found higher effects of sudden gains on primary outcomes than on secondary outcomes. That is, we can expect higher effects, when sudden gains and treatment outcome are assessed with the same measure or in our context when they are highly related to each other (BSI and HSCL-11 r = 0.91).

The present study shows that the rates of sudden gains as well as their associations with ultimate treatment outcome are comparable to those reported by Tang and DeRubeis (1999), when we control for confounding baseline variables. Thus, the HSCL-11 seems to perform similarly to the BDI when identifying sudden gains. This provides clinical practice and research with additional options due to the brevity of the HSCL-11. Although the HSCL-11 correlates substantially with the BDI (r = 0.7), the instrument can be considered a generic rather than a disease-specific measure. In this study, sudden gains do not solely represent symptom improvements of depression but improvements in more general psychopathology. Accordingly, we found a more pronounced effect of sudden gains on the generic outcome measure (BSI) than on the depression-specific measure (BDI-II). Unlike Tang and DeRubeis (1999), we used different instruments for the assessment of sudden gains and treatment outcome. Our assessment strategy may hamper comparability to the original study, though it has the advantage of reducing circularity between sudden gains and treatment outcome.

The closer the match between the RCT and the naturalistic sample, the more similar the association between sudden gains and treatment outcome. This points to the fact that baseline characteristics are crucial for the comparison of sudden gains across samples. The NN-matched sample revealed higher effects of sudden gains than the rest of the naturalistic sample. It was comprised of patients with a younger age, a lower rate of unemployment and higher intake symptom severity than other routine care patients. In line with Kessler, van Loo, Wardenaar, Bossarte, Brenner, Ebert et al. (2016) our results suggest that baseline characteristics have an impact on a patient’s ability to sustain a sudden gain and eventually to recover. Given that patients from the NN matched sample are characterized by a higher level of functioning, as indicated by their working capabilities, they may have more opportunity to leverage sudden improvements in terms of an upward spiral. Gains may help these patients enjoy activities more, which may increase the probability of these patients to also engage in other activities. Older patients with restricted work functioning may have more difficulties translating sudden symptom improvements into their everyday life. However, it is important to consider these explanations as preliminary. Future research must provide more insight into the therapeutic processes that potentially facilitate the effects of sudden gains on treatment outcome.

Replications are important to tackle the uncertainty of scientific evidence. Consequently, scientific credibility must be based on both the quality of the original study and its replication success. Our study provides supporting evidence that sudden gains have a significant impact on recovery rates, even in treatments under routine care. We were able to show that one important aspect of replication success is to control for confounding baseline covariates. In line with Ioannidis (2014), our study emphasizes the merits of a replication practice based on appropriate statistical methods such as PSM and on utilizing data and protocols from the original study (see Figure 1).

Our results are subjected to certain limitations. The ultimate goal of PSM is to generate a strong ignorability. The assumption holds if the treatment assignment is independent of the outcome. It is unlikely that we were able to accomplish a strong ignorability. Our selection of baseline covariates was limited to their availability in all samples. Our results may be biased due to further covariates not included in the PSM model. The PSM was based on a relatively small sample size, which makes it more difficult to find good matching partners, though we achieved sufficient balance between samples. Tang and DeRubeis (1999) findings are based on an RCT conducted in the US many years ago, whereas our recent data is drawn from an outpatient clinic in Germany. Such differences in culture and actuality of the data could hamper the comparability of our study. Furthermore, some differences between samples remained even after PSM. The treatment length in the matched naturalistic sample was about twice as long as in the RCT. This raises the question whether treatments under routine care are less efficient than in RCTs or whether they provide additional benefits other than symptom relief (Lutz, Jong, & Rubel, 2015). There are additional limitations common under routine care conditions: lack of adherence data, heterogeneity of therapists and only little information on treatment protocol. We are confident that therapists provided lege artis CBT for depression as licensed CBT clinicians supervised them and independent surveyors examined their CBT intervention strategies. Therapists in this study were however not constrained to follow a treatment protocol, which may hamper comparability to CBT commonly examined in RCTs.

In conclusion, our findings suggest that the association between sudden gains and outcome is comparable across different populations and therapeutic settings when based on a fair comparison. We were able to replicate the original findings under routine care conditions by applying PSM to reduce bias between samples.

6.6 Tables and Figures

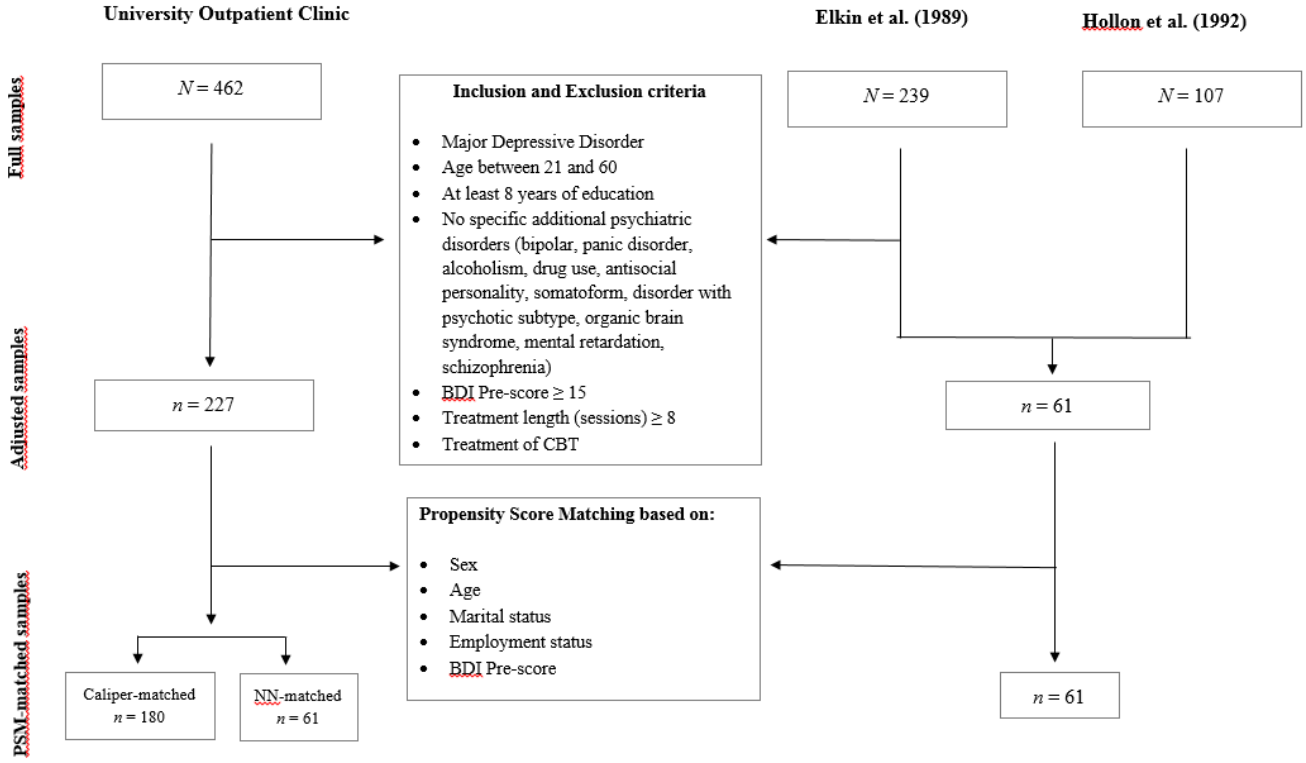
University Outpatient Clinic
Elkin et al. (1989)
Hollon et al. (1992)
Full samples
Adjusted samples
PSM-matched samples
N = 462
N = 239
N = 107
Inclusion and Exclusion criteria
Major Depressive Disorder
Age between 21 and 60
At least 8 years of education
No specific additional psychiatric disorders (bipolar, panic disorder, alcoholism, drug use, antisocial personality, somatoform, disorder with psychotic subtype, organic brain syndrome, mental retardation, schizophrenia)
BDI Pre-score ≥ 15
Treatment length (sessions) ≥ 8
Treatment of CBT
n = 227
n = 61
Propensity Score Matching based on:
Sex
Age
Marital status
Employment status
BDI Pre-score
Caliper-matched n = 180
NN-matched n = 61
n = 61

Figure 1. Flow chart of the full, adjusted and propensity score matched (PSM) samples from the University Outpatient Clinic and the randomized controlled trial (RCT) by Tang and DeRubeis (1999). BDIpre = Beck Depression Inventory initial patient score, CBT= cognitive behavioral therapy.

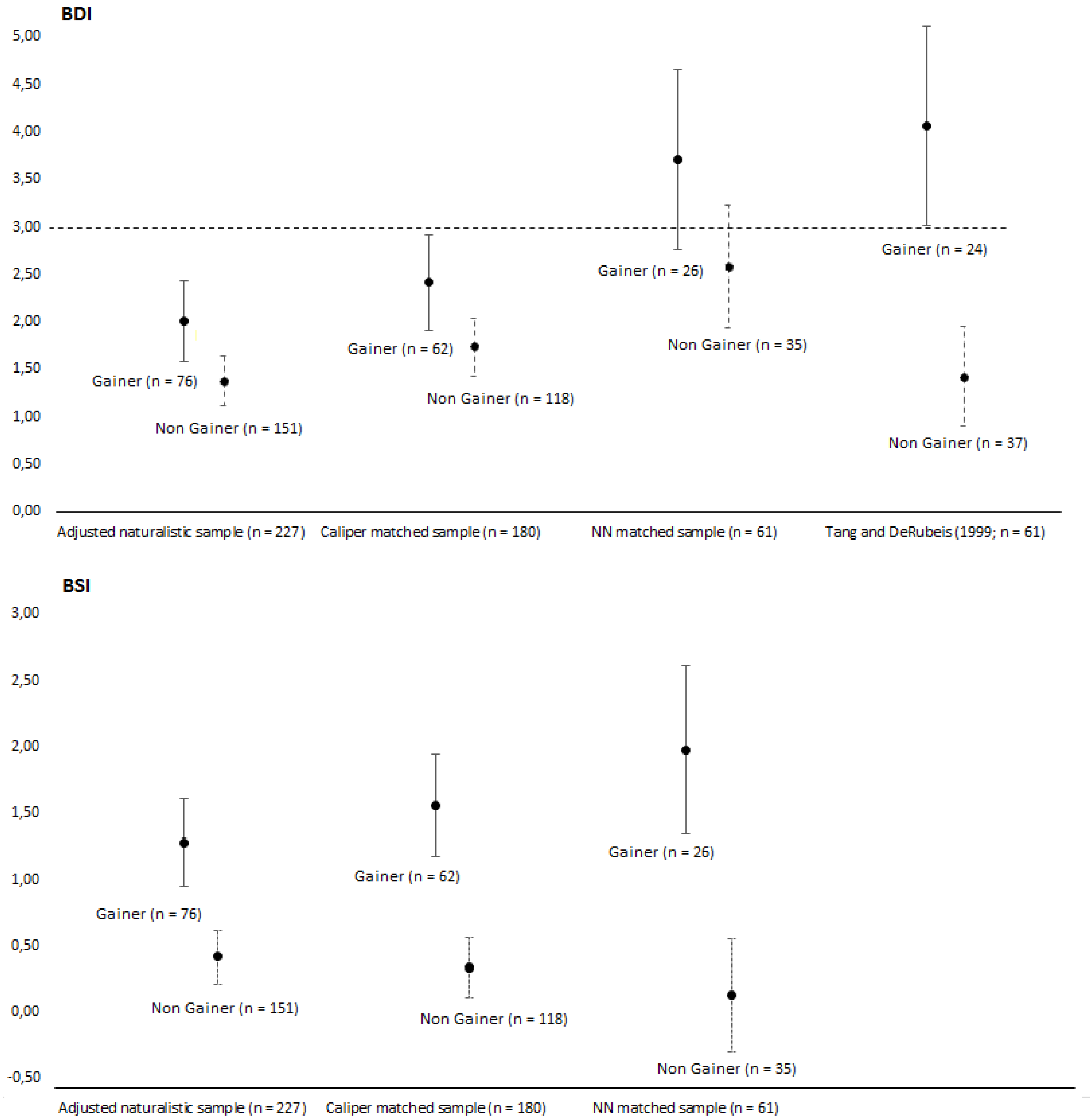

Figure 2. Effect size comparison between the naturalistic samples and the (RCT) by Tang and DeRubeis (1999) based on the BDI and BSI, respectively. NN matched = Nearest Neighbor matched; BDI = Beck Depression Inventory; BSI = Brief Symptom Inventory

Table 1

Standardized mean difference (smd) of the naturalistic samples matched to the RCT by Tang and DeRubeis (1999) across baseline covariates

	Naturalistic samples		
	Adjusted (*n* = 227)	Caliper-matched (*n* = 180)	NN-matched (*n* = 61)
Covariate	smd pre PSM	smd post PSM	smd post PSM
BDIpre	0.42	0.20	0.01
Sex	0.20	0.12	0.04
Age	0.58	0.24	0.03
Marital status	0.35	0.24	0.24
Employment status	0.38	0.20	0.02

Note. NN-matched = Nearest Neighbor matched; smd = standardized mean difference; PSM = propensity score matching; BDIpre = Beck Depression Inventory initial patient score

Table 2

Sample characteristics of the naturalistic sample and the RCT trial by Tang and DeRubeis (1999)

	Naturalistic sample			RCT
	Adjusted sample (*n* = 227)	Caliper-matched sample (*n* = 180)	NN-matched sample (*n* = 61)	Tang and DeRubeis (1999, *n* = 61)
	Mean (SD) or %	Mean (SD) or %	Mean (SD) or %	Mean (SD) or %
BDI PRE	26.49 (10.02)	27.11 (9.21)	29.33 (6.31)	28.26 (6.58)
BDI POST	11.39 (10.39)	11.40 (10.44)	11.51 (9.66)	12.21 (11.65)
BSI PRE	1.35 (0.62)	1.41 (0.60)	1.47 (0.48)	-
BSI POST	0.90 (1.08)	0.94 (1.14)	0.99 (1.28)	-
Sex (% female)	68.75	68.33	78.68	77.04
Age	38.67 (12.59)	36.88 (12.31)	33.41 (10.9)	33.16 (9.46)
Employment status (% unemployed or unskilled employee)	24.23	23.33	11.48	11.19
Marital status (% single)	48.46	52.22	49.18	36.06
Treatment length (sessions)	36.67 (17.32)	36.99 (17.38)	36.52 (18.02)	17.18 (3.44)

Note. RCT = Randomized controlled trial; NN-matched = Nearest Neighbor matched; BDI = Beck Depression Inventory (sum-score); BSI = Brief Symptom Inventory

Table 3

Treatment outcome of sudden gainers and non-gainers in the naturalistic samples and the RCT by Tang and DeRubeis (1999)

	Naturalistic sample			RCT
	Adjusted (*n* = 227)	Caliper-matched (*n* = 180)	NN-matched (*n* = 61)	Tang and DeRubeis (1999, *n* = 61)
	Mean (SD) or %	Mean (SD) or %	Mean (SD) or %	Mean (SD) or %
Gainer				
Patients with sudden gains (%)	76.00 (33.48%)	62.00 (34.44%)	26.00 (42.62 %)	24.00 (39.34%)
BDI PR	29.81 (9.60)	30.14 (8.39)	31.03 (6.17)	27.70 (5.80)
BDI POST	11.52 (10.39)	11.04 (10.09)	9.75 (8.12)	5.90 (5.60)
Recovery rate BDI (BDI Post < 10)	51.31%	51.61%	57.70%	79.00%
BSI PR	1.49 (0.62)	1.56 (0.56)	1.58 (0.50)	-
BSI POST	0.67 (0.56)	0.65 (0.57)	0.55 (0.54)	-
Recovery rate BSI (RCI 0.27 & Cut Off 0.61)	46.05%	53.23%	69.23%	-
Non-Gainer				
Patients without sudden gains (%)	151.00 (66.52%)	118.00 (65.56%)	35.00 (57.40%)	37.00 (60.70%)
BDI PR	24.82 (9.83)	25.52 (9.26)	28.06 (6.18)	27.90 (7.90)
BDI POST	11.32 (10.42)	11.60 (10.66)	12.81 (10.58)	16.80 (13.00)
Recovery rate BDI (BDI Post < 10)	52.98%	51.69%	34.28%	41.00%
BSI PR	1.28 (0.61)	1.33 (0.60)	1.39 (0.44)	-
BSI POST	1.01 (1.25)	1.10 (1.33)	1.33 (1.56)	-
Recovery rate BSI (RCI 0.27 & Cut Off 0.61)	41.06%	41.53%	45.71%	-

Note. RCT = Randomized controlled trail; NN-matched = Nearest Neighbor matched; BDI = Beck Depression Inventory (sum-score); BSI = Brief Symptom Inventory

6.7 Study II: References

Aderka, I. M., Anholt, G. E., van Balkom, Anton 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. doi:10.1159/000329995

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. doi:10.1037/a0026455

Barber, J. P. (2007). Issues and findings in investigating predictors of psychotherapy outcome: Introduction to the special section. *Psychotherapy Research*, *17*(2), 131–136. doi:10.1080/10503300601175545

Barkham, M., Rees, A., Stiles, W. B., Shapiro, D. A., Hardy, G. E., & Reynolds, S. (1996). Dose–effect relations in time-limited psychotherapy for depression. *Journal of consulting and clinical psychology*, *64*(5), 927–935. doi:10.1037/0022-006X.64.5.927

Beck, A. T., Steer, R. A., Ball, R., & Ranieri, W. (1996). Comparison of Beck Depression Inventories -IA and -II in psychiatric outpatients. *Journal of personality assessment*, *67*(3), 588–597. doi:10.1207/s15327752jpa6703_13

Beck, A. T., Steer, R. A., & Carbin, M. G. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review*, *8*(1), 77–100. doi:10.1016/0272-7358(88)90050-5

Busch, A. M., Kanter, J. W., Landes, S. J., & Kohlenberg, R. J. (2006). Sudden gains and outcome: a broader temporal analysis of cognitive therapy for depression. *Behavior therapy*, *37*(1), 61–68. doi:10.1016/j.beth.2005.04.002

Castonguay, L. G., Barkham, M., Lutz, W., & McAleavey, A. A. (2013). Practice oriented research: approaches and application. In M. J. Lambert (Ed.), *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change*. New York, NY: Wiley & Sons.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Dehejia, R. H., & Wahba, S. (1999). Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs. *Journal of the American Statistical Association, 94*(448), 1053–1062. doi:10.1080/01621459.1999.10473858

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. doi:10.1016/j.brat.2016.02.003

Derogatis, L. R. (1992). *The symptome checklist-90-revised*. Minneapolis, MN: NCS.

Derogatis, L. R., & Melisaratos, N. (1983). The Brief Symptom Inventory: An introductory report. *Psychological Medicine*, *13*(03), 595. doi:10.1017/S0033291700048017

DeRubeis, R. J., Gelfand, L. A., German, R. E., Fournier, J. C., & Forand, N. R. (2014). Understanding processes of change: how some patients reveal more than others-and some groups of therapists less-about what matters in psychotherapy. *Psychotherapy research : journal of the Society for Psychotherapy Research*, *24*(3), 419–428. doi:10.1080/10503307.2013.838654

Doane, L. S., Feeny, N. C., & Zoellner, L. A. (2010). A preliminary investigation of sudden gains in exposure therapy for PTSD. *Behaviour research and therapy*, *48*(6), 555–560. doi:10.1016/j.brat.2010.02.002

Elkin, I., Shea, M. T., Watkins, J. T., Imber, S. D., Sotsky, S. M., Collins, J. F.,. . . Parloff, M. B. (1989). National Institute of Mental Health Treatment of Depression Collaborative Research Program. *Archives of General Psychiatry*, *46*(11), 971. doi:10.1001/archpsyc.1989.01810110013002

Franke, G. (2000). *BSI. Brief symptome inventory: Deutsche version. Manual*. Göttingen: Beltz.

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. doi:10.1037/0022-006X.71.2.386

Guo, S., & Fraser, M. W. (2014). *Propenisty Score Analysis: Statisical methods and Apllication*. Thousan Oaks, CA: Sage Publications.

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. doi:10.1037/0022-006X.73.1.59

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analyis*. Orlando, FL: Academic Press.

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. doi:10.1037/0022-006X.74.4.687

Hollon, S. D., DeRubeis, R. J., Evans, M. D., Wiemer, M. J., Garvey, M. J., Grove, W. M., & Tuason, V. B. (1992). Cognitive Therapy and Pharmacotherapy for Depression. *Archives of General Psychiatry*, *49*(10), 774. doi:10.1001/archpsyc.1992.01820100018004

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), e124. doi:10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2014). How to make more published research true. *PLoS medicine*, *11*(10), e1001747. doi:10.1371/journal.pmed.1001747

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. doi:10.1037/0022-006X.59.1.12

Keller, S. M., Feeny, N. C., & Zoellner, L. A. (2014). Depression sudden gains and transient depression spikes during treatment for PTSD. *Journal of consulting and clinical psychology*, *82*(1), 102–111. doi:10.1037/a0035286

Kelly, K. A., Rizvi, S. L., Monson, C. M., & Resick, P. A. (2009). The impact of sudden gains in cognitive behavioral therapy for posttraumatic stress disorder. *Journal of traumatic stress*, *22*(4), 287–293. doi:10.1002/jts.20427

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. doi: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. doi:10.1016/j.brat.2004.06.002

Kessler, R. C., van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Ebert, D. D.,. . . Zaslavsky, A. M. (2016). Using patient self-reports to study heterogeneity of treatment effects in major depressive disorder. *Epidemiology and psychiatric sciences*, 1–15. doi:10.1017/S2045796016000020

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

Lutz, W., Ehrlich, T., Rubel, J., Hallwachs, N., Röttger, M.-A., Jorasz, C.,. . . Tschitsaz-Stucki, A. (2012). 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. doi:10.1080/10503307.2012.693837

Lutz, W., Jong, K. de, & 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. doi:10.1080/10503307.2015.1079661

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. doi:10.1016/j.jad.2015.08.072

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. doi: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. doi:10.1016/j.janxdis.2010.06.012

Present, J., Crits-Christoph, P., Connolly Gibbons, M. B., Hearon, B., Ring-Kurtz, S., Worley, M., & Gallop, R. (2008). Sudden gains in the treatment of generalized anxiety disorder. *Journal of clinical psychology*, *64*(1), 119–126. doi:10.1002/jclp.20435

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

Rubin, D. B. (2001). *Health Services and Outcomes Research Methodology*, *2*(3/4), 169–188. doi:10.1023/A:1020363010465

Schmidt, S. (2009). Shall we really do it again?: The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*(2), 90–100. doi:10.1037/a0015108

Shadish, W. R., Navarro, A. M., Matt, G. E., & Phillips, G. (2000). The effects of psychological therapies under clinically representative conditions: A meta-analysis. *Psychological Bulletin*, *126*(4), 512–529. doi:10.1037/0033-2909.126.4.512

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. doi:10.1177/0956797611417632

Spitzer, R. L., Gibbon, M., Skodol, A. E., Williams, J. B. W., & First, M. B. (2002). *DSM-IV-TR Casebook: A Learning Companion to the Diagnostic and Statistical Manual of Mental Disorders* (Vol. 1). Arlington, VA: American Psychiatric Publishing, Inc.

Stekhoven, D. J., & Bühlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics (Oxford, England)*, *28*(1), 112–118. doi: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. doi:10.1037/0022-006X.76.2.298

Stiles, W. B., Leach, C., Barkham, M., Lucock, M., Iveson, S., Shapiro, D. A.,. . . 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. doi:10.1037/0022-006X.71.1.14

Stiles, W. B., Startup, M., Hardy, G. E., Barkham, M., Rees, A., Shapiro, D. A., & Reynolds, S. (1996). Therapist session intentions in cognitive-behavioral and psychodynamic-interpersonal psychotherapy. *Journal of Counseling Psychology*, *43*(4), 402–414. doi:10.1037//0022-0167.43.4.402

Stulz, N., Lutz, W., Kopta, S. M., Minami, T., & Saunders, S. M. (2013). Dose–effect relationship in routine outpatient psychotherapy: Does treatment duration matter? *Journal of Counseling Psychology*, *60*(4), 593–600. doi:10.1037/a0033589

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. doi: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. doi: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. doi:10.1037//0022-006X.70.2.444

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. doi:10.1037/0022-006X.73.1.173

Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J., Balis, U.,. . . Higgins, P. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ open*, *3*(8). doi: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. doi:10.1037/a0036387

Wilson, G. T. (1999). Rapid Response to Cognitive Behavior Therapy. *Clinical Psychology: Science and Practice*, *6*(3), 289–292. doi:10.1093/clipsy.6.3.289

Zimmermann, D., Rubel, J., Page, A. C., & Lutz, W. (2016). Therapist Effects on and Predictors of Non-Consensual Dropout in Psychotherapy. *Clinical psychology & psychotherapy*. doi:10.1002/cpp.2022

7 Study III: Processes of Change After a Sudden Gain and Relation to Treatment Outcome - Evidence for an Upward Spiral

Wucherpfennig, F., Rubel, J.A., Hofmann, S.G., & Lutz, W. Mechanisms of Change After a Sudden Gain and Relation to Treatment Outcome – Evidence for an Upward Spiral. Submitted to *Journal of Consulting and Clinical Psychology*, December 14th, 2016

7.1 Abstract

Objective: Sudden gains are sudden symptom improvements from one psychotherapy session to the next. This study investigates the processes that may facilitate treatment outcome after a sudden gain occurred.

Method: A sample of 211 depressed patients who underwent cognitive-behavioral therapy was analyzed. Sudden gains were identified using a session-by-session self-report symptom measure. Patient ratings of general change factors (therapeutic alliance; coping skills) in the sessions before and after a sudden gain were investigated as predictors of outcome. Propensity score matching was used to compare sudden gain patients with similar patients who did not experience a sudden gain.

Results: Therapeutic alliance and coping skills increased in the post-gain sessions. There were no comparable levels of change factors among patients without sudden gains. The therapeutic alliance was found to moderate the association between sudden gains and treatment outcome.

Conclusion: Results suggest that sudden gains trigger change factors that facilitate the association between gains and treatment outcome. Patient-therapist dyads should work along sudden gains to consolidate symptom relief.

What is the public health significance of this article?

This study suggests that sudden symptom improvements (sudden gains) in one session are associated with higher levels of therapeutic alliance and decreased cognitive bias in subsequent sessions. Improvements in the therapeutic alliance were found to moderate the association between sudden gains and treatment outcome. Therapists should work along sudden gains to consolidate symptom relief and ultimately achieve good treatment outcomes.

7.2 Introduction

Tang and DeRubeis (1999) were the first to define and empirically test the concept of sudden gains. They demonstrated that several patients in cognitive-behavioral therapy (CBT) for depression improved suddenly between two consecutive sessions rather than gradually. Further, they showed that these rapid improvements predicted better treatment outcome at termination. They defined three criteria for the identification of a *sudden gain*: The between session symptom improvement must a) be large in absolute terms, b) represent a symptom reduction of at least 25 %, and c) be relatively stable (lower symptom severity in the three sessions following a sudden gain than in the three sessions preceding the gain).

There is accumulating evidence to suggest that sudden gains represent a common phenomenon among patients of different diagnosis and treatment modalities 016)(cf. Wucherpfennig, Rubel, Hollon, & Lutz, 2. Initially, sudden gains were investigated in CBT for depression (Hardy et al., 2005; Lutz et al., 2012; Tang & DeRubeis, 1999; Tang et al., 2005), subsequently in interpersonal psychotherapy (Kelly, Cyranowski et al., 2007; Lemmens et al., 2016), family therapy (Gaynor et al., 2003), group therapy (Kelly et al., 2005) and even pharmacotherapy (Vittengl et al., 2005). Sudden gains have been also found in various treatments for social anxiety disorder (Bohn et al., 2013; Hofmann et al., 2006), post-traumatic stress disorder (Aderka, Appelbaum-Namdar, Shafran, & Gilboa-Schechtman, 2011; Doane et al., 2010; Kruger et al., 2014) obsessive-compulsive disorder (Aderka, Anholt et al., 2012) and generalized anxiety disorder (Deschênes & Dugas, 2013).

Results suggest that sudden gains are meaningful changes in a between session interval, which are predictive of significant symptom improvements at the end of the treatment (Busch et al., 2006; Doane et al., 2010; Tang & DeRubeis, 1999; Tang et al., 2002). Accordingly, in their meta-analysis, Aderka, Nickerson et al. (2012) yielded a mean effect of sudden gains on treatment outcome of Hedges' $g_{gain\ vs.no\ gain}$ = 0.62 (range: 0.03 - 1.19), suggesting that patients who experience sudden gains achieve treatment outcomes superior to those of patients without sudden gains. This average effect is an aggregate of 19 individual study effects ranging from large (Doane et al., 2010; Hardy et al., 2005; Tang & DeRubeis, 1999) to small or even null (Kelly, Cyranowski et al., 2007; Present et al., 2008;

Stiles et al., 1996; Stiles et al., 2003). Not all patients, however, benefit from the experience of a sudden gain in the same manner. Not all patients, however, benefit from the experience of a sudden gain in the same manner. Aderka, Nickerson et al. (2012) found stronger effects of sudden gains in CBT interventions than in non-CBT interventions. Furthermore, some researchers suggest that sudden gains experienced early in treatment yield stronger effects than those experienced in later treatment sessions (Busch et al., 2006; Kelly et al., 2005; Stiles et al., 2003). Yet, the reliability of these findings is questionable as there is no consistent definition of "early sudden gains". For instance, Kelly et al. (2005) defined "early" as taking place between the first and 5th session (i.e. the 45.87th percentile of the treatment relative to the overall treatment length of M = 10.9 sessions), whereas Stiles et al. (2003) termed sudden gains before session 16 as "early" (i.e. the 76.55th percentile relative to the overall treatment length of M = 20.9 session). In addition, the meta-analysis by Aderka and colleagues (2012) suggests that the variation in time points at which sudden gains occur does not moderate the association between sudden gains and treatment outcomes.

Results of process-outcome research have shown that a significant proportion of variance in outcome can be explained by patient characteristics (Barber, 2007; Delgadillo et al., 2016; DeRubeis et al., 2014). Accordingly, there is a substantial variance across patients with regard to the degree to which they sustain a sudden gain. Some patients experience long lasting improvements, others only temporary improvements with a marginal effect on treatment outcome (Hardy et al., 2005; Stiles et al., 2003; Tang et al., 2002). It may be more likely for some patients to consolidate a sudden gain and eventually recover than for others (Wucherpfennig et al. 2016). Typically, research on the mechanisms involved in sudden gains has focused on the causes of these gains. Some studies provide evidence for a cognitive mediation of sudden gains. This hypothesis proposes that a sudden gain is triggered by patients' cognitive changes in the session preceding a gain (Norton et al., 2010; Tang & DeRubeis, 1999; Tang et al., 2005). On a related note Adler, Harmeling, and Walder-Biesanz (2013) found increased narrative meaning making processes in the pre-gain sessions. Recently, Abel et al. (2016) demonstrated that patients express more hope and emotional processing prior to a sudden gain and that these processes predict long term treatment outcome. However, other studies did not find support for an increase of cognitive changes or therapeutic processes in the pre-gain sessions (Andrusyna, Luborsky, Pham, &

Tang, 2006; Bohn et al., 2013; Hardy et al., 2005; Hofmann et al., 2006; Kelly, Roberts, & Bottonari, 2007; Vittengl et al., 2005). Only a few studies have focused on the sessions following a sudden gain. This lack of attention is puzzling, as the analysis of subsequent processes may help to understand how sudden gains can be leveraged during treatment. Tang and DeRubeis (1999) assumed that sudden symptom improvements spark an upward spiral by improving the therapeutic alliance and cognitive changes in the following therapy sessions. Furthermore, they hypothesized that this improved alliance quality and decreased cognitive bias sustain symptom relief and eventually lead to recovery. In support of their hypothesis, they found higher levels of therapeutic alliance in post-gain sessions than in pre-gain sessions. Similarly, two subsequent studies provided evidence for an increase of cognitive changes (Bohn et al., 2013) and of the therapeutic alliance (Lutz et al., 2012) in sessions following a sudden gain.

We assume that sudden gains are associated with changes in patients' perceptions of their ability to complete tasks and accomplish goals (Flückiger, Grosse Holtforth, Del Re, & Lutz, 2013). Sudden gains may elicit hope and other positive emotions that are key for the remoralization process and commitment to treatment (Howard, Moras, Brill, Martinovich, & Lutz, 1996). According to the broaden-and-build theory (Fredrickson, 2004), positive emotions broaden an individual's momentary thought-action repertoire. That is, a positive mindset promotes the discovery of novel ideas and actions and the consolidation of social bonds, which in turn strengthen one's ability to deal with adversities 010)(cf. Garland et al., 2. In the context of CBT for depression, we believe a sudden gain can be utilized by therapists to work on patients' generalized thoughts regarding future expectations and actions. Further, a gain is perhaps an opportunity to discriminate between negative cognitions and more constructive thoughts. Patient-therapist dyads may differ in their ability to build on a sudden gain, that is, to harness the potential of a sudden gain for the reinforcement of generalized self-efficacy (Flückiger, Grosse Holtforth, Del Re et al., 2013).

As sudden gains have been found in various treatment modalities, the phenomenon of an upward spiral may best be understood from a trans-theoretical perspective (Grawe, 1997; Stiles, 2001). Dual models of psychotherapy suggest that there are two types of general change factors that account for a substantial amount of variance in outcome across different treatment modalities (Flückiger, Grosse

Holtforth, Znoj, Caspar, & Wampold, 2013; Grawe, 2006; Schulte, 1996). Interpersonal therapeutic processes refer to processes between the patient and the therapist. A particularly important aspect is the therapeutic alliance, which comprises a responsive collaboration based on trust and openness and a mutual agreement on the tasks and goals of treatment (Bordin, 1979). On the other hand, intrapersonal processes refer to the clarification of meaning, mastery and working on patients' target complains. In CBT, clarification is related to processes that help improve patients' knowledge of their own cognitive and emotional schemata (i.e. cognitive changes). Mastery comprises the acquisition of social skills and problem solving skills that help the patient cope with adversities. Processes of mastery and clarification of meaning are linked in modern CBT. It has been repeatedly shown that improvements in patients' perceived therapeutic alliance, mastery and clarification predict symptom reduction at the end of the treatment (Mander et al., 2013; Rubel, Rosenbaum, & Lutz, 2017).

Tang and DeRubeis (1999) hypothesize that sudden gains trigger a long lasting upward spiral in the sessions following a gain. Accordingly, we expect patients with sudden gains to experience significant improvements of the therapeutic alliance and coping skills in the post-gain sessions. Further, we assume that some patients tend to maintain these high levels of general change factors over a meaningful duration of treatment. That is, in the post-gain sessions, we expect to find an increase in the magnitude of therapeutic alliance and coping skills (i.e. high in means), which is consolidated in the subsequent sessions (i.e. low in variance).

The present study aims to investigate therapeutic processes associated with sudden gains that are comparable to the phenomenon of an upward spiral. Patients may differ with regard to how they benefit from the experience of a sudden gain. In order to obtain patients who are comparable to those of the original study, we applied the same inclusion/exclusion criteria as Tang and DeRubeis (1999) to our routine care sample. We investigated processes of change in patients with sudden gains and patients without sudden gains respectively. Propensity score matching (PSM) was applied to ensure that inter-individual differences are attributable to the experience of sudden gains rather than to uncontrolled patient or treatment characteristics.

By doing so, we want to address two hypotheses. First, we expect to find a significant improvement of the patients' perceived therapeutic alliance and coping skills in the sessions following a sudden gain in comparison to sessions from the same patient before the sudden gain. If sudden gains set the stage for such improvements, we may not find comparable improvements among patients who did not experience a sudden gain. Second, we expect that the improvement of general change factors in the sessions following a sudden gain moderates the association between sudden gains and treatment outcome. Taking into account that not all patients benefit from the experience of a sudden gain in the same manner (Aderka, Nickerson et al., 2012), we expect that sudden gains are differentially effective for patients depending on how general change factors can be brought to work in the following sessions.

7.3 Methods

Setting and patients

The routine care sample was comprised of a total of 462 patients treated at the University of Trier's outpatient clinic between 2010 and 2015. According to the inclusion/exclusion criteria of the original study, eligible patients had to meet the following criteria: a) presence of a current episode of major depression, b) male or female outpatients aged between 21 and 60, c) at least eight years of education, d) no specific additional psychiatric disorders (bipolar I or II, psychotic disorder, alcoholism or other drug use disorder, antisocial personality, schizophrenia, organic brain syndrome), e) treatment length of at least 8 sessions, f) individual CBT treatment, and additionally g) no more than 20% of data missing on the session reports (see below).

According to these criteria , 211 patients with a primary diagnosis of major depression and who had received an average of 36.59 sessions (*SD* = 17.19, interquartile range = 24 – 45) of CBT were eligible. All patients participating in this study provided written informed consent. Treatment was provided by 89 therapists who took part in a three- (full-time) or five-year (part-time) postgraduate training program with a CBT focus. All therapists had received at least one year of training before entering the study and were supervised by licensed CBT clinicians. According to German healthcare requirements, therapists were obligated to provide case formulations at the beginning of treatment. All case formulations were examined by independent surveyors (commissioned by health insurance companies), who endorsed the suggested treatment strategies as lege artis CBT interventions. Therapists were familiar with treatment manuals, though not constrained to follow a strict protocol. Data collection was part of the outpatient clinic's routine outcome monitoring and took place before treatment, at each session and at termination. Diagnoses were based on the Structured Clinical Interview for Axis I DSM-IV Disorders (Spitzer et al., 2002), which was conducted before treatment 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.

Measures

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

Beck Depression Inventory II (BDI-II). The BDI-II is the revised version of the BDI and contains 21 items (Beck et al., 1996). It is a self-report instrument developed to assess depressed mood based on both mental (e.g. hopelessness, guilt, feelings of being punished) and physical symptoms (e.g. loss of libido or appetite). Items are based on a four-point Likert scale ranging from 0 to 3. Higher scores indicate higher symptom severity. This instrument has good psychometric properties (α = .76 - .95, r_{tt} = .90; Beck et al., 1996). The sum score of the BDI-II, assessed before treatment and at termination, was used as the outcome measure.

Bern post-session reports (BPSR-Patient). Bern post-session reports (BPSR-Patient). The BPSR (Flückiger, Regli, Zwahlen, Hostettler, & Caspar, 2010) captures therapeutic processes assessed by patients immediately after each session. These processes are based on Grawe's (1997) general change factors as proposed in his unified model of psychotherapy. Items are based on a seven-point Likert scale ranging from -3 (not at all) to 3 (yes, exactly).We used two scales of the BPSR, the therapeutic alliance and coping skills, that have been previously validated by Rubel et al. (2017). The therapeutic alliance scale is based on six items. It represents an alliance concept based on a trustful bond between the therapist and the patient and a mutual agreement on tasks and goals (Bordin, 1979). The coping skills scale is based on six items and assesses patients' experiences of clarification of meaning and problem solving. Clarification describes a process by which the patient perceives improved knowledge of his own cognitive and emotional schemata (i.e. cognitive changes). In CBT these processes are usually induced by means of Socratic questioning. Problem solving skills are intended to help the patient to find a more functional way to cope with adversities. Several interventions in CBT aim to improve patients' coping skills (e.g. social skills training). The items of the therapeutic alliance and coping skills scales are provided in the Appendix. The BPSR has good psychometric properties (Flückiger, Grosse Holtforth, Znoj et al., 2013; Grosse Holtforth et al., 2014; Rubel et al., 2017). The therapeutic alliance and coping skills scales were used to assess the processes of change in the sessions before and after the occurrence of a sudden gain.

Missing Data

Following recommendations by (Waljee et al., 2013), we used the missForest method implemented by the missForest package in R (Stekhoven & Bühlmann, 2012) for missing data imputation. The missForest method appears to be a robust imputation strategy for data missing at random. For each variable, the algorithm fits a random forest model using the rest of the variables in the data set. Eventually, the algorithm uses the model to predict the missing values for a given variable. MissForest provides two error estimates: The randomized root mean square error (NRMSE) for continuous variables and the proportion of falsely classified entries (PFC) for categorical variables. In both cases, good performance of missing value imputation lead to scores close to zero. The patients in our data set revealed no more than 20% missing values in the BPSR. Furthermore, there was no patient with more than two missing values on baseline variables. The error estimates NRMSE: 0.21 and PFC: 0.00 indicated good performance of the missForest implementation.

Identification of sudden gains

For the identification of sudden gains, we used the criteria developed by Tang and DeRubeis (1999) in order to ensure comparability. However, we had to alter the first criterion due to the use of the HSCL-11 rather than the BDI for the identification of sudden gains. In accordance with the suggestions provided by Stiles et al. (2003), we modified the first criterion (improvement of at least 7 BDI points) by using the reliable change index (RCI) to indicate a meaningful between-session improvement. 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. Following Tang and DeRubeis (1999), a sudden gain between the pre-gain session (N) and the post-gain session (N+1) occurred if:

d) the gain represented a difference between two subsequent sessions of at least 0.61 scores in the HSCL-11 ($\text{HSCL-11}_{N} - \text{HSCL-11}_{N+1} \geq 0.61$).

e) the gain represented at least 25% of the HSCL-11 score in the pre-gain session ($\text{HSCL-11}_{N} - \text{HSCL-11}_{N+1} \geq 0.25 \text{ x } \text{HSCL-11}_{N}$)

f) the mean score of the two or 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 the following critical t-values (5 % significance level): $t_{(4;97.5\%)} > 2.78$; $t_{(3;97.5\%)} > 3.18$; $\text{t}_{(2;97.5\%)} > 4.30$.

We used the criterion suggested by Tang and DeRubeis (1999) to identify reversals of sudden gains. A reversal was identified whenever a patient lost 50% of the symptom improvement resulting from the sudden gain. However, according to this criterion, a reversal is not necessarily a stable phenomenon. Reversals may reflect only short-term symptom fluctuations that are not meaningful. Consequently, we used a modified criterion in addition to the original definition of reversals. Whenever a patient with a sudden gain experienced a sudden loss in the following sessions (Lutz et al., 2012), we defined this as a stable reversal. A sudden loss is the reversed phenomenon of a sudden gain. Accordingly, a sudden loss occurs if: $\text{HSCL-11}_{N} - \text{HSCL-11}_{N+1} \geq - 0.61$, the loss represents 25 % of the HSCL score in the pre-loss session and the mean score of the three sessions before (sessions N-2, N-1 and N) and after the sudden loss (sessions N+1, N+2, N+3) are significantly different.

Within subject and between subject control sessions

We analyzed the magnitude and stability of general change factors in the post-gain sessions and compared these with control sessions from the same patient (pre-gain sessions). Furthermore, we compared the group of sudden gain patients with patients who did not experience sudden gains during their treatments. We expected a less pronounced change in therapeutic alliance and coping skills for patients who did not experience a sudden gain. For non-gainers we selected "pseudo gain" sessions. That is, a non-gain session comparable to a sudden gain session with regard to the time point at which it occurred during treatment. In order to ensure comparability, we used propensity score matching (PSM) to match each non-gainer to a patient with a sudden gain. For each non-gainer, we then selected a pseudo gain session that occurred at the same time point as the sudden gain of the matched counterpart. (eg. Session 14 was chosen as a pseudo gain session if the matched counterpart's sudden gain occurred in session 14).

Application of PSM

The merits of PSM in reducing bias by balancing two samples based on a range of pre-treatment differences have been demonstrated in previous studies (Lutz et al., 2016; Rosensbaum & Rubin, 1983; Wucherpfennig et al., 2016). We applied a method known as full matching, which was developed by Rosenbaum (1991) and has been shown to be effective at reducing bias between samples (Hansen, 2004; Stuart & Green, 2008). Like other matching approaches, full matching utilizes propensity scores to adjust for confounding baseline variables. We used logistic regressions to calculate propensity scores with the binary dependent variable sudden gainer (1) or non-gainer (0) and several baseline covariates as independent variables. Following the recommendations provided by West et al. (2014), we implemented all available baseline variables that potentially confound the comparison between gainers and non-gainers. These are the intake symptom severity (pre BDI-II), overall levels of therapeutic alliance and coping skills (BPSR scores averaged over all sessions), treatment length, sex, age, education status and marital status. Based on propensity scores, full matching provides a series of matched sets i.e. subclasses. Sudden gainers with many comparison individuals were grouped with several non-gainers, whereas sudden gainers with few comparison individuals were grouped with only one non-gainer. Hence, each subclass contained a match of one gainer and at least one non-gainer (there were up to four non-gainers per subclass).

The goodness of the match was scrutinized by standardized mean difference scores (smd). Smd scores were calculated for each covariate before and after the application of full matching. The smd is defined as the weighted difference (weights depend on the number of comparison individuals in each subclass) in means between sudden gainers and non-gainers, standardized by the standard deviation of the non-gainer sample. A covariate with an smd < 0.25 indicates an acceptable match between samples (Rubin, 2001).

General change factors associated with sudden gains

We investigated the effects of sudden gains (pseudo gains) on the magnitude and stability of the perceived therapeutic alliance and coping skills.

Magnitude (BPSR): Using paired t-tests we compared the mean score of patients assessment in up to three sessions (N-2, N-1, N) before a sudden gain (pseudo gain) with the mean score of up to three sessions (N+1, N+2, N+3) following a gain (pseudo gain). This procedure was chosen in accordance with the third criterion for the identification of sudden gains. We aggregated the scores to reduce random bias due to fluctuations in the assessment of the BPSR. If a patient experienced more than one sudden gain, we assessed only the first gain to ensure that each observation represented a unique patient.

Stability (BPSR): The stability of the experienced alliance and coping skills was assessed by the coefficient of variation (CV), which is defined as the standard deviation standardized by the mean. Small CV scores indicate a high consistency, that is, low fluctuation over time. The CV is particularly useful for the comparison of time series stemmed from different samples or individuals, because the standard deviation is sensitive to the sample mean (van Geert & van Dijk, 2002; Weber, Shafir, & Blais, 2004). In order to ensure sufficient variation, we compared the CV of up to five sessions preceding a sudden gain (pseudo gain) via a paired t-test with the CV of up to five sessions succeeding a sudden gain (pseudo gain).

A one-way multivariate analysis of variance (MANOVA) was performed to test for mean differences in magnitude and stability between sudden gainers and non-gainers.

General change factors as moderators for the association between sudden gains and treatment outcome

Initially, we tested for interindividual differences between patients with sudden gains (*n* = 69) who either recovered from depression or did not fully recover at treatment termination. Recovery was assessed according to the concept of clinical significance (Jacobson & Truax, 1991) using BDI-II sum scores. Patients with a pretreatment BDI-II > 11, a posttreatment score < 11 and an RCI score ≥ 7.83 were termed as recovered (Beck et al., 1988). Interindividual differences were assessed with respect to baseline variables, number of sudden gains, timing of sudden gains, reversal rates and therapeutic change factors in the sessions after a sudden gain.

In a preliminary analysis, we calculated intercorrelations between the therapeutic alliance and coping skills in critical sessions and depression at pretreatment and posttreatment. Subsequently, we used hierarchical multiple linear regression models to specify the association between sudden gains, general change factors and treatment outcome. The analysis included all patients in the sample (*N* = 211). Critical sessions were either marked by sudden gains or pseudo gains. Consistent with our second hypothesis, we based our analysis on a moderation model 004)(cf. Frazier, Tix, & Barron, 2 . We used intake symptom severity (BDI-II pre score), sudden gain pattern (0,1), general change factors in the critical sessions and the interaction between sudden gains and general change factors to predict posttreatment depression (BDI-II post-score). General change factors in the critical sessions were measured by the change in means (post-gain/ post-pseudo gain sessions $_{(N+1,\ N+2,\ N+3)}$ – pre-gain/pre-pseudo gain sessions $_{(N-2,\ N-1,\ N)}$) of the therapeutic alliance and coping skills, respectively. All continuous predictors were grand-mean centered to facilitate interpretation of the interaction results.

7.4 Results

Sudden gains

Out of 211 patients, we identified 90 sudden gains experienced by 69 (32.70%) patients. There were 52 patients with one gain, 14 patients with two gains, two patients with three gains, and one patient with four gains. Sudden gains occurred throughout therapy, the median was observed in the 8th session (interquartile range: 4th-22th session). According to the individual treatment length (*M* = 41.68), the majority of patients experienced a sudden gain relatively early, i.e. within the first half of treatment. We identified 51 patients (73.9 %) who showed a reversal of the sudden gain. However, there were only 26 patients (37.7 %) with a sudden loss after the gain or, in other words, with a stable reversal. Patients with sudden gains yielded a BDI-II pre-post effect size of *d* = 1.95 (95% CI [1.54,2.35]) and patients without sudden gains of *d* = 1.36 (95% CI [1.11,1.62]). This difference equals a medium differential effect size of Hedges' *g* = 0.50 (95% CI [0.20,0.79]), suggesting that sudden gainers revealed treatment outcomes superior to non-gainers.

Application of PSM

Sudden gainers (*n = 69)* and non-gainers (*n* = 142) differed substantially with respect to several baseline covariates, revealing smd scores > 0.25 (*M* = 0.21, range: 0.01-0.43, see Table 1). Patients with sudden gains revealed significantly higher BDI-II intake scores (*t*(137.63) = 2.90, *p* = .004) and longer treatment durations (*t*(120.12) = 2.63, *p* = .01). There was a trend toward higher overall levels of therapeutic alliance (*t*(185.58) = 1.57, *p* = .12) and coping skills (*t*(134.17) = 1.62, *p* = .11) among sudden gainers in comparison to patients, who did not experience a sudden gain. After the application of full matching, all baseline covariates were sufficiently well balanced. None of the weighted smd scores exceeded 0.25 (*M* = 0.07; range: 0.01-0.16).

For each non-gainer, we selected a pseudo gain session that occurred at the same time point as its matched counterpart's sudden gain. This procedure led to an allocation of pseudo gain sessions that resembles the allocation of sudden gains (pseudo gain session: *M* = 11.38, *SD* = 12.57; sudden gain sessions: *M* = 11.01, *SD* = 13.47; *U*(142,69) = 4962.5, *p* = .88).

General change factors associated with sudden gains

Figure 1 shows the average observed scores in the HSCL-11 and the general change factors for the three sessions before and after the gain. The magnitude of experienced alliance ($t(68) = -3.71$, $p < .001$, $d = 0.27$) and coping skills ($t(68) = -5.44$, $p < .001$, $d = 0.47$) increased significantly after the occurrence of a sudden gain (Table 2). The stability of patients' assessment indicated by the coefficient of variance (CV) was significantly higher in the sessions following a sudden gain than in the preceding sessions (therapeutic alliance: $t(68) = 3.39$, $p = .001$, $d = 0.47$; coping skills: $t(68) = 3.58$, $p < .001$, $d = 0.56$). There was no significant change in the experienced alliance after the occurrence of pseudo gain sessions ($p > .05$). The magnitude ($t(141) = -3.13$, $p = .002$, $d = 0.14$) and stability ($t(141) = 2.63$, $p = .009$, $d = 0.20$) of coping skills slightly increased in the sessions following a pseudo gain. The one-way MANOVA showed that the improvements of general change factors were significantly stronger for patients with sudden gains than for patients without sudden gains (Pillais Trace = 0.14, $F(1,209) = 5.40$, $p < .001$). Post hoc Bonferroni analysis revealed a significantly higher increase in the magnitude of the therapeutic alliance ($p < .01$) and coping skills ($p < .001$) in post-gain sessions than in sessions following pseudo gains

General change factors and treatment outcome

The were 37 sudden gainers who recovered at the end of the treatment and 32 patients with sudden gains who did not show recovery based on the criteria by Jacobson and Truax (1991). Neither the number of sudden gains per patient ($\chi^2(3, N = 69) = 1.62$, $p = .65$) nor the time points sudden gains occurred[7] (recovered: Median = 7, $SD = 14.7$; non-recovered: Median = 8.00, $SD = 11.95$; $t(64.39) = 0.36$, $p = .72$) differed significantly between recovered and non-recovered sudden gainers. There were significantly more stable reversals (sudden loss after a sudden gain) among non-recovered sudden gainers (47%, $n = 15$) than among recovered sudden gainers (29%, n =11; $\chi^2(3, N = 69) = 7.8$, $p = .03$). Recovered sudden gainers showed a stronger increase in the therapeutic alliance after the occurrence of

[7] Time point of sudden gain was standardized by individual treatment length. Additionally, we compared the outcome of patients with "early" sudden gains (before the fifth session) with the outcome of patients with "late" sudden gains. We did not find a significant difference in outcome between these patients (early gainer $n = 15$, $M_{\text{BDIdiff(PR,PO)}} = 19.71$; late gainer $n = 54$, $M_{\text{BDIdiff(PR,PO)}} = 18.13$; $t(26.28) = -6.50$, $p = .63$).

a sudden gain than non-recovered sudden gainers ($t(64.85) = 2.59$, $p = .01$). No other characteristics (BDI-II intake score, treatment length, sex, age, education status and marital status) differed significantly between these patients ($p > .05$).

Intercorrelations between treatment outcome and general change factors are shown in Table 3. The therapeutic alliance in the post-gain sessions and the change in the alliance were significantly correlated with less depression at posttreatment. There was no significant correlation between depression at posttreatment and coping skills in the critical sessions.

Subsequently, we used the entire sample ($N = 211$) to analyze whether the therapeutic alliance and coping skills moderate the association between sudden gains and treatment outcome. In the first step of the hierarchical multiple linear regression, pretreatment depression ($b = 0.41$, $t(209) = 6.65$, $p < .001$) and sudden gain pattern ($b = -2.67$, $t(209) = -2.01$, $p = .04$) were entered as predictors of posttreatment depression. In the second step, general change factors and the interaction between general change factors and sudden gains (general change factors x sudden gains) were added to the model. As depicted in Table 4, the interaction between sudden gains and the therapeutic alliance ($b = -7.38$, $t(209) = -2.77$, $p = .006$, $\Delta R^2 = .03$) uniquely predicted depression at posttreatment. This interaction resulted in different outcome predictions for sudden gainers and non-gainers. Sudden gainers ($n = 69$): Posttreatment Depression = 10.19 + 0.44 * Pretreatment Depression – 7.28 * Therapeutic Alliance $_{\text{(post-gain – pre-gain)}}$. Non-gainers ($n$ = 142): Posttreatment Depression = 11.86 + 0.44 * Pretreatment Depression + 0.10 * Therapeutic Alliance $_{\text{(post-gain – pre-gain).}}$ Accordingly, a posttreatment BDI-II score of 10.19 is to be expected for a patient with a sudden gain, an average pretreatment BDI-II score and average improvements in the therapeutic alliance in the post-gain sessions, whereas an average posttreatment BDI-II score of 11.86 is predicted for patients who did not experience a sudden gain. For patients with a sudden gain who experienced an increase in the therapeutic alliance of one standard deviation above average ($SD = 0.53$) during the post-gain sessions, an average decrease of 3.9 points on the posttreatment BDI-II is expected. The therapeutic alliance in the sessions following a pseudo gain did not predict posttreatment depression for non-gainers.

7.5 Discussion

Our findings provide support for Tang and DeRubeis' (1999) assumption that sudden gains spark an upward spiral in the sessions following a sudden gain. We identified sudden gains with a timing and prevalence (32.70%) comparable to previous studies (Aderka, Nickerson, et al., 2012, 37.4%). As to be expected in a routine care setting, the treatment effects associated with sudden gains (Hedges' *g* = 0.5) were somewhat smaller than in the original study (Tang and DeRubeis, 1999; Hedges' *g* = 0.98). In agreement with our first hypothesis, we found a significant improvement in patients' perceived therapeutic alliance and coping skills in the sessions following sudden gains. There were no comparable improvements among patients who did not experience a sudden gain. Finally, we confirmed our second hypothesis by demonstrating that the therapeutic alliance in the post-gain sessions moderates the association between sudden gains and treatment outcome.

In line with previous research (Bohn et al., 2013; Lutz et al., 2012), results suggested that sudden gains lead to an improvement in the magnitude and stability of patients' perceived therapeutic alliance and coping skills. Typically, the magnitude of general change factors increased after the occurrence of a sudden gain. This increase continued for about two to three sessions. Subsequently, patients with sudden gains tended to reach a plateau with high levels of alliance and coping skills. Once this plateau was reached, there was not much potential for additional improvement. This increase of magnitude and stability during the post-gain sessions suggests that some patients repeatedly experienced (over a meaningful duration of treatment) a high dosage of general change factors. This experience seems to reflect patients' higher opportunity to benefit from the treatment in the long run and eventually to consolidate symptom improvements.

It seems important to address the question of whether the observed processes are actually triggered by sudden gains or whether they are also observable among patients without sudden gains. Non-gainers also showed slightly higher levels of change factors in later therapy sessions. However, these improvements were comparatively smaller. The progress shown by patients who did not experience a sudden gain is more in line with the dose-response relationship of psychotherapy (Hansen et al., 2002; Howard, Kopta, Krause, & Orlinsky, 1986). This assumption of gradual treatment progress does not account for patients with sudden gains. Sudden symptom improvements between two consecutive sessions were associated with rapid improvements of general change factors. Previous results of process-outcome research have shown that a significant proportion of treatment response is explained by patient characteristics (Delgadillo et al., 2016; DeRubeis et al., 2014). In the present study, patients with sudden gains revealed a higher intake symptom severity and longer treatments than patients who did not experience a sudden gain. Our results suggest that the combination of a relatively high impairment at baseline and longer treatment duration may increase the likelihood for a routine care patient to experience a sudden symptom improvement 006)(cf. Hofmann et al., 2. Future research should provide more insight into patient characteristics that contribute to the development of sudden gains. Eventually, we applied full matching to ensure that sudden gainers and non-gainers did not differ with respect to any of the assessed baseline characteristics, treatment length, or time points of critical sessions. Our analytic strategy improved the odds that interindividual differences in patients' perceived therapeutic alliance and coping skills are attributable to sudden gains rather than to uncontrolled patient or treatment characteristics.

As sudden gains have been found in various treatment modalities (Aderka, Nickerson et al., 2012) we used an instrument (BPSR) that captures common factors of psychotherapy rather than specific factors. Our assessment strategy may strengthen the generalizability of our findings and contributes to a replication culture across different clinical settings (Ioannidis, 2014). However, the use of the BPSR hampers comparability to the original findings by Tang and DeRubeis (1999), which were based on ratings of cognitive changes and therapists' application of CBT techniques. However, we are confident that all therapists in this study provided lege artis CBT for depression as licensed CBT clinicians supervised them and independent surveyors examined their CBT intervention strategies. The coping skills scale in the BPSR represents patients' experiences of clarification of meaning and problem solving skills. According to treatment protocol, we assume that these processes are related to cognitive changes, mastery of (social) exposures and the acquisition of problem-solving skills. However, this is a tentative assumption only and the relation between common and specific factors must be addressed empirically. Future research should provide deeper insight into specific CBT intervention strategies in post-gain sessions and how these are related to the improvement of coping skills and the therapeutic alliance.

Given that sudden gains lead to an increase in therapeutic change factors, we assume that sudden symptom improvements trigger patients' hope and contentment with the therapy. Positive emotions have the potential to broaden an individual's momentary thought-action repertoire. That is, a positive mindset facilitates the discovery of novel ideas and actions and the consolidation of social bonds (Fredrickson, 2004). Accordingly, sudden gains may strengthen the therapeutic bond as well as the consensual agreement about treatment goals and tasks. In turn, the engagement in challenging cognitive work may increase and consequently the refinement of coping skills. However, it is important to consider these explanations as preliminary due to the lack of process and adherence data.

To our knowledge, this is the first study that investigated whether therapeutic processes in the sessions following sudden gains contribute to treatment outcome. Sudden gainers who recovered at the end of the treatment showed more pronounced improvements of the therapeutic alliance in the post-gain sessions than non-recovered sudden gainers. Non-recovered gainers seem to struggle to consolidate their sudden symptom improvements as we found significantly more stable reversals among these patients. Interestingly, neither the time points sudden gains occurred nor any of the assessed baseline characteristics differed between recovered and non-recovered gainers, respectively. Therefore, in line with Aderka, Nickerson et al. (2012), we did not find an association between sudden gains occurring early in treatment and superior treatment effects.

Our results suggest that the therapeutic alliance in the post-gain sessions moderates the association between sudden gains and ultimate treatment outcome. Sudden gains were differentially effective for patients with either a high or low increase in perceived therapeutic alliance. General change factors in post-gain sessions may help to harness sudden gains to their full potential. They seem to play a crucial role in the consolidation of sudden symptom improvements and may help patients to recover at treatment termination. Patient-therapist dyads may differ in their ability to work along a sudden gain. Interindividual differences may exist with regard to how patients attribute a sudden gain. A rapid symptom improvement triggers hope and increases commitment to treatment more when attributed to the patient's ability or to the intervention itself, rather than to random symptom fluctuation. This, in turn, may improve contentment with therapy and reinforce collaboration between the patient and therapist. Therapists may differ in their ability to recognize and work with sudden gains. That is, some may use a sudden gain to strengthen a patient's self-efficacy, which could result in an upward spiral. Therapists may explore situations in which improvements were experienced and may acknowledge explicitly that "things are changing". Further, they may explore helpful thoughts and positive self-verbalizations that are linked to sudden gains. Therapists may also differ in their ability to use sudden gains for further cognitive work or challenging exposures. Some may use Socratic methods to carve out the positive experience of a sudden gain and to attenuate patients' generalized negative thoughts regarding future expectations and actions. Finally, therapists may use the patient's positive mindset to discriminate between negative cognitions and more constructive thoughts (Flückiger, Grosse Holtforth, Del Re et al., 2013).

A number of limitations should be noted. First, the directionality and causality of these changes remain unknown. Moreover, the general change factors investigated in the current study are restricted to patients' self-assessment. Thus, we do not know whether alliance and coping skills improved or if these are only proxies for other factors such as patients' enthusiasm about therapy. Future research should provide process data (e.g. video tape ratings) to validate patients' assessments. On a related note, adherence data is needed to specify the effects of sudden gains on treatment interventions. Further insight into how sudden gains can be leveraged would provide clinicians with valuable information. Graphical highlighting and instant feedback of sudden gains as well as information on the individual importance of these sudden shifts could provide therapists with clinically relevant information (Lambert, 2007; Lutz, Jong et al., 2015). As a result, it could become easier for therapists to stabilize sudden symptom improvements.

Despite these limitations, our findings suggest that sudden gains are followed by an increase in the magnitude and stability of perceived alliance levels and cognitive changes. The therapeutic alliance in the post-gain sessions was found to moderate the association between sudden gain and treatment outcome. Consequently, therapists should seek to leverage sudden symptom improvements to improve the alliance and cognitive changes in post-gain sessions in order to maximize the potential benefits of sudden gains.

7.6 Tables and Figures

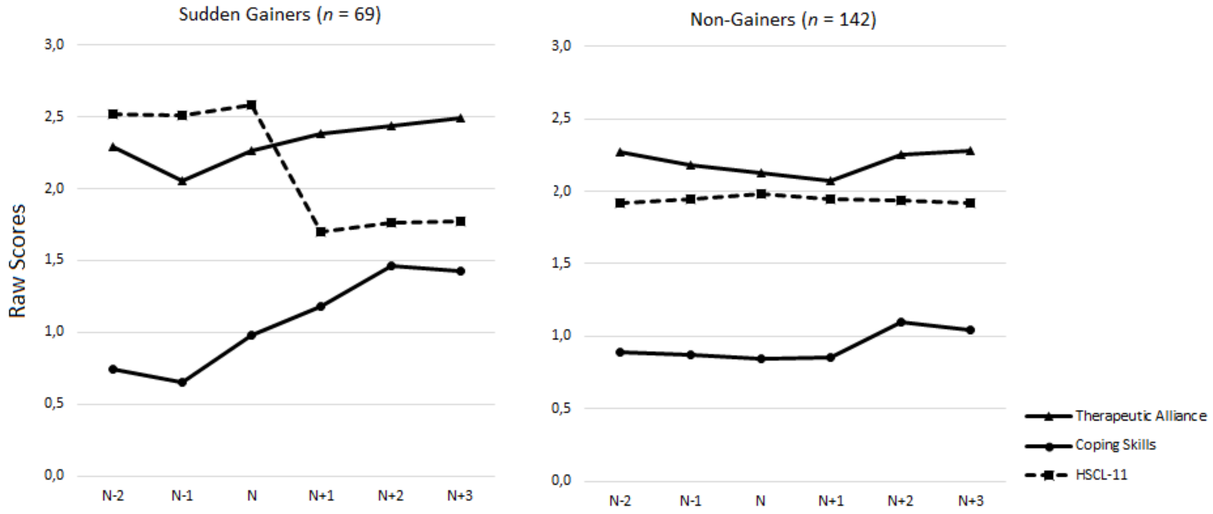

Figure 1. Trajectories of symptom severity (HSCL-11), Therapeutic alliance (BPSR) and Coping skills (BPSR) displayed for the three sessions (N-2,N-1,N) before a sudden gain/pseudo gain and the three sessions (N+1,N+2,N+3) after a sudden gain/pseudo gain.

Table 1

Sample characteristics by sudden gain status and standardized mean difference (smd) before and after propensity score matching (PSM)

	Gainers (*n* = 69)	Non-Gainers (*n* = 142)	smd	
Variable	*M* (*SD*) or %	*M* (*SD*) or %	pre PSM	post PSM
Treatment length (sessions)	41.18 (18.43)	34.36 (16.15)	0.37	0.13
Sex (% female)	73.91	66.66	0.17	0.01
Age	38.13 (13.06)	37.97 (12.42)	0.01	0.05
Education (% more than 12 years)	39.13	42.27	0.05	0.02
Marital status (% single)	52.17	49.64	0.09	0.01
Coping skills (overall mean)	1.31 (0.87)	1.10 (0.89)	0.24	0.14
Therapeutic alliance (overall mean)	2.41 (0.42)	2.29 (0.61)	0.28	0.07
Pretreatment BDI-II	29.24 (9.70)	24.91 (9.97)	0.43	0.16
Posttreatment BDI-II	10.36 (8.46)	11.25 (10.24)	-	-

Note. Mean scores of coping skills and therapeutic alliance were averaged over all sessions, BDI-II = Beck Depression Inventory, second edition.

Table 2

General change factors in the sessions before (PRE) and after (POST) a sudden gain/ pseudo gain

	Full sample *N* = 211					
	Gainers (*n* = 69)			Non-Gainers (*n* = 142)		
	PRE	POST	Difference	PRE	POST	Difference
	M or *CV* (*SD*)	*M* or *CV* (*SD*)	Cohen's *d*	*M* or *CV* (*SD*)	*M* or *CV* (*SD*)	Cohen's *d*
Therapeutic alliance						
Magnitude (*M*)	2.16 (0.79)	2.37 (0.74)	0.27*	2.14 (0.75)	2.19 (0.74)	0.06
Stability (*CV*)	0.16 (0.20)	0.09 (0.10)	0.47*	0.13 (0.14)	0.11 (0.15)	0.14
Coping skills						
Magnitude (*M*)	0.77 (1.26)	1.33 (0.98)	0.50*	0.76 (1.04)	0.91 (1.06)	0.14*
Stability (*CV*)	0.36 (0.29)	0.22 (0.21)	0.56*	0.31 (0.29)	0.26 (0.21)	0.20*

Note. CV = Coefficient of variance; * *p* < .05 based on paired t-test.

Table 3

Intercorrelations between depression severity at pretreatment and posttreatment and therapeutic alliance and coping skills in pre-gain and post-gain sessions and change scores (*n* = 69)

Variables	1	2	3	4	5	6	7
1. Pretreatment BDI-II	-						
2. Posttreatment BDI-II	.25*	-					
3. Therapeutic alliance pre-gain	-.21	-.06	-				
4. Therapeutic alliance post-gain	-.07	-.28*	.81***	-			
5. Coping skills pre-gain	-.20*	-.09	.70***	.53***	-		
6. Coping skills post-gain	-.20	-.13	.60***	.61***	.74***	-	
7. Change in alliance (post-gain - pre-gain)	.25*	-.33**	-.41***	.21	-.34**	-.05	-
8. Change in coping (post-gain - pre-gain)	.07	-.02	-.34**	-.08	-.63**	.06	.44***

Note. BDI-II = Beck Depression Inventory, second edition; * $p < .05$. ** $p < .01$. *** $p < .001$.

Table 4

Hierarchical multiple regression of pretreatment depression, sudden gains, therapeutic alliance and the interaction between sudden gains and alliance (*N* = 211)

	Step 1				Step 2			
Variables	B	*SE*	β	*t*	B	*SE*	β	*t*
Intercept	11.81***	0.74	.09	15.81	11.86***	0.74	.09	16.03
Pretreatment BDI-II	0.41***	0.06	.43	6.65	0.44***	0.06	.45	7.05
Sudden gain	-2.67*	1.33	-.28	-2.01	-1.67	1.34	-.17	-1.24
Therapeutic alliance					0.10	1.28	.01	0.08
Therapeutic alliance x Sudden gain					-7.38**	2.66	-.42	-2.77
R²	.18				.22			
F (change in R²)	13.71***				4.96**			

Note. BDI-II = Beck Depression Inventory, second edition; * *p* < .05. ** *p* < .01. *** *p* < .001.

7.7 Appendix

The therapeutic alliance scale of the BPSR is comprised of six items:

- “Today I felt comfortable with my therapist.”
- “The therapist helps me see where my strengths are.”
- “The therapist and me are getting along well.”
- “I believe that the therapist is truly interested in my wellbeing.”
- “At the moment I fell supported by the therapist in being the way I want to be.”
- “I feel that the therapist has real appreciation for me.”

The coping scale of the BPSR is comprised of six items:

- “Now I feel better up to situations, to which I have not felt up to until now.”
- “Now I’m more confident in my ability to solve problems by myself.”
- “I have the feeling that I better understand myself and my problems.”
- “Today we got closer to the core of my problems.”
- “Today I became aware why I react towards some people in a certain way and not in a different way.”
- “Now I know better what I want.”

7.8 Study III: References

Abel, A., Hayes, A. M., Henley, W., & Kuyken, W. (2016). Sudden gains in cognitive-behavior therapy for treatment-resistant depression: Processes of change. *Journal of Consulting and Clinical Psychology*, *84*(8), 726–737. https://doi.org/10.1037/ccp0000101

Aderka, I. M., Anholt, G. E., van Balkom, Anton 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

Adler, J. M., Harmeling, L. H., & Walder-Biesanz, I. (2013). Narrative meaning making is associated with sudden gains in psychotherapy clients’ mental health under routine clinical conditions. *Journal of Consulting and Clinical Psychology*, *81*(5), 839–845. https://doi.org/10.1037/a0033774

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

Barber, J. P. (2007). Issues and findings in investigating predictors of psychotherapy outcome: Introduction to the special section. *Psychotherapy Research*, *17*(2), 131–136. https://doi.org/10.1080/10503300601175545

Beck, A. T., Steer, R. A., Ball, R., & Ranieri, W. (1996). Comparison of Beck Depression Inventories -IA and -II in psychiatric outpatients. *Journal of Personality Assessment*, *67*(3), 588–597. https://doi.org/10.1207/s15327752jpa6703_13

Beck, A. T., Steer, R. A., & Carbin, M. G. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review*, *8*(1), 77–100. https://doi.org/10.1016/0272-7358(88)90050-5

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

Bordin, E. S. (1979). The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, Research & Practice*, *16*(3), 252–260. https://doi.org/10.1037/h0085885

Busch, A. M., Kanter, J. W., Landes, S. J., & Kohlenberg, R. J. (2006). Sudden gains and outcome: a broader temporal analysis of cognitive therapy for depression. *Behavior Therapy*, *37*(1), 61–68. https://doi.org/10.1016/j.beth.2005.04.002

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

DeRubeis, R. J., Gelfand, L. A., German, R. E., Fournier, J. C., & Forand, N. R. (2014). Understanding processes of change: how some patients reveal more than others-and some groups of therapists less-about what matters in psychotherapy. *Psychotherapy Research*, *24*(3), 419–428. https://doi.org/10.1080/10503307.2013.838654

Deschênes, S. S., & Dugas, M. J. (2013). Sudden Gains in the Cognitive-Behavioral Treatment of Generalized Anxiety Disorder. *Cognitive Therapy and Research*, *37*(4), 805–811. https://doi.org/10.1007/s10608-012-9504-1

Doane, L. S., Feeny, N. C., & Zoellner, L. A. (2010). A preliminary investigation of sudden gains in exposure therapy for PTSD. *Behaviour Research and Therapy*, *48*(6), 555–560. https://doi.org/10.1016/j.brat.2010.02.002

Flückiger, C., Grosse Holtforth, M., Del Re, A. C., & Lutz, W. (2013). Working along sudden gains: responsiveness on small and subtle early changes and exceptions. *Psychotherapy*, *50*(3), 292–297. https://doi.org/10.1037/a0031940

Flückiger, C., Grosse Holtforth, M., Znoj, H. J., Caspar, F., & Wampold, B. E. (2013). Is the relation between early post-session reports and treatment outcome an epiphenomenon of intake distress and early response? A multi-predictor analysis in outpatient psychotherapy. *Psychotherapy Research*, *23*(1), 1–13. https://doi.org/10.1080/10503307.2012.693773

Flückiger, C., Regli, D., Zwahlen, D., Hostettler, S., & Caspar, F. (2010). Der Berner Patienten- und Therapeutenstundenbogen 2000 [The Bern Post-Session Reports for Patients and Therapists 2000]. *Zeitschrift für Klinische Psychologie und Psychotherapie*, *39*(2), 71–79. https://doi.org/10.1026/1616-3443/a000015

Franke, G. (2000). *BSI. Brief symptome inventory: Deutsche version. Manual [German version]*. Göttingen: Beltz.

Frazier, P. A., Tix, A. P., & Barron, K. E. (2004). Testing Moderator and Mediator Effects in Counseling Psychology Research. *Journal of Counseling Psychology, 51*(1), 115–134. https://doi.org/10.1037/0022-0167.51.1.115

Fredrickson, B. L. (2004). The broaden-and-build theory of positive emotions. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *359*(1449), 1367–1378. https://doi.org/10.1098/rstb.2004.1512

Garland, E. L., Fredrickson, B., Kring, A. M., Johnson, D. P., Meyer, P. S., & Penn, D. L. (2010). Upward spirals of positive emotions counter downward spirals of negativity: insights from the broaden-and-build theory and affective neuroscience on the treatment of emotion dysfunctions and deficits in psychopathology. *Clinical Psychology Review*, *30*(7), 849–864. https://doi.org/10.1016/j.cpr.2010.03.002

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

Grawe, K. (2006). *Neuropsychotherapie [Neuropsychotherapy]*. Cambridge, MA: Hogrefe.

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

Grosse Holtforth, M., Altenstein, D., Krieger, T., Flückiger, C., Wright, A. G. C., & Caspar, F. (2014). Interpersonal differentiation within depression diagnosis: relating interpersonal subgroups to symptom load and the quality of the early therapeutic alliance. *Psychotherapy research : journal of the Society for Psychotherapy Research*, *24*(4), 429–441. https://doi.org/10.1080/10503307.2013.829253

Hansen, B. B. (2004). Full Matching in an Observational Study of Coaching for the SAT. *Journal of the American Statistical Association*, *99*(467), 609–618. https://doi.org/10.1198/016214504000000647

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

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

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

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

Ioannidis, J. P. A. (2014). How to make more published research true. *PLoS Medicine*, *11*(10), e1001747. https://doi.org/10.1371/journal.pmed.1001747

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

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., & Bottonari, K. A. (2007). Non-treatment-related sudden gains in depression: the role of self-evaluation. *Behaviour Research and Therapy*, *45*(4), 737–747. https://doi.org/10.1016/j.brat.2006.06.008

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

Kruger, 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

Lambert, M. (2007). Presidential address: What we have learned from a decade of research aimed at improving psychotherapy outcome in routine care. *Psychotherapy Research*, *17*(1), 1–14. https://doi.org/10.1080/10503300601032506

Lemmens, L. H., DeRubeis, R. J., Arntz, A., Peeters, F. P., & Huibers, M. J. (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., Ehrlich, T., Rubel, J., Hallwachs, N., Röttger, M.-A., Jorasz, C.,. . . Tschitsaz-Stucki, A. (2012). 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., Jong, K. de, & 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., 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., 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

Mander, J. V., Wittorf, A., Schlarb, A., Hautzinger, M., Zipfel, S., & Sammet, I. (2013). Change mechanisms in psychotherapy: multiperspective assessment and relation to outcome. *Psychotherapy research : journal of the Society for Psychotherapy Research*, *23*(1), 105–116. https://doi.org/10.1080/10503307.2012.744111

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

Present, J., Crits-Christoph, P., Connolly Gibbons, M. B., Hearon, B., Ring-Kurtz, S., Worley, M., & Gallop, R. (2008). Sudden gains in the treatment of generalized anxiety disorder. *Journal of Clinical Psychology*, *64*(1), 119–126. https://doi.org/10.1002/jclp.20435

Rosenbaum, P. (1991). A Characterization of Optimal Designs for Observational Studies. *Journal of the Royal Statistical Society*. (3), 597–610.

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

Rubel, J. A., Rosenbaum, D., & Lutz, W. (2017). Patients' in-session experiences and symptom change: Session-to-session effects on a within- and between-patient level. *Behaviour Research and Therapy*, *90*, 58–66. https://doi.org/10.1016/j.brat.2016.12.007

Rubin, D. B. (2001). *Health Services and Outcomes Research Methodology*, *2*(3/4), 169–188. https://doi.org/10.1023/A:1020363010465

Schulte, D. (1996). *Therapieplanung [Panning of Therapy]*. Göttingen: Hogrefe.

Spitzer, R. L., Gibbon, M., Skodol, A. E., Williams, J. B. W., & First, M. B. (2002). *DSM-IV-TR Casebook: A Learning Companion to the Diagnostic and Statistical Manual of Mental Disorders* (Vol. 1). Arlington, VA: American Psychiatric Publishing, Inc.

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. (2001). Assimilation of problematic experiences. *Psychotherapy: Theory, Research, Practice, Training*, *38*(4), 462–465. https://doi.org/10.1037/0033-3204.38.4.462

Stiles, W. B., Leach, C., Barkham, M., Lucock, M., Iveson, S., Shapiro, D. A.,. . . 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

Stiles, W. B., Startup, M., Hardy, G. E., Barkham, M., Rees, A., Shapiro, D. A., & Reynolds, S. (1996). Therapist session intentions in cognitive-behavioral and psychodynamic-interpersonal psychotherapy. *Journal of Counseling Psychology*, *43*(4), 402–414. https://doi.org/10.1037//0022-0167.43.4.402

Stuart, E. A., & Green, K. M. (2008). Using full matching to estimate causal effects in nonexperimental studies: examining the relationship between adolescent marijuana use and adult outcomes. *Developmental Psychology*, *44*(2), 395–406. https://doi.org/10.1037/0012-1649.44.2.395

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

van Geert, P., & van Dijk, M. (2002). Focus on variability: New tools to study intra-individual variability in developmental data. *Infant Behavior and Development*, *25*(4), 340–374. https://doi.org/10.1016/S0163-6383(02)00140-6

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.,. . . 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

Weber, E. U., Shafir, S., & Blais, A.-R. (2004). Predicting risk sensitivity in humans and lower animals: risk as variance or coefficient of variation. *Psychological Review*, *111*(2), 430–445. https://doi.org/10.1037/0033-295X.111.2.430

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

Wucherpfennig, F., Rubel, J. A., Hollon, S. D., & Lutz, W. (2016). 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

8 General Discussion

The three studies summarized in this dissertation may help to shed light on the transferability of findings from RCTs to routine care treatment settings. The first study compares the efficacy with the effectiveness of psychotherapy. The second and third studies expand this comparison to process-outcome research.

Study I examined whether the effects of routine care CBT for depression are similar to the effects in a high-quality RCT, if the naturalistic sample is adjusted for inclusion/exclusion criteria of the RCT and further baseline covariates that may affect treatment outcome. Study II investigated whether similar rates and effects of sudden gains can be expected under routine care conditions when patients are comparable to those in the RCT examined by Tang and DeRubeis (1999). Study III assessed whether sudden gains led to improved levels of therapeutic alliance and coping skills and whether these improvements predict treatment outcome. In the following, some general conclusions and limitations drawn from the three studies will be discussed.

8.1 General Conclusion

All three studies suggest that findings from RCTs can be generalized to routine care treatment settings if samples are adjusted for pretreatment differences. The closer the match between patients from efficacy and effectiveness studies the more similar the treatment effects of CBT for depression. This finding resembles Shadish and colleagues (2000) assumption that effect sizes vary along a continuum of clinical representativeness. Study designs with a close proximity on this continuum, meaning that they are subjected to a comparable selection of eligible patients, reveal comparable treatment effects.

In line with McEvoy and Nathan (2007), we found larger standard deviations (pre- and post-treatment symptom severity) in the unadjusted naturalistic sample than in the RCT. Given the considerable heterogeneity among routine care patients, larger changes are required to achieve effect sizes equivalent to RCTs. Previous results have repeatedly shown that a significant proportion of variance in outcome is explained by the variance attributable to patient characteristics (Barber, 2007; Delgadillo et al., 2016; DeRubeis et al., 2014; Kessler, van Loo, Wardenaar, Bossarte, Brenner, Ebert et al., 2016). The simple application of RCT inclusion/exclusion criteria to naturalistic data seems not to be sufficient for a fair comparison. Before the application of PSM, the majority of baseline covariates differed significantly between samples. In all three studies, baseline covariates were adequately well balanced after the application of PSM, as indicated by smd scores below 0.25. According to Kessler, van Loo, Wardenaar, Bossarte, Brenner, Ebert et al. (2016), baseline variables such as intake symptom severity, number of comorbid disorders, age, education, employment status and marital status have been found to predict treatment response. Consequently, if these covariates are not controlled by adjustment, they may introduce bias to the comparison of treatment effects in RCTs and naturalistic studies. However, some differences between study designs remained even after PSM. In all three studies, the treatments under routine care lasted about twice as longs as the treatments in the RCTs. This raises the question whether treatments under routine care are less efficient than in RCTs or whether they provide additional benefits other than symptom relief (Lutz, Jong et al., 2015). On a related note, we cannot rule out the possibility that treatment quality differed significantly within the naturalistic sample, as we did not control for treatment adherence. That is, a therapist in the PSM adjusted sample may not represent the average routine care therapist. Although not intended, the application of PSM may have selected a set of overachieving therapists who are more experienced and knowledgeable than their colleagues.

This dissertation's findings are in contrast to the results of Johnsen and Friborg (2015). Their meta-analysis suggests that the effects of CBT have declined steadily over the last decades. The comparison from Study I was based on a high quality RCT (Elkin et al., 1989) that was conducted in the US more than 20 years ago and recent data from a routine care university outpatient clinic in Germany. Results suggest that the effects of CBT for depression in clinical practice are equally as effective as in RCTs, when applied to comparable patients. That is, after PSM adjustment we did not find a significant decline in effect sizes over time. Results are in accordance with the meta-analysis by Cristea et al. (2017), which demonstrated that neither the country of origin nor the year of publication appears to be a reliable and independent moderator of the effectiveness of CBT for depression.

Patients in the RCTs tend to reveal a younger age, a lower rate of unemployment and higher levels of education than routine care patients. In Study I, 80% of the patients in the RCT had finished more than 12 years of school education. The level of education was comparatively lower in the routine care sample with only 40.25 % of patients with higher education. After the application of PSM, there were 72.70 % (caliper matching) and 85.5 % (NN matching) of routine care patients with more than 12 years of education. Given that RCT patients are characterized by a higher level of functioning than routine care patients, they may have had more opportunity to benefit from CBT interventions with regard to cognitive work and the acquisition of coping skills. Ultimately, these patients did reveal treatment outcomes superior to the unadjusted naturalistic sample. However, after PSM adjustment, we found routine care patients with comparable levels of functioning and similar treatment effects. This emphasizes that raw effect sizes do not allow conclusions to be drawn about the relative effectiveness of a treatment setting.

Study II and III expanded the comparison between treatments in RCTs and naturalistic studies to process-outcome research. Study II revealed similar rates and effects of sudden gains under routine care conditions, when patients were comparable to those in the RCT examined by Tang and DeRubeis (1999). Sudden gains seem to have a significant impact on recovery rates even in treatments in routine care. Study III provides supporting evidence for Tang and DeRubeis’ (1999) assumption that sudden gains spark an upward spiral by improving the therapeutic alliance and cognitive changes in the following therapy sessions. This improved alliance quality and decreased cognitive bias predicted treatment success at termination. Findings suggest that patient-therapist dyads differ in their ability to leverage a sudden gain. Some patients with sudden gains experienced long lasting improvements, others only temporary improvements. One explanation for these inter-individual differences may be differences in how patients attribute a sudden gain. A rapid symptom improvement triggers hope and increases the therapeutic bond more when attributed to the patient’s ability or to the intervention itself, rather than to random symptom fluctuation.

All three studies presented in this dissertation emphasize that patients differ in how they respond to psychotherapy. Replication is the attempt to recreate the conditions believed sufficient to obtain a previously observed finding. This dissertation provides evidence that one important aspect of recreating these conditions is to balance samples on a range of pretreatment differences such as intake symptom severity, number of comorbid disorders, education and employment status. By doing so, we demonstrated the merits of a replication practice, which is based on appropriate statistical methods such as PSM and on utilizing data and protocols from the original studies.

8.2 General Limitations and Future Research

There are some limitations that must be addressed. The ultimate goal of PSM is to generate a strong ignorability. The assumption holds if there are no unmeasured confounders that influence the association between treatment and outcome. However, it is unlikely that a strong ignorability was achieved through the application of PSM. The selection of covariates was limited to their availability in all samples. Symptomatic distress was exclusively assessed by self-report measures. In the naturalistic sample, additional treatments (e.g. medication) were not controlled for and there was a lack of adherence data. Consequently, our results may be biased due to further covariates not included in the PSM model.

There are indicators that therapists in the routine care sample provided CBT for depression as licensed CBT clinicians supervised them and independent surveyors examined their CBT intervention strategies. Therapists were, however, not constrained to follow a treatment protocol. Routine care CBT may differ substantially from CBT commonly examined in RCTs. Due to the lack of adherence data, we do not know whether treatment quality in the PSM adjusted sample (NN matched, caliper matched) differed significantly from the treatments in the unadjusted naturalistic sample. Accordingly, we do not know whether a therapist in the NN match sample represents the average routine care therapist. This points to the fact that therapist differences were completely neglected in the current studies. Therapists have been repeatedly found to explain about 5 % of the differences in treatment outcome (Baldwin & Imel, 2013; Lutz, Rubel et al., 2015). Some clinical researchers found poorer treatment effects for therapists that depart from treatment protocol than therapists who strictly follow treatment manuals such as the Beck manual for CBT (Luborsky, 1985; Luborsky, McLellan, Diguer, Woody, & Seligman, 1997; Shafran et al., 2009). Future comparisons between different treatment settings should consider this factor. It raises the question whether treatment adherence in naturalistic studies is equally predictive of treatment outcome as in RCTs.

Future research should investigate the transferability of findings from RCTs to routine care settings based on even larger sample sizes with more information available regarding patient, therapist and treatment characteristics. This would allow the improvement of bias reduction by means of a more efficient application of PSM approaches.

8.3 Concluding Remarks

Despite these limitations, the present dissertation demonstrates the merits of PSM adjustment for a sound comparison across different populations and therapeutic settings. When all three studies are taken together, results suggest that routine care CBT for depression is equally effective and associated with comparable individual change patterns as in RCTs, when applied to comparable patients.

Do we know what we think we know? This question minds us to acknowledge some degree of uncertainty to scientific evidence. We were able to reproduce findings under routine care conditions that were previously observed in RCTs. However, we cannot infer that the original findings are necessarily correct or generalizable to different treatment settings. Alternative explanations for the original findings can likewise account for the replications. Nevertheless, accurate replications by independent researchers based on appropriate statistical methods and the utilization of data and protocols from the original studies are key to increase scientific credibility. In other words, replications help to increase certainty about what we think we already know.

9 References

Abel, A., Hayes, A. M., Henley, W., & Kuyken, W. (2016). Sudden gains in cognitive-behavior therapy for treatment-resistant depression: Processes of change. *Journal of consulting and clinical psychology*, *84*(8), 726–737. https://doi.org/10.1037/ccp0000101

Aderka, I. M., Anholt, G. E., van Balkom, Anton 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

Adler, J. M., Harmeling, L. H., & Walder-Biesanz, I. (2013). Narrative meaning making is associated with sudden gains in psychotherapy clients' mental health under routine clinical conditions. *Journal of consulting and clinical psychology*, *81*(5), 839–845. https://doi.org/10.1037/a0033774

American Psychiatric Association. (2000). *Diagnostic criteria from dsm-iv-tr*: American Psychiatric Pub.

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

Augurzky, B., & Schmidt, C. M. (2001). The Propensity Score: A Means to An End. *IZA Discussion paper series*. (271).

Babcock, L., & Loewenstein, G. (1997). Explaining Bargaining Impasse: The Role of Self-Serving Biases. *The Journal of Economic Perspectives*. (11(1)), 109–126.

Baldwin, S. A., & Imel, Z. E. (2013). Therapist effects: Findings and methods. In M. J. Lambert (Ed.), *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change* (pp. 258–297). New York, NY: Wiley & Sons.

Barabas, J. (2004). How Deliberation Affects Policy Opinions. *American Political Science Review*, *98*(04), 687–701. https://doi.org/10.1017/S0003055404041425

Barber, J. P. (2007). Issues and findings in investigating predictors of psychotherapy outcome: Introduction to the special section. *Psychotherapy Research*, *17*(2), 131–136. https://doi.org/10.1080/10503300601175545

Barkham, M., Rees, A., Stiles, W. B., Shapiro, D. A., Hardy, G. E., & Reynolds, S. (1996). Dose–effect relations in time-limited psychotherapy for depression. *Journal of consulting and clinical psychology*, *64*(5), 927–935. https://doi.org/10.1037/0022-006X.64.5.927

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

Beck, A. T., Steer, R. A., Ball, R., & Ranieri, W. F. (1996). Comparison of Beck Depression Inventories-IA and-II in psychiatric outpatients. *Journal of personality assessment*, *67*(3), 588–597.

Beck, A. T., Steer, R. A., & Carbin, M. G. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical psychology review*, *8*(1), 77–100. https://doi.org/10.1016/0272-7358(88)90050-5

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of general psychiatry*, *4*(6), 561–571.

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

Bordin, E. S. (1979). The generalizability of the psychoanalytic concept of the working alliance. *Psychotherapy: Theory, Research & Practice*, *16*(3), 252–260. https://doi.org/10.1037/h0085885

Boyd, C. L., Epstein, L., & Martin, A. D. (2010). Untangling the causal effects of sex on judging. *American journal of political science*, *54*(2), 389–411.

Busch, A. M., Kanter, J. W., Landes, S. J., & Kohlenberg, R. J. (2006). Sudden gains and outcome: a broader temporal analysis of cognitive therapy for depression. *Behavior Therapy*, *37*(1), 61–68. https://doi.org/10.1016/j.beth.2005.04.002

Castonguay, L. G., Barkham, M., Lutz, W., & McAleavey, A. A. (2013). Practice oriented research: approaches and application. In M. J. Lambert (Ed.), *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change.* New York, NY: Wiley & Sons.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1995). The earth is round (p <.05): Rejoinder. *American Psychologist*, *50*(12), 1103. https://doi.org/10.1037/0003-066X.50.12.1103

Cristea, I. A., Stefan, S., Karyotaki, E., David, D., Hollon, S. D., & Cuijpers, P. (2017). The effects of cognitive behavioral therapy are not systematically falling: A revision of Johnsen and Friborg (2015). *Psychological Bulletin*, *143*(3), 326–340. https://doi.org/10.1037/bul0000062

Cuijpers, P., Karyotaki, E., Weitz, E., Andersson, G., Hollon, S. D., & van Straten, A. (2014). The effects of psychotherapies for major depression in adults on remission, recovery and improvement: a meta-analysis. *Journal of affective disorders, 159*, 118–126. https://doi.org/10.1016/j.jad.2014.02.026

Cuijpers, P., van Straten, A., Andersson, G., & van Oppen, P. (2008). Psychotherapy for depression in adults: a meta-analysis of comparative outcome studies. *Journal of consulting and clinical psychology*, *76*(6), 909–922. https://doi.org/10.1037/a0013075

Cuijpers, P., van Straten, A., & Warmerdam, L. (2007). Behavioral activation treatments of depression: a meta-analysis. *Clinical psychology review*, *27*(3), 318–326. https://doi.org/10.1016/j.cpr.2006.11.001

Dawson, E., Gilovich, T., & Regan, D. T. (2002). Motivated Reasoning and Performance on the was on Selection Task. *Personality and Social Psychology Bulletin*, *28*(10), 1379–1387. https://doi.org/10.1177/014616702236869

Dehejia, R. H., & Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, *94*(448), 1053–1062.

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. (1992). *The symptome checklist-90-revised*. Minneapolis, MN: NCS.

Derogatis, L. R. (1977). Administration, scoring, and procedures manual for the SCL-90-R. *Baltimore: Clinical Psychometrics Research*.

Derogatis, L. R., & Melisaratos, N. (1983). The Brief Symptom Inventory: An introductory report. *Psychological Medicine*, *13*(03), 595. https://doi.org/10.1017/S0033291700048017

DeRubeis, R. J., Gelfand, L. A., German, R. E., Fournier, J. C., & Forand, N. R. (2014). Understanding processes of change: how some patients reveal more than others-and some groups of therapists less-about what matters in psychotherapy. *Psychotherapy Research*, *24*(3), 419–428. https://doi.org/10.1080/10503307.2013.838654

Deschênes, S. S., & Dugas, M. J. (2013). Sudden Gains in the Cognitive-Behavioral Treatment of Generalized Anxiety Disorder. *Cognitive Therapy and Research*, *37*(4), 805–811. https://doi.org/10.1007/s10608-012-9504-1

Doane, L. S., Feeny, N. C., & Zoellner, L. A. (2010). A preliminary investigation of sudden gains in exposure therapy for PTSD. *Behaviour Research and Therapy*, *48*(6), 555–560. https://doi.org/10.1016/j.brat.2010.02.002

Easterbrook, P., Gopalan, R., Berlin, J., & Matthews, D. (1991). Publication bias in clinical research. *The Lancet*, *337*(8746), 867–872. https://doi.org/10.1016/0140-6736(91)90201-Y

Elkin, I., Parloff, M. B., Hadley, S. W., & Autry, J. H. (1985). NIMH treatment of Depression Collaborative Research Program: Background and research plan. *Archives of general psychiatry*, *42*(3), 305–316.

Elkin, I., Shea, M. T., Watkins, J. T., Imber, S. D., Sotsky, S. M., Collins, J. F.,. . . 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.

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

Finger, M. S., & Rand, K. L. (2003). Addressing validity concerns in clinical psychology research. *Handbook of research methods in clinical psychology*, 13–30.

Floyd, M., Scogin, F., & Chaplin, W. F. (2004). The Dysfunctional Attitudes Scale: factor structure, reliability, and validity with older adults. *Aging & mental health*, *8*(2), 153–160.

Flückiger, C., Grosse Holtforth, M., Del Re, A. C., & Lutz, W. (2013). Working along sudden gains: responsiveness on small and subtle early changes and exceptions. *Psychotherapy (Chicago, Ill.)*, *50*(3), 292–297. https://doi.org/10.1037/a0031940

Flückiger, C., Grosse Holtforth, M., Znoj, H. J., Caspar, F., & Wampold, B. E. (2013). Is the relation between early post-session reports and treatment outcome an epiphenomenon of intake distress and early response? A multi-predictor analysis in outpatient psychotherapy. *Psychotherapy research : journal of the Society for Psychotherapy Research*, *23*(1), 1–13. https://doi.org/10.1080/10503307.2012.693773

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. (2000). *BSI. Brief symptome inventory: Deutsche version. Manual*. Göttingen: Beltz.

Frazier, P. A., Tix, A. P., & Barron, K. E. (2004). Testing Moderator and Mediator Effects in Counseling Psychology Research. *Journal of Counseling Psychology*, *51*(1), 115–134. https://doi.org/10.1037/0022-0167.51.1.115

Fredrickson, B. L. (2004). The broaden-and-build theory of positive emotions. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *359*(1449), 1367–1378. https://doi.org/10.1098/rstb.2004.1512

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

Garland, E. L., Fredrickson, B., Kring, A. M., Johnson, D. P., Meyer, P. S., & Penn, D. L. (2010). Upward spirals of positive emotions counter downward spirals of negativity: insights from the broaden-and-build theory and affective neuroscience on the treatment of emotion dysfunctions and deficits in psychopathology. *Clinical psychology review*, *30*(7), 849–864. https://doi.org/10.1016/j.cpr.2010.03.002

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

Gibbons, C. J., Fournier, J. C., Stirman, S. W., DeRubeis, R. J., Crits-Christoph, P., & Beck, A. T. (2010). The clinical effectiveness of cognitive therapy for depression in an outpatient clinic. *Journal of affective disorders*, *125*(1-3), 169–176. https://doi.org/10.1016/j.jad.2009.12.030

Grawe, K. (1992). Psychotherapieforschung zu Beginn der neunziger Jahre [Psychotherapyresearch at the beginning of the nineties]. *Psychologische Rundschau*. (43(3)), 132–162.

Grawe, K. (2006). *Neuropsychotherapie [Neuropsychotherapy]*. Cambridge, MA: Hogrefe.

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

Green, L. W., & Glasgow, R. E. (2006). Evaluating the relevance, generalization, and applicability of research: issues in external validation and translation methodology. *Evaluation & the health professions*, *29*(1), 126–153.

Grosse Holtforth, M., Altenstein, D., Krieger, T., Flückiger, C., Wright, A. G. C., & Caspar, F. (2014). Interpersonal differentiation within depression diagnosis: relating interpersonal subgroups to symptom load and the quality of the early therapeutic alliance. *Psychotherapy research : journal of* *the Society for Psychotherapy Research*, *24*(4), 429–441. https://doi.org/10.1080/10503307.2013.829253

Guo, S., & Fraser, M. W. (2014). *Propenisty Score Analysis: Statisical methods and Apllication*. Thousan Oaks, CA: Sage Publications.

Hamilton, M. A. (1967). Development of a rating scale for primary depressive illness. *British journal of social and clinical psychology*, *6*(4), 278–296.

Hansen, B. B. (2004). Full Matching in an Observational Study of Coaching for the SAT. *Journal of the American statistical Association*, *99*(467), 609–618. https://doi.org/10.1198/016214504000000647

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

Harder, V. S., Stuart, E. A., & Anthony, J. C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological methods*, *15*(3), 234–249. https://doi.org/10.1037/a0019623

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

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analyis*. Orlando, FL: Academic Press.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: nonparametric preprocessing for parametric causal inference. *J Stat Soft*, *42*, 1–28.

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

Hollon, S. D., DeRubeis, R. J., Evans, M. D., Wiemer, M. J., Garvey, M. J., Grove, W. M., & Tuason, V. B. (1992). Cognitive Therapy and Pharmacotherapy for Depression. *Archives of General Psychiatry*, *49*(10), 774. https://doi.org/10.1001/archpsyc.1992.01820100018004

Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: A program for missing data. *Journal of statistical software*, *45*(7), 1–47.

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

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

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, *86*(1), 4–29.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology (Cambridge, Mass.)*, *19*(5), 640–648. https://doi.org/10.1097/EDE.0b013e31818131e7

Ioannidis, J. P. A. (2014). How to make more published research true. *PLoS medicine*, *11*(10), e1001747. https://doi.org/10.1371/journal.pmed.1001747

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

Johnsen, T. J., & Friborg, O. (2015). The effects of cognitive behavioral therapy as an anti-depressive treatment is falling: A meta-analysis. *Psychological Bulletin*, *141*(4), 747–768. https://doi.org/10.1037/bul0000015

Keitner, G. I., Posternak, M. A., & Ryan, C. E. (2003). How many subjects with major depressive disorder meet eligibility requirements of an antidepressant efficacy trial? *The Journal of clinical psychiatry*, *64*(9), 1091–1093.

Keller, S. M., Feeny, N. C., & Zoellner, L. A. (2014). Depression sudden gains and transient depression spikes during treatment for PTSD. *Journal of consulting and clinical psychology*, *82*(1), 102–111. https://doi.org/10.1037/a0035286

Kelly, K. A., Rizvi, S. L., Monson, C. M., & Resick, P. A. (2009). The impact of sudden gains in cognitive behavioral therapy for posttraumatic stress disorder. *Journal of traumatic stress*, *22*(4), 287–293. https://doi.org/10.1002/jts.20427

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., & Bottonari, K. A. (2007). Non-treatment-related sudden gains in depression: the role of self-evaluation. *Behaviour Research and Therapy*, *45*(4), 737–747. https://doi.org/10.1016/j.brat.2006.06.008

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., van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Cai, T.,. . . 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., Brenner, L. A., Ebert, D. D.,. . . Zaslavsky, A. M. (2016). Using patient self-reports to study heterogeneity of treatment effects in major depressive disorder. *Epidemiology and psychiatric sciences*, 1–15. https://doi.org/10.1017/S2045796016000020

Kessler, R. C., Chiu, W. T., Demler, O., & Walters, E. E. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of general psychiatry*, *62*(6), 617–627.

Klerman, G. L., Weissman, M. M., Rounsaville, B. J., & Chevron, E. S. (1984). *Interpersonal psychotherapy of depression*. New York: Basic Books.

Kruger, 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

Kurth, T., Walker, A. M., Glynn, R. J., Chan, K. A., Gaziano, J. M., Berger, K., & Robins, J. M. (2006). Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American journal of epidemiology*, *163*(3), 262–270. https://doi.org/10.1093/aje/kwj047

Lambert, M. J. (2013). The Efficacy and Effectiveness of Psychotherapy. In M. J. Lambert (Ed.), *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change* (pp. 169–218). New York, NY: Wiley & Sons.

Lambert, M. (2007). Presidential address: What we have learned from a decade of research aimed at improving psychotherapy outcome in routine care. *Psychotherapy Research*, *17*(1), 1–14. https://doi.org/10.1080/10503300601032506

Lemmens, L. H., DeRubeis, R. J., Arntz, A., Peeters, F. P., & Huibers, M. J. (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

Lipman, R. S., Covi, L., & Shapiro, A. K. (1979). The Hopkins Symptom Checklist (HSCL): factors derived from the HSCL-90. *Journal of affective disorders*, *1*(1), 9–24.

Luborsky, L. (1985). Therapist Success and Its Determinants. *Archives of General Psychiatry*, *42*(6), 602. https://doi.org/10.1001/archpsyc.1985.01790290084010

Luborsky, L., McLellan, A. T., Diguer, L., Woody, G., & Seligman, D. A. (1997). The Psychotherapist Matters: Comparison of Outcomes Across Twenty-Two Therapists and Seven Patient Samples. *Clinical Psychology: Science and Practice*, *4*(1), 53–65. https://doi.org/10.1111/j.1468-2850.1997.tb00099.x

Lutz, W., Ehrlich, T., Rubel, J., Hallwachs, N., Röttger, M.-A., Jorasz, C.,. . . Tschitsaz-Stucki, A. (2012). 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.,. . . 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.

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

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., Stulz, N., & Köck, K. (2009). Patterns of early change and their relationship to outcome and follow-up among patients with major depressive disorders. *Journal of affective disorders*, *118*(1), 60–68.

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

Mander, J. V., Wittorf, A., Schlarb, A., Hautzinger, M., Zipfel, S., & Sammet, I. (2013). Change mechanisms in psychotherapy: multiperspective assessment and relation to outcome. *Psychotherapy research : journal of the Society for Psychotherapy Research*, *23*(1), 105–116. https://doi.org/10.1080/10503307.2012.744111

McEvoy, P. M., & Nathan, P. (2007). Effectiveness of cognitive behavior therapy for diagnostically heterogeneous groups: a benchmarking study. *Journal of consulting and clinical psychology*, *75*(2), 344–350. https://doi.org/10.1037/0022-006X.75.2.344

Merrill, K. A., Tolbert, V. E., & Wade, W. A. (2003). Effectiveness of cognitive therapy for depression in a community mental health center: A benchmarking study. *Journal of consulting and clinical psychology, 71*(2), 404.

Minami, T., Wampold, B. E., Serlin, R. C., Hamilton, E. G., Brown, G. S. J., & Kircher, J. C. (2008). Benchmarking the effectiveness of psychotherapy treatment for adult depression in a managed care environment: a preliminary study. *Journal of consulting and clinical psychology*, *76*(1), 116–124. https://doi.org/10.1037/0022-006X.76.1.116

Murray, C. J. L., & Lopez, A. D. (1996). Evidence-based health policy--lessons from the Global Burden of Disease Study. *Science, 274*(5288), 740.

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

Open Science Collaboration. (2015). PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science (New York, N.Y.), 349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Orlinsky, D. E., Grawe, K., & Parks, B. K. (1994). Process and outcome in psychotherapy - noch einmal. In A. Bergin & S. L. Garfield (Eds.), *Handbook of psychotherapy and behavior change* (4th ed., pp. 270–376). New York, NY: John Wiley & Sons, Inc.

Popper, K. R. (1983). *The logic of scientific discovery* (11th impr). London: Hutchinson.

Present, J., Crits-Christoph, P., Connolly Gibbons, M. B., Hearon, B., Ring-Kurtz, S., Worley, M., & Gallop, R. (2008). Sudden gains in the treatment of generalized anxiety disorder. *Journal of Clinical Psychology, 64*(1), 119–126. https://doi.org/10.1002/jclp.20435

Rosenbaum, P. (1991). A Characterization of Optimal Designs for Observational Studies. *Journal of the Royal Statistical Society*. (3), 597–610.

Rosensbaum, 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

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.

Rubel, J. A., Rosenbaum, D., & Lutz, W. (2017). Patients' in-session experiences and symptom change: Session-to-session effects on a within- and between-patient level. *Behaviour Research and Therapy*, *90*, 58–66. https://doi.org/10.1016/j.brat.2016.12.007

Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology*, *2*(3), 169–188.

Schindler, A. C., Hiller, W., & Witthöft, M. (2011). Benchmarking of cognitive-behavioral therapy for depression in efficacy and effectiveness studies—How do exclusion criteria affect treatment outcome? *Psychotherapy Research*, *21*(6), 644–657.

Schmidt, S. (2009). Shall we really do it again?: The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*(2), 90–100. https://doi.org/10.1037/a0015108

Schulte, D. (1996). *Therapieplanung [Panning of Therapy]*. Göttingen: Hogrefe.

Seligman, M. E. P. (1995). The effectiveness of psychotherapy: The Consumer Reports study. *American Psychologist*, *50*(12), 965.

Shadish, W. R. (2013). Propensity score analysis: promise, reality and irrational exuberance. *Journal of Experimental Criminology*, *9*(2), 129–144.

Shadish, W. R., Matt, G. E., Navarro, A. M., Siegle, G., Crits-Christoph, P., Hazelrigg, M. D.,. . . Robinson, L. (1997). *Evidence that therapy works in clincally representative conditions*: American Psychological Association.

Shadish, W. R., Navarro, A. M., Matt, G. E., & Phillips, G. (2000). The effects of psychological therapies under clinically representative conditions: A meta-analysis. *Psychological Bulletin*, *126*(4), 512–529. https://doi.org/10.1037//0033-2909.126.4.512

Shafran, R., Clark, D. M., Fairburn, C. G., Arntz, A., Barlow, D. H., Ehlers, A.,. . . Wilson, G. T. (2009). Mind the gap: Improving the dissemination of CBT. *Behaviour Research and Therapy*, *47*(11), 902–909. https://doi.org/10.1016/j.brat.2009.07.003

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*. (32 (9)), 752–760.

Spitzer, R. L., Endicott, J., & Robins, E. (1978). Research diagnostic criteria: rationale and reliability. *Archives of general psychiatry*, *35*(6), 773–782.

Spitzer, R. L., Gibbon, M., Skodol, A. E., Williams, J. B. W., & First, M. B. (2002). *DSM-IV-TR Casebook: A Learning Companion to the Diagnostic and Statistical Manual of Mental Disorders* (Vol. 1). Arlington, VA: American Psychiatric Publishing, Inc.

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. (2001). Assimilation of problematic experiences. *Psychotherapy: Theory, Research, Practice, Training*, *38*(4), 462–465. https://doi.org/10.1037/0033-3204.38.4.462

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.,. . . 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

Stiles, W. B., Startup, M., Hardy, G. E., Barkham, M., Rees, A., Shapiro, D. A., & Reynolds, S. (1996). Therapist session intentions in cognitive-behavioral and psychodynamic-interpersonal psychotherapy. *Journal of Counseling Psychology*, *43*(4), 402–414. https://doi.org/10.1037//0022-0167.43.4.402

Stuart, E. A., & Green, K. M. (2008). Using full matching to estimate causal effects in nonexperimental studies: examining the relationship between adolescent marijuana use and adult outcomes. *Developmental psychology*, *44*(2), 395–406. https://doi.org/10.1037/0012-1649.44.2.395

Stulz, N., Lutz, W., Kopta, S. M., Minami, T., & Saunders, S. M. (2013). Dose–effect relationship in routine outpatient psychotherapy: Does treatment duration matter? *Journal of Counseling Psychology, 60*(4), 593–600. https://doi.org/10.1037/a0033589

Stulz, N., Lutz, W., Leach, C., Lucock, M., & Barkham, M. (2007). Shapes of early change in psychotherapy under routine outpatient conditions. *Journal of consulting and clinical psychology, 75*(6), 864.

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

Taylor, S., & Asmundson, G. J. G. (2008). Internal and external validity in clinical research. *Handbook of research methods in abnormal and clinical psychology*. *Sage Publications, Los Angeles*, 23–34.

van Geert, P., & van Dijk, M. (2002). Focus on variability: New tools to study intra-individual variability in developmental data. *Infant Behavior and Development*, *25*(4), 340–374. https://doi.org/10.1016/S0163-6383(02)00140-6

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.,. . . 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

Warden, D., Rush, A. J., Trivedi, M. H., Fava, M., & Wisniewski, S. R. (2007). The STAR*D project results: A comprehensive review of findings. *Current Psychiatry Reports*, *9*(6), 449–459. https://doi.org/10.1007/s11920-007-0061-3

Weber, E. U., Shafir, S., & Blais, A.-R. (2004). Predicting risk sensitivity in humans and lower animals: risk as variance or coefficient of variation. *Psychological Review*, *111*(2), 430–445. https://doi.org/10.1037/0033-295X.111.2.430

Weissman, A. N. (Ed.) 1980. *Assessing depressogenic attitudes: A validation study.*

Weisz, J. R., Weiss, B., & Donenberg, G. R. (1992). The lab versus the clinic: Effects of child and adolescent psychotherapy. *American Psychologist*, *47*(12), 1578–1585. https://doi.org/10.1037/0003-066X.47.12.1578

West, S. G., Cham, H., & Thoemmes, F. (2015). Propensity score analysis. In R. L. Cautin & S. O. Lilienfeld (Eds.), *The encyclopedia of clinical psychology* (pp. 1–10). New York: John Wiley & Sons.

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

Wilcox, R. (2011). *Modern statistics for the social and behavioral sciences: A practical introduction*. Boca Raton, FL: CRC press.

Wilson, G. T. (1999). Rapid Response to Cognitive Behavior Therapy. *Clinical Psychology: Science and Practice*, *6*(3), 289–292. https://doi.org/10.1093/clipsy.6.3.289

Wucherpfennig, F., Rubel, J. A., Hollon, S. D., & Lutz, W. (2016). 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

Zetin, M., & Hoepner, C. T. (2007). Relevance of exclusion criteria in antidepressant clinical trials: a replication study. *Journal of clinical psychopharmacology*, *27*(3), 295–301.

Zimmermann, D., Rubel, J., Page, A. C., & Lutz, W. (2016). Therapist Effects on and Predictors of Non-Consensual Dropout in Psychotherapy. *Clinical psychology & psychotherapy*. Advance online publication. https://doi.org/10.1002/cpp.2022

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

Nachname: Wucherpfennig; Vorname: Felix

Unterschrift: ___________________