From THE DEPARMENT OF CLINICAL NEUROSCIENCE
Karolinska Institutet, Stockholm, Sweden

ADAPTIVE TREATMENT STRATEGIES IN INTERNET-DELIVERED COGNITIVE BEHAVIOR THERAPY: PREDICTING AND AVOIDING TREATMENT FAILURES

Erik Forsell

Stockholm 2020

Published by Karolinska Institutet.
Printed by Eprint AB, 2020
Cover art by Arvid Kalmaru

ISBN 978-91-7831-776-9

Adaptive Treatment Strategies in Internet-delivered Cognitive Behavior Therapy: Predicting and avoiding treatment failures

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Erik Forsell

Principal Supervisor:
Professor Viktor Kaldo
Karolinska Institutet
Department of Clinical Neuroscience
Centre for Psychiatry Research
and
Linnaeus University
Department of Psychology

Co-supervisor(s):
Dr. Susanna Jernelöv
Karolinska Institutet
Department of Clinical Neuroscience
Division of Psychology
and
Centre for Psychiatry Research

Dr. Kerstin Blom
Karolinska Institutet
Department of Clinical Neuroscience
Centre for Psychiatry Research

Professor Nils Lindefors
Karolinska Institutet
Department of Clinical Neuroscience
Centre for Psychiatry Research

Opponent:
Dr. Johanna Boettcher
Freie Universität Berlin
Department of Psychology
Division of Clinical Psychology and psychotherapy

Examination Board:
Professor Matteo Bottai
Karolinska Institutet
Institute of Environmental Medicine
Unit of Biostatistics

Dr. Ida Flink
Örebro University
School of Law, Psychology and Social Work
Center for Health and Medical Psychology

Dr. Fredrik Falkenström
Linköping University
Department of Behavioural Sciences and Learning
Division of Psychology

To my family Josefin, Selma and Oskar.

Not only did they put up with me during the entire PhD journey, but they also spent the last two weeks of it in quarantine with me because of the coronavirus pandemic of 2020.

ABSTRACT

Background: Internet-delivered Cognitive Behavior Therapy (ICBT) is efficacious for a number of psychiatric disorders and can be successfully implemented in routine psychiatric care. Still, only about half of patients experience a good enough treatment outcome. Using data from the early part of treatment to identify patients with high risk of not benefitting from it, and target them with additional resources to prevent the predicted failure is a potential way forward. We call this an Adaptive Treatment Strategy, and a very important part of it is the ability to predict the outcome for a specific patient.

Aims: To establish a proof of concept for an Adaptive Treatment Strategy in ICBT, and explore outcome prediction further by evaluating the accuracy of an empirically supported classification algorithm, the time point in treatment when acceptable accuracy can be reached, and the accuracy of ICBT-therapists' own predictions. Preliminary benchmarks regarding the clinical usefulness of prediction will be established.

Studies: Four studies were performed: Study I was a randomized controlled trial (RCT; *n*=251) where patients' risk of treatment Failure (Red=high risk of failure, Green=low risk) was predicted during week 4 out of 9 in ICBT for Insomnia. Red patients (*n*=102) were then randomized to either continuing with standard treatment (*n*=51) or having their treatment individually adapted (*n*=51). In Study II, the classification algorithm from Study I was evaluated in terms of classification accuracy and the contribution of the different predictors used. In Study III, data from 4310 regular care ICBT-patients having received treatment for either Depression, Social anxiety disorder or Panic disorder were analyzed in a series of multiple regression models using weekly observations of the primary symptom measure as predictors to classify risk of Failure. As a contrast, Study IV examines ICBT therapists' own predictions on both categorical and continuous treatment outcomes, as they made predictions for each of their patients (*n*=897) during week 4 in the same three treatments as in Study III.

Results: The RCT was successful in that Red patients receiving Adapted treatment improved significantly more than Red patients receiving standard treatment, and their odds of failure were nearly cut in half. Green patients did better than Red patients, indicating that the accuracy of the classification algorithm was clinically useful. Study II showed that the balanced accuracy of the classifier was 67% and that only 11 of 21 predictors correlated significantly with Failure. Notable predictors were symptom levels as well as different markers of treatment engagement. Study III and IV showed that acceptable predictions could be made halfway through treatment using only symptom scores and basic statistics, and that ICBT-therapists predicted outcomes better than chance but on average 9.5 % less accurate than the statistical models. Therapist predictions reached the clinical acceptance benchmark only for remission in Social anxiety disorder. At treatment week four, therapist could predict on average 16% of the variance in continuous outcomes, compared to a statistical model explaining 39%.

Conclusions: We find support for the clinical usefulness of an Adaptive Treatment Strategy in ICBT for insomnia, and establish a preliminary benchmark that a classification algorithm with at least 67% balanced accuracy should be sufficient for clinical purposes. Simple statistical models using only symptom scores can reach clinically acceptable levels of accuracy halfway through 12-week ICBT-programs. Previous findings that therapists' predictions are less accurate than statistical models seem to hold also for therapists providing ICBT. However, it was also indicated that clinicians' ratings of adherence and activity do add unique information to prediction algorithms. In line with previous findings, the vast majority of useful prediction variables were found during early treatment, rather than before treatment start.

LIST OF SCIENTIFIC PAPERS

I. **Forsell, E.**, Jernelöv, S., Blom, K., Kraepelien, M., Svanborg, C., Andersson, G., Lindefors, N. & Kaldo, V. (2019). Proof of Concept for an Adaptive Treatment Strategy to Prevent Failures in Internet-Delivered CBT: A Single-Blind Randomized Clinical Trial With Insomnia Patients. *American Journal of Psychiatry*, *176*(4), 315-323.

II. **Forsell, E.**, Jernelöv, S., Blom, K. & Kaldo, V. (submitted manuscript). Clinically sufficient classification accuracy and key predictors of treatment failure in a randomized controlled trial of Internet-delivered Cognitive Behavior Therapy for Insomnia

III. **Forsell, E.**, Isacsson, N., Blom, K., Jernelöv, S., Ben Abdesslem, F., Lindefors, N., Boman, M. & Kaldo, V. (2019). Predicting treatment failure in regular care Internet-Delivered Cognitive Behavior Therapy for depression and anxiety using only weekly symptom measures. *Journal of Consulting and Clinical Psychology*, [online ahead of print]. doi:10.1037/ccp0000462

IV. **Forsell, E.**, Mattsson, S. & Kaldo, V. (submitted manuscript). Accuracy of therapists' predictions of outcome in Internet delivered Cognitive Behavior Therapy for depression and anxiety in routine psychiatric care.

CONTENTS

1 Introduction 1
1.1 The state of psychological treatment of common psychiatric disorders 1
1.2 Defining success: dichotomous clinical outcomes 2
1.2.1 Dichotomizing is problematic but useful 2
1.2.2 Approaches to dichotomizing continuous treatment outcomes 2
1.3 Stepped Care 5
1.3.1 Accelerated Care: making more of entry level steps 5
1.4 Outcome prediction in psychotherapy 8
1.4.1 Everyday clinical predictions are common but unreliable 8
1.4.2 Predictors at baseline 8
1.4.3 Grouping cases and using group membership as the predictor 11
1.4.4 The big leap to big data and big models 12
1.4.5 Moving away from baseline and into treatment 13
1.5 How accurate do predictions need to be within an adaptive treatment strategy? 16
1.5.1 Predictability of psychotherapy and ideal levels of accuracy 16
1.5.2 The Decision-Threshold Approach in clinical decision making 17
1.5.3 High vs low-risk decisions 18
1.6 The advantages of ICBT for research on prediction and Adaptive treatment strategies 18
2 The Aims of the thesis 19
3 The Empirical studies 19
3.1 Methods 19
3.1.1 Participants, recruitment and assessments 19
3.1.2 The Internet-delivered Cognitive Behavior Therapies 19
3.1.3 Symptom measures 20
3.1.4 Definition of Success and Failure 21
3.1.5 Classification accuracy measure 21
3.1.6 Ethical considerations 22
3.2 Study I: Proof of Concept for an Adaptive Treatment Strategy in ICBT: A Single-Blind RCT With Insomnia Patients 23
3.2.1 Aim and hypothesis 23
3.2.2 Methods 23
3.2.3 Results 23
3.2.4 Conclusions 24
3.3 Study II: Clinically sufficient classification accuracy and key predictors of treatment failure from the RCT 24
3.3.1 Aim and hypothesis 24
3.3.2 Methods 24
3.3.3 Results 25
3.3.4 Conclusions 25

3.4 Study III: Predicting treatment failure in regular care ICBT for depression and anxiety using weekly symptom measures 25
3.4.1 Aim and hypothesis 25
3.4.2 Methods 26
3.4.3 Results 26
3.4.4 Conclusions 26
3.5 Study IV: Accuracy of therapists' predictions in ICBT for depression and anxiety in routine psychiatric care 27
3.5.1 Aims and hypothesis 27
3.5.2 Methods 27
3.5.3 Results 28
3.5.4 Conclusions 28
3.6 Summary table regarding predicting failure during week 4 29
4 Discussion 31
4.1 Main findings 31
4.1.2 Causal inferences based on this thesis 39
4.2 Adaptive treatment and its effects 39
4.3 Ethical considerations of Adaptive Treatment strategies 40
4.4 Limitations 41
4.4.1 Dropout 41
4.4.2 Variations in outcomes across studies 41
4.4.3 The definition of Success as either response or remission 42
4.4.4 The clinical implications of prediction that focuses on identifying cases rather than supporting theory 43
4.5 Directions for future research 44
5 Conclusions 45
6 Acknowledgements 47
7 References 49

LIST OF ABBREVIATIONS

ICBT	Internet-delivered Cognitive Behavior Therapy
ICBT-i	ICBT for Insomnia
RCI	Reliable Change Index
RCT	Randomized Controlled Trial
BACC	Balanced Accuracy
PHQ-9	Patient Health Questionnaire 9-items
GAD-7	Generalized Anxiety Disorder 7-items questionnaire
MDD	Major Depressive Disorder
SAD	Social Anxiety Disorder
PD	Panic Disorder
ISI	Insomnia Severity Index
MADRS-S	Montgomery-Åsberg Depression Rating Scale-Self report
LSAS-SR	Liebowitz Social Anxiety Scale-Self report
PDSS-SR	Panic Disorder Severity Scale-Self report
IAPT	Improving Access to Psychological Therapies

1 INTRODUCTION

1.1 THE STATE OF PSYCHOLOGICAL TREATMENT OF COMMON PSYCHIATRIC DISORDERS

There is ample evidence that psychiatric disorders can be treated using psychological interventions and, in Sweden at least, psychological interventions have taken the place as first choice interventions for a number of conditions (Socialstyrelsen, 2017). Cognitive Behavior Therapy (CBT) is perhaps the most thoroughly researched psychotherapy today, and has positive effects on a wide variety of psychiatric disorders (Hofmann, Asnaani, Vonk, Sawyer, & Fang, 2012) ranging from depression (Cuijpers et al., 2013; Hedman et al., 2014), insomnia (Blom, Jernelov, Rück, Lindefors, & Kaldo, 2016; Siebern & Manber, 2011) and social phobia (El Alaoui et al., 2015; El Alaoui, Hedman, Ljótsson, & Lindefors, 2015) all the way to obsessive compulsive disorder (Andersson et al., 2012; Foa, 2010), and psychosis (Turkington, Wright, & Tai, 2013). The last 20 years of research also shows that delivering CBT in the form of guided self-help via the internet (ICBT) can be as effective as face-to-face CBT (Carlbring, Andersson, Cuijpers, Riper, & Hedman-Lagerlof, 2018; Cuijpers, Donker, van Straten, Li, & Andersson, 2010; Hedman, Ljotsson, & Lindefors, 2012; Karyotaki et al., 2017). ICBT for common mental disorders has been successfully implemented in routine psychiatric care in several countries around the world (Titov et al., 2018) and is rapidly expanding in Sweden and elsewhere.

The evidence for CBT is, like all evidence for treatment of psychiatric disorders, almost exclusively based on group level effect sizes and statistically significant mean differences, failing to address one key factor: not all patients get better, and quite a few actually worsen. In fact, 5-10% of patients both in psychological treatments in general as well as in ICBT leave treatment worse off than when they started (Rozental et al., 2014; Slade, Lambert, Harmon, Smart, & Bailey, 2008) and, depending on definitions used, 25-65% of patients do not achieve a satisfactory treatment outcome. This is quite consistent across many forms of face-to-face psychotherapy, as well as CBT, guided self-help and internet delivered CBT (ICBT) for a wide range of disorders (Andersson, Carlbring, & Rozental, 2019; Lambert, 2015; Rozental, Andersson, & Carlbring, 2019; Slade et al., 2008). Even more worrying is that and clinicians are generally unable to predict this (Hannan et al., 2005; Lambert, 2015, 2017). While CBT and other psychotherapies are constantly being evaluated in new contexts, versions, and populations there is a paucity of research into what works for whom and how failing treatment attempts can be salvaged, leaving a gap between what we know of group level versus individual effectiveness.

1.2 DEFINING SUCCESS: DICHOTOMOUS CLINICAL OUTCOMES

Despite many decades of research and clinical practice in psychiatry and psychotherapy, there is still no definitive answer to how we should decide when an individual patient has been adequately treated. We know that, on average, patients who receive these treatments get better, compared to if they had not received the treatment. The pervasive measure of an effective treatment is whether it has a large effect size based on mean symptom change compared to no treatment, but how do we decide when better becomes well, and how much better is enough for the individual?

1.2.1 Dichotomizing is problematic but useful

The gold standard of quantitative analysis is to use variables that are continuous, and preferably on ratio scale. It gives high statistical power and enables many different kinds of analyses. If given the choice, thoughtful researchers will opt for continuous rather than categorical data. A perfect example would be weight in grams. In our field however, things are rarely that simple. A continuous measure in our field if often a symptom questionnaire on which a patient scores their symptoms and receives a sum reflecting how severe their symptoms are, which is already far less exact and reliable than ordinary kitchen scales.

In the medical and health sciences, researchers often add some form of categorical outcome, so that we can, as mentioned earlier, decide when someone has been successfully treated. This is often achieved by applying cutoffs to the continuous scale and thereby making it dichotomous. This is a gross simplification of reality and can create false dichotomies (Naudet, Millet, Michel Reymann, & Falissard, 2014). A real life human being will not feel markedly different when scoring one point above a cutoff as opposed to one point below, even though as far as psychiatric research is concerned, that person suddenly snapped out of the clutches of depression. Nevertheless, for publicly funded clinical practice at least, we do still have to decide at what point the suffering of one individual should be considered pathological and be treated. Oddly, it is not agreed upon exactly how this should be done.

1.2.2 Approaches to dichotomizing continuous treatment outcomes

Jacobson and Truax (1991) aimed to account for pre-treatment symptoms by defining a good clinical outcome as a statistically significant decrease in symptoms combined with a post-treatment score below a certain cutoff. One problem with this is that in clinical reality that combination is not necessarily relevant. Treatments are given to patients based on their current symptom level; regardless of what it was several weeks or months ago. There is a meaningful scientific reason to include significant change in the definition. This assures that a treatment is not evaluated as powerful based on trials having patients with low symptoms enter, and consequently exit, treatment with low scores. In our clinical reality however, that problem should be addressed before treatment, so that treatment is never offered to patients who do not need it. Most symptom measures have an established cutoff for clinical versus non-clinical symptom levels, but Jacobson and Truax's definition also requires a definition of a reliable change in symptoms from pre- to post-treatment.

1.2.2.1 The Reliable Change Index (RCI) and its limitations

The Reliable Change Index (RCI) is defined as the difference between two scores divided by the standard error of difference between scores (Jacobson, Follette, & Revenstorf, 1984). It is a way to quantify the minimal difference between two measurements taken with the same measure that are beyond the measurement error of that measure. In other words, the minimal difference that can be confidently considered real from a statistical point of view. Originally, the RCI produces a number that corresponds to the standard normal distribution where 1.96 or more refers to the tails of a 95% confidence interval. It can however, and is most often, translated in to a measurement specific score by using the formula $1.96*SD1*\sqrt{2}*\sqrt{(1-rel)}$ (Wise, 2004) where SD1 refers to the standard deviation of the measurement at some point (usually pre-treatment in your sample, but more on that later) and rel refers to the test-retest reliability of the measurement.

One problem with the RCI is that is has always been somewhat unclear from where the different estimates should be drawn when calculating it. The formula requires a standard deviation and a test-retest reliability for the scale. However, those estimates can be drawn from many different sources, for example from psychometric studies, prevalence studies or more typically, the sample being studied, yielding quite different results. This makes the RCI a relative measure even within a specific symptom measure, influenced by the variability of the sample being studied, when it is clear that the intention is for it to be a static characteristic of a scale. Illustrating this, Hiller, Schindler, and Lambert (2012) showed that for the Beck Depression Inventory, the calculated RCI's differed between 6 and 14 points among the 11 published randomized controlled trials (RCTs) being compared, which has a massive impact on the overall evaluations of these treatments. Using the RCI based on sample estimates is advantageous if your sample has low variance at baseline but hindering if your sample has a large range of symptom scores at baseline. It is quite unreasonable to think that the meaningfulness of a change in an individual's symptom scores should be affected by the spread of symptom scores among other patients being treated at the same time.

Furthermore, most users erroneously use the Cronbach's α as the test-retest reliability. Cronbach's α measures reliability between *items* within a scale not between *timepoints*. A high Cronbach's α indicates that each item contributes relatively uniformly to the sum of the scale. It has absolutely nothing to do with the stability of that scale over time in unchanging individuals; how could it? They may be forgiven, since using Cronbach's α is also what is most often recommended when reading about how to use the RCI, but it is still wrong.

Another important distinction to make is that, as the name suggests, the Reliable Change Index is concerned with statistically reliable rather than clinically relevant change. Concluding that a patient's post-treatment levels of depression are not identical to their pre-treatment levels is of limited interest in clinical reality. Jacobson & Truax suggested a combination of being reliably improved and moving from the clinical to the non-clinical range, which is a reasonable combination. It does however create problems with patients with high baseline scores. A patient moving from 54 points on the Montgomery-Åsberg Depression Rating Scale-Self report

(MADRS-S) to 15 points would be considered a failed treatment (the cutoff for remission is 10), while many clinicians and patients would likely view such an outcome as quite an achievement. The RCI is biased towards labeling regression to the mean and proportionally small improvements as response for initially severe patients, and vice versa for patients with mild initial symptoms. Remission on its own has the opposite problem.

1.2.2.2 Proportional Improvement

Another way to define the magnitude of a change is to look at it as a proportion of a patient's pre-treatment score. This would circumvent the problems with the RCI, namely labeling small changes in severe patients as good and relatively large changes in mild patients as no change at all. Regarding this, Karin, Dear, Heller, Gandy, and Titov (2018) found two critical pieces of evidence in their data of over 1000 ICBT-patients compared to untreated controls. Firstly, patients' change in raw scores were very much dependent on where they started, with each category of more severe patients (patients were divided into minimal, mild, moderate, moderately severe and severe) showing larger improvements than the last. However, when change was expressed as a proportion of each patient's baseline score, that relationship completely disappeared and all clinical severity groups improved, on average 50-55% from pre- to post-treatment (the minimal symptoms group did not). Secondly, they found that this change in statistical relationship was actually not the case for untreated controls. This means that expressing change as a proportion, both compensates for initial severity and results in clearer differentiation between treatment-specific and non-specific effects. This is highly valuable in an uncontrolled setting such as routine psychiatric care.

Statistically, proportional improvement should be used only as a qualifier for a dichotomous outcome (i.e. a patient is either above or below a certain percentage of improvement) and not as a change-score per se in statistical analyses as this produces biased and underpowered results (Vickers, 2001). As a basis for the dichotomous outcomes response and remission however, percentage change is already widely used in psychopharmacological research and has been shown to be reliable and less affected by variability in the sample and baseline severity (Hiller et al., 2012). In depression, 50% or more is typically the percentage used in research to define response (Hiller et al., 2012; Kuk, Li, & Rush, 2010; Leucht et al., 2017). This mirrors what was found by Karin et al. (2018) in a mixed ICBT-sample with depression and anxiety. Rather than the spread in the sample or population as the RCI is, percentage improvement is affected by the individual patient's pre-treatment severity (Hiller et al., 2012).

A proportional improvement of 50% compensates for the problems with the RCI and provides a way to dichotomize improvement that considers the individual and differentiates more between specific and non-specific effect than other alternatives. Proportional improvement as the sole criterion of success has the problem that it is possible to reach remission without having a large proportional improvement, and such patients should not be viewed as in need of more treatment. Therefore, we believe that the best choice is to combine the two dichotomous outcomes remission and response (50% reduction) with an either or criteria to define Success and conversely Failure in ICBT.

1.3 STEPPED CARE

In clinical practice, the gap between having effective treatments and successfully treating all patients can be handled in different ways. It seems that historically, a failed psychotherapy is followed by trying something else, be it another therapist or school of psychotherapy, trying medications or physiological interventions, or simply moving on to counseling and care taking rather than treatment. This is likely costly and disorderly. One attempt to remedy this is the principle of stepped care (Bower & Gilbody, 2005).

Stepped care means organizing this sequence of evidence-based treatment attempts in a standardized way so that the cheapest, least invasive, and most accessible treatment is offered first and patients who turn out to need more then move up a step in intensity and cost (Bower & Gilbody, 2005). This creates a sort of standardized pyramid where most patients will get the cheap and accessible treatments quickly and only the more problematic cases go on to long and costly treatments. While the idea may be that the moving from one step to the next could be done during, rather than after treatment, in reality that does not seem to be the case, at least in psychotherapy (Richards et al., 2012), perhaps due to a lack of methods for accurately identifying failing treatments. The idea of stepping itself indicates that one form of treatment will be abandoned rather than enhanced even if one does not wait until a planned treatment has run its course. Therefore, in stepped care the patients who do end up climbing several of the stepped care steps will likely have long treatment periods that will include a number of failed treatments attempts. A recent study found that while stepped care was non-inferior to initiating the high intensity treatment right away, patients who received stepped care were less satisfied than were those receiving the high-intensity treatment right away (Mohr et al., 2019).

Stepped care may be a step in the right direction, but there is still a dearth of research into its effectiveness (Bower & Gilbody, 2005; van Straten, Hill, Richards, & Cuijpers, 2015). Gyani, Shafran, Layard, and Clark (2013) performed a large analysis (n>11 000) of the Stepped Care-based Improving Access to Psychological Therapies program (IAPT). They found that the strongest predictor of reliable recovery was having been moved up the steps (i.e. abandoned low intensity treatments) followed by the place of treatment having a high average number of sessions (again, not indicative of low intensity). This indicates that in the IAPT context at least, the entry-level steps are not really doing enough for the patients.

1.3.1 Accelerated Care: making more of entry level steps

I believe it is important to focus our energy on making the first step, whatever that may be for an individual, more efficacious. Focus should not only be on how intensive it is from a caregiver perspective but also factor in how long it takes a specific patient to get well. Starting at the appropriate, rather than lowest, step and having the ability to intensify efforts within a step can make for a smoother healthcare process. Rather than Stepped Care we should strive to create systems for Accelerated Care. There are several areas where we need to improve in order to make Accelerated Care possible. One area of improvement is pre-treatment matching of individual patients to the most suitable level of treatment so that the individual gets offered the

treatment that is most likely to succeed rather than what's cheapest or quickest. Another area of improvement is during-treatment identification of failing treatments followed by adaptations or alterations of ongoing treatments to prevent a failure from happening. Doing this, rather than "stepping up" to a different treatment will henceforth be referred to as having an Adaptive Treatment Strategy. Obviously, these approaches are not mutually exclusive and ideally both should be developed and used in the future. However, both approaches must rely heavily on making accurate predictions about outcomes in psychotherapy. In this thesis, we focus exclusively on Adaptive Treatment Strategies and predicting outcomes during ongoing treatments, but I will present both components briefly.

1.3.1.1 Accelerated care component 1: Choosing the right treatment right away

A relatively new concept, personalized medicine is concerned with making clinical choices based on predictions of treatment suitability and chances of success, or risk of harm for each specific individual (Auffray & Hood, 2012). The idea is to select a treatment that has not only showed group level efficacy, but that has showed efficacy in patients resembling the current patient as closely as possible. So far, personalized medicine has been primarily used in somatic rather than psychological treatments. There are however some examples of matching patients to treatments in psychological treatments.

DeRubeis et al. (2014) did a retrospective leave-one-out predictive model to classify optimal and non-optimal treatments for patients with depression that had undergone pharmacotherapy or CBT. The analysis meant creating n-1 models that were each used to predict the outcomes for the one patient left out of each model. They found that about 60% of patients had a distinguishable optimal treatment out of the two options, and that if they had been randomly assigned to the optimal treatment they fared significantly better then had they not been, although with a wide confidence interval (*d*=.58, 95% CI =.17-1.01). Similarly, Nadine et al. (2020) calculated a personalized advantage index for 251 patients, who had received either blended face-to-face and internet- CBT or treatment as usual in psychiatric care. They found that 29% of the sample had a discernible superior treatment option, though it is unclear how often, if ever, treatment as usual was superior for patients. This begs the question if the finding is about differential responses to treatment or about treatment-specific versus non-specific effects, as treatment as usual is likely, at least for some patients, analogous to a waiting list.

It is important to note that DeRubeis et al. (2014) predicted suitability to two different treatments that are generally found to be equally effective. Treatment matching could theoretically also include deciding not to treat at all, or to suggest a treatment with smaller group level effects but predicted to be highly suitable for a specific patient. Such decisions would however require even stronger predictions.

For meaningful patient-treatment matching, we need very strong predictors that all have to be collected at baseline and there needs to be several effective treatment options readily available to which patients can be matched. Such high demands do not fit very well with the current

accuracy of our best predictions nor with the way psychological treatments are organized in our society today. Adaptive Treatment Strategies, starting a low cost, effective treatment right away and then predicting and adapting as you go along seems more implementable, providing it works.

1.3.1.2 Accelerated care component 2: Identifying failing treatments and adapting to prevent failure (Adaptive Treatment Strategies)

Michael J Lambert and his colleagues have done several studies using models in which patients are compared to previously found trajectories on a key outcome (OQ-45) and then labeled as on track or not on track (Lambert, 2015). Based on five previous studies with the OQ-measure Lambert found that 85-100% of patients who would go on to deteriorate could be identified during treatment, whereas in one of these studies less than 1% of deteriorating patients were identified by their therapist (Lambert, 2015). Hannan and colleagues found that using what they call lab-test data (questionnaire instead of clinician's judgement) for following treatment progression they could identify 85% of treatments that would go on to fail by treatment session three (Hannan et al., 2005). In a systematic review of 24 studies, Lambert, Whipple, and Kleinstäuber (2018) found that in two thirds of the studies, having an adaptive treatment strategy in place was superior to the normal psychotherapy process delivered by the same therapists.

Lambert and colleagues have found that monitoring patient progress and giving feedback to therapists about progression does not really enhance treatments for patients who are on track but rather that the enhancement effect was isolated to those who were not on track (i.e. failing treatments; Lambert, 2015; 2017).

These classifications based on the OQ-measure show great predictive power even though Schibbye et al. (2014) did show that, at least in ICBT, disorder specific symptom scales were superior to the OQ-45. The OQ measures are also commercial products and as such, the extent of their use might not accurately reflect the usefulness of the scales. It is also important to define what we want to measure and why. Using either measures as the OQ scales or disorder specific ones as the PHQ-9 will lead to different definitions of not only how much change we want to see but also what parameter we are actually interested in changing in psychotherapy. Monitoring and assessing progress on a different measure than the one that will later be used to define the outcome is likely inferior to monitoring the actual outcome of interest.

1.4 OUTCOME PREDICTION IN PSYCHOTHERAPY

1.4.1 Everyday clinical predictions are common but unreliable

In a clinical setting, predictions are made every day and critical clinical decisions are made based upon these predictions. In most cases however, these predictions are made based on clinical expertise and intuition and rely far less on statistical methods (Dawes, Faust, & Meehl, 1989). In some ways it can be argued that the 5-10% deterioration and 30-60% non-response rates seen in psychotherapy research (Andersson et al., 2019; Lambert, 2017) are likely actually based on data where patients have already been clinically matched to treatments, and likely had some adaptive elements to their care. Studies have inclusion and exclusion criteria that, while still trying to preserve generalizability, at least partly reflect our best attempts at maximizing treatment gains and minimizing unsuccessful treatment attempts and harm. An assessing clinician will probably not recommend a treatment that they see as unlikely to succeed. Most would also agree that therapists do already try to gauge how treatments are progressing and do try to avoid failures if they see them coming. The problem is that we as clinicians do not know what works for whom, and we do not see treatment failures coming. At least not well enough.

In a large review of almost 60 years of research in comparing clinical versus actuarial prediction on a wide variety of tasks ranging from future academic performance to risk of future violent acts, Ægisdóttir et al found that using statistical methods rather than clinical judgement was consistently more accurate (Ægisdóttir et al., 2006). Our intuition as clinicians seems to be flawed and often outperformed by more data-driven methods, yet we are often unwilling to choose actuarial predictions over our own judgement (Dawes et al., 1989). Dawes argues that we currently know enough about the fallibility of our clinical expert judgement to warrant a serious discussion on whether or not it is ethically defensible to make clinical decisions based on expertise and intuition alone (Dawes, 2005).

1.4.2 Predictors at baseline

Most research in this field is focused on identifying several relatively independent individual predictors (e.g. age, gender, income) and almost exclusively at baseline. While often theoretically reasonable, such predictors often turn out to be irrelevant in ICBT when examined (Andersson, 2018). For instance, the idea that ICBT is for non-complex or mild psychiatric cases does not seem to hold up. Edmonds, McCall, Dear, Titov, and Hadjistavropoulos (2020) found that concurrent psychotropic medication did not predict worse outcomes and Flygare et al. (2019) found that comorbidities, apart from personality disorders, did not make for worse outcomes in typical Swedish ICBT for depression. It is even unclear if being severely ill or mildly ill to begin with makes for better or worse outcomes.

1.4.2.1 Baseline severity as a predictor of treatment outcome

Baseline severity of the disorder is often described as showing inconsistent results, sometimes predicting better outcomes, sometimes worse outcomes and sometimes not predicting anything (El Alaoui et al., 2016). However, that is affected greatly by the definitions used for success

(for instance responder vs remitter, clinical or statistical significance and whether you combine several criteria or not), and is perhaps much less inconsistent than one might think. Table 1 describes a number of studies assessing how baseline severity of symptoms relates to various positive outcomes depending on symptoms. Based on the table, it is possible to draw the conclusion that more severely ill patients are in fact easier to treat and respond better to CBT and ICBT. However, this is biased by the definition of response, which is either using the Reliable Change Index or using raw scores on a symptom measure.

Instead, these findings could be understood as an artefact of the outcomes being used. High scores before treatment, means having many points to lose (and to regress towards the mean) but is naturally also associated with still having many points after treatment. Therefore, responder should be positively associated with high intake severity, and remitter should be positively associated with low initial severity. There are of course exceptions. Rozental, Andersson, and Carlbring (2019) found the opposite; that high initial severity was in fact associated with non-responder status in an individual patient data meta-analysis of 2866 patients receiving ICBT in 29 different RCTs. This means that, in the contexts that the sample comes from, high initial severity was altogether a bad thing. This could however, be biased by type of symptom, as the sample consisted of patients with many different disorders or problems, perhaps with different quality of severity measure. For example, having an anxiety disorder, which was likely associated with coming from the studies in the sample that addressed anxiety disorders, had a much stronger predictive value for non-responder. Similarly, Furukawa et al. (2017) found that rate of improvement did not interact with baseline severity in an individual patient data meta-analysis of CBT for depression versus pill-placebo (n=509), and that result does not share the same potential sources of bias.

Table 1: Studies finding that initial severity is both negatively and positively associated with "good outcomes" depending on the outcome

Study	Disorder	*Large improvement* is more likely when initial symptoms were high	*Low symptoms after treatment* is more likely when initial symptoms were low
Catarino et al (2018)	Depression & Anxiety	X	X
Edmonds et al (2018)	Depression & Anxiety	X	
Andersson et al (2015)	OCD	X	X
El Alaoui et al (2016)	Depression	X	X
Hadjistavropoulos, Pugh, Hesser, and Andersson (2016)	Depression & Anxiety	X	
Stjerneklar, Hougaard, and Thastum (2019)	Anxiety	X	
El Alaoui, Hedman, et al. (2015)	Social anxiety	X	
Nordgreen et al. (2012)	Social anxiety	X	X
Bower et al. (2013)	Depression	X	
Mojtabai (2017)	Depression		X
Karin et al. (2018)	Depression & Anxiety	X	

Notes: OCD= Obsessive Compulsive Disorder, Stjenreklar, Hougaard and Thastrum (2019) studied adolescents, Bower et al (2013) is an individual patient-data meta-analysis.

1.4.2.2 Predictors found at the Internet Psychiatry Clinic where the studies in this thesis were conducted

El Alaoui et al. (2016) used multilevel modeling to identify predictors of rate of recovery and endpoint levels of depression in a large cohort sample from the Internet Psychiatry Clinic's Depression treatment (n=1738). They found that treatment credibility and adherence were important predictors as well as having a full time employment, both in terms of rate of recovery and endpoint symptom scores. Higher levels of depression at intake predicted higher rates of recovery but also higher endpoints (perhaps due to regression to the mean). History of psychotropic medication was a negative predictor of outcomes in terms of both rate and endpoint.

In Social phobia treatment (n=764) high levels of adherence and credibility ratings predicted faster rates of recovery whereas high over-all functioning predicted slower improvements. Family history of social phobia symptoms was a predictor of higher adherence to treatment (El Alaoui, Ljótsson, et al., 2015). In a long-term analysis of predictors in social phobia treatment (n=446), a family history of social phobia was a predictor of poor outcomes whereas high initial severity predicted higher rates of change (El Alaoui, Hedman, et al., 2015). It is possible that in the short term, identifying strongly with the descriptions of social phobia is good for treatment engagement but in the long-term, patients with strong predispositions for social anxiety tend to relapse more.

In treatment of Panic disorder, having a high age of onset and having low levels of work life impairment predicted better outcomes, whereas the previous findings that comorbid Depression or Generalized anxiety disorder would impair outcomes were not found in this context (El Alaoui et al., 2013).

1.4.3 Grouping cases and using group membership as the predictor

Another way to create predictors is to use various forms of statistical grouping methods to identify subgroups or phenotypes of individuals and then use that group membership as the predictor of outcome.

Deckersbach et al. (2016) used agglomerative hierarchical cluster analyses and k-means clustering to identify subgroups in a sample of patients with bipolar disorder. They found two distinct groups (More chronic & less chronic) that had different likelihoods of recovery. Bucholz, Hesselbrock, Heath, Kramer, and Schuckit (2000) used latent class analysis to attempt to identify subgroups or phenotypes in antisocial personality disorder. What they found was that the main feature of their classes were that each class tended to be more severe than the next rather than distinct in any other meaningful way.

Lutz et al. (2017) identified classes of change patterns in the first four weeks of ICBT using piecewise growth mixture modeling, and found that class membership increased variance explained (adj. r^2) from 21.5% to 34% over baseline predictors. The regression analysis did account for baseline severity, but it was not compared to using the first four weeks' of change

in severity as the only predictors. The predictive value of this early-pattern-of-change class-membership beyond simply adding the extra measurement points on the outcome variable is unknown.

Saunders, Cape, Fearon, and Pilling (2016) used latent profile analysis on intake data for over 16 000 outpatients with depression and anxiety, and identified eight replicable latent profiles of patients using a number of key baseline characteristics. These classes then showed significant associations with response and remission after treatment. However, baseline symptom severity on depression (PHQ-9) and anxiety (GAD-7) were among the characteristics used to build the profiles. If one only orders the profiles found according to their baseline PHQ-9 severity, and orders their outcome graph accordingly, a distinct linear slope appears. High baseline severity means that the patient is less likely to recover during treatment. That is an important predictor, but it is lost in the mix, and furthermore if there actually is an additional value of the other variables in the profiling model, we cannot see it.

While theoretically enticing, grouping approaches have some drawbacks. For one: grouping methods do not always automatically indicate how "real" a solution is (often there is no p value or simple effect size to guide the researcher). There is also a possible problem in clustering cases based on continuous variables as such variables should theoretically be evenly spread and not have clear jumps in levels, thereby creating the same problem as with dichotomizing the outcomes.

Clustering studies, as well as the more traditional prediction studies, indicate that symptom severity could be the one of the stronger starting points for making meaningful individual level predictions.

1.4.4 The big leap to big data and big models

Previous research has not been able to find strong and consistent baseline predictors that have a large impact on whether or not a patient will benefit from treatment. This could be because of two main problems: 1) we are not using analyses that are sophisticated enough to find the signal 2) we are looking in the wrong places, i.e. at irrelevant or too weak predictors. The problem could certainly be a combination of the two.

Some ways are being used to try to solve this. One way is by measuring complex things like genetic markers (Andersson et al., 2019) and brain imaging (Månsson et al., 2015). Another is by analyzing the data in more complex ways, such as using machine learning (Lenhard et al., 2018), growth mixture modeling (Lutz et al., 2017) or latent profile or class analysis (Saunders et al., 2016). There is evidence of data mining and machine learning (i.e. completely data-driven) analyses accurately predicting future occurrence of quite complex psychological phenomena such as suicidal ideation (Niculescu et al., 2015) and psychosis (Koutsouleris et al., 2009). The main drawback of the idea of measuring something complex is of course that it is time consuming and expensive. Another drawback is that it often requires a lot of preprocessing and can create overfitted models that are far less robust when samples are as small as they usually are (Flint et al., 2019). The idea of analyses that are more complex

however, might be more achievable in routine care as by now, computational power is so high compared to the complexity of the data, and that data is mostly collected and stored digitally.

There is also a third potential solution. 3) We are looking at the wrong time. Perhaps it is not possible to know if a psychotherapy will help a patient before treatment even starts, or perhaps the possibility of knowing that is not close enough within reach to be viable compared to changing *when* we look instead.

1.4.5 Moving away from baseline and into treatment

1.4.5.1 Early Change, Sudden gains or Rapid Response

A notable finding in psychotherapy research is that somewhere between 60 and 75 % of symptomatic improvements in psychotherapy happen within the first 3-4 weeks of treatments (Wilson, 1999) and that such gains can be maintained through treatment (Haas, Hill, Lambert, & Morrell, 2002). This is sometimes referred to as rapid response. Also of importance is the phenomenon of sudden gains, defined first by Tang and DeRubeis (1999) as a change between one session and the next that was:

> "… large (a) in absolute terms, (b) relative to depressive symptom severity before the gain, and (c) relative to symptom fluctuations preceding and following the gain."(Tang & DeRubeis, 1999)(p 895 §9).

They suggest that a gain should be of a certain absolute magnitude (they used 7 points on the BDI somewhat arbitrarily but one could make the argument that a Reliable Change Index score could be used) and of a certain relative magnitude (25 % of the patients previous score). They use a third criterion, which is stability, defined as the mean of the scores from the three sessions prior to the gain be significantly higher than the mean of the three sessions after the gain using simple inferential statistics. A sudden gain does not have to be early.

O'Mahen, Wilkinson, Bagnall, Richards, and Swales (2017) found that experiencing a sudden gain was associated with greater improvements during online behavioral activation treatment for perinatal women suffering from depression. Lewis, Simons, and Kim (2012) also found that change in depression over the first five sessions predicted better outcomes overall. Mulder, Joyce, Frampton, Luty, and Sullivan (2006) found that even over the course of six months, initial rapid response predicted recovery in the long term for depression. Schlagert and Hiller (2017) also found that those who had an early response to treatment for depression had better post-treatment outcomes.

There are however, some potential problems with the concept of sudden gains. For one, measuring the level of change between before, through well after, the sudden gain itself does confound the conclusions that sudden improvements are good with the fact that any improvement is good, and sudden improvements are improvements. A patient who improves a lot at some point will have improved a lot by the end of treatment, provided they do not relapse after that sudden improvement. Additionally, O'Mahen et al. (2017) found that 33% of those

with a sudden gain hade more than one sudden gain. This begs the question of how sudden a sudden gain is. Are these patients really improving suddenly or just improving steeply? This is an important distinction since the theoretical underpinnings of the sudden gains phenomena was originally, and is still the idea of a breakthrough in psychotherapy. Rapid and consistent improvements do not really signify a "breakthrough". These studies do not actually show that the early response or sudden gain is associated with any increases in improvement rates above-and-beyond that early response in itself, which had already happened. Is that what we are looking for? If we want to theorize around the idea of a breakthrough in psychotherapy, then probably not. However, if we only want to make guesses about future symptom levels, then absolutely. They show that early response seems stable and is not associated with later relapse, which would have indicated fluctuations in symptoms rather than a true improvement. In contrast, Forand and Derubeis (2013) found that high anxiety at intake was associated with rapid improvement in depression but actually not to overall improvement in depression, perhaps indicating that rapid improvements can happen without continued improvements. Most data does however indicate that an early and sudden or large improvement is not often followed by increases in symptoms.

Another problem with using sudden gains as a predictor in an Adaptive treatment strategy is that, by the definition by Tang and DeRubeis, a minimum of seven sessions needs to go by before a sudden gain can be recognized. Three sessions before and after, as well as the sudden-gain-session. Considering the length of many CBT and ICBT protocols, sudden gains is not very useful within an Adaptive treatment strategy unless the stability-criterion is changed.

Based on data from the Internet Psychiatry Clinic, Schibbye et al. (2014) aimed to clarify the general finding that early change is a predictor of outcome. General outcomes (CORE-10 and OQ-45) were compared to disorder specific outcomes (MADRS-S, LSAS-SR and PDSS-SR) in regular care ICBT patients being treated for depression, social anxiety or panic disorder. They estimated that four weeks in to treatment seemed the optimal time point in terms of balancing predictive power against the time left in treatment to have any use of the prediction. They also found that disorder specific measures outperformed general ones. Specifically, Schibbye et al. (2014) found that when predicting themselves, the OQ-45 did worse than the CORE-10, which in turn did worse than all the disorder specific measures, and that disorder specific measures can explain 34-43% of outcomes at week 4.

1.4.5.2 Treatment activity and on-boarding

Besides assessing early changes in symptoms, moving away from baseline and into the treatment itself gives us the opportunity to examine how the patient interacts with, and perceives, the treatment. This is both mostly impossible to examine before treatment starts, and at the same time strongly believed to be an important factor for success; we all <u>believe</u> that *doing* the treatment *causes* the results. One way to capture a precursor to on-boarding before treatment could be through an attitudes questionnaire about the treatment in question. Schroder et al. (2015) developed an attitudes questionnaire regarding online psychological treatments,

and ratings on this from baseline was later found to moderate outcome (Schroder, Jelinek, & Moritz, 2017). Still, most indicators of on-boarding will be from during the treatment.

There are several examples of adherence predicting good outcomes (El Alaoui, Ljótsson, et al., 2015; El Alaoui et al., 2016; Mojtabai, 2017; Nordgreen et al., 2012; Salomonsson et al., 2019), though adherence is often operationalized with a rough proxy such as number of modules completed or messages sent. In stark contrast, Kraepelien, Blom, Lindefors, Johansson, and Kaldo (2019) did a high resolution operationalization where they assessed both quantity and quality of component specific compliance and overall compliance. They found that overall compliance with treatment predicted symptom reduction in depression and worry, panic, social anxiety, stress, or insomnia in a comorbid sample being treated with a tailored ICBT-intervention. They also found that component specific compliance predicted reduction in targeted symptoms in social anxiety, stress, and insomnia, but not for stress and pain, while component specific compliance with stress and insomnia components actually predicted further reductions in depression. Thus, in some conditions, doing the specific interventions that are believed to be directly targeting the mechanism of change is driving the improvements whereas in other conditions, such as depression, just "doing treatment" may be as effective as doing behavioral activation specifically according to the findings by Kraepelien et al. (2019). Regardless, whether or not a patient is doing the treatment as prescribed, or at least being actively engaged with the treatment should be an important predictor, at least theoretically. Conversely, Stjerneklar et al. (2019) found that none of their proposed therapy process variables (such as, alliance, number and length of contacts etc.) was associated with outcome.

A basic problem with adherence or process-variables in psychotherapy research has to do with timing. Adherence has often been measured at post treatment, simultaneously with outcomes, meaning that even a strong association would not be able to provide insight into causality, or be used as predictors for new patients. Completers may improve, but improvers may also complete.

1.4.5.3 Deciding the timing of outcome predictions

There are several factors to take into account when choosing a time-point for the prediction of outcome if the goal is to intervene with at-risk patients. If predictions are made too soon, they will be weak, whereas if they are made too late, they will be useless, as there is no time left to intervene.

In most prediction models (i.e. where only baseline parameters are used), the value of adding a parameter can be mathematically justified if it significantly increases predictive power, which the researcher must weigh against the risk of overfitting and the burden of collecting the data. In the context of this thesis however, added parameters sometimes represents time lost in treatment and therefore a decrease in time left to help the patient. Schibbye et al. (2014) found a decrease in variance explained by their model after week 4, which was a sound reason to make prediction at week 4 rather than later. However, that decrease may have been a chance finding. If predictive power simply increases in a linear fashion, then using that as a guide for

deciding when to predict is only helpful if combined with some form of minimally acceptable accuracy in order to identify the earliest possible time-point for a meaningful prediction.

When predicting how a patient will fare within a certain treatment, one could take into account whether or not the patient has had time to be exposed to said treatment. One way of doing this is to make sure that you allow for all patients working at the prescribed pace to be exposed to the treatment component that is assumed to be the main mechanism of change. Most protocols include some basic orientation, rational and goal setting in the very beginning of treatment, so the prediction should be made after this is out of the way. How many weeks after into treatment this will be will vary from protocol to protocol, but will likely be a few weeks into treatment. It is also important that there is adequate time left in treatment to intervene once a patient is identified as at-risk. If the key component of the treatment takes several weeks to work, then there must several weeks left in treatment when predictions are made.

Another important consideration is how long unsuccessful patients stay in treatment before they drop out. That is, if there is a substantial increase in patients dropping out at a certain point in treatment, then one can argue that prediction needs to happen before this, to enable clinicians to intervene before it is too late.

1.5 HOW ACCURATE DO PREDICTIONS NEED TO BE WITHIN AN ADAPTIVE TREATMENT STRATEGY?

1.5.1 Predictability of psychotherapy and ideal levels of accuracy

There is very little to go on when assessing if the accuracy obtained by a classifier is good or bad. Better than chance is a very low bar to set, and perfect is probably impossible not only from a computational point of view, but also from a theoretical point of view. How much of the outcome of the treatment of a patient do we really believe is already set in stone less than halfway through that treatment? The diversity of all of the things that might happen between the time a patient starts treatment and the outcome is measured at the end of it is too large for it to be reasonable to assume that we could get anywhere near 100% accurate in our predictions. Especially in ICBT, where there is a huge gap between the supposed mechanisms of the treatment and what is actually being delivered to the patient. For instance, we believe that a patient suffering from depression will improve if they increase their daily positively reinforced activities, reduce negatively reinforced activities and ascribe less credence to their negative thoughts. We then try to achieve this by having the patient read about it, hoping they will a) believe us and b) act accordingly. Of course, there are infinite ways these scenarios can play out, not to mention that the patient can also improve or deteriorate regardless of the treatment (Bruce et al., 2005). How could we possibly know for sure what the end will look like a third, of half way through the therapy? Thus, an observed accuracy somewhere in the 90-100% range should be viewed with skepticism, as it is probably a computational error or incorrectly designed model, much like if you find a person that is three meters tall, it is more likely that you measured that person incorrectly, than it is that the person is actually three meters tall. For example, Hannan et al. (2005) state in the abstract that their lab-test identified 100% of the

patients who would deteriorate. This is however only equivalent to 100% sensitivity and not at all analogous with 100% accuracy. The balanced accuracy of that lab test was actually about 65% based on the confusion matrix in Table 1 in that article.

1.5.2 The Decision-Threshold Approach in clinical decision making

In medical research, some attempts have been made to define criteria or principles for deciding to prescribe treatments, to go on with further testing, or not offer any health care. This is based largely on simple probability theory. A patient may or may not have a disease, and there are associated benefits and risks with 1) running additional tests, 2) treating a disease that isn't there and, 3) not treating a disease that is there. In the late 70's, Pauker and Kassirer (Pauker & Kassirer, 1975; Pauker & Kassirer 1980) attempted to formalize the decision making process in medical treatment. They suggest a "decision-threshold approach" where a dichotomous decision should be made based on whether a threshold probability of the disease has been reached. They define the "indifference point" as the point in probability where the expected value of treatment is the same as not treating. They suggest the formula $T = 1/(B/C +1)$ where T is the threshold probability and B is the benefit of treating a diseased patient minus treating a non-diseased person, and C is the cost of not treating a diseased patient minus the cost of not treating a non-diseased person. Cost here refers to potential harm, but therapist time or monetary expense can also be taken into account. I believe that psychotherapy research is quite far away from being able to apply this reasoning. This is because we, as mentioned before do not really know much about the harms of treating (Rozental et al., 2014) or of not treating (though, we do know that waiting is in itself bad (Cunningham, Kypri, & McCambridge, 2013; Furukawa et al., 2014) and, rarely actually have several treatment options available. Instead, we have the option of attempting to treat now with the treatment we have or putting the patient on a waitlist (i.e. sending a referral) for maybe getting a more high intensity treatment somewhere else, probably much later.

Another factor to consider is how accurate the clinician wants the prediction to be before they are willing to act. Eisenberg and Hershey (1983) conducted an empirical study where clinicians where asked to decide on treatment, further tests, or neither based on differing ranges of probabilities that a patient had the disease. The best estimation of this range found in their diverse sample of patients and diseases being assessed lied in the 65-70% range (Eisenberg & Hershey, 1983) (i.e. clinicians want to be about 65-70% sure that the treatment is appropriate in order to prescribe it). These results, although preliminary and dated, give some idea about what should be the minimal accuracy required from a classification algorithm used for deciding on offering or denying treatment. On the other hand, those ratings were from a range of tests and treatments of different diseases. As such, many of those tests were likely more harmful, if given needlessly, than we assume psychotherapy to be if intensified needlessly. For instance, giving chemotherapy to a patient without cancer is probably worse than if a therapist decides to give extra attention to a patient who did not need it. Therefore, in psychotherapy we might dare to be a bit more generous with intensifying treatments as long as the costs in terms of therapist time are not too great.

1.5.3 High vs low-risk decisions

One can argue that the types of clinical decisions that we want to make should affect the certainty that we try to achieve in order to make such a decision. For instance, it would be far bolder to suggest that a non-gold-standard treatment is a *better choice* for a specific patient than to suggest that a specific patient needs extra help in an ongoing treatment. However, we must still ascertain that the cost of intensifying treatment has some value (i.e. that treatment effects are in fact enhanced by more time spent). This poses a very salient problem with patient-to-treatment matching using predictions. Namely, what is the alternative? Are there more than one treatment available to the patients, from which to choose? Can the models tell us that the patient would be better off not getting any treatment at all rather than getting an efficacious treatment for which they seem unsuitable? In this thesis, we focus entirely on what I would argue is a very low-risk decision, namely allocating some extra resources to patients who seem to be struggling with an ongoing, gold-standard treatment.

1.6 THE ADVANTAGES OF ICBT FOR RESEARCH ON PREDICTION AND ADAPTIVE TREATMENT STRATEGIES

There are a few commonly cited advantages of ICBT over CBT and other evidence based treatments. One is time spent by therapists where, generally, a therapy session is 45-60 minutes whereas the average ICBT therapist spends markedly less time per week on a patient (Andersson et al., 2008). This is especially important since there is a general lack of qualified CBT therapists, making efficiency critical not only to save money but to make treatment accessible at all (Larsson, 2009). This makes guided ICBT a natural early step in a stepped care model (i.e. evidence based and low intensity from the care providers perspective; Bower & Gilbody, 2005; Gyani et al., 2013; Mohr et al., 2019; Richards et al., 2012) but could also evolve into an ideal area for Accelerated Care. The format also means that there is variation between the time spent by therapists on their different patients. Patients who are easy going maybe only get five minutes a week, whereas patients who are struggling might get up to 40 minutes. It is hard to imagine a therapist in face-to-face treatment sending their patient home after five minutes because they are doing so well in treatment and giving the saved up 40 extra minutes to their most difficult patient. In ICBT, this is possible!

ICBT also has an advantage when it comes to quantitative analyses since ICBT, even in routine care, tends to have highly structured formats and contents, as well as a myriad of structured data, collected from both patients and therapists. ICBT also has the potential of including automated measures, assessments and even predictions since all the data being generated is already electronic and contained within the treatment database.

2 THE AIMS OF THE THESIS

The aims of the thesis were:

a) To establish a proof of concept for an Adaptive Treatment Strategy in ICBT by
    1) Providing evidence of successful predictions of treatment Failure in ongoing treatments
    2) Providing evidence of successfully implemented Adaptive Treatment through a randomized controlled experiment where half of those predicted to fail receive additional support and adaptations by their therapits in order to prevent the predicted Failure
b) To establish three, much needed, baselines for the accuracy of outcome predictions in Adaptive Treatment Strategies
    1) How accurate are therapists themselves if asked to make these kinds of predictions?
    2) How soon in treatment could we make clinically acceptable predicitons based solely on weekly symptom measures and simple statistical models?
    3) What is the minimal clinically supported accuracy for predicting treatment Failure in an Adaptive Treatment Strategy?

3 THE EMPIRICAL STUDIES

3.1 METHODS

This section will cover some of the methods used across more than one study in this thesis.

3.1.1 Participants, recruitment and assessments

While Study I and Study II report on a randomized controlled trial whereas Studies 3 and 4 report data from regular care at the Internet Psychiatry Clinic, in Stockholm, Sweden (the clinic), the recruitment and assessment procedures were still essentially the same across all four studies. Patients self-refer by filling out an online questionnaire at the Internet Psychiatry Clinic's public website, after which they are invited to a face-to-face assessment at the clinic. The assessment is usually conducted by a specialist physician in training, supervised by a psychiatrist. It includes psychiatric history, and a MINI psychiatric interview (Sheehan et al., 1997), current medications and assessment of inclusion and exclusion criteria, which are specified in each paper. All participants were adults, fluent in Swedish, able and willing to partake actively in the treatment during the coming months and did not have a severe psychiatric or somatic comorbidity that needed to be given precedence.

3.1.2 The Internet-delivered Cognitive Behavior Therapies

In total, there are four different treatments used in this thesis, all of which are now implemented in regular care at the Internet Psychiatry Clinic. They are all diagnosis specific and target one (1) diagnosis each: In studies 1 and 2, Insomnia is treated in the context of an RCT and in

studies 3 and 4; Major depressive disorder, Social anxiety disorder and Panic disorder are treated in regular care. All of the treatments are delivered via the same internet platform. They are all based on written texts, worksheets and questionnaires that are divided into modules (chapters). These modules are usually meant to take one week to complete, but the therapist is always the one that unlocks the next module whenever a patient has sent in a homework report and received feedback. Patients have a specific psychologist assigned to them, who monitors and communicates with the patient via written text messages in an e-mail-like system within the platform. Communication is asynchronous, and therapists respond within 48 hours. Patient fill out weekly online assessments that includes questions about suicidal ideation. If a patient reports suicidal ideation, they are flagged within the system and is contacted via telephone as soon as possible for assessment and if needed, crisis referral.

3.1.3 Symptom measures

For Insomnia, Primary outcome was the Insomnia Severity Index (ISI), a 7 item scale ranging from 0 to 28 points (Bastien, Vallières, & Morin, 2001). The scale is reliable and sensitive to change (Bastien et al., 2001; Morin, Belleville, Belanger, & Ivers, 2011). Response was defined as a decrease in ISI score of at least 8 points rather than using the RCI as this is a commonly used convention for this measure (Morin et al., 2011). The cut-off for remission was less than 8 points (Morin et al., 2011).

For Depression, we used the Montgomery-Åsberg Depression Rating Scale Self-report version (MADRS-S; Montgomery & Asberg, 1979; Svanborg & Asberg, 1994, 2001). Scores range from 0-54 points. Test-retest reliability is high (ICC=.78; Fantino & Moore, 2009) and the RCI in our sample comes out at 8 or more points change and the cut-off for remission is a score of 10 or less at post treatment (Fantino & Moore, 2009).

For Social anxiety disorder, we used the Leibowitz Social Anxiety Scale-Self report version (LSAS-SR; Baker, Heinrichs, Kim, & Hofmann, 2002; Fresco et al., 2001). Scores range from 0 to 144 points. Test-retest reliability is high (r=.83), but there is high variation in the sample and the range of the measure is large, making the RCI 28 or more points change. We used a score of 34 or less at post treatment as the remission cut-off (von Glischinski et al., 2018).

For Panic disorder, we use the Panic Disorder Symptom Scale-Self Report (PDSS-SR). PDSS-SR has 7 items and scores range from 0 to 28 points. Test-retest reliability is high (ICC=.81; Houck, Spiegel, Shear, & Rucci, 2002) and the RCI came to 6 or more points. A score of 6 or less was used for defining remission (Monkul et al., 2004).

3.1.4 Definition of Success and Failure

In this thesis, the criteria for a successful treatment is that a patient's post-treatment symptom scores in their self-rated primary outcome measure **EITHER** falls below the clinical threshold for remission **OR** that the reduction in symptoms is of a certain magnitude. In studies 1 and 2, this reduction is 8 points on the Insomnia Severity Index and in Study III this is a 50% reduction compared to pre-treatment. In Study IV, both the RCI-criterion and the 50% reduction-criterion are tested, and the criteria for remission for each diagnosis, but Success/Failure is not defined because therapists did not answer that specific question.

3.1.5 Classification accuracy measure

Table 2: Confusion matrix

		Actual	
		Failure	Success
Predicted	Failure	True Positive	False Positive
	Success	False Negative	True Negative

To understand classification accuracy, we must consider the aptly named confusion matrix (Table 2). If the outcome and the prediction are both dichotomous/binary (Yes/No), then there are four possible outcomes, illustrated in a 2x2 table (Table 2). If a classifier predicts that the patient will fail and then the patient actually does fail, that is a True positive. What is Positive or Negative is completely dependent of what outcome is considered the one (1) in the equation. In this thesis, Failure is always the one (1), so even though Failure is a bad thing it is a Positive in the analyses.

Accuracy, defined as the ratio of correct classifications, is a commonly used performance metric in classification analyses, but it is misleading as it does not account for classes of differing sizes. For instance if Deterioration happens in 5% of the sample, then never predicting Deterioration will yield a 95% accuracy, even though you miss all actual cases of Deterioration.

$$Accuracy = \frac{True\ Positives + True\ Negatives}{True\ Positives + False\ Positives + True\ Negatives + False\ Negatives}$$

The primary accuracy measure used in this study is instead Balanced Accuracy (BACC), which does account for the distribution of the classes.

$$Balanced\ Accuracy = \left(\frac{True\ Positives}{True\ Positives + False\ Positives} + \frac{True\ Negatives}{True\ Negatives + False\ Negatives}\right)/2$$

Balanced accuracy ranges from 0 to 1 where .50 is completely random and 1 is perfect. This is used because a) it is a single value, b) it is expressible as a percentage, c) it is clinically easy to

understand, and d) it compensates for uneven class distribution. If using the same example as with Accuracy above, the Balanced Accuracy would be 47.5% rather than 95%, which is more fare considering, again that not a single deteriorating patient was found. If for instance Failure happens exactly 50% of the time, then Accuracy and Balanced Accuracy will be identical and both are valid measures.

3.1.6 Ethical considerations

All studies were approved by the Regional Ethics Board in Stockholm, Sweden. In Study I and II, patients provide written informed consent and in Study III and IV were data from regular care was used, patients are informed that their results may be used for research purposes and are given the option to opt-out.

In Study I, we identified patients who we believed would have a poor treatment outcome, and then we did nothing in half of those cases. Doing this is ethically problematic. We still considered this defensible for a number of reasons: First, we did not know that our classifier was correct and that those predicted to fail really would fail, nor did we know that the adapted treatment would help the patient if provided, and the only way to answer that question was through performing this experiment. Additionally, we did not remove or deny anything that would otherwise have been given, but gave all patients gold-standard treatments and level of care for their Insomnia. In fact, at least when the trial was initiated, evidence based treatment for Insomnia was largely unavailable to the public in Sweden, so conducting the study was overall likely a positive thing for persons suffering for insomnia and even for patients in the Red-Standard group (i.e. predicted to fail, but not given adapted treatment).

Therapists in Study IV were of course informed of what was going on when they made their predictions, but not asked to actively consent as the study was deemed part of their duties to develop and improve the clinical procedures that we use. However, in consideration of their integrity, we did not perform any analyses on treatment effects for different therapists or prediction accuracy for individual therapists, but only on their confidence in their predictions, and therapists identities are masked in the manuscript. All other tests were on group level.

3.2 STUDY I: PROOF OF CONCEPT FOR AN ADAPTIVE TREATMENT STRATEGY IN ICBT: A SINGLE-BLIND RCT WITH INSOMNIA PATIENTS

3.2.1 Aim and hypothesis

In this study, we aimed to provide a proof of concept for an Adaptive Treatment Strategy in ICBT-i. The Adaptive Treatment Strategy had two parts: a) predict which patients were likely to fail treatment and b) adapt the treatment for the patients predicted to fail.

3.2.2 Methods

The Adaptive Treatment Strategy was tested in a randomized controlled trial, where a prediction was performed for all patients, and the patients predicted to fail were randomized to continued standard treatment, or to an adapted treatment. The study included 251 adult patients with diagnosed Insomnia and at least 11 points on the Insomnia Severity Index who started ICBT-i with standard treatment for three weeks. During week 4, a semi-automated classification algorithm was used to classify patients as Green (likely Success) or Red (likely Failure). Green patients continued with standard treatment, whereas Red patients were randomized either to continue the Standard treatment or to receive an Adapted treatment for the remainder of the treatment period. The adaptation of the treatment started with a telephone assessment to analyze the patients' difficulties and needs, after which the therapist presented the patient with a plan for the remainder of the treatment period. Adaptations were strictly Insomnia focused but could otherwise entail anything from extra telephone calls or in-person appointments, to reordering modules or sending daily SMS reminders.

The primary outcome was change over time in insomnia symptoms measured with the Insomnia Severity Index. Two latent growth curve analyses were performed, the first assessing the difference between Green and Red-Standard groups during the whole treatment, and the second assessing the difference between the Red-Standard and Red-Adapted groups during the time after randomization. Secondary outcomes were response, remission, deterioration and, of course, Failure.

3.2.3 Results

About 40% of patients were classified as Red (102 Red and 149 Green) and 51 of these received the Adapted treatment. On average, the adapted treatment took an additional 14 minutes per week for the therapist to deliver for the remaining five weeks of treatment. This includes the initial telephone assessment (usually about 30 minutes). Green patients were more satisfied with the treatment ($p<.001$) but the Red groups were similarly satisfied and the overall satisfaction was good to excellent.

Green patients improved more than did Red-Standard patients ($p<.001$), but not more than Red-Adapted patients once their treatment had been adapted. Once treatment was adapted, Red-Adapted patients improved significantly more than Red-Standard patients ($p<.001$) and actually even more than Green patients during that period ($p<.001$). Fewer Red-Adapted

patients than Red-Standard patients became Failed treatments (37 % vs 64 % failures, OR=.33, $p<.01$).

3.2.4 Conclusions

We found proof for the concept of an Adaptive Treatment Strategy in ICBT-I since a) treatment Failures could be predicted and b) predicted treatment failures could be avoided if treatment was adapted. We also found that this adaptation did not require an inordinate amount of time from the therapist. However, not all those predicted to succeed actually did, nor did all those predicted to fail actually do so, and nor did all those who received adapted treatment actually benefit. This means that predictions were far from perfect and that the adaptations used in this study could not always avoid a poor treatment outcome.

3.3 STUDY II: CLINICALLY SUFFICIENT CLASSIFICATION ACCURACY AND KEY PREDICTORS OF TREATMENT FAILURE FROM THE RCT

3.3.1 Aim and hypothesis

Aim to establish a clinically sufficient level of accuracy by performing an in-depth analysis of the classification algorithm used in the RCT (Study I), and to explore if alternative classifiers that would be easier to implement might be as accurate or better. We also wanted to explore the 21 input predictors used in the RCT-algorithm to see if, and how strongly, they really relate to Failure and if they could be combined in ways that make predictions stronger or require fewer predictors to reach the same results.

3.3.2 Methods

The algorithm in the RCT (Study I) classified 149 patients as Green and 102 patients as Red, out of which 51 were randomized to adapted treatment. The 51 patients randomized to adapted treatment were manipulated after classification and cannot be included in this study. The 200 patients included in this study (149 Green + 51 Red) should all have received the same level of care in the same treatment. If the classifier was 100% accurate, all 149 Green patients and none of the 51 Red patients should have a successful treatment outcome.

The RCT-classifier had several separate steps before the final classification which were all evaluated using the Balanced accuracy and compared to two alternative classifiers that would be simpler to implement, but had not been validated in the RCT. One alternative classifier used only the patient ratings from the RCT-classifier and the other instead used linear regression and symptom measures from the first three weeks of treatment (that is the same method as we used in Study III).

The 21 predictors were first examined with correlations to Failure, and then used in four logistic regression models to predict Failure (the absence of both responder and remitter status). The first model used all patient rating from the RCT-classifier (12 predictors). The second model used only the clinician ratings (9 predictors). The third model used all predictors (patient +

clinician ratings = 21 predictors). After that, fourth model was built using only predictors that had a significant correlation to the outcome Failure (11 predictors), which should be more parsimonious and require less data collection. For comparison, we then built a model using only the observations on the ISI from screening until week three (the same type of model as in Study III).

3.3.3 Results

The Balanced Accuracy of the RCT-classifier was 67% (95% CI: .61-.74), whereas the prospective alternative classifiers produced a Balanced accuracy of 62% (95% Confidence intervals did however overlap). Out of the 21 predictors, only eleven correlated significantly with Failure. The logistic regression model using all predictors could explain 56% of the outcome variance, whereas the various simpler models explained between 16-47%. Important predictors were patient rated stress from baseline, treatment credibility from week 3, depression change by the third week, and insomnia symptoms at week 3 as well as clinician rated attitudes towards homework and sleep medication that were made after week 3.

3.3.4 Conclusions

We found that a Balanced Accuracy of 67% could serve as a preliminary benchmark for clinical usefulness of predictions in an Adaptive Treatment Strategy. Simpler models, using less data, generally performed slightly worse, but do have the advantage of being easier to implement, and should be developed further, perhaps with more advanced statistical models. Most of the important predictors were collected during rather than before treatment, highlighting the advantage of letting patients start treatment and monitor early progress. Relevant predictors were, of course symptom severity, but also depressive symptoms, and various markers of treatment engagement such as early activity, attitudes towards the rational of treatment homework and treatment credibility ratings, indicating that patients who are not on-board with the treatment after three weeks probably need extra support in order to benefit.

3.4 STUDY III: PREDICTING TREATMENT FAILURE IN REGULAR CARE ICBT FOR DEPRESSION AND ANXIETY USING WEEKLY SYMPTOM MEASURES

3.4.1 Aim and hypothesis

We aimed to assess at what treatment week a clinically acceptable level of accuracy that clinicians would be willing to act on could be reached using only basic statistics and weekly symptom measurements. This could then be used for prediction, as it is simple to implement, as well as serve as a benchmark for more complicated prediction paradigms, which must be better than this one to be worth extensive additional data collection and/or complex data management and modeling procedures.

3.4.2 Methods

Data from the Internet Psychiatry Clinics three main treatments, Depression (n=2052), Social anxiety disorder (n=1103), and Panic disorder (n=1155) was used. Patient rated symptom scores on their respective primary outcome measure (MADRS-S for Depression, LSAS-SR for Social anxiety disorder, and PDSS-SR for Panic disorder) was used as predictors of post-treatment symptom scores in a series of multiple linear regression models. Models were built consecutively, adding one weekly observation at a time (i.e. Screening, Screening + Pre, Screening + Pre + Week 1 etc.).

In the first step, we compared using the raw sums of the measures as the predictors compared to using a change score each week (calculated from Screening) to see which could explain more end-point variance for each treatment. Then, the best version for each treatment was used to predict final symptom scores for each patient in hold-out test-samples consisting of random subsets of patient that were left out of the model building datasets to be used later for validation (Depression, n=390, Social Anxiety, n=281, Panic, n=292).

The predicted score was then compared to cutoffs for remission and response and calculated into a predicted Failure (neither remitter nor responder) or success. Predicted Failure was compared to the observed outcome for each patient, resulting in a Balanced accuracy, which was compared to pure chance. As a benchmark for Clinical acceptability, we used information from an old experiment by (Eisenberg & Hershey, 1983) where they essentially found that clinicians became more likely to act according to a prediction once it became at least 65% accurate. We compared against this benchmark by declaring that the lower bound of a 95% confidence interval around the Balanced accuracy should start on at least 65%.

3.4.3 Results

On average, half of the 4310 treatments were Failures according to our definitions (50.5% in Depression, 77.3% in Social anxiety disorder and 24.3% in Panic disorder), whereas only just over 4% deteriorated, which is less than in previous research. The Balanced accuracy indicated that predictions were better than chance even before treatment started. The Clinical Acceptability benchmark (95% CI min $\geq$ 65%) was reached at week five for Social anxiety disorder (95% CI low bound=.67) while Depression and Panic disorder were only close at that time (95% CI low bound =.645 and .642 respectively). By week six, Depression and Panic disorder reached the benchmark (95% CI BACC low bound = .66 & .65). Point estimates of Balanced accuracy at week 6 varied between 72-75%.

3.4.4 Conclusions

Failure can be predicted with a clinically acceptable level of accuracy (65% or above) after 5-6 weeks in 12 week long ICBT for Depression, Social anxiety disorder and Panic disorder using only weekly measurements on the respective primary outcome measure and a basic linear multiple regression model. Predictions quickly become better than chance but further increases

are not very big using this model. Further improvement might require using more data as predictors, more sophisticated models, or both.

3.5 STUDY IV: ACCURACY OF THERAPISTS' PREDICTIONS IN ICBT FOR DEPRESSION AND ANXIETY IN ROUTINE PSYCHIATRIC CARE

3.5.1 Aims and hypothesis

We aimed to assess the accuracy of ICBT-therapists own guesses when asked to make predictions on several relevant clinical outcomes, both categorical and continuous, and see if they are overly optimistic when making predictions. We also wanted to see if they differ in how confident they are in such predictions and if confidence in the prediction relates to how likely it is to be correct.

3.5.2 Methods

During about one year, fourteen different therapists at the Internet Psychiatry Clinic filled out a prediction questionnaire during the fourth week of treatment for all of their patients, which was in total 897 consecutive patients. The predictions were made within a questionnaire in the treatment platform at the clinic. Therapists answered a number of questions such as:

Will the patient be "cured" from his/her [depression, social phobia, panic disorder] - that is to say do you think that the patient's symptoms post treatment will be at a level comparable to a person without [depression, social phobia, panic disorder]?

☐ *Yes*
☐ *No*

How do you think the patient will have changed in their [depression/social anxiety/panic disorder] measured with [measure] from pre-measurement to post-measurement?

☐ *Deteriorated in a clinically meaningful manner*
☐ *No change so large that it can be considered clinically meaningful*
☐ *Improved in a clinically meaningful manner*

Exactly how many points on [symptom measure] do you guess the patient will be improved or changed from pre-measurement to post-measurement?

☐ *Improved, number of points:_____*
☐ *Deteriorated, number of points:_____*

Remission was defined as a post-treatment score below the clinical cutoff (Section 3.1.3), and responder was defined in two ways, using the Reliable Change Index (RCI) and using 50% reduction as a criterion. Deterioration was defined using the RCI*.84 rather than 1.96 and non-response was the absence of all of these. Therapists' predictions were then compared with the observed outcomes for these patients using confusion matrix statistics and the Balanced accuracy for categorical outcomes and simple regression for continuous outcomes. Categorical

predictions were compared to two simple benchmarks, apart from comparing to pure chance (i.e. 50%): the Clinically acceptable benchmark that we derived from (Eisenberg & Hershey, 1983) and used in Study III, as well as the Balanced accuracies of a predictive statistical model using symptom measures and multiple regression (those observations are from the appendix for Study III).

3.5.3 Results

For the categorical outcomes, therapists were more accurate than chance when predicting remitters. For responders (both RCI and 50%), they were better than chance in Depression and Social anxiety disorder. Overall, therapists were on average 9.5% less accurate than statistical models using only symptom ratings across remission and response-predictions. Amongst predictions that could be compared to the statistical benchmark based on Study III, the point estimates of Balanced accuracy of the therapists ranged from 53-73%, whereas the statistical benchmark had point estimates ranging from 66-81%. The clinical acceptance benchmark of a 95% confidence interval starting at 65% Balanced accuracy was only reached by therapists for predicting remission in social anxiety treatment, but was reached by the statistical model in all treatments. Deterioration was never correctly predicted by therapists.

For continuous symptom outcomes, therapists' predictions could explain 13%, 16 % and 18 % of the variance in the outcome in Depression, Panic disorder and Social anxiety disorder respectively which can be compared to previous findings with simple regression explaining 41%, 34% and 43% of the variance respectively for the same treatments at the same clinic (Schibbye et al., 2014).

Regarding optimism, therapists predicted positive categorical outcomes about twice as often as they occurred, with one notable exception being remission in Social anxiety disorder, which was predicted as often as it occurs. Additionally, they predicted larger mean changes in continuous outcomes than were observed, though again not for Social anxiety disorder.

Therapist varied quite a lot in how confident they were in their predictions. Both within therapists and between therapists, with some therapists being consistently very confident or doubtful while others had hugely different confidence levels between their different patients. How confident the therapist was in their prediction did however not correlate with how correct they were neither for categorical nor continuous outcomes.

3.5.4 Conclusions

Therapists delivering ICBT can often predict outcomes for their patients better than chance but most likely not as accurately as statistical models. As has been found in other contexts, therapists are overly optimistic about the prognosis their patients. ICBT therapists differ in how confident they are when making predictions, but this does not seem to be related to how correct those predictions are. Whether therapists are better at categorical or continuous predictions is still unclear, though their predictions of categorical outcomes were closer to statistical models.

3.6 SUMMARY TABLE REGARDING PREDICTING FAILURE DURING WEEK 4

We decided to make predictions during week 4 in Study I (and II) based on previous research and clinical judgement regarding time left in treatment vs. predictive accuracy. Table 3 summarizes what we find in this thesis regarding predicting Failure during week 4 in treatment (i.e. using data collected during week 3 or earlier), and how this relates to the two benchmarks we used/created in this thesis. The Clinically sufficient benchmark is what we found in Study II. Therefore, we have access to a 95% CI around the Balanced accuracy to use instead of the more arbitrary decision that the low bound of a CI should exceed the point estimate required by the benchmark as we do with the Clinically acceptable benchmark in Study III. For one thing, doing that would mean that Study II both creates and then fails to meet the Clinically sufficient benchmark.

Table 3: Balanced accuracy when predicting Failure by Week 4

Study and disorder	Method for prediction	BACC	CI Low	CI High	Clinical acceptability benchmark (CI low ≥65%)	Clinically sufficient benchmark (CI low ≥61%)
Study I & II Insomnia	Multicomponent classification algorithm	67%	61	74	No	Yes
Study I & II Insomnia	If automated (all patient ratings)	62%	55	69	No	No
Study I & II Insomnia	Linear regression with only ISI	62%	55	69	No	No
Study III MDD	Linear regression with only MADRS-S	67%	62	72	No (not until week 6)	Yes
Study III SAD	Linear regression with only LSAS-SR	70%	64	75	No (not until week 5)	Yes
Study III PD	Linear regression with only PDSS-SR	63%	57	69	No (not until week 6)	No

Notes: BACC= Balanced accuracy, MDD= Major depressive disorder, SAD= Social anxiety disorder, PD= Panic disorder, ISI= Insomnia Severity Index, MADRS-S= Montgomery-Åsberg Depression Rating Scale-Self report, LSAS-SR= Liebowitz Social Anxiety Scale-Self report, PDSS-SR= Panic Disorder Severity Scale- Self Report

4 DISCUSSION

This thesis had two main aims. Firstly, to establish a proof of concept for an Adaptive Treatment Strategy in ICBT where failing treatment attempts are identified early and treatment is then adapted in order to avoid the predicted failure. Secondly, to establish some preliminary minimal benchmarks regarding the timing and accuracy of such outcome predictions. These were the following: How accurate is an empirically supported classification algorithm and are all of its constituent parts important? How early in treatment can an accuracy that might be accepted by a therapist be reached with very basic prediction procedures? How accurate are therapists themselves are when making such predictions?

4.1 MAIN FINDINGS

In Study I, we found that once adaptations had started, Red patients receiving adapted treatment improved significantly more than those Red patients who did not receive adapted treatment, and in fact even more than Green patients, who had already improved a lot more by the time of randomization. The odds of Failure amongst those receiving adapted treatment was only a third of the odds amongst those predicted to fail who did not get adapted treatment, and the proportion of failures amongst Red patients was close to be cut in half by adding the adapted treatment. The Green group were still more likely to become remitters than Red patients with adapted treatment, but in other respects the groups were equal.

This proof of concept was still only in Insomnia treatment. Insomnia is a relatively specific problem where we have relatively substantial evidence that engaging in the proposed mechanism of change (i.e. sleep restriction and stimulus control) produces positive outcomes (Blom, Jernelöv, et al., 2015; Harvey, Dong, Bélanger, & Morin, 2017; Kaldo, Ramnerö, & Jernelöv, 2015; Kraepelien et al., 2019). This could make it easier for therapists to make adapted treatment efficient. In treatment of depression or anxiety, it is perhaps more likely that the adapted treatment diverts on a tangent from the main focus of the therapy and thus produces less convincing results, while perhaps being helpful in an area that we are not measuring as the main outcome, though that remains to be tested.

We still do not know if all Red cases could have been helped, if the adapted treatment had been better. It is unlikely that every patient will benefit from treatment even with perfect predictions and exceptional treatment adaptations. For one thing, this would imply that every patient was assessed perfectly in the first place, which is unlikely.

I mention in the introduction that before one considers an Adaptive treatment strategy, it is important that the average effect of the treatment is acceptable. If patients do not do well overall, then the treatment need to be changed overall, not just for some. Was the treatment in the RCT good enough overall to begin with? This was not directly examined, but by averaging the within group effect sizes, we can get an idea of this. The within group Cohen's *d* was 1.24 for the Red-Standard group and 2.05 for the Green group. Given that there should have been twice as many Red-Standard patients we can take (1.24+1.24+2.05)/3 and get a quite

speculative within group effect size of 1.51 for the whole group if no-one had received the adapted treatment. This is well within the bounds of a large effect size. Similarly, if we account for the missing Red-Adapted group and assume that as many of those would have Failed treatment we would get 32+32+34 Failures (Table 3 in Study I) out of 251 patients, which means a 61% Success rate for the treatment if no-one had received the adapted treatment. That is certainly good enough according to current standards. This indicates that we did not need to overhaul the entire treatment, but that the treatment was good enough on it's own to warrant the use of an Adaptive Treatment Strategy instead. The reason this was not examined as part of Study I is that we already had good reason to believe that the treatment was good enough based on previous RCTs of the same treatment (Blom et al., 2016; Blom, Tillgren, et al., 2015; Kaldo, Jernelöv, et al., 2015).

Let us speculate even further and ask how good the treatment might be if we implemented this and every Red patient receives the adapted treatment. Using the same, slightly flawed, logic as above we would get an effect size of Cohen's *d = 1.81* and a 72 % success rate for the treatment overall. An effect size increase of .30 and an 11 % reduction in treatment Failures overall is quite good considering it only took another 14 minutes per remaining week per Red patient, and only 40% become Red. That means less than six minutes of extra work per patient-week for the remaining weeks across all patients, or just under four more minutes of work per treatment week averaged on all patients and the whole treatment.

In Study II, we found that a Balanced Accuracy of 67 % is potentially clinically sufficient for predicting treatment failure in an Adaptive treatment strategy based on it having been used in a fruitful way in a randomized trial of an Adaptive treatment strategy for ICBT-i. Each step in the RCT-classifier seems to function well enough not to warrant omission of steps in future applications of the algorithm. The simpler alternative classifiers did not do quite as well as the original RCT-classifier. However, they would require far less input data.

The alternative classifier using only symptom scores was close to being as accurate as the full RCT-version but in the logistic regressions, using only symptom scores seemed markedly worse than using more of the carefully selected predictors used in the trial. Ideally, symptom scores and trajectories can be combined with more process-related variables in even more fruitful ways. The most influential predictors of Failure was symptom severity and psychological distress on one hand and several indicators of treatment engagement on the other. I will now discuss the various indicators of engagement or on-boarding with treatment, as that is less obviously connected to the outcome than symptom severity is.

Our finding that alliance was correlated with positive outcome is in line with previous research (Flückiger, Del Re, Wampold, & Horvath, 2018). However, it is important to note that working alliance did not remain a significant predictor in any of the logistic regression models when other, potentially more specific, factors such as attitudes to CBT homework, willingness to taper sleep medication and treatment credibility ratings were controlled for. This could also reflect that the working alliance inventory measures things that may be less salient in ICBT, such as bonding with the therapist (Berger, 2017). Perhaps we need to measure alliance in a

different way in ICBT, when one can argue that the treatment content is its own entity and that the patient most likely separates this from the therapist, making a therapeutic triad rather than dyad.

The finding that treatment credibility was a relevant predictor that added unique input in all regression models in which it was included is well in line with previous findings from Social anxiety disorder and Depression (Boettcher, Renneberg, & Berger, 2013; El Alaoui, Ljótsson, et al., 2015; El Alaoui et al., 2016; Hedman et al., 2012). An interesting note is that in face-to-face psychotherapy, the results have been the opposite; that therapeutic alliance has a stronger influence on outcomes than expectations (Constantino et al., 2017; Flückiger et al., 2018).

We found a significant association between weekly ratings on changes in knowledge or understanding and attitudes towards your symptoms predicting better outcomes. Berg et al. (2019) on the other hand, found that baseline declarative knowledge of CBT and depression actually predicted lower treatment gains in ICBT, even controlling for previous psychotherapy. While knowledge increased during treatment, increased knowledge did not correlate with treatment gains. Knowledge in Berg et al (2019) is not completely analogous to our findings since we did not measure declarative knowledge about CBT-i, but rather the personal experience of "having learned something new" or "changed my perspective". Perhaps Berg et al.'s (2019) findings can be understood as an effect of either knowing but not doing, or that the proposed therapeutic mechanisms are not functional for that patient. Theoretically, we aim to teach the patients how to be their own therapist. This is something that is often said about CBT. If that truly is what prevents relapse, then perhaps it is not surprising that the combination of knowing a lot about CBT but still having significant symptoms and seeking treatment is a bad sign. Either the patient is not doing as CBT prescribes, or the solution proposed by CBT does not work for that patient. This might be reflected in our data as well if low ratings on the Knowledge-item perhaps reflects the patient "having heard it all before", which begs the same question: why then, is the patient not sleeping? Whether this would be because the proposed mechanism is not functional for that patient, or that the patient is very unlikely to try, is an empirical question in need of further examination.

Recently, similar findings have been made on a larger sample than in Study II, in the context of ICBT for depression. Zagorscak, Heinrich, Schulze, Bottcher, and Knaevelsrud (2020) found that the strongest predictor of later improvements amongst over a thousand patients treated with ICBT for depression were, by far, early improvements in symptoms. However just like we did, they also found associations with stress at baseline and early usage of treatment components and mid-treatment working alliance (specifically for task/goal ratings) albeit weaker than for symptom severity. This mirrors the results from Study II closely, which is especially promising considering the treatment was for Depression and not Insomnia, which is not only a different condition, but also arguably the most heterogeneous psychiatric disorder. Zagorscak et al. (2020) also found that expectations before treatment were not directly associated with symptom developments, but only if they were followed by high alliance ratings at mid-treatment (specifically task/goal alliance).

If early on-boarding in treatment, reflected by high activity, acceptance of the rationale and willingness to try the main treatment components, is predictive of outcomes when measured a few weeks into treatment, perhaps we could look for predictors of this at baseline. In line with this, Levallius, Clinton, Hogdahl, and Norring (2020) recently found that the personality traits Openness to Experience and Conscientiousness were associated with more overall improvement and higher remission rates in ICBT for bulimic eating disorders. It is easy to imagine that early adherence to treatment components and rationale could be associated with both of these traits.

Interestingly, Probst et al. (2020) found that factors that were external to psychotherapy (such as not feeling able to talk to loved ones about your problems or feeling betrayed or rejected) were more prominent predictors of not being on track in treatment than were for instance working alliance. Perhaps this external load on the patients would also be strong predictors in our studies, but we do not measure this in any sophisticated way. In Study II, there are two clinician rated domains that come close to this, where clinicians indicated if they had knowledge of any external or internal factors in the patient's life right now that might interfere with either their sleep or their ability to adhere to treatment. That was however, rather a blunt instrument considering most other predictors used were validated self-report measures. Perhaps adding a general measure of current psychosocial chaos could improve the classifiers used in this thesis substantially, as well as inform clinicians about what the patient is currently dealing with.

In Study II, clinicians rated various indicators of adherence or on-boarding with treatment. Lenhard et al. (2019) have constructed a clinician-assessment (The Internet Intervention Patient Adherence Scale) which consist of items very similar to the domains rated by clinicians in Study I & II, such as whether the patient in on schedule, is willing to try the interventions, is actively engaged, is communicating frequently and is motivated. They find that this scale was predictive of outcome whereas other proxies for adherence such as number of logins, login time, how much the patient had written, and how many modules they had completed were not. Our and their findings indicate that clinicians know a lot more about whether the patient is actively engaged in treatment or not than we can capture with any of these rough proxies, at least so far. One possible future avenue for this is natural language processing or other text/language based machine learning methods, where the contents of patient and therapist messages and well as written work sheets could be mined for compliance/adherence/on-boarding indicators (Boman et al., 2019; Nadkarni, Ohno-Machado, & Chapman, 2011). The findings by Kraepelien et al. (2019), where messages and work sheets were rated by experts, suggest that such indicators can be gleaned from those sources and that they do predict outcomes. Having machines do this instead could save a lot of time and resources.

Establishing this benchmark for a classifier with potential to fuel a successful Adaptive treatment strategy is sorely needed. Take for instance a recent example of predicting psychiatric outcomes using machine learning and routine data from intake where the machine learning algorithm could predict non-response with an area under the curve of .65 (Wolff et al., 2020).

The authors unfortunately conclude that this was a poor accuracy, but based on Study II in this thesis, we might say that this is likely good enough to be usable in an Adaptive treatment strategy that could reduce non-response without overburdening the clinicians. Yes, we need to aspire to make even better predictions, but we also need to move forward with deploying the predictions that we already can make in a clinically meaningful way.

It is possible that a classifier with the same accuracy or higher could be built using less input data if it combines a linear model of the symptom trajectory with the other important factors from the RCT-classifier. Of note is that the full RCT-version of the classifier was better as a classifier and using all predictors produced the strongest logistic regression model, even though almost half of the input variables did not have a statistically significant relationship with Failure. Furthermore, those that did correlated with Failure, did so weakly to moderately. These findings point to statistical significance as a hard rule for selecting predictors might leave out predictors that can contribute in a meaningful way when combined with several others. This must be weighed against the burden of collecting the data and the risk of overfitting the model.

In Study III, clinically acceptable predictions, i.e. the benchmark based on Eisenberg and Hershey (1983), could be made at week 5 in Social anxiety disorder treatment and one week later in treatment for Depression and Panic disorder using only their respective primary symptom measures and a basic multiple regression analysis. At the time these acceptable predictions could be made, the adjusted r^2 of the regression models were around .60-.70, in contrast to the full logistic regression model in Study II which had a pseudo r^2 of .56 and was the strongest model in that study.

Regrettably, Study III was carried out before we had the benchmark of clinically sufficient established in Study II. However, it is simple to compare our findings from Study III to the benchmark found in Study II. The benchmark used in Study III was that the lower bound of the confidence interval should be at least 65%. In Study II however, the observed Balanced Accuracy of 67% refers to the point estimate. That confidence interval actually started at 61 %. Starting with just matching Study II with a point estimate of 67% or above, the model in Study III reached this after three weeks in Depression (67%, CI=62-72) and Social Anxiety disorder (70%, CI=64-75), and after four weeks in Panic disorder (67%, CI=61-73). If we instead say that the lower bound of the confidence interval needs to be 67%, which is just above the clinical acceptability benchmark that we actually used in Study III, we have to wait another four weeks in Depression, another two weeks in Social anxiety disorder, and another three weeks in Panic disorder treatment. This would place the prediction during the sixth week of treatment in Social anxiety disorder and as late as during the eighth week of treatment in Depression and Panic disorder. Doing this is however, maybe not that reasonable since this would mean that Study I & II, where this benchmark comes from, would fail to meet it. It also highlights that data from the first few weeks of treatment brings a lot of useful information for prediction, but that future weeks do not add quite as much, especially considering they represent time lost to intervene.

Regarding the idea of using these predictions in an Adaptive treatment strategy, we might again ask if the treatments were good enough as a whole to begin with. Based on the criteria of

Success/Failure that we used, perhaps that should not be said for the treatment for Social anxiety disorder, which had a Success rate of only 23%. While this treatment has a smaller effect size than Depression and Panic disorder, I do not believe that the treatment is very bad. The within group effect size found in the effectiveness trial was large (d = 0.86) as post-treatment and maintained at six months follow-up (d = 1.15) after all (El Alaoui et al., 2015). A reasonable explanation is a combination of a perhaps slightly less effective treatment and an outcome measure (LSAS-SR) that does not function well with our definitions of response (50% reduction) and where achieving remission is relatively unlikely. Since all of these findings are from an uncontrolled setting, we cannot assume that the treatment for Social anxiety really is any less effective than the other two. In fact, Social anxiety disorder has been found to have lower rates of spontaneous remission than Depression and Panic disorder (Bruce et al., 2005), which could very well account for this discrepancy in effects. The MADRS-S and the PDSS-SR are more similar to each other than either one is to the LSAS-SR in terms of score range, number of items and factor structure (the LSAS has been found to have anywhere between 2-8 factors; see Caballo, Salazar, Arias, Hofmann, & Curtiss, 2019 for a brief and up to date summary). While the LSAS-SR is widely used, and can works well for an individual clinician to get a comprehensive picture of a specific patient, it might be less suitable for the type of quantification applied in this thesis.

In Study IV, we found that therapists are often better than chance when predicting categorical outcomes and definitely better than chance for continuous outcomes. However, we also find that whenever we can compare, statistical models seem to do better. When the balanced accuracies of therapists are better than chance, these are clearly driven by high sensitivity and lowered by a low specificity (Table 1 in Study IV). i.e. therapists assume that most patients will do well, and are often right, but this is perhaps mostly because the treatments are good.

We make the point in the manuscript that therapists are not necessarily doing their best because they have no incentive. This still means that their performance reflects how they are thinking about each patient at the time. If anything, performing these predictions likely made therapists more aware of the progress of their current patients rather than less. It is still important to note that this idea that therapists are blinded by optimism and have no idea what is going to happen with their clients is not at all fare. Looking back to the introduction of this thesis, (Section 1.4.1) I make the point that patients have already been selected into treatment by being thoroughly assessed before starting treatment. According to quarterly reports, just over 50% of all who sign up at the clinic actually start treatment. They are also almost entirely self-referred. What we are really examining is the ability of therapists to guess which patients will not benefit, among patients who were all deemed likely to benefit, and who all wanted this specific treatment. Additionally, all of these patients were retained in treatment and did not drop out or deteriorate to the point where the clinic aborted the treatment and referred them elsewhere. All of these patients continued the full course of a treatment into which they had been carefully selected. If therapists saw the failures coming, they would intervene, and we may not have the data from such cases. Perhaps this can partially explain that deterioration was never correctly

predicted by therapists, if deteriorating patients are in fact referred elsewhere if they are detected.

What this means for the clinic is that, despite out best efforts, many patients steadily go through treatment as planned, flying under the radar, and end up with arguably unsatisfactory outcomes, at least compared to the lofty standards for success that we use in this thesis.

This study is also important as a benchmark to which we can hold up more advanced decision support tools, statistical models or machine learning algorithms. This is important in the same way as it is important to evaluate a therapy against an active control rather than a waiting list. For instance, Kessler, Chalker, Luedtke, Sadikova, and Jobes (2020) used a rather advanced machine-learning ensemble to allocate optimal treatment for suicidal ideation among 148 U.S. soldiers. They found that this could increase 3-month remission of suicidal ideation with 13.6 % (albeit with a wide confidence interval between 0.9 - 26.3%). This was however, 13.6 % better than random assignment to treatments, and random assignment to treatments is hardly a gold standard comparator, and unlikely to be the norm in clinical settings today. This makes those 13.6 % hard to interpret.

Based on previous research, we had reason to believe that severity at baseline and early symptom changes was going to be the most powerful and consistent predictors of treatment outcome. The results from the studies in this thesis generally support this, though Study II shows some other potentially important predictors apart from symptom severity. Firstly, that severity of other related symptoms such as and general psychological distress, stress, and depressive symptoms are relevant predictors. The second major group of relevant predictors were all some form of indicators of what I like to think of as "on-boarding" with treatment. Factors such as early activity, attitudes towards and willingness to adhere to central treatment components and rationale, finding the treatment credible and reporting a good working alliance as well as reporting that the treatment so far has altered your thoughts or knowledge about your symptoms. However, neither the classifier nor the logistic regression model in Study II take full advantage of the available primary symptom ratings (ISI) at the time of prediction, and the model that used ISI-data was not that much less powerful than were the far more complicated models that accounted for all these other variables.

It is important to remember that the symptom rating not only correlates to the outcome; it is the outcome, just at an earlier point in time. Therefore, all of these findings need to be taken together with careful consideration of what our outcome definitions are. If we were trying to predict something other than the outcome on the primary symptom measure, perhaps such as absenteeism or satisfaction ratings, the situation would be completely different. Based on the studies in this thesis, there is no reason to believe that early observations on the primary symptom measure would be optimal predictors of anything other than late observations on the exact same measure. However, Schibbye et al. (2014) did actually find that disorder specific symptom measures were better predictors than the generalized OQ-45 and CORE-10 measures even when predicting themselves.

4.1.1.1 Summary of main findings

Is the concept of Adaptive treatment strategies in ICBT viable, based on the results of this thesis? Yes, at least in Insomnia treatment!

Can clinically acceptable predictions of treatment Failure in treatment for Depression and Anxiety be made with only symptom scores and simple statistics? Yes, halfway through treatment, but the benchmark for this is old and very preliminary

Can clinically sufficient (based on the RCT for Adaptive treatment) predictions of treatment Failure be made during Week 4 in treatment for Insomnia using basic data and simple models? Yes, and some predictors could probably be excluded, while some factors might be expanded!

Can clinically sufficient (based on the RCT for Adaptive treatment) predictions of treatment Failure be made during Week 4 in treatment for Depression and Anxiety using only symptom scores and simple models? Yes, probably, but it is unclear what is the best way of comparing to a benchmark for Balanced accuracy!

Can therapists themselves make clinically sufficient (based on the RCT for Adaptive treatment) predictions of treatment Failure during Week 4 in treatment for Depression and Anxiety? No, probably not, but therapist predictions do add unique information that could be used together with other data!

Is it important to leave baseline and instead make predictions a few weeks into treatment? Yes, no predictors from baseline were strong enough to make these predictions without adding more data from the treatment period!

4.1.1.2 What could we have done differently in the RCT based on our findings?

Based on our findings, it is possible that the classification algorithm in Study I could be improved in some ways before we use it again:

1) By adding a linear function of all available measurements on the ISI to the classifier.

2) By having therapists provide their own best guess for change score, response and remission for the patient.

3) By ignoring some of the seemingly unimportant clinician rated domains and instead perhaps having clinicians make more detailed ratings on the domains that were important. For example by doing as in Kraepelien et al. (2019) and separating adherence and homework into how much the patient has done on one hand and how well the patient has done it on the other.

4) By removing some seemingly unimportant patient-rated variables (such as beliefs and attitudes about sleep at baseline) and perhaps replace with other variables such as the personality traits conscientiousness and openness to experiences, and some indicator of current psychosocial chaos as well as attitudes towards the treatment format. It would also

likely be possible to quantify early activity in treatment using activity-logs in the database and the contents of worksheets etc.

4.1.2 Causal inferences based on this thesis

Causality is important to keep in mind then interpreting the results and their implications as we are using several causally colored words throughout the studies in this thesis even when the design of the investigation does not permit causal inferences. Specifically the word Response or Responder is problematic in this context. The only place in this thesis where we can make, or have any interest in making, causal inferences is in Study I, and even then only specifically about the effect of the adapted treatment (i.e. the differences between the two Red groups after randomization). There, we want to be able to say that the adapted treatment *caused* a steeper improvement trajectory and *caused* more patients to have a Successful outcome. The design of Study I, it being a randomized controlled trial with adequate power, successful randomization, and good adherence to the study protocol, thankfully allows for such causal inferences. Anywhere else in this thesis and the papers, the word Reponder is used because it is a commonly used term to describe a certain magnitude of change over the course of treatment, and we believe that we are introducing enough new terms as it is. A more stringent word might be Improver as the only thing we really know is that their symptoms are improving over time. Similarly, the word success or failure also implies that we know what *the treatment* did to the patient, which we mostly do not. What we know is what happened to their symptoms. Along that line, the word Failed treatment could possibly have been replaced with something along the lines of "still not well enough" or "not improved enough" or "non-recovered" but all of those have their own problems and we believe that the impact of the word Failure is useful when we use it to study treatment outcomes that we produce. I would urge the reader to think "good enough" when thinking about the definition of Success used here. I should mention that there are ways to use correlational data to examine causality, but it is very, very difficult and requires very bold assumptions and really good measurements of potential confounders, colliders, mediators and covariates as well as preferably no attrition and perfectly representative data (Rohrer, 2018). Considering the state of psychological science regarding the empirical evidence of our theoretical models, exactness of measurements and understanding of psychotherapy mechanisms, making causal inferences based on correlational data should probably not be attempted yet.

4.2 ADAPTIVE TREATMENT AND ITS EFFECTS

As I mentioned earlier, we only tested applying an Adaptive treatment strategy on Insomnia patients, and therefore wave not yet provided any evidence that this would work in other treatments. The idea that allocating more time and effort to patients who are struggling is not controversial, and it is easy to imagine that patients in treatment for depression or anxiety would also benefit from getting an Adaptive treatment, but that need to be empirically examined in future studies. We have also not yet examined exactly what therapists did in the adapted treatment and which adaptations produced positive changes or did not help at all. Previous research suggests that support might lead to greater improvements via increased engagement

in core treatment activities (i.e. adherence/compliance; Kaldo, Ramnerö, et al., 2015), which was the main reasoning behind the way we carried out the adaptive treatment strategy in Study I. However, it is possible that in some circumstances, the best thing to do could be to address a comorbid problem, or help the patient communicate with social services and so on in order to increase the likelihood of a good treatment outcome.

In Study I, we accounted for how much time therapists spent and gave some examples of things they did, but we have also collected detailed data on what they did differently with each patient, which is yet to be analyzed. Doing this could not only help guide future therapists trying to provide an adapted treatment. It is also possible that we can find some adaptations that could have been automated or solved beforehand. For instance, some patients felt that calculating new sleep windows was too difficult and confusing. This could be built into the program instead of being something that the therapist steps in and does only if an adapted treatment is initiated.

It is also possible that some patients simply need more time, rather than a big change to the treatment. In fact, Dunlop et al. (2019) found that when sequentially adding medication or CBT after a treatment with medication or CBT, patients who had responded to the first treatment but not achieved remission were the most likely to respond to the second treatment as well, regardless of which treatment came first and second. On the other hand, those who had not responded to one type of treatment did not tend to respond to the other. Therefore, perhaps changing treatment did not really help, but those who had started to respond in one treatment continued to do so when given more treatment.

How to plan, and what to do in, an adapted treatment is beyond the scope of this thesis but a crucial question for future research.

4.3 ETHICAL CONSIDERATIONS OF ADAPTIVE TREATMENT STRATEGIES

Turning over the decision making to mathematical models in such a sensitive area as psychiatry is not without some ethical dilemmas, though some would argue that not doing it is more ethically questionable (Dawes et al., 1989). As I mentioned in the introduction, it would be an entirely different thing to let an algorithm decide which patients to abandon or to deny treatment in the first place, rather than decide which patients should get extra support. This means that the results and preliminary benchmarks that we present here are inextricably linked with the type of decision that we made, and cannot be translated directly into for instance be used to deny patients treatment without serious ethical consideration. Yes, letting machines into clinical decision-making can be scary. But left to their own devices, humans come up with silly ideas about causation that never hold up to any empirical scrutiny. Humans did however also invent empirical scrutiny. Before disparaging the idea of letting machines influence decisions in psychiatry, we should remember that we are the ones who made these machines, and at least in this thesis, we are not suggesting actually letting the machines make decisions without clinician oversight.

In this thesis, we used the Balanced accuracy and did not weight or prioritize one type of correct classification relative to another. Moving forward, this field will likely mature to the point

where the decision-threshold or another similar approach can be used fruitfully. This will bring with it a lot of ethical dilemmas, when risks and poor treatment outcomes can be weighed against costs and resources in a more exact way. The benchmark from Eisenberg & Hershey (1983) was based on clinicians having reservations about acting on uncertain predictions mainly out of concern for the patient as there were clear risks and side-effects of proceeding with treatment. In our case, it is almost impossible to imagine any downside to the individual patient of being misclassified as Red and getting extra help, especially considering the patient has a lot of influence of what that help will be. In Adaptive treatment strategies in ICBT as we have used them here, the deciding factor is not risk to the patient, but a lack of resources, where therapists reallocate their time to those who need it most. In this thesis, we focus on having an optimal balance between correct Red and Green classifications, but at some point, we should instead ask ourselves how many patients we can afford to classify as Red before we are no longer able to provide the adapted treatment in a useful way.

4.4 LIMITATIONS

The limitations of each study is reported in their respective manuscripts. Here I will discuss some of the more broad limitations that encompass the entire thesis, or limitations that arise between the different studies themselves.

4.4.1 Dropout

Attrition was not a major problem in any of the studies in this thesis, but it can still bias the results, especially when we are looking at Failure, which is often a minority outcome, and we are not accounting for dropout, which is another minority, but one that could easily be related.

In the context of adaptive treatment strategies, dropout should be considered its own outcome, and is something very relevant to predict and avoid (Karin, Dear, Heller, Crane, & Titov, 2018). If a patient is best off dropping out of an ongoing treatment, then clearly something is very wrong with the treatment or the patient was not properly diagnosed/assessed beforehand. That is unlikely in our treatments, since we have rigorous assessment procedures and rather well evaluated treatments. Thus, dropout should probably always be seen as a kind of Failure, something to be avoided. Therefore, using the logic in this thesis, dropout should be predicted. We do not do this at all in any of the studies in this thesis, which is a gap that needs to be filled by future research. An upside with trying to avoid dropout is that, in contrast to a post-treatment symptom score, it is possible to stop a patient from dropping out or to get them back into treatment, at least in some cases. As I mentioned in the introduction (Section 1.4.5.3), it can be useful just to assess when during treatment dropouts tend to happen, so that this can be used to weight predictive accuracy against, in order to make predictions when they are as accurate as possible but before it is too late to catch those who will otherwise drop out.

4.4.2 Variations in outcomes across studies

The cutoff for remission on the PDSS-SR is different by 1 point between Study III and Study IV (7 points or less in Study III and 6 points of less in Study IV). This is because I found better

evidence along the way and made sure that each manuscript was as accurate as possible, even though this does affect comparability between these studies. That effect is likely negligible considering Study III only uses this as one of the criteria for Failure whereas Study IV uses this cutoff to look directly at remission. In addition, a difference of one point is in itself probably not very consequential for the proportions of remitters in Panic disorder treatment. Regardless of which cutoff is used, they are all based on empirical data on the clinician rated version of the scale, so neither of them is ideal for my purposes.

Another variation in outcomes that limits comparisons between studies is that therapists in Study IV were not asked to predict Failure defined in the way that we do in this thesis. The reason for this is that when the questionnaire for the therapists was designed, we had not finished deciding on our main definition of failure. The separate categories such as remission and response have been used for many years by countless researchers so those were the only things we asked therapists to guess. Study III and Study IV can still be compared via the appendix from Study III where those categories are also predicted.

4.4.3 The definition of Success as either response or remission

We made the decision not to use Jacobson and Truax's definition of clinically relevant change when defining Success in this thesis, which may be controversial. However, Jacobson and Truax's criteria was built to include a sort of guarantee of a causal effect of the actual treatment, which a) it does not and b) is not important in this context. I believe that our definition, where a patient is EITHER below the remission cutoff OR has a substantial, rather than statistically significant, reduction in symptoms is currently the best way to classify success for ICBT for a mood disorder, when a few basic assumptions or criteria are taken into consideration:

1) **We already know that the treatment is efficacious and are not concerned with proving that the treatment has an effect per se** neither on a group nor individual level. In treatment research in general, it is important to prove that there is a statistically significant change over time, and therefore the "either or" statement used in this thesis is controversial. A remitter status at post-treatment or a 50% reduction can either or both occur without meeting criteria for statistical difference if the patient has low scores at pre-treatment. In the context of this thesis however, the concept of interest is whether the individual patients expected outcome needs to be boosted to be satisfactory post treatment. A patient that is nearly sub-clinical now, does not need to be boosted for the remainder of treatment and I would argue nor does a patient that will likely end up with a very large, much larger than average, proportional reduction in symptoms. Those patients should not be given extra support or adapted treatment. If we believed that the *average* or above average effect of the treatment or remission at post was an unsatisfactory result on the individual level, we should be overhauling the entire treatment rather than building systems for helping the ones with the worst outcomes.
2) **The patients were not already healthy when treatment started. The prediction occurs during treatment, not before.** Along the same line, it is important to remember that this criterion of success is applied during treatment and must not influence the sample starting treatment in any way. It is obvious that a treatment's success-rate could be dramatically changed by changing the inclusion criteria. Only include very healthy

individuals and the success-rate will increase substantially if success is defined as in this thesis (this is what Jacobson and Truax wanted to avoid). The concept under study is meant to be added to a clinical context that should already have sound practices for avoiding offering treatments to healthy individuals. Therefore, an important assumption is that the treatment was considered necessary by a mental health professional to begin with.

One might also argue that by using the 50% reduction instead of the RCI we are opening ourselves up to labeling measurement error as a success for milder cases, as we are not including the measurement error of the scale in the equation. As a thought experiment, let us use the MADRS-S depression measure. It has a remission cutoff of 10 points or less (Fantino & Moore, 2009) and in our sample, the RCI comes out at 8 or more points change. This means that for any patient to have a 50% improvement that does not also include at least the RCI (and thereby covering for the measurement error), they have to have a pre-treatment score of 15 or less. For them, a 50% reduction, which does no longer include the RCI, will instead put them below the cutoff for remission. All instances of a 50% reduction not being a reliable change will be among remitters (this is true for LSAS-SR and the PDSS-SR as well. This is one of the advantages of using either response or remission as the rule.

4.4.4 The clinical implications of prediction that focuses on identifying cases rather than supporting theory

In this thesis, we primarily focus on discerning the *who* (i.e. identifying cases who will end up with a certain outcome). Only in Study II do we examine specific factors that could help explain *why* a patient is heading towards failure. Still, the why and the how is largely beyond the scope of this thesis. There is a very important, but sometimes overlooked, implication of this individual-identification-paradigm as opposed to identifying important success-factors from which to learn. The prediction algorithms used in this thesis can only tell the user who will fail. This means that the only real way of applying such findings in clinical practice is to take the same observations off each patient and put them into these exact, or very similar, algorithms in order to get the prediction for that patient. This actually means that perhaps the most important barrier to implementation of such findings is how difficult, or expensive, it is to collect the input data. Complex prediction models based on routine data may be time consuming to develop, but once developed, they can produce predictions for new patients quickly and with no additional data collection.

4.5 DIRECTIONS FOR FUTURE RESEARCH

- By using the criteria that a 50% reduction in symptom severity is a good enough clinical outcome regardless of the severity of remaining symptoms, I am touching upon an idea that proportional differences in symptoms over time within an individual are as important as absolute severity levels after treatment. That perhaps a patient who has halved her symptoms might feel cured even if her symptom level is still in the clinical range according to our cutoffs. This is a very interesting empirical question. The data we have today guiding our cutoffs for remission are usually based on comparing individuals to the average score in healthy populations. It is possible that patients vary in how high or low scores they would give themselves even when they are feeling well, perhaps based on personality, social comparison or previous experiences. It would be interesting to assess patients after treatment and ask them if they feel cured or successfully treated or treated well enough and see if those ratings are more associated with current symptom scores, absolute changes, or proportional changes in symptom severity.
- The classifier used in Study I and II was cumbersome to use and could perhaps be automated. The findings that many input predictors did not have a meaningful relationship to the outcome indicates that the classifier could be made simpler by omitting some data, though this would need to be empirically tested.
- We suggest that matching patients to treatments demand things that we do not have (such as several treatment alternatives and very strong predictions). However, another avenue of research could be lowering the standard level of support in ICBT within an adaptive treatment strategy. We started with normal levels of support (about 15 minutes per week) and intensified for patients who were not responding. Some emerging evidenve suggests that ICBT can work without therapist support, at least for some (Karyotaki et al., 2017). Therefore, it is possible that the avaliability of treatment could be increased if treatments were initiated without any therapist support, and then intensifications could start from there. The first step being no therapist, followed by minimal therapist support and so on.
- We still no not know much about what kind of treatment adaptations are helpful for patients, and if this differs depending on why the patient was struggling in the first place. We will investigate this based on data from Study I, but considering the relevant sample consists of only 50 people, larger studies are required to investigate which adapations works for whom in more detail. Also, this is still limited to Insomnia and needs to be tested in other conditions as well.
- We still do not know what works for whom, but considering we still do not even really know why CBT works at all (Cuijpers, Reijnders, & Huibers, 2019), matching every patient to their optimal treatment is perhaps still too lofty a goal to have. What we have proposed here is a much more low-hanging fruit; using the best and most available treatments we have, focusing on what happens with patients during those treatments,

and helping those who are struggling. This need to be replicted in other settings and other conditions.

- An important avenue for future research based on this thesis is to use more advanced prediction models to examine if equal of greater accuracy could be reached using only data that is routinely collected, so that no additional burden is laid upon patient or therapists just to make the prediction. Ideally, this would be done in an automated manner that could be integrated into the ICBT-platform, allowing therapists to have quick access to reliable predictions about their ongoing patients.

5 CONCLUSIONS

In conclusion, we find support for the concept of an Adaptive Treatment Strategy in Internet-delivered Cognitive Behavior Therapy for Insomnia, where failing treatment attempts can be identified three weeks into treatment and where these poor trajectories can be changes for the better by adapting their treatments for the remaining weeks of treatment. We also find that a classification algorithm with at least 67% balanced accuracy should be clinically sufficient for such purposes if the context is similar in terms of the risks and benefits of classification. Furthermore, we find that it is unclear if such an accuracy can be reached confidently after just three weeks of treatment if only primary symptom measures are used as predictors in simple statistical models, though it is possible that we can. Either way, it is possible that even slightly less accurate predictions would be preferable considering how much easier they would be to implement. Additional direct comparisons are warranted to explore this further. Simple statistical models using only symptom scores can however, reach clinically acceptable levels of accuracy halfway through 12-week interventions for Depression, Social anxiety and Panic disorder. Previous findings that therapists themselves are less accurate than statistical models seem to be true also for therapists in Internet-delivered interventions, though we do find that clinician input such as rating adherence and activity does add unique information to prediction algorithms and that therapists own predictions are better than chance and could add to predictive models. The vast majority of useful information for making the predictions in this thesis were collected during treatment rather than before, indicating how this approach stands apart from the more common practice of predicting based on pre-treatment data.

Big data, machine learning, biomarkers and other sophisticated methods for data collection and prediction analysis should, and almost definitely will, be able to outperform the methods used in this thesis. Still, it is important to remember that such methods are not magic, that we can already achieve clinically useful predictions if we just start making them, and that successful Adaptive Treatment Strategies are possible even with simple statistical models based on non-complex data. The question is not if methods that are more complex can be more accurate than the ones used in this thesis. The question is if they can achieve this in a way that is practical enough, and economical enough, to be implementable in routine psychiatric care.

Otherwise, what is the point?

6 ACKNOWLEDGEMENTS

"The point of the PhD is for you to become and independent scholar.
At the same time, I have to show you the right way to do things.
Therefore, the only solution is for you to do what I tell you to do,
but don't assume that I'm right,
so you can learn for yourself that I am".

-Professor Smith, PhD Comics
by Jorge Cham.

Main supervisor **Viktor Kaldo**, my master. Thank you for your unmatched ability to be simultaneously incredibly encouraging and still have literally hundreds of notes. You have taught me countless things, but teaching me to actually revise and edit my work was, surprisingly, something relatively new to me, as many-a previous teacher will attest.

Co-supervisor **Kerstin Blom**, for showing me the ropes, being my drill sergeant and getting me ready to watch over your RCT-babies. Thanks for calling me out on my beige foods and for all the chats about cooking, baking, hunting, fishing and general bad-assery. You truly are the coolest.

Co-supervisor **Susanna Jernelöv,** for teaching me about sleep and insomnia and especially for teaching me to smile in the face of overwhelming ~~stress~~ hurry, and enjoy your day even if things are chaotic. Your enthusiasm is highly contagious. One day, we will run a simple study! One day…

Co-supervisor **Nils Lindefors,** the big man. Your vision and dedication in moving ICBT from a purely research focused endeavor into implementation in routine psychiatric care is the reason we can all say with such confidence that ICBT *really* works, in *real life*. I am also grateful for the opportunity to move forward and explore just how far ICBT can go in our new self-care project.

Per Carlbring, my mentor and first ever PI. You brought me into this ICBT-world back when I was writing my master thesis. I feel lucky and privileged to have crossed paths with you when I did, and certainly hope to continue doing so in the future.

Gerhard Andersson, my doctoral grandfather and the godfather of research on internet interventions in Sweden, and the world, making every Swedish ICBT-researcher cool by proxy.

My fellow research group members **Ann Rosén**, **Berkeh Nasri**, **Pontus Bjurner**, **Nils Isacsson**, **Cecilia Svanborg**, **Marie Bendix** and **Sandra Tamm** as well as my new project group members for the self-care project **Martin Kraepelien, Amira Hentati** and **Dorian Kern** and Viktor's lovely Linnaeus University-gang **Anneli Farnsworth von Cederwald** and **Gustav Nilsson**. Research is infinitely more fun when surrounded by such excellent people!

All the students I have supervised. Your ambitious and dedicated work is what keeps our research machine running. If you were among the first: Sorry I hadn't really cracked the whole supervisor thing... Particularly, thank you to **Ekaterina Ivanova** who helped me run a sudden extra-RCT when things got a bit ~~stressful~~ busier than intended, and **Simon Mattsson** who got me started with R-statistics and co-authored Study IV.

My collaborators from KTH, Magnus Boman and **Fehmi Ben Abdesslem**. I know I haven't performed one fifth of one full set of appropriate analyses for what I'm trying to do with our data, but bear with me; I'm learning.

Thank you to the IPSY-crew **Nina Lind**, **Monica Hellberg** and **Eila Johansson**, for putting up with me during all the RCT-chaos, and to all my new IPSY-psychologist co-workers. I am so grateful for the opportunity to develop our clinic further and to do research *in the real world.*

Also, thank you to **Christian Rück** and his research team next door for doing such cool and important research and for being a lot of fun!

Alexander Rozental, who gave me my first taste of research and my first taste of internet treatments. Your supervision and encouragements sent me merrily on my way towards becoming a PhD student. Not a word of a warning…

Andreas Svensson, my first research partner and fellow procrastinator. Thanks for sharing my love for nature and the fundamentals in life and for your curiosity in trying to understand the human condition.

Nick Titov, **Blake Dear**, **Eyal Karin** and the **MindSpot** and **eCentreClinic** crews down under, for being such admirable researchers and wonderful hosts and guests. I look forward to many years of collaboration!

To "the band" and my brothers from other mothers **Arvid Kalmaru** (who also made the illustration for the cover), **Joakim Jalap** and **Johan Mucchiano**. Thanks for rocking out together for many years and maybe years to come. Who knows? Patience *is* a virtue after all.

To **Carina** and **Gunnar Forsell**. Mom and dad. Thanks for teaching me almost everything I know in terms of volume. Invaluable education, supervision and support in truly applicable skills, from driving, cooking and cleaning (yes, I *can* clean) all the way back to using the toilet. If you made it here without skipping anything, then *that* is what I've been doing for six years.

Selma, my daughter and eldest child. Thank you for being both though and kind and simply awesome. Also, thanks for making people think that I had anything to do with that. **Oskar**, my son, five years junior to Selma. You have an innate ability to bring us all down to your level. Then again, being a toddler is a lot of fun. You guys are, quite literally, the reason I get up in the morning.

Josefin Rydberg, mother of my children. Where would I be without you? Thank you for sharing almost half of my life (so far) with me and making everything worthwhile.

7 REFERENCES

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., . . . Rush, J. D. (2006). The Meta-Analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction. *The Counseling Psychologist, 34*(3), 341-382. doi:10.1177/0011000005285875

Andersson, E., Crowley, J. J., Lindefors, N., Ljótsson, B., Hedman-Lagerlöf, E., Boberg, J., . . . Rück, C. (2019). Genetics of response to cognitive behavior therapy in adults with major depression: a preliminary report. *Molecular Psychiatry, 24*(4), 484-490. doi:10.1038/s41380-018-0289-9

Andersson, E., Enander, J., Andren, P., Hedman, E., Ljotsson, B., Hursti, T., . . . Ruck, C. (2012). Internet-based cognitive behaviour therapy for obsessive-compulsive disorder: a randomized controlled trial. *Psychological Medicine, 42*(10), 2193-2203. doi:10.1017/s0033291712000244

Andersson, G. (2018). Internet interventions: Past, present and future. *Internet Interventions, 12*, 181-188. doi:https://doi.org/10.1016/j.invent.2018.03.008

Andersson, G., Bergström, J., Buhrman, M., Carlbring, P., Holländare, F., Kaldo, V., . . . Waara, J. (2008). Development of a New Approach to Guided Self-Help via the Internet: The Swedish Experience. *Journal of Technology in Human Services, 26*(2-4), 161-181. doi:10.1080/15228830802094627

Andersson, G., Carlbring, P., & Rozental, A. (2019). Response and Remission Rates in Internet-Based Cognitive Behavior Therapy: An Individual Patient Data Meta-Analysis. *Frontiers in Psychiatry, 10*, 13. doi:10.3389/fpsyt.2019.00749

Auffray, C., & Hood, L. (2012). Editorial: Systems biology and personalized medicine - the future is now. *Biotechnology Journal, 7*(8), 938-939. doi:10.1002/biot.201200242

Baker, S. L., Heinrichs, N., Kim, H.-J., & Hofmann, S. G. (2002). The Liebowitz social anxiety scale as a self-report instrument: a preliminary psychometric analysis. *Behaviour Research and Therapy, 40*(6), 701-715. doi:http://dx.doi.org/10.1016/S0005-7967(01)00060-2

Bastien, C. H., Vallières, A., & Morin, C. M. (2001). Validation of the Insomnia Severity Index as an outcome measure for insomnia research. *Sleep Medicine, 2*(4), 297-307. doi:http://dx.doi.org/10.1016/S1389-9457(00)00065-4

Berg, M., Rozental, A., Johansson, S., Liljethörn, L., Radvogin, E., Topooco, N., & Andersson, G. (2019). The role of knowledge in internet-based cognitive behavioural therapy for adolescent depression: Results from a randomised controlled study. *Internet Interventions, 15*, 10-17. doi:https://doi.org/10.1016/j.invent.2018.10.001

Berger, T. (2017). The therapeutic alliance in internet interventions: A narrative review and suggestions for future research. *Psychotherapy Research, 27*(5), 511-524. doi:10.1080/10503307.2015.1119908

Blom, K., Jernelov, S., Ruck, C., Lindefors, N., & Kaldo, V. (2016). Three-Year Follow-Up of Insomnia and Hypnotics after Controlled Internet Treatment for Insomnia. *Sleep, 39*(6), 1267-1274. doi:10.5665/sleep.5850

Blom, K., Jernelöv, S., Kraepelien, M., Bergdahl, M. O., Jungmarker, K., Ankartjärn, L., . . . Kaldo, V. (2015). Internet Treatment Addressing either Insomnia or Depression, for Patients with both Diagnoses: A Randomized Trial. *Sleep, 38*(2), 267-277. doi:10.5665/sleep.4412

Blom, K., Tillgren, H., Wiklund, T., Danlycke, E., Forssén, M., Söderström, A., . . . Kaldo, V. (2015). Internet-vs. group-delivered cognitive behavior therapy for insomnia: A randomized controlled non-inferiority trial. *Behaviour Research and Therapy, 70*, 47.

Boettcher, J., Renneberg, B., & Berger, T. (2013). Patient expectations in internet-based self-help for social anxiety. *Cognitive Behavior Therapy, 42*(3), 203-214. doi:10.1080/16506073.2012.759615

Boman, M., Ben Abdesslem, F., Forsell, E., Gillblad, D., Görnerup, O., Isacsson, N., . . . Kaldo, V. (2019). Learning machines in Internet-delivered psychological treatment. *Progress in Artificial Intelligence*. doi:10.1007/s13748-019-00192-0

Bower, P., & Gilbody, S. (2005). Stepped care in psychological therapies: access, effectiveness and efficiency: narrative literature review. *British Journal of Psychiatry, 186*. doi:10.1192/bjp.186.1.11

Bower, P., Kontopantelis, E., Sutton, A., Kendrick, T., Richards, D. A., Gilbody, S., . . . Liu, E. T.-H. (2013). Influence of initial severity of depression on effectiveness of low intensity interventions: meta-analysis of individual patient data. *BMJ : British Medical Journal, 346*, f540. doi:10.1136/bmj.f540

Bruce, S. E., Yonkers, K. A., Otto, M. W., Eisen, J. L., Weisberg, R. B., Pagano, M., . . . Keller, M. B. (2005). Influence of psychiatric comorbidity on recovery and recurrence in generalized anxiety disorder, social phobia, and panic disorder: a 12-year prospective study. *American Journal of Psychiatry, 162*(6), 1179-1187. doi:10.1176/appi.ajp.162.6.1179

Bucholz, K. K., Hesselbrock, V. M., Heath, A. C., Kramer, J. R., & Schuckit, M. A. (2000). A latent class analysis of antisocial personality disorder symptom data from a multi-centre family study of alcoholism. *Addiction, 95*(4), 553-567. doi:10.1046/j.1360-0443.2000.9545537.x

Caballo, V. E., Salazar, I. C., Arias, V., Hofmann, S. G., & Curtiss, J. (2019). Psychometric properties of the Liebowitz Social Anxiety Scale in a large cross-cultural Spanish and Portuguese speaking sample. *Brazilian Journal of Psychiatry, 41*, 122-130.

Carlbring, P., Andersson, G., Cuijpers, P., Riper, H., & Hedman-Lagerlof, E. (2018). Internet-based vs. face-to-face cognitive behavior therapy for psychiatric and somatic disorders: an updated systematic review and meta-analysis. *Cognitive Behavior Therapy, 47*(1), 1-18. doi:10.1080/16506073.2017.1401115

Constantino, M. J., Coyne, A. E., Luukko, E. K., Newkirk, K., Bernecker, S. L., Ravitz, P., & McBride, C. (2017). Therapeutic alliance, subsequent change, and moderators of the alliance-outcome association in interpersonal psychotherapy for depression. *Psychotherapy (Chic), 54*(2), 125-135. doi:10.1037/pst0000101

Cuijpers, P., Berking, M., Andersson, G., Quigley, L., Kleiboer, A., & Dobson, K. S. (2013). A meta-analysis of cognitive-behavioural therapy for adult depression, alone and in comparison with other treatments. *Canadian Journal of Psychiatry, 58*(7), 376-385.

Cuijpers, P., Donker, T., van Straten, A., Li, J., & Andersson, G. (2010). Is guided self-help as effective as face-to-face psychotherapy for depression and anxiety disorders? A systematic review and meta-analysis of comparative outcome studies. *Psychological Medicine, 40*(12), 1943-1957. doi:10.1017/s0033291710000772

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

Cunningham, J. A., Kypri, K., & McCambridge, J. (2013). Exploratory randomized controlled trial evaluating the impact of a waiting list control design. *BMC Medical Research Methodology, 13*, 150. doi:10.1186/1471-2288-13-150

Dawes, R. M. (2005). The ethical implications of Paul Meehl's work on comparing clinical versus actuarial prediction methods. *Journal of Clinical Psychology, 61*(10), 1245-1255. doi:10.1002/jclp.20180

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*(4899), 1668-1674.

Deckersbach, T., Peters, A. T., Sylvia, L. G., Gold, A. K., Magalhaes, P. V. D., Henry, D. B., . . . Miklowitz, D. J. (2016). A cluster analytic approach to identifying predictors and moderators of psychosocial treatment for bipolar depression: Results from STEP-BD. *Journal of Affective Disorders, 203*, 152-157. doi:10.1016/j.jad.2016.03.064

DeRubeis, R. J., Cohen, Z. D., Forand, N. R., Fournier, J. C., Gelfand, L. A., & Lorenzo-Luaces, L. (2014). The Personalized Advantage Index: Translating Research on Prediction into Individualized Treatment Recommendations. A Demonstration. *PLoS ONE, 9*(1). doi:10.1371/journal.pone.0083875

Dunlop, B. W., LoParo, D., Kinkead, B., Mletzko-Crowe, T., Cole, S. P., Nemeroff, C. B., . . . Craighead, W. E. (2019). Benefits of Sequentially Adding Cognitive-Behavioral Therapy or Antidepressant Medication for Adults With Nonremitting Depression. *American Journal of Psychiatry, 176*(4), 275-286. doi:10.1176/appi.ajp.2018.18091075

Edmonds, M., McCall, H., Dear, B. F., Titov, N., & Hadjistavropoulos, H. D. (2020). Does concurrent medication usage affect patient response to internet-delivered cognitive behaviour therapy for depression and anxiety? *Internet Interventions, 19*, 100302. doi:https://doi.org/10.1016/j.invent.2019.100302

Eisenberg, J. M., & Hershey, J. C. (1983). Derived thresholds. Determining the diagnostic probabilities at which clinicians initiate testing and treatment. *Medical Decision Making, 3*(2), 155-168.

El Alaoui, S., Hedman, E., Kaldo, V., Hesser, H., Kraepelien, M., Andersson, E., . . . Lindefors, N. (2015). Effectiveness of Internet-based cognitive-behavior therapy for social anxiety disorder in clinical psychiatry. *Journal of Consulting and Clinical Psychology, 83*(5), 902-914. doi:10.1037/a0039198

El Alaoui, S., Hedman, E., Ljótsson, B., Bergström, J., Andersson, E., Rück, C., . . . Lindefors, N. (2013). Predictors and Moderators of Internet- and Group-Based Cognitive Behaviour Therapy for Panic Disorder. *PLoS ONE, 8*(11), e79024. doi:10.1371/journal.pone.0079024

El Alaoui, S., Hedman, E., Ljótsson, B., & Lindefors, N. (2015). Long-term effectiveness and outcome predictors of therapist-guided internet-based cognitive–behavioural therapy for social anxiety disorder in routine psychiatric care. *BMJ Open, 5*(6). doi:10.1136/bmjopen-2015-007902

El Alaoui, S., Ljótsson, B., Hedman, E., Kaldo, V., Andersson, E., Rück, C., . . . Lindefors, N. (2015). Predictors of Symptomatic Change and Adherence in Internet-Based Cognitive Behaviour Therapy for Social Anxiety Disorder in Routine Psychiatric Care. *PLoS ONE, 10*(4), e0124258. doi:10.1371/journal.pone.0124258

El Alaoui, S., Ljotsson, B., Hedman, E., Svanborg, C., Kaldo, V., & Lindefors, N. (2016). Predicting Outcome in Internet-Based Cognitive Behaviour Therapy for Major Depression: A Large Cohort Study of Adult Patients in Routine Psychiatric Care. *PLoS ONE, 11*(9), e0161191. doi:10.1371/journal.pone.0161191

Fantino, B., & Moore, N. (2009). The self-reported Montgomery-Åsberg depression rating scale is a useful evaluative tool in major depressive disorder. *BMC psychiatry, 9*(1), 1-6. doi:10.1186/1471-244x-9-26

Flint, C., Cearns, M., Opel, N., Redlich, R., Mehler, D., Emden, D., . . . Kircher, T. (2019). Systematic Overestimation of Machine Learning Performance in Neuroimaging Studies of Depression. *arXiv preprint arXiv:1912.06686*.

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

Flygare, A.-L., Engström, I., Hasselgren, M., Jansson-Fröjmark, M., Frejgrim, R., Andersson, G., & Holländare, F. (2019). Internet-based CBT for patients with depressive disorders in primary and psychiatric care: Is it effective and does comorbidity affect outcome? *Internet Interventions*, 100303. doi:https://doi.org/10.1016/j.invent.2019.100303

Foa, E. B. (2010). Cognitive behavioral therapy of obsessive-compulsive disorder. *Dialogues in Clinical Neuroscience, 12*(2), 199-207.

Forand, N. R., & Derubeis, R. J. (2013). Pretreatment anxiety predicts patterns of change in cognitive behavioral therapy and medications for depression. *Journal of Consulting and Clinical Psychology, 81*(5), 774-782. doi:10.1037/a0032985

Fresco, D., Coles, M., Heimberg, R. G., Liebowitz, M., Hami, S., Stein, M., & Goetz, D. (2001). The Liebowitz Social Anxiety Scale: a comparison of the psychometric properties of self-report and clinician-administered formats. *Psychological Medicine, 31*(06), 1025-1035.

Furukawa, T. A., Noma, H., Caldwell, D. M., Honyashiki, M., Shinohara, K., Imai, H., . . . Churchill, R. (2014). Waiting list may be a nocebo condition in psychotherapy trials: a contribution from network meta-analysis. *Acta Psychiatrica Scandinavica, 130*(3), 181-192. doi:10.1111/acps.12275

Furukawa, T. A., Weitz, E. S., Tanaka, S., Hollon, S. D., Hofmann, S. G., Andersson, G., . . . Cuijpers, P. (2017). Initial severity of depression and efficacy of cognitive-behavioural therapy: individual-participant data meta-analysis of pill-placebo-controlled trials. *British Journal of Psychiatry, 210*(3), 190-196. doi:10.1192/bjp.bp.116.187773

Gyani, A., Shafran, R., Layard, R., & Clark, D. M. (2013). Enhancing recovery rates: Lessons from year one of IAPT. *Behaviour Research and Therapy, 51*. doi:10.1016/j.brat.2013.06.004

Haas, E., Hill, R. D., Lambert, M. J., & Morrell, B. (2002). Do early responders to psychotherapy maintain treatment gains? *Journal of Clinical Psychology, 58*(9), 1157-1172. doi:10.1002/jclp.10044

Hadjistavropoulos, H. D., Pugh, N. E., Hesser, H., & Andersson, G. (2016). Predicting Response to Therapist-Assisted Internet-Delivered Cognitive Behavior Therapy for Depression or Anxiety Within an Open Dissemination Trial. *Behavior Therapy, 47*(2), 155-165. doi:10.1016/j.beth.2015.10.006

Hannan, C., Lambert, M. J., Harmon, C., Nielsen, S. L., Smart, D. W., Shimokawa, K., & Sutton, S. W. (2005). A lab test and algorithms for identifying clients at risk for treatment failure. *Journal of Clinical Psychology, 61*(2), 155-163. doi:10.1002/jclp.20108

Harvey, A. G., Dong, L., Bélanger, L., & Morin, C. M. (2017). Mediators and Treatment Matching in Behavior Therapy, Cognitive Therapy and Cognitive Behavior Therapy for Chronic Insomnia. *Journal of Consulting and Clinical Psychology, 85*(10), 975-987. doi:10.1037/ccp0000244

Hedman, E., Ljótsson, B., Kaldo, V., Hesser, H., El Alaoui, S., Kraepelien, M., . . . Lindefors, N. (2014). Effectiveness of Internet-based cognitive behaviour therapy for depression in routine psychiatric care. *Journal of Affective Disorders, 155*, 49-58. doi:http://dx.doi.org/10.1016/j.jad.2013.10.023

Hedman, E., Ljotsson, B., & Lindefors, N. (2012). Cognitive behavior therapy via the Internet: a systematic review of applications, clinical efficacy and cost-effectiveness. *Expert Review of Pharmacoeconomics and Outcomes Research, 12*(6), 745-764. doi:10.1586/erp.12.67

Hiller, W., Schindler, A. C., & Lambert, M. J. (2012). Defining response and remission in psychotherapy research: A comparison of the RCI and the method of percent improvement. *Psychotherapy Research, 22*(1), 1-11. doi:10.1080/10503307.2011.616237

Hofmann, S. G., Asnaani, A., Vonk, I. J. J., Sawyer, A. T., & Fang, A. (2012). The Efficacy of Cognitive Behavioral Therapy: A Review of Meta-analyses. *Cognitive therapy and research, 36*(5), 427-440. doi:10.1007/s10608-012-9476-1

Houck, P. R., Spiegel, D. A., Shear, M. K., & Rucci, P. (2002). Reliability of the self-report version of the panic disorder severity scale. *Depression and anxiety, 15*(4), 183-185. doi:10.1002/da.10049

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy, 15*(4), 336-352. doi:https://doi.org/10.1016/S0005-7894(84)80002-7

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

Kaldo, V., Jernelöv, S., Blom, K., Ljótsson, B., Brodin, M., Jörgensen, M., . . . Lindefors, N. (2015). Guided internet cognitive behavioral therapy for insomnia compared to a control treatment – A randomized trial. *Behaviour Research and Therapy, 71*(Supplement C), 90-100. doi:https://doi.org/10.1016/j.brat.2015.06.001

Kaldo, V., Ramnerö, J., & Jernelöv, S. (2015). Involving Clients in Treatment Methods: A Neglected Interaction in the Therapeutic Relationship. *Journal of Consulting and Clinical Psychology, 83*(6), 1136-1141. doi:10.1037/ccp0000039

Karin, E., Dear, B. F., Heller, G. Z., Crane, M. F., & Titov, N. (2018). "Wish You Were Here": Examining Characteristics, Outcomes, and Statistical Solutions for Missing Cases in Web-Based Psychotherapeutic Trials. *JMIR Mental Health, 5*(2), e22. doi:10.2196/mental.8363

Karin, E., Dear, B. F., Heller, G. Z., Gandy, M., & Titov, N. (2018). Measurement of Symptom Change Following Web-Based Psychotherapy: Statistical Characteristics and Analytical Methods for Measuring and Interpreting Change. *JMIR Mental Health, 5*(3), e10200. doi:10.2196/10200

Karyotaki, E., Riper, H., Twisk, J., Hoogendoorn, A., Kleiboer, A., Mira, A., . . . Cuijpers, P. (2017). Efficacy of Self-guided Internet-Based Cognitive Behavioral Therapy in the Treatment of Depressive Symptoms: A Meta-analysis of Individual Participant Data. *Jama Psychiatry, 74*(4), 351-359. doi:10.1001/jamapsychiatry.2017.0044

Kessler, R. C., Chalker, S. A., Luedtke, A. R., Sadikova, E., & Jobes, D. A. (2020). A Preliminary Precision Treatment Rule for Remission of Suicide Ideation. *Suicide and Life-Threatening Behavior, n/a*(n/a). doi:10.1111/sltb.12609

Koutsouleris, N., Meisenzahl, E. M., Davatzikos, C., Bottlender, R., Frodl, T., Scheuerecker, J., . . . Gaser, C. (2009). Use of neuroanatomical pattern classification to identify subjects in at-risk mental states of psychosis and predict disease transition. *Archives of General Psychiatry, 66*(7), 700-712. doi:10.1001/archgenpsychiatry.2009.62

Kraepelien, M., Blom, K., Lindefors, N., Johansson, R., & Kaldo, V. (2019). The effects of component-specific treatment compliance in individually tailored internet-based treatment. *Clinical Psychology & Psychotherapy, 26*(3), 298-308. doi:10.1002/cpp.2351

Kuk, A. Y. C., Li, J. L., & Rush, A. J. (2010). Recursive Subsetting to Identify Patients in the STAR*D: A Method to Enhance the Accuracy of Early Prediction of Treatment Outcome and to Inform Personalized Care. *Journal of Clinical Psychiatry, 71*(11), 1502-1508. doi:10.4088/JCP.10m06168blu

Lambert, M. J. (2015). Progress feedback and the OQ-system: The past and the future. *Psychotherapy (Chic), 52*(4), 381-390. doi:10.1037/pst0000027

Lambert, M. J. (2017). Maximizing Psychotherapy Outcome beyond Evidence-Based Medicine. *Psychotherapy and Psychosomatics, 86*(2), 80-89. doi:10.1159/000455170

Lambert, M. J., Whipple, J. L., & Kleinstäuber, M. (2018). Collecting and delivering progress feedback: A meta-analysis of routine outcome monitoring. *Psychotherapy, 55*(4), 520.

Larsson, B. P. M., Kaldo, V., & Broberg, A.G. . (2009). Similarities and Differences Between Practitioners of Psychotherapy in Sweden: A Comparison of Attitudes Between Psychodynamic, Cognitive, Cognitive-Behavioral, and Integrative Therapists. *Journal of Psychotherapy Integration, 19*(1), 34-66. doi:10.1037/a0015446

Lenhard, F., Mitsell, K., Jolstedt, M., Vigerland, S., Wahlund, T., Nord, M., . . . Högström, J. (2019). The Internet Intervention Patient Adherence Scale for Guided Internet-Delivered Behavioral Interventions: Development and Psychometric Evaluation. *Journal of Medical Internet Research, 21*(10), e13602. doi:10.2196/13602

Lenhard, F., Sauer, S., Andersson, E., Månsson, K. N., Mataix-Cols, D., Rück, C., & Serlachius, E. (2018). Prediction of outcome in internet-delivered cognitive behaviour therapy for paediatric obsessive-compulsive disorder: A machine learning approach. *International Journal of Methods in Psychiatric Research, 27*(1), n/a-n/a. doi:10.1002/mpr.1576

Leucht, S., Fennema, H., Engel, R. R., Kaspers-Janssen, M., Lepping, P., & Szegedi, A. (2017). What does the MADRS mean? Equipercentile linking with the CGI using a company database of mirtazapine studies. *Journal of Affective Disorders, 210*, 287-293. doi:http://dx.doi.org/10.1016/j.jad.2016.12.041

Levallius, J., Clinton, D., Hogdahl, L., & Norring, C. (2020). Personality as predictor of outcome in internet-based treatment of bulimic eating disorders. *Eating Behavior, 36*, 101360. doi:10.1016/j.eatbeh.2019.101360

Lewis, C. C., Simons, A. D., & Kim, H. K. (2012). The role of early symptom trajectories and pretreatment variables in predicting treatment response to cognitive behavioral therapy. *Journal of Consulting and Clinical Psychology, 80*(4), 525-534. doi:10.1037/a0029131

Lutz, W., Arndt, A., Rubel, J., Berger, T., Schröder, J., Späth, C., . . . Moritz, S. (2017). Defining and Predicting Patterns of Early Response in a Web-Based Intervention for Depression. *Journal of Medical Internet Research, 19*(6), e206. doi:10.2196/jmir.7367

Mohr, D. C., Lattie, E. G., Tomasino, K. N., Kwasny, M. J., Kaiser, S. M., Gray, E. L., . . . Schueller, S. M. (2019). A randomized noninferiority trial evaluating remotely-delivered stepped care for depression using internet cognitive behavioral therapy (CBT) and telephone CBT. *Behaviour Research and Therapy, 123*, 103485. doi:https://doi.org/10.1016/j.brat.2019.103485

Mojtabai, R. (2017). Nonremission and time to remission among remitters in major depressive disorder: Revisiting STAR*D. *Depression and Anxiety, 34*(12), 1123-1133. doi:10.1002/da.22677

Monkul, E. S., Tural, U., Onur, E., Fidaner, H., Alkin, T., & Shear, M. K. (2004). Panic Disorder Severity Scale: reliability and validity of the Turkish version. *Depression and Anxiety, 20*(1), 8-16. doi:10.1002/da.20011

Montgomery, S. A., & Asberg, M. (1979). A new depression scale designed to be sensitive to change. *British Journal of Psychiatry, 134*. doi:10.1192/bjp.134.4.382

Morin, C. M., Belleville, G., Belanger, L., & Ivers, H. (2011). The Insomnia Severity Index: psychometric indicators to detect insomnia cases and evaluate treatment response. *Sleep, 34*(5), 601-608.

Mulder, R. T., Joyce, P. R., Frampton, C. M., Luty, S. E., & Sullivan, P. F. (2006). Six months of treatment for depression: outcome and predictors of the course of illness. *American Journal of Psychiatry, 163*(1), 95-100. doi:10.1176/appi.ajp.163.1.95

Månsson, K. N. T., Frick, A., Boraxbekk, C. J., Marquand, A. F., Williams, S. C. R., Carlbring, P., . . . Furmark, T. (2015). Predicting long-term outcome of Internet-delivered cognitive behavior therapy for social anxiety disorder using fMRI and support vector machine learning. *Translational psychiatry, 5*(3), e530-e530. doi:10.1038/tp.2015.22

Nadine, F., Tobias, K., Karine, C., Jean Baptiste, H., Jérôme, H., Mark, H., . . . Thomas, B. (2020). Using the Personalized Advantage Index for Individual Treatment Allocation to Blended Treatment or Treatment as Usual for Depression in Secondary Care. *Journal of Clinical Medicine, 9*(2), 490. doi:10.3390/jcm9020490

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011). Natural language processing: an introduction. *Journal of the American Medical Informatics Association, 18*(5), 544-551. doi:10.1136/amiajnl-2011-000464

Naudet, F., Millet, B., Michel Reymann, J., & Falissard, B. (2014). The fallacy of thresholds used in defining response and remission in depression rating scales. *International Journal of Methods in Psychiatric Research, 23*(4), 469-473. doi:10.1002/mpr.1393

Niculescu, A. B., Levey, D. F., Phalen, P. L., Le-Niculescu, H., Dainton, H. D., Jain, N., . . . Salomon, D. R. (2015). Understanding and predicting suicidality using a combined genomic and clinical risk assessment approach. *Molecular Psychiatry, 20*(11), 1266-1285. doi:10.1038/mp.2015.112

Nordgreen, T., Havik, O. E., Ost, L. G., Furmark, T., Carlbring, P., & Andersson, G. (2012). Outcome predictors in guided and unguided self-help for social anxiety disorder. *Behaviour Research and Therapy, 50*(1), 13-21. doi:10.1016/j.brat.2011.10.009

O'Mahen, H. A., Wilkinson, E., Bagnall, K., Richards, D. A., & Swales, A. (2017). Shape of change in internet based behavioral activation treatment for depression. *Behaviour Research and Therapy, 95*, 107-116. doi:https://doi.org/10.1016/j.brat.2017.05.011

Pauker, S. G., & Kassirer, J. P. (1975). Therapeutic decision making: a cost-benefit analysis. *New England Journal of Medicine, 293*(5), 229-234. doi:10.1056/nejm197507312930505

Pauker , S. G., & Kassirer , J. P. (1980). The Threshold Approach to Clinical Decision Making. *New England Journal of Medicine, 302*(20), 1109-1117. doi:doi:10.1056/NEJM198005153022003

Probst, T., Kleinstäuber, M., Lambert, M. J., Tritt, K., Pieh, C., Loew, T. H., . . . Delgadillo, J. (2020). Why are some cases not on track? An item analysis of the Assessment for Signal Cases during inpatient psychotherapy. *Clinical Psychology & Psychotherapy, n/a*(n/a). doi:10.1002/cpp.2441

Richards, D. A., Bower, P., Pagel, C., Weaver, A., Utley, M., Cape, J., . . . Vasilakis, C. (2012). Delivering stepped care: an analysis of implementation in routine practice. *Implement Science, 3*.

Rohrer, J. M. (2018). Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data. *Advances in Methods and Practices in Psychological Science, 1*(1), 27-42. doi:10.1177/2515245917745629

Rozental, A., Andersson, G., Boettcher, J., Ebert, D. D., Cuijpers, P., Knaevelsrud, C., . . . Carlbring, P. (2014). Consensus statement on defining and measuring negative effects of Internet interventions. *Internet Interventions, 1*(1), 12-19. doi:http://dx.doi.org/10.1016/j.invent.2014.02.001

Rozental, A., Andersson, G., & Carlbring, P. (2019). In the Absence of Effects: An Individual Patient Data Meta-Analysis of Non-response and Its Predictors in Internet-Based Cognitive Behavior Therapy. *Frontiers in Psychology, 10*, 15. doi:10.3389/fpsyg.2019.00589

Rozental, A., Andersson, G., & Carlbring, P. (2019). In the Absence of Effects: An Individual Patient Data Meta-Analysis of Non-response and Its Predictors in Internet-Based Cognitive Behavior Therapy. *Frontiers in Psychology, 10*(589). doi:10.3389/fpsyg.2019.00589

Salomonsson, S., Santoft, F., Lindsäter, E., Ejeby, K., Ingvar, M., Öst, L.-G., . . . Hedman-Lagerlöf, E. (2019). Predictors of outcome in guided self-help cognitive behavioural therapy for common mental disorders in primary care. *Cognitive Behaviour Therapy*, 1-20. doi:10.1080/16506073.2019.1669701

Saunders, R., Cape, J., Fearon, P., & Pilling, S. (2016). Predicting treatment outcome in psychological treatment services by identifying latent profiles of patients. *Journal of Affective Disorders, 197*, 107-115. doi:10.1016/j.jad.2016.03.011

Schibbye, P., Ghaderi, A., Ljótsson, B., Hedman, E., Lindefors, N., Rück, C., & Kaldo, V. (2014). Using Early Change to Predict Outcome in Cognitive Behaviour Therapy: Exploring Timeframe, Calculation Method, and Differences of Disorder-Specific versus General Measures. *PLoS ONE, 9*(6), e100614. doi:10.1371/journal.pone.0100614

Schlagert, H. S., & Hiller, W. (2017). The predictive value of early response in patients with depressive disorders. *Psychotherapy Research, 27*(4), 488-500. doi:10.1080/10503307.2015.1119329

Schroder, J., Jelinek, L., & Moritz, S. (2017). A randomized controlled trial of a transdiagnostic Internet intervention for individuals with panic and phobias - One size fits all. *Journal of Behavior Therapy and Experimental Psychiatry, 54*, 17-24. doi:10.1016/j.jbtep.2016.05.002

Schroder, J., Sautier, L., Kriston, L., Berger, T., Meyer, B., Spath, C., . . . Moritz, S. (2015). Development of a questionnaire measuring Attitudes towards Psychological Online Interventions-the APOI. *Journal of Affective Disorders, 187*, 136-141. doi:10.1016/j.jad.2015.08.044

Sheehan, D. V., Lecrubier, Y., Harnett Sheehan, K., Janavs, J., Weiller, E., Keskiner, A., . . . Dunbar, G. (1997). The validity of the Mini International Neuropsychiatric Interview (MINI) according to the SCID-P and its reliability. *European Psychiatry, 12*(5), 232-241. doi:10.1016/S0924-9338(97)83297-X

Siebern, A. T., & Manber, R. (2011). New developments in cognitive behavioral therapy as the first-line treatment of insomnia. *Psychology research and behavior management, 4*, 21-28. doi:10.2147/PRBM.S10041

Slade, K., Lambert, M. J., Harmon, S. C., Smart, D. W., & Bailey, R. (2008). Improving psychotherapy outcome: the use of immediate electronic feedback and revised clinical support tools. *Clinical Psychology & Psychotherapy, 15*(5), 287-303. doi:10.1002/cpp.594

Socialstyrelsen. (2017). *Nationella riktlinjer för vård vid depression och ångestsyndrom : stöd för styrning och ledning*. Stockholm: Socialstyrelsen.

Stjerneklar, S., Hougaard, E., & Thastum, M. (2019). Guided internet-based cognitive behavioral therapy for adolescent anxiety: Predictors of treatment response. *Internet Interventions, 15*, 116-125. doi:https://doi.org/10.1016/j.invent.2019.01.003

Svanborg, P., & Asberg, M. (1994). A new self-rating scale for depression and anxiety states based on the Comprehensive Psychopathological Rating Scale. *Acta Psychiatrica Scandinavica, 89*(1), 21-28.

Svanborg, P., & Asberg, M. (2001). A comparison between the Beck Depression Inventory (BDI) and the self-rated version of the Montgomery-Asberg Depression Rating Scale (MADRS). *Journal of Affective Disorders, 64*. doi:10.1016/s0165-0327(00)00242-1

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

Titov, N., Dear, B., Nielssen, O., Staples, L., Hadjistavropoulos, H., Nugent, M., . . . Kaldo, V. (2018). ICBT in routine care: A descriptive analysis of successful clinics in five countries. *Internet Interventions, 13*, 108-115. doi:https://doi.org/10.1016/j.invent.2018.07.006

Turkington, D., Wright, N. P., & Tai, S. (2013). Advances in Cognitive Behavior Therapy for Psychosis. *International Journal of Cognitive Therapy, 6*(2), 150-170. doi:10.1521/ijct.2013.6.2.150

van Straten, A., Hill, J., Richards, D. A., & Cuijpers, P. (2015). Stepped care treatment delivery for depression: a systematic review and meta-analysis. *Psychological Medicine, 45*(2), 231-246. doi:10.1017/s0033291714000701

Vickers, A. J. (2001). The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study. *BMC Medical Research Methodology, 1*(1), 6. doi:10.1186/1471-2288-1-6

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

Wise, E. A. (2004). Methods for analyzing psychotherapy outcomes: a review of clinical significance, reliable change, and recommendations for future directions. *Journal of Personality Assessment, 82*(1), 50-59. doi:10.1207/s15327752jpa8201_10

Wolff, J., Gary, A., Jung, D., Normann, C., Kaier, K., Binder, H., . . . Franz, M. (2020). Predicting patient outcomes in psychiatric hospitals with routine data: a machine learning approach. *BMC Medical Informatics and Decision Making*, *20*(1), 9. doi:10.1186/s12911-020-1042-2

von Glischinski, M., Willutzki, U., Stangier, U., Hiller, W., Hoyer, J., Leibing, E., . . . Hirschfeld, G. (2018). Liebowitz Social Anxiety Scale (LSAS): Optimal cut points for remission and response in a German sample. *Clinical Psychology and Psychotherapy*, *25*(3), 465-473. doi:10.1002/cpp.2179

Zagorscak, P., Heinrich, M., Schulze, J., Bottcher, J., & Knaevelsrud, C. (2020). Factors contributing to symptom change in standardized and individualized Internet-based interventions for depression: A randomized-controlled trial. *Psychotherapy (Chic)*. doi:10.1037/pst0000276