TREATMENT SELECTION: UNDERSTANDING WHAT WORKS FOR WHOM IN MENTAL HEALTH

Zachary Daniel Cohen

A DISSERTATION

in

Psychology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2018

Supervisor of Dissertation	Graduate Group Chairperson
_____	_____
Robert DeRubeis, Professor of Psychology	Sara Jaffee, Professor of Psychology

Dissertation Committee:

Dianne Chambless, Professor of Psychology

Robert DeRubeis, Professor of Psychology

Sara Jaffee, Professor of Psychology

TREATMENT SELECTION: UNDERSTANDING WHAT WORKS FOR WHOM IN MENTAL HEALTH

DEDICATION

To those who made it possible for me to pursue what I love, and who were patient with and believed in me for these (too) many years: my parents, grandparents, mentors, and advisor. And to my amazing wife, who inspires me to work hard and try to succeed.

ACKNOWLEDGMENT

As this is an academic work, I will begin by acknowledging the person without whom it would not exist, and who has shaped the vast majority of whatever positive qualities I have as a researcher and clinician: my advisor, mentor, and friend, Rob DeRubeis. Although his impact on clinical psychology has been recognized by many of the field's most prestigious academic awards, and most researchers dream of producing even a single of the many field-defining papers that Rob has produced across several different areas over the past three decades, if you ask Rob what he is most proud of, he will tell you that it's the award he received for mentorship. Those of us lucky enough to be trained by Rob have the gift and the burden of carrying on his legacy of critical, thoughtful, and innovative scientific research. For most of my seven years in graduate school, I spent more time in a room with Rob than with anyone else. He always treated me like family, and our travels around the world, from apartments in Amsterdam to camper-vans in Iceland, will remain some of my favorite graduate school memories. That level of face-time speaks not only to Rob's incredible dedication, but also to his patience, generosity, humor, camaraderie, and passion for science. Rob's unwavering support over my long and unconventional graduate career allowed me to never once lose faith in myself or my future. Duderino. Bossman. Couldn't have done it without you. Wouldn't have wanted to. Keep up the good stuff.

At the orientation on the first day of graduate school, they told us that academia was, at its core, a solitary and monastic pursuit. As is so often true in my life, I missed that memo, and have been truly blessed to have been constantly surrounded by the most wonderful and inspirational colleagues and teachers throughout my journey. Thank you

to my committee for helping guide me through this important process, and thank you Steve for setting me on this path and sending me to Rob. To my lab-mates, especially Lou – you always made it fun to be at "work" and I am constantly in awe of how much you all have accomplished. To my many collaborators and research assistants – I literally could not have accomplished anything that we have achieved over the last 5+ years without all of you. The fact that the lines between friends and co-workers have blurred beyond the point of recognition makes it easy to forget that this is technically a "job," and for that I am indescribably grateful.

I was extremely lucky to be born into the most loving, supportive family imaginable. My parents and sister have encouraged me to pursue my passions wherever they have taken me, and for that I will be forever grateful.

Finally, it is my greatest joy to get to thank my two ladies: my wife, Emily, and our dog, Samantha. Everyone should have a dog – they are the world's most fool-proof mental health treatment. The stress of graduate school melts away when you have a happy, furry, four-legged friend with you in the trenches. Although Emily and I met in UPenn's graduate program (where Emily was studying dogs), it took Samantha to get her to finally talk with me and agree to go out on a date. My amazing, rock-star academic wife is without question the best thing to have come out of my graduate school experience. You inspire me daily, and you make the hard work, long-hours, and sacrifice all worth it.

ABSTRACT

TREATMENT SELECTION: UNDERSTANDING WHAT WORKS FOR WHOM IN MENTAL HEALTH

Zachary Daniel Cohen

Robert DeRubeis

Individuals seeking treatment for mental health problems often have to choose between several different treatment options. For disorders like depression and PTSD, many of the available treatments have been found to be, on average, equally effective. Research on precision medicine aims to identify the most effective treatment for each patient. This work is based on the idea that individuals respond differently to treatment, and that these differences can be studied and characterized. The push for personalized and precision approaches in mental health involves identifying moderators - variables that predict differential response into treatment recommendations. Unfortunately, there has been little real-world application of these findings, in part due to the lack of systems suited to translating the information in actionable recommendations. This dissertation will review the history of treatment selection in mental health, and will present specific examples of treatment selection models in depression and PTSD. Differences between treatment selection in the context of two equivalently effective interventions and stratified medicine applications in which goal is to optimize the allocation of stronger and weaker interventions will be discussed. Methodological challenges in building (e.g., variable selection) and evaluating (e.g., cross-validation) treatment selection systems will be explored. Approaches to precision medicine being used by different groups will be compared. Finally, recommendations for future directions will be made.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: Treatment Selection in Depression

This work was originally published in Annual Review of Clinical Psychology:

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

Abstract

Mental health researchers and clinicians have long sought answers to the question "What works for whom?" The goal of precision medicine is to provide evidence-based answers to this question. Treatment selection in depression aims to help each individual receive the treatment, among the available options, that is most likely to lead to a positive outcome for them. Although patient variables that are predictive of response to treatment have been identified, this knowledge has not yet translated into real-world treatment recommendations. The Personalized Advantage Index (PAI) and related approaches combine information obtained prior to the initiation of treatment into multivariable prediction models that can generate individualized predictions to help clinicians and patients select the right treatment. With increasing availability of advanced statistical modeling approaches, as well as novel predictive variables and big data, treatment selection models promise to contribute to improved outcomes in depression.

Keywords: treatment selection, precision medicine, personalized medicine, stratified medicine, depression, mental health treatment

Introduction

Depression is the world's leading cause of disability (World Health Organization 2017). Despite the existence of a variety of evidence-based interventions for major depressive disorder (MDD), response rates in the treatment of depression remain approximately 50% (National Health Service 2016, Papakostas & Fava 2010). The pursuit of novel neurological (e.g., deep brain stimulation; Mayberg et al. 2005), pharmacological (e.g., ketamine; McGirr et al. 2015), and psychological (e.g., positive affect treatment; Craske et al. 2016) treatments is one avenue through which researchers are attempting to improve treatment outcomes (Holmes et al. 2014). This review focuses on an alternative approach: treatment selection, the aim of which is to provide for each individual the treatment, among the available options, that is most likely to lead to a positive outcome for them.

Half a century ago, Gordon Paul (1967) stated, in a paper that has been cited more than 1,000 times: "[i]n all its complexity, the question towards which all outcome research should ultimately be directed is the following: What treatment, by whom, is most effective **for this individual** with that specific problem, and under which set of circumstances?" The spirit of this passage—the question "What works for whom?"—has been invoked in countless discussions of evidence-based practices in clinical psychology.

The idea is a good one, recognizing that no single treatment is likely to be the best for everyone. How to address this issue, however, has not been obvious or simple. In recent years, researchers have developed and tested the utility of multivariable prediction models to address the "What works for whom?" question. The promise of this work lies in the ability of such models to integrate multiple sources of information, rather than to

rely on a single feature to inform treatment selection. In other areas of medicine, the effort to match individuals to their indicated treatments is called precision medicine (Hamburg & Collins 2010), which has largely replaced the term personalized medicine (Katsnelson 2013, Schleidgen et al. 2013). Precision medicine[1] has afforded major advances in cancer treatment (National Research Council 2011, Schwaederle et al. 2015). For example, chemotherapy is the standard treatment for non–small-cell lung carcinoma (NSCLC). Early trials of the drugs erlotinib and gefitinib found little to no benefit of these drugs alone or in combination with chemotherapy (Pao & Miller 2005). However, recent clinical trials have found significantly improved outcomes for these drugs, relative to chemotherapy, in a specific subset of NSCLC patients with tumor mutations linked to the mechanisms of action of erlotinib and gefitinib (Paez et al. 2004, Rosell et al. 2012). We believe that similar approaches can help improve outcomes in mental health.

In this review, we describe several approaches to selecting the right treatment for an individual with depression. A striking feature of efforts in this area is the heterogeneity of the statistical approaches that are employed (Petkova et al. 2017, Weisz et al. 2015).

Variables that predict outcome are of two kinds: prescriptive or prognostic. Prescriptive variables, often referred to as moderators, affect the direction or strength of

[1]As defined by a National Research Council report, precision medicine "refers to the tailoring of medical treatment to the individual characteristics of each patient. It does not literally mean the creation of drugs or medical devices that are unique to a patient, but rather the ability to classify individuals into subpopulations that differ in their susceptibility to a particular disease, in the biology and/or prognosis of those diseases they may develop, or in their response to a specific treatment. Preventive or therapeutic interventions can then be concentrated on those who will benefit, sparing expense and side effects for those who will not. Although the term 'personalized medicine' is also used to convey this meaning, that term is sometimes misinterpreted as implying that unique treatments can be designed for each individual." (National Research Council 2011, p. 125).

the differences in outcome between two or more treatments (Baron & Kenny 1986), and thus can help predict whether a patient will benefit more from one treatment relative to another. Cronbach (1957) described prescriptive relationships as "aptitude-by-treatment" interactions, which have typically been explored through subgroup or subset analysis (Doove et al. 2014, Wang & Ware 2013). **Figure 1** displays a variety of types of prescriptive relationships, which can be ordinal (sometimes called quantitative interactions) or disordinal (sometimes called qualitative interactions; involving a full crossover) (Gail & Simon 1985, Gunter et al. 2011a, Wellek 1997, Widaman et al. 2012). Fournier et al. (2008) reported an example of a disordinal moderator in depression: The presence of a comorbid personality disorder (PD) predicted better response with antidepressant medication (ADM), relative to cognitive therapy (CT), and its absence predicted a better response to CT than to ADM.

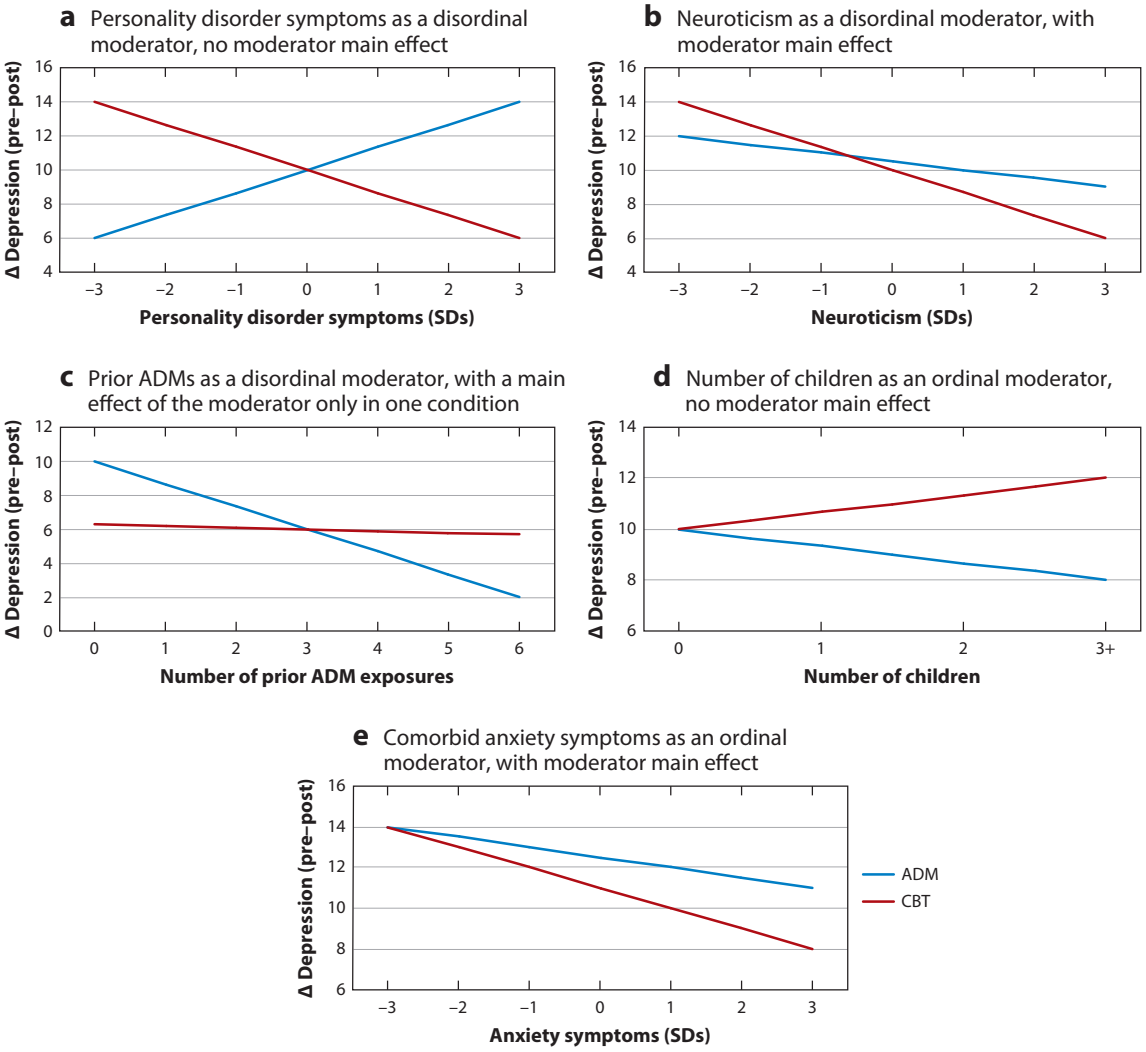


Figure 1. Five ways in which the differential effects of two treatments can vary as a function of a continuous moderator variable, and in which the interactions between treatment and moderator are linear. The relationships shown are for illustrative purposes only, but they draw on observations in the relevant empirical literatures. In *a–c*, at high levels of the moderator, one treatment is expected to produce stronger effects, whereas at low levels the other treatment is expected to be superior (disordinal interactions). In *d* and *e*, one of the two treatments is superior, on average, but the degree of superiority is expected to vary with the level of the moderator (ordinal interactions). In *b*, *c*, and *e*, the moderator is also a prognostic variable, such that a score on the moderator predicts outcome, independent of treatment (moderator main effects). In *a* and *d*, there is no moderator main effect. The moderator is predictive only in concert with treatment. Higher change scores on the y-axis indicate more improvement. Abbreviations: ADM, antidepressant medication; CBT, cognitive-based therapy; SD, standard deviation

A variable is prognostic if it predicts response in a single treatment, or irrespective of treatment condition. If only one intervention is being analyzed, only prognostic relationships can be inferred. Although a predictor[2] may appear to be prognostic in a single-treatment analysis, it might predict differential treatment response in a study that compares two or more treatments. Additionally, a variable can function as a prognostic predictor in one context and as a prescriptive predictor in another. For example, higher depression severity is associated with worse outcomes in depression. In comparisons of medication to CT, baseline severity is prognostic because it has the same relationship to outcome in both treatments (Weitz et al. 2015). However, in comparisons of medication to placebo (Ashar et al. 2017) or of psychotherapies to control conditions, higher baseline severity predicts a larger advantage of the active treatment over the control, making severity prescriptive in these contexts (Driessen et al. 2010, Fournier et al. 2010).

Therapists select treatments for patients as a matter of course in every day practice. A clinician who attends to information about a specific client's presentation will likely generate hypotheses about the client's expected responses to potentially available treatments (Lorenzo-Luaces et al. 2015). These predictions may draw on a variety of sources, including a clinician's history with clients with similar features, their experiences in training and supervision, reasoning based on theory, and the empirical literature on treatment response (Raza & Holohan 2015). However, there is limited

[2]The term predictor is sometimes used to refer specifically to prognostic relationships, but it can also be used to refer broadly to both prognostic and prescriptive variables, which is the way we use it here.

empirical literature to guide personalized, or precision, selection of treatments. Clinicians are therefore forced to practice what Perlis (2016) has dubbed "artisanal medicine."

Artisanal medicine is the practice of making treatment decisions in an idiosyncratic or unsystematic manner, or in a manner guided by theory and experience but largely uninformed by empirical evidence or feedback. Unfortunately, the lack of standardization that defines artisanal medicine limits the validity and utility of such approaches for decision making (Dawes 1979, 2005; Dawes et al. 1989; Tversky & Kahneman 1983).

In treatment contexts, statistical decision making, also called actuarial decision making, relies on predictions made with the use of algorithms, in a reproducible way. Grove et al. (2000) detail the ways in which actuarial approaches to decision making can overcome limitations and biases prevalent in human judgment (Dawes et al. 1989, Pauker & Kassirer 1980). By and large, empirical tests of clinical versus actuarial prediction (Grove & Meehl 1996) have revealed the superiority of actuarial methods. More than 60 years ago, Paul Meehl (1954) published his seminal monograph on this topic, titled *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. The field of mental health treatment has only just begun to apply Meehl's line of thinking to precision mental health.

For much of the twentieth century, evidence-based practice in mental health has largely concerned the provision of a specific treatment to patients based on a specific DSM-defined disorder. Evidence to guide such decisions has come from randomized clinical trials (RCTs), in which active treatments are compared to control conditions or to other active treatments (Chambless & Hollon 1998). For example, on the basis of positive

findings in RCTs, CT (Beck et al. 1979) and interpersonal therapy (IPT; Klerman & Weissman 1994) are each considered evidence-based psychotherapies for MDD. Similarly, among the psychoactive medications, specific classes of drugs have been studied under the assumption that they have differential efficacy with specific disorders (Fineberg et al. 2012).

However, as has been widely discussed in the literature, the library of empirically supported treatments (ESTs) is insufficient to address clinician and client needs (Hollon et al. 2002). The average treatment effect (ATE) is the extent to which, for the clients in the sample, a given intervention leads to more (or less) symptom improvement, on average, relative to comparator conditions. The main findings from RCTs refer to effects of treatments, on average, and not to potentially important sources of variability in treatment response (Imai & Ratkovic 2013, Kessler et al. 2017). Consider, for example, a case in which an ATE of 10 points on change in the Beck Depression Inventory (BDI) is estimated in a comparison of a strong treatment versus a weaker intervention. This might reflect an average change of 20 in the strong condition and 10 in the weak condition. It would be a mistake to assume that, even with such a large average change in the strong treatment condition, every client benefited substantially more from it than they would have from the weak condition. In fact, it is typical in such studies that reductions in the BDI observed in some clients are near the group average, whereas in others the reductions will be quite large, and in still others there will be little or no change observed. Understanding the heterogeneity of treatment effects could facilitate treatment selection by identifying individuals for whom more than a 10-point advantage of the stronger

treatment would be expected, as well as individuals for whom the weaker treatment might be equally, or even more, effective (Kessler et al. 2017).

INTRODUCTION TO RESEARCH ON PREDICTIVE VARIABLES

Overview

Variables examined in research on the prediction of mental health outcomes in depression have been drawn from a variety of sources, including routinely assessed domains such as demographic, environmental, and diagnostic information. A recent emphasis on neurobiological variables (Gabrieli et al. 2015, Jollans & Whelan 2016, Leuchter et al. 2009, Pizzagalli 2011, Stephan et al. 2017) has begun to reveal the potential of the inclusion of neurocognitive (Gordon et al. 2015) and biomarker-based assessments as predictors (Uher et al. 2014). For example, McGrath et al. (2013) measured pretreatment brain activity using positron emission tomography in a depression RCT and identified right-anterior-insula metabolism as a disordinal moderator: Insula-hypometabolism was associated with better outcomes in CT and worse outcomes with ADM, while insula-hypermetabolism was associated with the opposite pattern. In **Supplemental Table 1** we list recent reviews of predictors in depression, the results of which could inform future treatment selection investigations.

By far, the most common approach to the prediction of treatment response in mental health is to take advantage of the information captured in prognostic relationships (e.g., Rubenstein et al. 2007). Prognostic statements regarding response to intervention are of the following form: A client with characteristic X, in a given context (e.g., with

any intervention, or with a specific treatment[3]), has a Y% chance of experiencing symptom remission. Prognostic information can be used to provide realistic expectations to the treating clinician, as well as the client and their family (Kessler et al. 2016). This includes expectations concerning the rapidity and extent of response to the treatment that will be provided, as well as whether special attention should be paid to the client's progress (Hunter et al. 2011, Lutz et al. 2014).

The information conveyed in a prognostic statement does not inform directly the following question: "What is the best available option for this client at this time?" A common mistake in the interpretation of a prognostic finding is to conclude that clients found to have a poor prognosis in a given treatment will have a better prognosis in a different treatment (Simon & Perlis 2010). Consider the finding that, in CT, patients with chronic depression exhibit lower recovery rates than those with nonchronic depression (Fournier et al. 2009). This might indicate that other interventions (e.g., ADMs) or treatments created specifically for chronic depression, such as Cognitive Behavioral Analysis System of Psychotherapy (CBASP; McCullough Jr 2003), should be preferred over CT for individuals with chronic depression. However, it could instead be that CT is as effective as (or even more effective than) other available interventions for such individuals (Cuijpers et al. 2017). Indeed, evidence from an RCT comparing CT to ADM suggested that chronicity is prognostic, in that it was associated with similarly lower response rates in both treatments (Fournier et al. 2009), and an RCT comparing CBASP

[3]We are referring to a case in which a characteristic predicts outcome in studies of a single treatment, and in which its predictive value is unknown in other contexts. Note that if a factor predicts outcome in one treatment but does not predict outcome in a second, it could be prescriptive in that context (see Figure 1*c* for an example).

to ADM in individuals with chronic depression found no difference in response rates (Nemeroff et al. 2003). The only type of investigation that can directly address the prescriptive question (i.e., "Which treatment is likely to be most effective for a client with X, Y, and Z characteristics?") is one that focuses on moderation[4]. Unfortunately, analyses of this type are much less frequently conducted (see **Supplemental Example-1** for an early example of the single moderator approach, with a twist, from Beutler et al. 1991).

Studies in which pretreatment variables are found to predict treatment response can provide clues about treatment mechanisms (typically identified in efforts to find variables that mediate a treatment effect) (MacKinnon et al. 2007), and thus can help distinguish between compensation and capitalization models of the effects of psychotherapies. The compensation model is that individuals with deficits in areas targeted by a therapy will benefit the most from it. An example of this is the hypothesis that CT, which targets dysfunctional cognitions, would be preferred over IPT for individuals high on cognitive dysfunction and low on interpersonal functioning, and vice versa. The support for this hypothesis is equivocal. Capitalization models, which propose that therapies work best when they build on clients' strengths, have received some support (Barber & Muenz 1996, Cheavens et al. 2012).

[4]Prescriptive questions can be investigated through the simultaneous use of two or more prognostic models in the same sample (e.g., see Kessler et al. 2017)

Understanding Moderator Relationships

Given the observed heterogeneity in the presentation, history, and prognosis of depression, it is unlikely that any single variable in isolation will have clinically useful predictive utility (Simon & Perlis 2010). Nonetheless, considering how a single moderator would guide treatment selection is a useful exercise for enhancing one's understanding of how multivariable treatment selection algorithms work. To that end, we created plots (see **Figure 1**) depicting hypothetical examples[5] of prescriptive relationships that could be observed with continuous moderators. In **Supplemental Figure 1**, we also discuss the application to clinical decision making of findings of prescriptive relationships when the moderators are binary, as well as when prognostic predictors are identified.

In empirical reports of moderator findings, the distinctions between different types of prescriptive relationships illustrated in **Figure 1** are rarely made, and when the details of these relations are implied they can be inconsistent, misleading, or incorrect. Issues with data processing and the behavior of regression coefficients can make interpreting and describing moderator relationships difficult even for the individuals who perform the analyses (Kraemer & Blasey 2004). To learn more about these topics, we refer the reader to Kuhn & Johnson (2013).

Consider the following statement: "Clients high on characteristic Z experienced superior outcomes with ADM, relative to CT." It is tempting to infer that those who are

[5]These examples are for illustrative purposes only. We drew upon patterns that have been observed in empirical studies, but the figures do not represent empirical findings, per se. We followed the structure used by Kraemer (2013) and Schneider et al. (2015) in creating these figures.

low on characteristic Z would respond better to CT than to ADM, but there is nothing in the statement about such individuals. Thus, the statement could describe any of a variety of relationships, including those that **Figure 1*a*,*b*,*e*** depicts. If **Figure 1*e*** were true, a clinician should encourage individuals with high levels of Z to pursue ADM, whereas individuals with low levels of Z should be informed that there is no indication of a meaningful difference between the two treatments. However, the relationship could also be characterized by the pattern in **Figure 1*a*,*b***. If this were true, individuals high on Z would receive the same recommendation (choose ADM), but individuals low on Z should be steered away from ADM and toward CT. In the case of **Figure 1*a***, an individual with an average level of Z would be informed that the two treatments are expected to be similarly effective for him or her, and the expected size of the advantage of one treatment over the other is similar at each end of the spectrum. In the case of **Figure 1*b***, the expected advantage of ADM over CBT for those high on Z is larger than the expected advantage of CBT over ADM for those low on Z. This example illustrates one of the many ways in which the translation from analysis to interpretation to implementation can result in either optimal, suboptimal, or even harmful application of prescriptive information to clinical decision making.

The importance of evaluating the evidence for predictors prior to utilizing them in clinical settings deserves special emphasis in treatment selection (Howland 2014, Perlis 2016). When reading an empirical investigation of individual differences in treatment response, one must identify the population from which the sample was drawn. Although a paper may describe its findings as pertaining to "treatment response in depression," it is necessary to attend to specific features of the sample (e.g., inclusion/exclusion criteria,

range of depression severity, extent of comorbidity, treatment history) to determine the pertinence of the evidence to a specific client. For example, depressive symptom severity has been reported to predict differential response to ADMs versus placebos, with ADMs evidencing superiority over placebo for moderate to high severity, and little to no differences seen at the lower end of the severity spectrum (Barbui et al. 2011, Fournier et al. 2010, Khan et al. 2002, Kirsch et al. 2008). However, for most trials, entry criteria include moderate or greater symptom severity (Zimmerman et al. 2015, 2016), thus restricting the range of severity that can be investigated, and constraining the applicability of many positive moderator findings to a subset of the population of patients with MDD. There are many examples of predictive algorithms built using data from a sample of clients treated with one antidepressant that have failed to generalize to a different antidepressant (Chekroud et al. 2016, Iniesta et al. 2016a, Perlis et al. 2010). Similarly, models predicting the onset of major depressive episodes in European primary care (King et al. 2008) have not generalized to the US general population (Nigatu et al. 2016).

Reliance on tests of significance can result in misleading impressions about the importance of predictive variables (Nuzzo 2014, Wasserstein & Lazar 2016). For example, if a variable selection approach relies on p-values (often with $p < .05$ as a threshold) to assess statistical significance, a variable could miss a predetermined cut-off by a small margin, leading to a report that the variable is not predictive (Bursac et al. 2008). However, the difference in the predictive utility of an excluded variable that "just missed" (e.g., $p = .06$) and an identified predictor that is "barely" significant (e.g., $p = .04$) is, of course, trivial (Mickey & Greenland 1989). Most RCTs are powered to detect

main effects, and therefore are powered to detect only very strong interactions. Complicating matters further, different analytic approaches can identify different variables, even when applied to the same data (Cohen et al., *under review*). Additionally, variables that were not assessed, or that were assessed and not analyzed, could also be important predictors. Finally, statistically significant results are not necessarily clinically significant, if effect sizes are small (Meehl 1978). In the context of a large sample, small or weak relationships can be identified as statistically significant. However, statistically significant variables are not always good predictors (Lo et al. 2015). More relevant are metrics that can characterize the importance of the relationship and can therefore quantify and translate the clinical meaning of the findings (Bossuyt & Parvin 2015). For example, Janes et al. (2011) developed a statistical method for evaluating treatment selection markers that went beyond the classic approach of testing for a statistical interaction between a predictor and treatment to answer four important questions: "1) Does the marker help patients choose among treatment options?; 2) How should treatment decisions be made that are based on a continuous marker measurement?; 3) What is the impact on the population of using the marker to select treatment?; and 4) What proportion of patients will have different treatment recommendations following marker measurement?" (p. 253). Moving beyond statistics, consideration of factors such as cost, feasibility, and client burden should be weighed against the additive predictive power of variables that exceed those routinely collected in clinical settings (Perlis et al. 2009).

Two Frequently Cited Treatment Selection Variables

In real-world contexts, two variables often influence treatment selection in depression. The first is client preference (McHugh et al. 2013): Many treatment

guidelines (Hollon et al. 2014) specify the importance of attending to clients' preferences. However, studies of the predictive utility of client preference include positive (Kocsis et al. 2009, Mergl et al. 2011, Swift & Callahan 2009), mixed (Dunlop et al. 2017, Group 2008, McHugh et al. 2013), and negative (Dunlop et al. 2012b, Leykin et al. 2007b, Renjilian et al. 2001, Winter & Barber 2013) findings. Seemingly, contrary to lay intuition, preference is not a reliable indicator of treatment response. What's more, patients' preferences might shift when given individualized information about expected outcomes.

Second, a client's experience with previous treatments for depression can serve a prognostic or prescriptive function, as suggested by findings from several outcome studies. Prior exposure to ADMs and history of nonresponse to ADMs have each been found consistently to predict poor response to future courses of antidepressants (Amsterdam et al. 2009, 2016; Amsterdam & Shults 2009; Byrne & Rothschild 1998. Moreover, there is evidence that the number of prior ADM exposures can provide prescriptive information. For example, Leykin et al. (2007a) found that multiple previous ADM exposures predicted a poorer response to ADM, but not to CT, such that a client with two or more prior exposures was significantly more likely to benefit from CT than ADM. Clearly, assessing treatment history is important and could be used to inform treatment selection.

THE PERSONALIZED ADVANTAGE INDEX APPROACH

In 2011, we, along with other members of our research team, began to explore the possibility that machine learning[6] (Iniesta et al. 2016b, Passos et al. 2016) or multivariable regression modeling approaches could be brought to bear on problems in precision mental health. We initiated our journey with a specific goal in mind: to find or develop an approach that could identify clients with MDD for whom antidepressants are likely to be more beneficial than CT, and vice versa. Two findings from our lab prompted our interest. First, in a sample of clients with moderate to severe MDD, ADM and CT had produced nearly identical group-average effects on depressive symptoms over the course of a 16-week RCT (DeRubeis et al. 2005). Second, we had discovered five variables (marital status, employment status, PD comorbidity, antidepressant treatment history, and number of recent stressful life events) that served as moderators of symptom change in this sample (Fournier et al. 2009).

What was striking about the variables that moderated the effects of ADM versus CT was that no single one dominated the differential predictions. To survive the variable selection procedure, each variable had to make an independent contribution to the statistical model. As a result, the variables needed to be relatively uncorrelated with each other in the sample, such that they could not be used to define a factor, per se. Rather, we had identified five vectors, represented by the five variables, any one of which could be used to point a client to either ADM or CT as their preferred treatment, although there was not an especially strong predictor in the bunch. We understood this to indicate that

[6] Gillan & Whelan (2017) explain the following: "Machine-learning (essentially synonymous with 'data-mining' or 'statistical learning') refers to a class of approaches that focus on prediction rather than interpretation or mechanism." (p. 35)

there are many "reasons" one treatment may be more effective than another for a given person.

This posed a challenge for selecting treatments for patients with contradicting indications. For example, as noted above, clients with comorbid PD improved more with ADM than they did in CT, whereas clients without comorbid PD improved more in CT than with ADM (Fournier et al. 2008). It was also the case that clients who were unemployed improved more in CT than with ADM. How is a clinician to use this information in recommendations to a client with comorbid PD (indicating ADM) who was unemployed (indicating CT)? How does the clinician integrate information when considering different recommendations from the other three variables when forming a treatment recommendation? The implication of the literature on actuarial versus clinical decision making, which has focused on prognosis, is that outputs from a well-constructed statistical method should be able to provide useful information in treatment selection contexts, as well.

We also reasoned that effective guidance for clinicians and clients would, ideally, be "graded," to reflect the likelihood that for some clients differential benefit would be expected to be quite substantial, for others it would be negligible, and for others in between. To address these challenges, we developed the Personalized Advantage Index (PAI) approach (DeRubeis et al. 2014a), which has since been featured in work both internal and external to our lab (Huibers et al. 2015, Vittengl et al. 2017, Zilcha-Mano et al. 2016; Cohen et al., *under review;* Keefe et al., 2018; Webb et al., 2018).

Essential to the PAI approach is the identification of variables in a dataset that predict differential response to two or more treatments. Once the variables have been

identified, a multivariable statistical model that includes interaction terms representing the prescriptive variables[7] is constructed. A PAI for a given client is then calculated as the difference between their predicted outcomes in two treatments (treatment A and treatment B). To generate the prediction for treatment A, the client's values on the baseline variables, as well as the value representing treatment A, are inserted into the model. This is repeated, with the value of treatment B inserted into the model. The predicted value with treatment A in the model is compared with the predicted value under treatment B. The sign of the difference reflects the model-indicated treatment, and the magnitude of the difference reflects the magnitude, or strength, of the predicted difference. We return to a focus on the PAI approach more specifically in a discussion of issues of broader importance in treatment selection, following a review of literatures on a variety of prognostic and prescriptive multivariable approaches in mental health.

REVIEW OF THE LITERATURE ON MULTIVARIABLE PREDICTION MODELS

Overview

If a single predictive variable with a very large effect can be identified in a treatment context, application to practice is likely to be straightforward. In depression, however, such variables have not been identified consistently. In part for this reason, single variables have not found widespread use in treatment selection contexts. One exception to this is baseline symptom severity, which has been included in many practice

[7]Some of the machine-learning models we have constructed do not include interaction terms per se, but they perform the same task of modeling differential response.

guidelines as an indication that stronger treatments, or the combination of ADMs and psychotherapy, are to be preferred over lower-intensity interventions (American Psychiatric Association 2010, National Institute for Health and Clinical Excellence 2009). The status of baseline severity as a prescriptive variable has been supported primarily in comparisons of an active treatment with a control (Driessen et al. 2010, Fournier et al. 2010), but not in comparisons of two active treatments (Vittengl et al. 2016, Weitz et al. 2015).

Multivariable models are more likely to yield powerful predictions (Perlis 2013), and they comport with our understanding of psychopathology and treatment response as complex, multiply determined phenomena (Drysdale et al. 2017). Unfortunately, the interpretation and application of multivariable models is less straightforward for the clinician than are single-variable approaches. To further complicate matters, it may be important not only to consider multiple variables simultaneously, but also to consider potential interactions among multiple variables (Tiemens et al. 2016). As new, more powerful modeling approaches become available (Kapelner & Bleich 2013, Luedtke & van der Laan 2016, Ma et al. 2016), researchers must weigh the increased flexibility and predictive power of such approaches against the interpretability (Hastie et al. 2009, James et al. 2013) of simpler models (Green & Armstrong 2015), especially insofar as the goal is to disseminate the models in ways that are acceptable to clinicians and clients (Delgadillo et al. 2016).

Prognostic Models

Recent multivariable modeling efforts (Chekroud et al. 2017) highlight the potential for these advanced approaches to improve prognostic prediction in mental health (see

Gillan & Whelan 2017 for an extensive review). For example, Chekroud et al. (2016) used archival data from the Sequenced Treatment Alternatives to Relive Depression (STAR*D) study (Trivedi et al. 2006) to identify predictors of response to acute selective serotonin reuptake inhibitor (SSRI) treatment of depression. They validated their model in an external sample from a separate study, the Combining Medications to Enhance Depression Outcomes trial (Rush et al. 2011). Interestingly, although they found that it had acceptable predictive power in the study's two SSRI conditions, it was not significantly predictive for the validation sample's non-SSRI ADM condition, suggesting that the model may not work outside of the drug class on which it was developed.

It is understandable that more progress has been made to date with prognostic models, relative to prescriptive models, but the former have less relevance to a core question in mental health treatment: "Which treatment should client X pursue to have the greatest likelihood of having a positive outcome?" In many medical contexts, treatment recommendations follow from accurate diagnosis. However, for many mental health diagnoses, especially depression, there exists an abundance of ESTs that a client could potentially receive. Thus, in mental health, using person characteristics to help guide individuals to their optimal treatment is especially important. Uher et al. (2012) noted the limitation of their prognostic model, which predicts response to ADM in MDD, relative to a prescriptive model: "Clinical application of this finding will require identification of a treatment that is effective in individuals [identified as less likely to respond to ADM**]**" (p. 976).

TREATMENT SELECTION APPROACHES

Overview

Byar & Corle (1977) published an early example of a multivariable treatment selection model in medicine. Working with longitudinal data from a sample of men who were randomized to one of two treatments for prostate cancer, they explored whether, for each man, the more promising treatment of the two could be identified using a set of characteristics ascertained prior to random assignment. At the time, the field's emphasis had been on discovering, for all patients with a given diagnosis, "which treatment is best." Byar & Corle capitalized on advances in statistical methodology that allowed for survival modeling with multiple covariates and used the heterogeneity of patients to develop a rubric that could, in principle, inform individual treatment recommendations. Byar (1985) later applied this general approach to the differential prediction of survival in response to two dosage levels of chemotherapy for prostate cancer. Surprisingly, not until 1994 was any of Byar's work cited by others in the context of actuarial modeling in treatment selection. Yakovlev et al. (1994) applied a similar methodology to a treatment selection problem in cervical cancer, but from 1994 until 2011 (Gunter et al. 2011b) none of these works was cited in a publication that applied or extended the differential prediction methods described by Byar or Yakovlev.

The past half-decade has witnessed a surge of interest in optimizing treatment selection using multivariable predictive models, and much of this work is focused on treatments for depression. When moving beyond prognostic prediction into treatment selection, several additional considerations come into play. The first factor to consider is whether the treatment decision is between two or more equivalent interventions or,

instead, between a stronger versus a weaker intervention, as this distinction has implications for how one builds and evaluates the models. We begin our discussion focusing on contexts in which the decision is between equally effective treatments, when the question truly is "What will work best for each given patient?" We follow this with a review of the special case of stratified medicine (stepped-care), in which at least one of the candidate interventions results in greater improvement than a comparison condition, on average.

One of the earliest examples of multivariable treatment selection in mental health came from Barber & Muenz's (1996) reanalysis of the Treatment of Depression Collaborative Research Program (Elkin et al. 1989) study, which compared CT to IPT for MDD. The authors built a "matching factor" that combined the prescriptive value of marital status, avoidance, obsessiveness, and baseline severity in a linear model predicting symptom change. They also tested the prescriptive value of two personality disorder diagnoses, avoidant PD and obsessive-compulsive PD, and proposed that the models including these factors could be used to match patients to CT or IPT.

Lutz et al. (2006) employed a statistical technique called "nearest-neighbors" to predict differential outcomes between two variations of CT. In the nearest-neighbors method, each client's outcome in each treatment is predicted from the average observed outcomes in the respective treatments of groups of clients who are most similar to the index client on a set of features.

Kraemer (2013) proposed a method that involves the creation of a single variable (termed M^*) that represents a weighted combination of multiple moderators. This approach was demonstrated using data from a randomized comparison of IPT versus

ADM (Wallace et al. 2013). The statistical approach behind the *M** method excludes any consideration of main effects in an attempt to maximize the power of the differential prediction of outcome (Kraemer 2013). Thus, two clients with identical *M** scores could have very different prognoses, but this information is not given by the method. Recently, the *M** approach has been used to analyze data from a comparison between aripiprazole augmentation and placebo augmentation for ADM-treatment-resistant late-life depression (Smagula et al. 2016), and between two psychological treatments for clients with anxiety disorders (Niles et al. 2017a,b).

A series of papers by Uher et al. demonstrates the evolution of treatment selection from single to multivariable approaches. Using data from the Genome-Based Therapeutic Drugs for Depression study (Uher et al. 2009), they tested the prognostic and prescriptive utility of three symptom clusters (factors) and the six symptom dimensions that made up the factors (Uher et al. 2012). They examined the predictive power of each of these nine variables in isolation and found evidence for only the anxiety symptom dimension as a moderator. Recently, they returned to the question of treatment selection in this sample, using a multivariable approach with an expanded set of potential variables (Iniesta et al. 2016a). They found that a model that simultaneously included the effects of multiple variables could predict differential response to antidepressants with clinically meaningful accuracy, thus demonstrating the potential of multivariable approaches for treatment selection.

Other groups have used variants of the methods already described to address treatment selection questions (Cloitre et al. 2016, Westover et al. 2015). In **Supplemental Table 2** we contrast some of the approaches used in the multivariable prediction work

referenced in this review. Although this abundance demonstrates the strong interest in precision medicine, the heterogeneity of methods (Doove et al. 2014) contributes to difficulties in detecting consistencies and inconsistencies in predictors, and creates a barrier to identifying "best practices."

To date, most attempts to build prescriptive models for treatment selection have utilized data from RCTs. Future efforts, exemplified by the ongoing work of Gillan et al. to collect mental health treatment outcome data online (Gillan & Daw 2016), will likely also rely on nonrandomized data. The potential influence of unknown confounds is a limitation of treatment selection efforts outside the context of RCT data. The bias in predictions in such studies can derive from "selection effects," which result when clients with a given feature (e.g., history of nonresponse to ADMs) are provided with one of the treatments preferentially (e.g., CT). In these contexts, approaches such as propensity score analysis (d'Agostino 1998) can be employed to mitigate the effects of confounds.

Extending the Personalized Advantage Index to Stratified Medicine

When a model is developed to guide treatment decisions in health care contexts in which the available interventions differ in terms of strength, cost, availability, or risk, the question "Which treatment is predicted to be most effective for each individual?" may be moot. The treatment with the strongest effect on average is likely to be predicted to be the most effective one for most or all of the clients. In these contexts, the more relevant question is often "What is the best way to allocate the stronger/costlier/less available/riskier (hereafter 'stronger') treatment?" The practical goals of predictive models in stratified medicine are to enhance the efficient allocation of scarce or costly resources, as well as to limit patients' unnecessary exposure to treatments that require

substantial time commitments or are associated with heightened side effect risk (Hingorani et al. 2013).

Considerations of treatment allocation for stronger versus weaker interventions, including part-whole comparisons (e.g., combined ADM + CT versus ADM alone) should address the distinction between two ways in which the stronger treatment produces superior average change. One possibility is that every client is expected to benefit more—and by a similar amount—from the stronger treatment. In such cases, decisions about who should be provided the stronger treatment will not be based on clients' predictive features, except insofar as the clients with the worst prognoses might be provided the strongest treatment, for ethical reasons. However, it may be that individuals vary in regard to how much more they will benefit from the stronger treatment, relative to the weaker one. In such cases, it becomes important to identify client characteristics that predict differential response to the stronger versus the weaker treatment.

Patient Subtypes

To better describe the patient types on whom treatment selection might be tested, we propose an adaptation of DeRubeis et al.'s (2014b) conceptualization of client types. In an attempt to highlight the relationship between therapy quality and patient improvement, they posited five exemplar types meant to represent the spectrum of potential associations that could be expected between therapy quality and improvement: spontaneous remitters, easy patients, pliant patients, challenging patients, and intractable patients. For spontaneous remitters, any level of therapy quality (from the best to the worst) would lead to high levels of improvement. For patients at the other end of the spectrum (the

intractable patients), little to no improvement would be expected, regardless of the level of therapy quality. In the middle of this spectrum are pliant patients, defined as those patients whose improvement would vary as a function of therapy quality, such that with very poor quality therapy or no therapy, no improvement would be expected, and with the highest quality therapy possible, complete improvement would be expected to result. For the purpose of treatment selection, the pliant patient category can be broken down into two subgroups: individuals who would improve insofar as they receive quality treatment of any type and other individuals for whom the correlation between quality and improvement would be moderated by the "match" between patient and treatment (see **Figure 2**, types 3 and 4a/4b). These latter individuals (**Figure 2**, type 4a/4b,), who will respond well to—but only to—a specific treatment, are the individuals for whom PAI-type treatment selection will be most important.

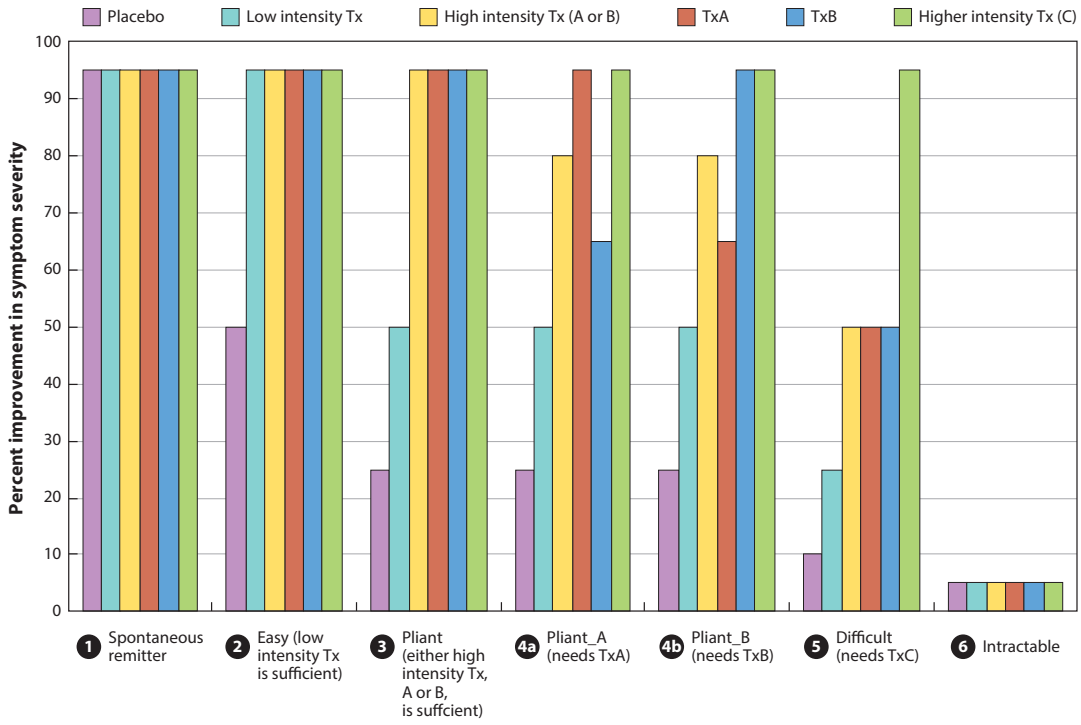


Figure 2. The figure depicts the expected improvement for different patient prototypes in

different treatment contexts. The treatment contexts range from lowest to highest intensity (colored bars). Patient prototypes, which range from spontaneous remitters to intractable patients, are labeled on the x-axis. As shown with the colored bars, spontaneous remitters would be expected to show the same high level of response (95%) in any treatment context. Similarly, intractable patients would be expected to show the same low level of response (5%) irrespective of the treatment provided to them. Prototypes 2, 3, 4a, 4b, and 5 would be expected to show different levels of response depending on the treatment provided. Prototypes 3, 4a, and 4b are all "pliant," but they differ in regard to the expected responses to the two high intensity treatments (TxA and TxB). Patients represented by prototypes 4a and 4b differ from those represented by prototype 3 in that they require a *specific* high intensity treatment, whereas prototype 3 patients would be expected to evidence a high level of response to either high intensity treatment. This distinction is also depicted by the heights of the yellow bars (unspecified high intensity treatment), which represent the averages of the expected responses to TxA and TxB within each prototype.

The analytical tools used to construct PAI models can be adapted to inform decisions in stratified medicine, where the choice is often between a high- versus low-intensity treatment, and where the high-intensity treatment is more effective, on average. In such cases, the goal is to distinguish between individuals who are likely to benefit much more from the high-intensity treatment than from the low-intensity treatment, versus those for whom the expected differential benefit is small. As with the PAI approach, a continuous index is created (Forand et al. 2017), but in this case its purpose is to array patients along a continuum from those who are most likely to experience a positive response irrespective of treatment to those for whom the expected outcome is poor (**Figure 2**, types 2/3/5). Lorenzo-Luaces et al. (2017) implemented such an approach as a proof of concept, with data from a randomized comparison of a high-intensity treatment (CT) with two lower-intensity treatments. On average, as described in the main outcome paper from the trial, the differences between CT and each of the two comparison conditions were small (van Straten et al. 2006). Lorenzo-Luaces et al. constructed a multivariable

prognostic index[8] as described above. Following DeRubeis et al. (2014b), they predicted that, for clients with poorer prognoses, the provision of CT would lead to a higher likelihood of response, relative to the lower-intensity conditions. Between-treatment comparisons were not expected to reveal differences in response rates in the subset of clients with scores indicating better prognoses. Findings were consistent with these predictions, suggesting that the application of these principles in stratified medicine could substantially increase the efficiency of mental health treatment systems. Gunn et al.'s (2017) recently initiated RCT tests a symptom-based depression clinical prediction tool called Target-D for stepped-care in primary care.

Two recently published works using data from the National Health Service's Improving Access to Psychological Therapy (IAPT) program also highlight ways in which multivariable models could be used to guide stratified medicine in mental health. Saunders et al. (2016) used latent profile analysis to create eight profiles that described sets of baseline demographic data and symptom features that defined patient clusters. One of their goals was to identify subsets of clients (those with profiles similar to each other) for whom differential predictions could be made between high-intensity treatment and low-intensity psychological treatment. In a different sample of clients treated for mood and anxiety disorders in the IAPT services, Delgadillo et al. (2016) explored the

[8]A prescriptive index could also be used in the context of such a comparison. Use of a prescriptive model in this context would identify patients for whom the stronger treatment is expected to lead to better outcomes than the weaker treatment, patients for whom less advantage of the stronger treatment is expected, and perhaps a subset of patients for whom no advantage of the stronger treatment is expected, or even a subset for whom the weaker treatment is predicted to be better than the stronger treatment.

potential utility of treatment selection models. The authors created an index that generated predictions as to which clients were likely to achieve reliable and clinically significant improvement in depression or anxiety symptoms. Recent work using a prognostic index of case complexity yielded similar results in a separate sample of IAPT patients (Delgadillo et al. 2017).

RECOMMENDATIONS FOR BUILDING TREATMENT SELECTION MODELS

In what follows, we review the major steps involved in constructing and evaluating a treatment selection approach from a dataset that includes values, for each client, on variables that reflect pretreatment characteristics, the treatment provided to the client, and the client's observed outcome in that treatment. Understanding these steps is critical for the clinical researcher who wants to conduct PAI analyses, as well as for the clinician who wants to interpret and evaluate the utility of findings from treatment selection studies.

The first step is to identify and prepare the candidate predictor variables. Good candidate predictor variables are those that are measured prior to the point of treatment assignment and that plausibly could be related to outcome, either in general (prognostic) or differentially between treatments (prescriptive). If prior research has indicated that a variable predicts outcome, then it should be included as a potential predictor, but as the literature on predictors (and especially on moderators) in mental health is still relatively sparse, including other putative variables is recommended. Variables must not have significant missingness, and tests for systematic missingness should be performed to inform the appropriateness of imputation (Jamshidian & Jalal 2010). Variables should also exhibit sufficient variability. For example, it makes little sense to include gender if

95% of the sample were female. Many variable selection and modeling techniques used in prediction are sensitive to situations in which the set of predictors has high collinearity, and thus it is wise to examine the covariance structure of the potential predictors, and to take steps to reduce high collinearity (Kraemer 2013). Other considerations for preparing potential predictors include dealing with outliers/leverage points, making categorical variables binary (where indicated), and transforming variables for theoretical reasons or to deal with problematic distributions (e.g., those with high skew). Finally, centering variables can help avoid inferential errors and increase stability when using regression-based approaches (Kraemer & Blasey 2004).

The choice of variable selection and modeling approaches can be constrained by the nature of the outcome variable. Although many approaches can accommodate both binary and continuous outcomes, the use of categorical outcomes, or longitudinal and survival-type outcomes, is limited to a select subset of the available approaches.

Once potential predictor and outcome variables have been selected, the next step is to build the prediction model. This is typically a two-step process comprising variable selection and model-weight specification. Many different variable selection approaches have been proposed for treatment selection, all of which attempt to identify which variables, among the potential predictors, contribute meaningfully to the prediction of outcome. Gillan & Whelan (2017) provide an excellent discussion of theory-driven versus data-driven approaches to model specification. Classic approaches rely on parametric regression models [e.g., forward or backward stepwise regression; see Fournier et al. (2009) for a worked example] that select only those variables with statistically significant relations with outcome. A subset of these approaches includes

penalties that aim to create parsimonious models by limiting the number of variables selected (Tibshirani 1996). Others employ bootstrapping procedures or shrinkage parameters that seek to maximize the stability and generalizability of the models (Austin & Tu 2004, Garge et al. 2013, Zou & Hastie 2005). Advances in statistical modeling and computational resources have led to feature selection approaches, many of which are based on machine learning, that can flexibly model and identify predictors with nonlinear and higher-order interactions (Bleich et al. 2014). The line between variable selection and model weight specification is not always clear, as some modeling approaches combine the two in one step. Gillan & Whelan (2017) provide an in-depth review of the merits of machine learning in mental health; interested readers can also consult books focused on applied clinical predictive modeling (Chakraborty & Moodie 2013, Parmigiani 2002, Steyerberg 2008).

Cohen et al. proposed a new variable selection approach that combines the outputs of multiple procedures with the aim of generating robust predictors (Cohen et al., *under review*). It also allows for the inclusion of complex relations that often exist between predictors and outcome in treatment selection contexts that are often overlooked in classic regression-based approaches. We performed four different variable selection approaches in seven mental health RCTs and found both consistencies and inconsistencies in which variables were identified in each dataset across the different approaches. Some variables were identified consistently as important, some variables were identified consistently as unimportant, and other variables had mixed indications, depending on the variable selection method. We can have increased confidence in the importance of variables that are consistently identified as important, and similarly, that

those variables rejected consistently should be considered unimportant. We also believe that we can use our understanding of the different methodologies to determine whether those variables that are identified in some approaches but not in others are inconsistently identified due to weak or noisy effects, and thus should be considered poor predictors, or whether this pattern can be attributed to shortcomings of specific approaches. For example, a variable might be selected by one approach that can flexibly model higher-order interactions (Bleich et al. 2014) but excluded by a second that cannot (Austin & Tu 2004) if that variable's predictive relationship to outcome involves a three-way or nonlinear interaction.

Once the variables that have prognostic or prescriptive relationships to outcome have been identified, the model weights are specified. Model weights determine how much, and in what direction, each variable contributes to the prediction of outcome. Although the specifics of how a modeling approach characterizes these relationships can differ (e.g., parametric approaches, which might use linear regression, versus nonparametric machine-learning approaches, which might utilize tree-based modeling approaches), any of these approaches can generate predictions for new clients for each treatment condition for which the prediction is to be made. Both variable selection and weight setting should be performed using techniques that maximize the stability and generalizability of the model (Gillan & Whelan 2017).

EVALUATING TREATMENT RECOMMENDATION APPROACHES

As described in previous sections, once a model is built it can be used to generate predictions for each patient's outcomes. The utility of the model can then be evaluated on the basis of comparisons of the predictions with observed outcomes. This can be done

either within the dataset that was used to generate the predictions or with a new sample of clients who are randomized to receive treatment A or treatment B. When the same dataset is used both to generate the model and to test its utility, special care must be taken to avoid a situation in which the model is fit specifically to the sample and is therefore unlikely to generalize in an independent application (Collaboration 2015, Ioannidis 2005).

To estimate the expected utility of the prediction-based recommendations without bias, data from the to-be-predicted patient cannot be included in the course of development of the algorithm (Hastie et al. 2009). This can be accomplished in model development with the use of bootstrapping or internal cross-validation methods. Ongoing efforts to refine feature selection and weight setting with cross-validation focus on ways of identifying robust feature sets and robust means of determining the weights that will be applied to those features. Well-constructed models are built with the aim of avoiding both underfitting and overfitting at both the feature-selection and weight-setting stages. The procedures for maximizing power (avoiding underfitting) and generalizability (by avoiding overfitting) are in continuous development.

Although there are many ways one could test a PAI prospectively, the most straightforward approach would be to randomize a new sample of clients to each treatment. A test of the utility of the model can then be derived from a comparison of the outcomes of those individuals who happen to be randomized to the intervention that was identified by the model as more likely to have a positive outcome, versus the outcomes of those who get randomized to their nonindicated intervention. In the context of equivalent average outcomes for the two treatments, if the average response of those who receive

their indicated treatment is (statistically significantly) superior to the average response of those who receive their nonindicated treatment, this can be taken as evidence that the model has predictive power. Further examination of the size of this benefit, in the context of other relevant factors (e.g., cost of administering the required assessments) would inform a judgment concerning the clinical utility of a model (Huang et al. 2012, 2015). Another approach to a prospective study would be to randomize participants to allocation-as-usual (AAU; for example, patient preference or clinical judgment) versus model-guided allocation. Although attractive for its comparison to a real-world treatment allocation strategy, this approach reduces the sample size available for comparison, as the only patients that can be used to compare the utility of the model are those for whom the AAU and model-based assignments disagree.

Careful consideration of the distinction between the different patient types reviewed earlier is important when evaluating treatment selection models. Indexes such as the PAI yield binary recommendations (A versus B), but they also contain information about the strength of the recommendation. When used to inform treatment selection in the context of two treatments with equivalent average effects, many individuals can be expected to have PAIs close to zero, indicating that little to no difference in outcomes is predicted between the treatments. For these individuals, one implication is that either treatment could be recommended, as would be so for a type-3 pliant patient from **Figure 2**, who will respond to any treatment according to its strength. However, an individual with a PAI near zero might instead be a spontaneous remitter (type 1), an easy patient (type 2), a difficult patient (type 5), or an intractable patient (type 6). An examination of the within-treatment prognostic predictions will provide an indication of which profile best

describes such an individual. Predictions of roughly equally poor outcomes in both treatments might indicate a challenging or intractable patient, whereas predictions of full symptom resolution in both treatments might indicate a spontaneous remitter, an easy patient, or a type-3 pliant patient. A patient with poor predicted outcomes in both treatments under consideration would tentatively be categorized as intractable (type 6), but it is possible that such a patient (type 5) might benefit from a treatment not included in the comparison, such as the combination of the two treatments studied. Identifying these individuals and recommending a stronger treatment could reduce the number of exposures to ineffective treatments.

A recommendation that treatment A is to be preferred over treatment B could arise from a PAI that is very large, in which case a clinician might strongly advise a client to pursue treatment A. However, if the predicted advantage is so small as to be clinically meaningless (e.g., a PAI close to zero), then the clinician would communicate this information to the client, and other factors would play a larger role in selecting treatment. Evidence for the importance of attending to recommendation strength can be found in the results of contexts in which greater expected benefit of treatment selection was observed for individuals with larger PAIs compared to those whose PAIs were smaller (DeRubeis et al. 2014a; Huibers et al. 2015; Cohen et al., *under review;* Keefe et al., 2018).

DISCUSSION

Clinical practitioners and researchers have long sought knowledge about what works for whom. This knowledge matters. Many stakeholders would benefit from improvements in our ability to identify, for each individual, the intervention among those under consideration that is most likely to yield the best response. The implications for

individuals are obvious. People suffering from depression want interventions that will work. Limiting the number of individuals exposed to ineffective first-line treatments and reducing the average time to recovery will not only reduce suffering from the symptoms of depression, but will increase economic productivity inasmuch as symptoms of depression interfere with a person's ability to perform work functions at a high level (Layard et al. 2007). Intelligent allocation of limited or costly resources has relevance for any class of treatment, including psychotherapy—the availability of which is often limited by the availability of trained clinicians—and pharmacotherapy, in which associated risks should be minimized.

Success in efforts to match individuals to treatments has been elusive. Historical attempts to use research findings to promote propitious matches of clients to treatments have relied on analyses that take into account a single feature of the client. Work with single features has been attractive in part due to its simplicity, and because of the ease with which a theory-based interpretation can be applied to the findings to support or understand the resulting recommendations. Unfortunately, the vast majority of this research on individual differences in treatment response (e.g., project MATCH; Allen et al. 1997) has failed to have a meaningful impact on client care (Simon & Perlis 2010).

Modern multivariable treatment selection approaches can overcome many of the shortcomings that have hindered progress and therefore hold great promise for the future of precision mental health. Part of this future will require a resolution of the tension between the statistical methodology of explanatory approaches that have dominated psychology and the predictive approaches the will power precision medicine going forward (Yarkoni & Westfall 2017).

Although it was not the focus of this review, we want to emphasize that we believe the treatment selection process should be an open and shared decision-making process between patients and clinicians. Treatment selection tools should be viewed as providing useful information that helps inform this collaborative decision-making process.

FUTURE DIRECTIONS

Research that informs treatment selection will continue to include analyses of data from RCTs, but it should and likely will also be conducted with large treatment databases (Kessler 2018), collected online or through electronic medical records (Perlis et al. 2012). The designs of RCTs will also be better tuned to the goals of precision mental health. Recent work has demonstrated the potential for dynamic assessment in precision mental health (Fernandez et al. 2017, Fisher & Boswell 2016). Modular psychotherapies that can be accessed online are fertile grounds for future efforts to personalize treatment for depression (Watkins et al. 2016). The pretreatment assessments that provide grist for treatment selection models will include biomarkers and other measures that promise to reveal prescriptive relationships, in addition to the self-report, demographic and clinical variables that have fueled most treatment selection findings reported to date. There are several ongoing studies, designed specifically to generate knowledge relevant to outcome prediction in depression treatment, that feature potential biomarkers, including information from neuroimaging and genetic testing (Brunoni et al. 2015, Lam et al. 2016, Williams et al. 2011). Two such trials are the Establishing Moderators and Biosignatures of Antidepressant Response for Clinical Care for Depression study (Trivedi et al. 2016), which focuses on two antidepressants (sertraline and bupropion) in the context of early-onset, recurrent MDD, and the recently completed Predicting Response to Depression

Treatment study, which compared CBT to ADM in treatment-naive adults with moderate to severe MDD (Dunlop et al. 2012a). Lutz et al. (2017) have recently initiated an RCT that tests personalized psychotherapy prediction and adaptation tools in a real-world clinic. The exploratory nature of many of the existing prediction models increases the importance of external validation and tests of generalizability. To realize the promise of precision mental health, existing models as well as those that are being developed will need to be validated prospectively against standard allocation schemes (Kingslake et al., 2017). Moreover, it will be important for all stakeholders, including providers and patients, to be involved in shaping the tools that will translate the findings into practice.

Acknowledgments

The authors acknowledge with gratitude the extensive and challenging feedback on previous drafts of this article by Sona Dimidjian, Yoni Ashar, Rob Saunders, Paul Andrews, and Steve Hollon. We thank Thomas Kim for his help preparing the manuscript, including his assistance with the literature reviews that informed **Supplemental Table 1**. We are grateful to the MQ Foundation, whose support has made this work possible. Finally, we acknowledge the extensive network of colleagues and collaborators who have contributed to this emerging area. To witness the eagerness of members of this network to work together and to share ideas, expertise and data, with the common goal of improving outcomes in mental health, has been most gratifying.

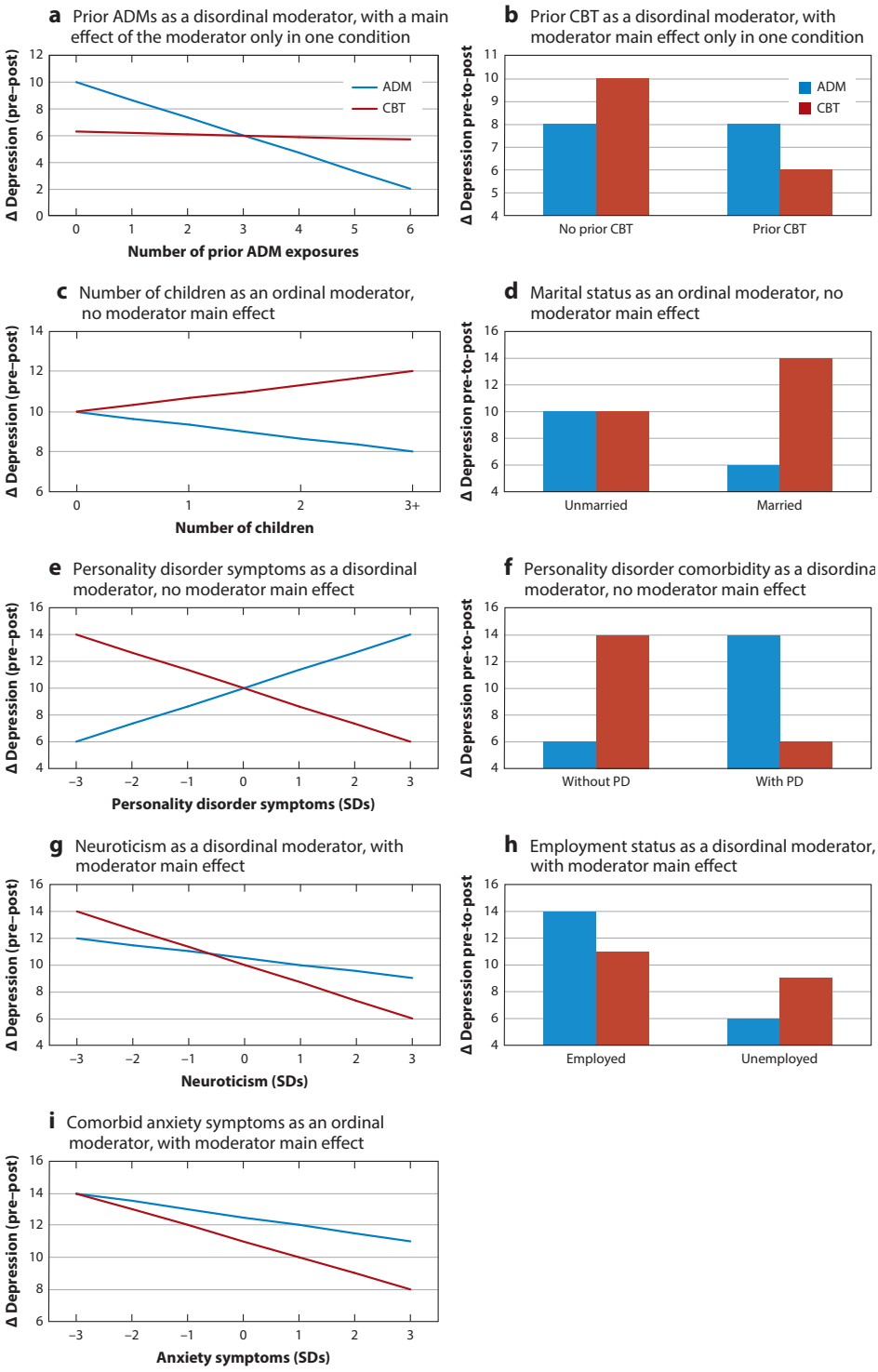
Supplemental Material: Supplemental Example 1

An example of the single moderator approach, with a twist. Prior to randomizing 63 patients with MDD to one of three treatment conditions, Beutler et al. (1991) assessed them on two dimensions hypothesized to be differentially predictive of outcomes in the conditions. Specifically, the investigators predicted that "the degree to which patients characteristically use externalization as a coping style (i.e., acting out, projection, etc.) would be positively associated with improvement in a treatment that focuses on behavioral change (CT) but would be negatively associated with improvement in insight-oriented (FEP and S/SD) treatments…(and that) level of patient preassessed resistance potential would be positively related to patient response in self-directed treatment (S/SD) but would be negatively related to improvement in authority directed (FEP and CT) therapies." (p. 334). This study possessed several admirable features, and the findings were impressive. The dimensions and treatments were selected based on clinical theory, and the investigators specified the directions of the associations of each of the two dimensions in each of the three treatments. Of the six directional predictions they made, four of them were borne out, with the absolute values of the correlations between the relevant predictor and outcome ranging from 0.47 to 0.60 (Beutler et al., 1991). A fifth correlation was in the predicted direction (0.17) and a sixth was in the direction opposite of the prediction, but only slightly (-0.10).

Unfortunately, their conclusions illustrated a common limitation on inferences from findings regarding the differential prediction of outcomes in two or more treatments. When more than one variable exhibits a predictive relation to outcome, the translation of prediction findings to sound clinical judgment is difficult. In the Beutler et

al. (1991) study, two variables were identified (externalization and resistance potential), each of which was used to make independent differential predictions of outcome. As has been true of other investigators who have identified multiple prescriptive variables in a dataset, the authors did not provide guidance as to how to combine or integrate information from the two predictors (see a paper from our group for a more recent example of this common shortcoming; Fournier et al. 2009). Without an integration of the variables, the conclusions from this kind of predictive work will be as limited as those from studies of single prescriptive variables. Indeed, recent publications by Beutler and colleagues speak about the importance of coordinating multiple sources of information when making treatment decisions (Beutler et al., 2016), and Beutler et al.'s Systematic Treatment Selection (Beutler & Clarkin, 1990) and Prescriptive Psychotherapy (Beutler & Harwood, 2000) represent attempts to formalize the application of empirical findings of the sort described above to clinical practice.

Supplemental Figure 1



a) This shows a disordinal moderator relationship between a continuous predictor (# of prior antidepressant exposures) and outcome. For those who received CT, there is no relationship between # of prior antidepressants and outcome. For those who received ADMs, the greater the

number of prior ADM exposures, the less change is expected in symptoms of depression over the course of treatment. People with two or few prior ADM exposures are expected to experience more change in ADM treatment than in CT, and individuals with large numbers (4 or more) of prior ADM exposures are predicted to experience greater change in CT than in ADM. For individuals with 3 prior ADMs, there is no predicted difference in outcomes between the two treatments. This moderator shows a main effect. **b)** This shows a disordinal moderator relationship between a categorical predictor (prior CT) and outcome. For those who receive ADM (in blue), there is no relation between prior CT and outcome. For those treated with CT (in red), individuals who have never had a course of CT are expected to benefit more with CT than are those with a history of CT. Looking within CT-history subgroups, individuals with no prior CT are expected to experience more symptom change in CT than with ADM, and within the subgroup of individuals who had previously received CT there is the opposite expectation. This moderator has no main effect. **c)** This shows an ordinal moderator relationship between a continuous variable (# of children) and outcome. For those treated with CT, there is a positive relationship between # of children and symptom change, such that the more children, the more symptom improvement could be expected. For those treated with ADM, the opposite relationship is observed. For people with no children, there is no expected difference between the two treatments in terms of change in symptoms. But for those with children, the more children a patient has, the larger the advantage the expected advantage of CT over ADM. This moderator has no main effect. **d)** This shows an ordinal moderator relationship between a categorical variable (marital status) and outcome. There is no difference between CT and ADM for unmarried people, but for married people there is a large advantage of CT over ADM. Married people are expected to do better than unmarried people in CT (red bars). Unmarried people are expected to do better than married people in ADM. This moderator has no main effect. **e)** This shows a disordinal moderator relationship between continuous predictor (personality disorder symptoms) and outcome. In ADM, there is a positive relationship between PD symptoms and outcome: the more PD symptoms, the more symptom change is expected. In CT, the opposite (a negative) relationship is observed. People with fewer PD symptoms are expected to experience more change in CT treatment than in ADM, and individuals with more PD symptoms experienced greater change in ADM than in CT. For individuals with average levels of PD symptoms, there is no difference expected in outcomes between the two treatments. There is no main effect of the number of personality disorder symptoms. **f)** This figure shows the same disordinal moderator relationship as in figure e, but for a categorical version of the personality disorder predictor (diagnosis yes/no). On average, patients with a PD experience more change in ADM than those without a PD. For those who got CT, individuals without a PD diagnosis experience, on average, more change than those with a PD diagnosis. Thus, ADM is expected to be better than CT for those with a PD, and CT is expected to be better than ADM for those without a PD. There is no main effect of having a PD. **g)** This shows a disordinal moderator relationship between continuous predictor (neuroticism) and outcome. For those who receive ADM, there is a negative relationship between neuroticism and outcome: the more neurotic a patient is, the less symptom change should be expected over treatment. For those who receive CT, a stronger relationship in the same direction is expected. There is a main effect of the moderator (such that in both conditions, more neuroticism is associated with less symptom change), but the nature of these relationships involves a crossover around the mean level of neuroticism for the same. Thus, for

people with very low levels of neuroticism, CT is expected to be superior to ADM, and for those who with high levels of neuroticism, ADM is preferred to CT. **h)** This illustrates the same disordinal moderator relationship as figure g but between a categorical predictor (employed vs. unemployed) and outcome. There is a main effect of employment, such that people who are unemployed experience less improvement than people who are employed. However, the extent to which unemployment predicts poorer response differs by treatment. The decrease in expected response comparing employed to unemployed individuals is larger for ADM than for CT. Practically, ADM is preferred to CT for people who are employed, and CT is preferred to ADM for people who are unemployed. **i)** This shows an ordinal moderator relationship between a continuous variable (anxiety symptoms) and outcome. There is a main effect of anxiety symptoms, such that more anxiety is related to less change in depression across treatment. For those with the fewest anxiety symptoms, there is no difference between the two treatments. For the rest of the sample, ADM is associated with more symptom change than CT, and the size of this predicted advantage of ADM grows as individuals have increasingly high levels of anxiety symptoms.

Supplemental Table 1. Review of reviews and meta-analyses of predictors in depression.

Focus	Focus	Predictor domains	Type	Reference
MDD	IPT	All	Systematic review	Bernecker et al. (2017)
MDD	All	Sociodemographic, clinical, personality, stress and adversity, cognitive	Review	Kessler et al. (2017)
MDD	ADM	Biomarkers	Review	Fabbri et al. (2017)
MDD	Exercise	Demographic, biological, clinical, psychosocial	Systematic review	Schuch et al. (2016)
MDD	ADM (withdrawal)	Demographic, clinical	Systematic review	Berwian et al. (2016)
MDD	Psychotherapies	Sociodemographic, clinical, environmental	Meta-analytic review	Cuijpers et al. (2016)
MDD, SZ, bipolar, substance abuse	ADM	Biomarkers (pharmacogenetics)	Review	El-Mallakh et al. (2016)

MDD	ADM, CBT	Demographic, clinical	Individual patient data analysis	Vittengl et al. (2016)
MDD and ADHD	EEG	Biomarkers	Review	Olbrich et al. (2015)
MDD	ADM	Biomarkers (pharmaco-epigenetics)	Review	Lisoway et al. (2017)
MDD	all	Biomarkers	Review	Kemp et al. (2015)
MDD	ADM	Biomarkers (neuroimaging)	Review	Phillips et al. (2015)
MDD	all	Biomarkers (neuroimaging)	Review	Lener and Iosifescu (2015)
MDD	ADM, ECT and TMS	Biomarkers	Systematic review	Dichter et al. (2015)
MDD	ADM and psychotherapy	Biomarkers (inflammation)	Meta-analysis	Strawbridge et al. (2015)
MDD	ADM	Biomarkers (neuroimaging)	Review	Chi et al. (2015)
MDD	ADM	Demographic, clinical, psychosocial	Review	Hirschfeld (2000)
TRD	ADM	Demographic, clinical, biomarker	Systematic review	Bennabi et al. (2015)
MDD	ADM	Biomarkers pharmacogenetics	Meta-analysis	Biernacka et al. (2015)
Mood disorders	ADM	Biomarker (BDNF)	Systematic and quantitative meta-analysis	Polyakova et al. (2015)
MDD	ADM	Pharmacogenetics	Review	Perlis (2014)
MDD	ADM	Biomarkers (EEG)	Review	Olbrich and Arns (2013)
MDD and ALZ	ADM	Biomarker (pharmacogenetics/pharmacodynamics)	Review	Souslova et al. (2013)
MDD and anxiety	All treatments	Biomarkers (neuroimaging)	Review	Jappe et al. (2013)

TRD	ADM	Biomarkers	Review	Smith (2013)
Late-life depression	ADM	Demographic, Clinical	Patient-level meta-analysis	Nelson et al. (2013)
MDD	Physical exercise	Demographic, clinical	Systematic review and meta-analysis	Silveira et al. (2013)
Mood disorders	Medication adherence	Demographics, clinical, psychosocial	Review	Pompili et al. (2013)
Observational studies in MDD	Medication adherence	Sociodemographic and clinical	Systematic review	Rivero-Santana et al. (2013)
MDD	All	Clinical (personality disorder)	Systematic review and meta-analysis	Newton-Howes et al. (2013)
MDD	ADM, psychotherapy	Demographic, clinical	Systematic review and meta-analysis	Cuijpers et al. (2012)
MDD	All treatments	Biomarkers (neuroimaging)	Meta-analysis and review	Pizzagalli (2011)
MDD	ADM	Demographic	Meta-regression	Naudet et al. (2011)
MDD	ADM	Demographic, clinical	Meta-regression	Serretti et al. (2011)
Mood and anxiety disorders	ADM	Demographic, clinical, psychosocial, biomarkers	Review	Serretti et al. (2009)
MDD, anxiety	Internet-based interventions	Demographic, Clinical, Psychosocial, Biomarkers	Systematic review	Christensen et al. (2009)
MDD	ADM	Demographic, clinical, psychosocial, biomarkers	Review	Kemp et al. (2008)
MDD	Psychotherapy	Demographic, clinical, environmental	Meta-regression	Cuijpers et al. (2008)
MDD	CBT	Clinical	Meta-regression	Haby et al. (2006)
MDD	ADM	Personality (personality disorder)	Systematic review and meta-analysis	Kool et al. (2005)
MDD	ADM, placebo	Demographic, biomarkers, clinical, personality, environmental	Selective review	Dodd and Berk (2004)

MDD	ADM	Demographic, clinical, psychosocial	Review	Bagby et al. (2002)
MDD	ADM	Clinical, biomarkers	Review	Joyce and Paykel (1989)

Abbreviations: ADHD, attention deficit hyperactivity disorder; ADM, antidepressant medication; ALZ, Alzheimer's disease; BDNF, brain-derived neurotropic factor; CBT, cognitive behavioral therapy; ECT, electroconvulsive therapy; EEG, electroencephalography; IPT, interpersonal therapy; MDD, major depressive disorder; PD, personality disorder; RCTs, randomized clinical trials; SSRI, serotonin-selective reuptake inhibitor; SZ, schizophrenia; TMS, transcranial magnetic stimulation; TRD, treatment-resistant depression.

Supplemental Table 2. Comparison of treatment selection methodology showing heterogeneity

Reference	Comparison	Variable Selection	Modeling	Testing	Approach
Barber and Muenz (1996)	CT vs. IPT	backwards stepwise elimination	Linear regression	within sample	"matching factor"
Lutz et al. (2006)	CT vs. iCBIT	nearest neighbor	nearest neighbor and ETR - tested with logistic regression	LOO	Nearest neighbors
Wallace et al. (2013)	IPT vs. ADM	*M** approach + PCA	linear regression	within sample	*M* * approach
McGrath et al. (2013)	CT vs. ADM	2-way ANOVA	ANOVA	within sample	TSB
DeRubeis, Cohen, et al. (2014)	CT vs. ADM	Domain Stepwise[a]	linear regression	LOO	PAI
Huibers et al. (2015)	CT vs. IPT	Domain Stepwise[a]	linear regression	LOO	PAI

Zilcha-Mano et al. (2016)	SET vs. ADM vs. PBO	mobForest	logistic regression	LOO	PAI
Delgadillo et al. (2016)	Step-2 vs. Step-3 in IAPT	backwards stepwise[b] elimination, bootstrapping, split-halves validation	logistic regression, simplified risk weighting scheme	within sample	Leeds Risk Index
Smagula et al. (2016)	Augmentation with aripiprazole vs. placebo for venlafaxine non-response	*M**approach + lasso	logistic regression	within sample	*M** approach
Saunders et al. (2016)	Step-2 vs. Step-3 in IAPT	none	LPA, split-halves, logistic regression	held-out validation sample	n/a
Iniesta, Malki, et al. (2016)	SRI ADM vs. NRI ADM	previous single-variable moderator analyses from 6 papers and ENRR, in four (inclusive) sets of variables	linear and logistic ENRR	10-fold CV with resampling, permutation test	n/a
Cloitre et al. (2016)	STAIR/EXP vs. STAIR/SupC vs. SupC/EXP	single-variable moderator analyses	mixed effects modeling	within sample, permutation test	GEM
Koutsouleris et al. (2016)	Antipsychotic medication[c]	4x5-fold CV, Stepwise forward selection using RBF-SVM[d]	Ensemble prediction	leave-site-out CV	n/a
Chekroud et al. (2016)	citalopram	10-fold CV, ENRR	GBM[e]	external sample[f]	n/a
Chekroud et al. (2017)	4 ADM conditions[g]	10-fold CV, ENRR	GBM	external sample	n/a
Vittengl et al. (2017)	C-CT vs. C-ADM	Single variable, backwards and forwards stepwise regression	Cox regression model	LOO	PAI
Niles, Loerinc, et al. (2017)	CALM vs. UC	*M** approach + stepwise[h]	linear regression	within sample	*M** approach

		regression with 5-fold CV			
Niles, Wolitzky-Taylor, et al. (2017)	CT vs. ACT	*M** approach + OLS stepwise[h] regression with 3-fold CV	logistic regression	within sample	*M** approach
Delgadillo et al. (2017)	Step-2 vs. Step-3 in IAPT	Lasso and the .632 bootstrap resampling method	(CATREG-Lasso)	Held-out validation sample	Prognostic-index of case-complexity
Kapelner et al. (*under review*)	CT vs. ADM	Theoretical / Prior Literature	linear regression	Robust bootstrap CV	PTE / PAI
Keefe et al. (2018)	CPT vs. PE	mobForest, bootStepAIC	logistic regression	5-fold CV	PAI
Webb et al. (2018)	ADM vs. PBO	mobForest, BART, ENRR, bootStepAIC	linear regression	10-fold CV	PAI[i]
Deisenhofer et al. (2018)	Tf-CBT vs. EMDR	Genetic model selection alogrithm	logistic regression	LOO	PAI / HTE[j]
Cohen et al. (*under review*)	CT vs. SPSP	mobForest, BART, ENRR, bootStepAIC	linear regression	10-fold CV	PAI
Kim et al. (*submitted*)	lithium vs. quetiapine	mobForest, BART, ENRR, bootStepAIC	linear regression	10-fold CV[k]	PAI
Schweizer et al. (*submitted*)	C-ADM vs. MBCT-TS	mobForest, BART, ENRR, bootStepAIC	logistic regression	10-fold CV[k]	PAI / HTE[j]

CT = Cognitive Therapy
IPT = Interpersonal Therapy
iCBIT = integrated cognitive-behavioral interpersonal therapy
LOO = Leave-one-out cross-validation
ADM = Antidepressant Medication
*M** = Combined moderator approach presented by Kraemer (2013)
PCA = principal-component analysis
ANOVA = Analysis of variance

TSB = treatment-specific biomarker

[a] = stepwise variable selection based on Fournier et al. (2009)

PAI = Personalized Advantage Index

SET = Supportive Expressive Therapy

PBO = Placebo

mobForest = bootstrap-aggregation of model-based recursive partitioning by the random forest algorithm

IAPT = Improving Access to Psychological Therapies

Step-2 in IAPT = Low intensity treatments (e.g., brief psychoeducational interventions based on cognitive therapy principles)

Step-3 in IAPT = High intensity treatments (e.g., cognitive therapy, interpersonal therapy)

[b] = stepwise variable selection based on Mick and Ratain (1994)

LPA = Latent Profile Analysis

SRI = serotonin-reuptake-inhibiting antidepressant, specifically escitalopram

NRI = norepinephrine-reuptake-inhibiting antidepressant, specifically nortriptyline

ENRR = Elastic Net Regularized Regression

CV = cross-validation

STAIR = Skills Training in Affective and Interpersonal Regulation

EXP = modified form of prolonged exposure

SupC = supportive counseling

[c] = The 5 treatment groups (haloperidol, amisulpride, olanzapine, quetiapine, and ziprasidone) were combined and analyzed together.

RBF = non-linear radial basis function kernel

SVM = Support Vector Machine

[d] = (also tested linear SVM, univariate logistic regression, L2-regularized multivariate regression, decision tree ensembles)

GBM = Gradient Boosting Machine

[e] = (also tested naive Bayes classifier, Linear Discriminant Analysis, and radial or 'Gaussian' SVM)

[f] = validation sample had three treatment conditions: Escitalopram + Placebo vs. Escitalopram + Bupropioin vs. Venlafaxine + Mirtazapine

[g] = Citalopram vs. Escitalopram + Placebo vs. Escitalopram + Bupropioin vs. Venlafaxine + Mirtazapine

C-CT = Continuation Cognitive Therapy

C-ADM = Continuation Antidepressant Medication

CALM = patient choice of computer-assisted CBT (*CALM Tools for Living*) and/or psychotropic medications

UC = Usual Care (any treatment administered by primary care provider)

[h] – stepwise variable selection based on James et al. (2013)

ACT = Acceptance and Commitment Therapy

OLS = Ordinary Least Squares

CATREG-Lasso = penalized categorical regressions with optimal scaling

PTE = Personalized Treatment Evaluator

CPT = Cognitive Processing Therapy

PE = Prolonged Exposure

bootStepAIC = bootstrapped variant of an AIC-based backward selection model
BART = Bayesian Additive Regression Trees
Tf-CBT = Trauma Focused Cognitive Behavioral Therapy
EMDR = Eye Movement Desensitization and Reprocessing
HTE = Heterogeneity of Treatment Effect, following recommendations by Kessler et al. (2017).
[i] = Webb and colleagues also examined a model in which the variables were selected a priori used previous findings in the literature or theory (uninformed by data-driven variable selection)
[j] = the HTE adaptation involved creating two separate prognostic models (one for each treatment condition) instead of a single model with interactions.
SPSP = Short Psychodynamic Supportive Psychotherapy
MBCT-TS = Mindfulness-Based Cognitive Therapy with support for medication tapering
[k] = these two studies employed a "full" 10-fold CV, in which both variable selection *and* weight setting were performed in the training samples

CHAPTER 2: Recommending cognitive-behavioral versus psychodynamic therapy for mild to moderate adult depression: A demonstration of a new variable selection approach for treatment selection

This work is under review as:

Cohen, Z. D., Kim, T. K., Van, H. L., Dekker, J. J. M., & Driessen, E. (*under review*). Recommending cognitive-behavioral versus psychodynamic therapy for mild to moderate adult depression: A demonstration of a new variable selection approach for treatment selection.

Abstract

Objective: We use a new variable selection procedure for the Personalized Advantage Index approach to generate treatment recommendations based on pre-treatment characteristics for adults with mild-to-moderate depression deciding between cognitive behavioral (CBT) versus psychodynamic therapy (PDT).

Method: Data are drawn from a randomized comparison of CBT versus PDT for depression (N=167, 71%-female, mean-age=39.6). The approach combines four different statistical techniques to identify patient characteristics associated consistently with differential treatment response. Variables are combined to generate predictions indicating each individual's optimal-treatment. The average outcomes for patients who received their indicated treatment versus those who did not were compared retrospectively to estimate model utility.

Results: Of 49 predictors examined, depression severity, anxiety-sensitivity, extraversion, and psychological-treatment-needs were included in the final model. The average post-treatment Hamilton-Depression-Rating-Scale score was 1.6 points lower (95%CI=[0.5:2.8]; *d*=0.21) for those who received their indicated-treatment compared to

non-indicated. Among the 60% of patients with the strongest treatment recommendations, that advantage grew to 2.6 (95%CI=[1.4:3.7]; *d*=0.37).

Conclusions: Variable selection procedures differ in their characterization of the importance of predictive variables. Attending to consistently-indicated predictors may be sensible when constructing treatment selection models. The small-N and lack of separate validation sample indicate need for prospective tests before this model is used.

Keywords: precision medicine; depression; cognitive behavioral therapy; psychodynamic therapy; treatment selection, variable selection

Significance Statement

Adults seeking treatment for mild to moderate depression have a large variety of psychological and pharmacological treatment options available to them, and clinicians helping clients decide which treatment to pursue could use client-factors associated with differential response to improve their ability to determine which treatment, among the available options, would be most likely to result in a positive response. The process of determining which factors to use, and how to synthesize the available information into a clear, actionable recommendation could be improved through the use of treatment selection approaches based on statistical prediction models. Variable selection, an essential step in the construction of treatment selection models, can be stabilized by attending to those variables that are consistently indicated across several different variable selection approaches.

Introduction

Major depressive disorder is a highly prevalent, debilitating mental disorder that is currently ranked as the single largest contributor to global disability (World Health Organization, 2017). Among the most frequently utilized psychotherapies for depression are cognitive behavioral therapy (CBT) and psychodynamic therapy (PDT). Two randomized clinical trials have found PDT noninferior to CBT in the outpatient treatment of depression (Driessen et al., 2013; Gibbons et al., 2016). These results are in line with meta-analytic findings reporting no significant differences between CBT and PDT for depression (Barth et al., 2013; Driessen et al., 2015). These minimal efficacy differences, along with differential therapeutic theories used in CBT and PDT (Hoffart & Johnson, 2017), raise the question whether individual patients can be identified that might benefit more from one of these treatments than the other. If so, treatment selection could improve outcomes in depression by helping individuals select the specific intervention that is most likely to be successful (Cohen & DeRubeis, 2018).

DeRubeis and colleagues (2014) developed a treatment selection approach that can be used to identify each individual's optimal treatment based on multiple patient characteristics, using data from a randomized clinical trial. This approach is called the Personalized Advantage Index (PAI). The core concept behind the PAI approach is to identify pre-treatment patient characteristics that are associated with differential response to treatment (so-called moderators) and, using these variables, to build a statistical model that can generate predictions for an individual in two (or more) treatments. For each individual, the treatment with the best predicted outcome is defined as the indicated treatment. In the case of a two-treatment comparison, an individual's PAI is a single

number derived through the subtraction of their predictions in one treatment from the other. The PAI provides a directional indication of which treatment the individual should receive, as well as information about the strength of the recommendation, represented by the size (in absolute value) of the PAI.

In their initial demonstration, DeRubeis et al. (2014) analyzed data from a randomized comparison of antidepressant medication and CBT and found that, for patients with large predicted advantages in one treatment over the other (60% of sample), those who received their PAI-indicated treatment had superior outcomes relative to patients who received the non-indicated treatment, with an effect size (Cohen's *d*=0.58) larger than that reported in a recent systematic review of drug-placebo differences (Turner et al., 2008). Huibers and colleagues (2015) published similar findings applying the PAI approach to a comparison of cognitive therapy versus interpersonal therapy for adult outpatient depression. Related efforts based on the PAI approach have generated models aimed at differentiating placebo and antidepressants responders (Webb et al., 2018), minimizing risk of dropout (Zilcha-Mano et al., 2016) and relapse (Schweizer et al., *submitted;* Vittengl et al., 2017),. The principles on which the PAI is based have also been applied to treatment selection in post-traumatic stress disorder (Deisenhofer et al., 2018; Keefe et al., 2018). These studies represent but one strand of research on treatment selection. Other approaches include the M* approach (Niles, Loerinc, et al., 2017; Niles, Wolitzky-Taylor, et al., 2017; Smagula et al., 2016; Wallace et al., 2013) initially introduced by Kramer (Kraemer, 2013), and a series of efforts by Uher, Iniesta and colleagues (Iniesta et al., 2018; Iniesta et al., 2016; Uher et al., 2012). For a

comprehensive review of this literature, we suggest recent reviews by Cohen and DeRubeis (2018), Gillan and Whelan (2017), and Kessler (2018).

Different statistical methods can be used to select the patient characteristics included in the statistical models that generate treatment recommendations. For instance, the initial applications (DeRubeis et al., 2014; Huibers et al., 2015) relied on a domain-based backwards stepwise-regression (Fournier et al., 2009) to build the statistical model. A recent PAI-based treatment selection effort by Vittengl and colleagues (2017) used a series of single-variable models to establish the statistical significance of independent moderators, and then used backwards and forwards stepwise variable selection procedures to reduce the set. Another recent PAI effort relied upon a machine-learning approach called random forests (RF) for variable selection (Zilcha-Mano et al., 2016). Building on this work, Keefe et al., (2018) used a two-stage variable selection approach in which they used RF followed by a stepwise AIC-penalized bootstrapped method. This variability is discussed in Cohen and DeRubeis' (2018) review of treatment prediction reports in depression. They noted that there is very little consistency in the variable selection approaches that have been employed in this area.

This heterogeneity is problematic because different variable selection approaches applied to the same dataset can lead to different conclusions about variable importance (Bleich et al., 2014), and treatment recommendations can vary based on which variables are included in the model. In their review, Cohen and DeRubeis (2018) identified 43 reviews (and meta-analyses) of predictors of treatment response in depression but, as many of these reviews noted, no coherent picture of which predictors are most important has emerged. The variability and lack of replicability in efforts to identify predictors of

response in depression may be partially explained by the heterogeneity of the statistical approaches that are used to identify predictive variables.

This methodological heterogeneity also makes it difficult to determine *which* variable selection procedure should be used in the context of treatment selection. The strengths and weaknesses of different approaches might make them more or less attractive for a given purpose. For example, variable selection with elastic net regularization (ENR) can handle high numbers of potential predictors and can overcome issues of high correlations between baseline variables (Friedman et al., 2010; Zou & Hastie, 2005). However, it does not have the capability to account for unspecified non-linear relationships in the way that is possible when using Random Forest (RF; (Garge et al., 2013). It is unlikely that any single variable selection procedure will be optimal for all situations. It will also be difficult to identify *which* single approach one should use without a sufficiently large dataset to allow for a training sample (in which one could try every approach and see which one appears to work best) and a held-out test sample (in which to show that the results hold, and are not due to chance findings). Unfortunately, RCT samples with relevant treatment comparisons are rarely large enough to support these efforts (Kessler, 2018), and the use of large non-randomized datasets risks potential confounds (e.g., selection effects) that could bias treatment selection efforts (Cohen & DeRubeis, 2018)(c.f. (Kessler, 2018)).

When looking across multiple studies with or within one study using multiple approaches, one can have increased confidence in the importance of variables that are consistently selected by different techniques (Kuhn & Johnson, 2013). Similarly, variables that are consistently rejected can likely be considered unimportant. Knowledge

of the different methodologies could be used to understand whether those variables that are identified in some approaches but not in others are inconsistently identified due to weak or noisy effects, and thus should be considered poor predictors, or whether this pattern can be attributed to shortcomings of specific approaches (Kuhn & Johnson, 2013). For example, a variable might be selected by an ensemble-of-trees approach (e.g., RF) but excluded by another approach based on classic regression (e.g., ENR) because it involves a three-way interaction or non-linear interaction that was not considered in the latter classic approach (Kuhn & Johnson, 2013).

Using these principles, we aimed to demonstrate an improved PAI approach by generating individual treatment recommendations for adult outpatients with depression deciding between CBT versus PDT. We introduce a novel selection process that synthesizes the results of four different variable selection techniques (RF, ENR, Bayesian Additive Regression Trees and the AIC-penalized bootstrapped approach) by selecting the patient characteristics that are consistently identified as associated with differential response.

Method

Design and participants. This paper draws on data from a randomized clinical trial comparing CBT and PDT in the outpatient treatment of depression (Driessen et al., 2013), which included 341 patients who met DSM-IV criteria for a major depressive episode and scored 14 or higher on the Hamilton Rating Scale for Depression (HAM-D; (Hamilton, 1960). The Dutch Union of Medical-Ethic Trial Committees for mental health organizations approved the study design and the study protocol was published (Driessen et al., 2007). Efficacy results of this study are reported elsewhere, with no significant

treatment differences found on any of the outcome measures (Driessen et al., 2013; Driessen et al., 2015; Driessen et al., 2017). Two prior efforts examined which subgroups of patients in this trial might benefit more from one of the treatments than the other. One (Kikkert et al., 2016) was a replication study examining obsessive-compulsive and avoidant personality disorder traits as potential moderators of treatment efficacy that failed to replicate previous findings in that regard (Barber & Muenz, 1996), while the other applied model-based recursive partitioning to 23 potential moderators to identify subgroups of patients that might benefit specifically from one of the two treatments (Driessen et al., 2016). However, these studies used different patient subsamples, examined only a subset of the potential predictors, and were not designed to produce a model that could be used to generate treatment recommendations for individual patients.

As part of the trial protocol, severely depressed (HAM-D > 24) patients at baseline were offered adjunctive antidepressant medication (n=129). As the treatment effects observed in these individuals could have been a result of the psychotherapy, the medication, or both, they were excluded from the current analyses. Thus, this report relates to the patients with moderately severe depressive symptoms (baseline HAM-D = 14 to 24) who were treated with psychotherapy only (n=212). Of these 212 individuals, 17 were removed for having too much missing baseline data (≥ 20% missing baseline predictors). Finally, an additional 28 individuals who dropped out before attending at least 4 sessions were excluded. As our goal was to build a model to answer how individuals who *received* a meaningful course of CBT or PDT fared, we felt that individuals who dropped out very early in treatment would not be informative for our models. In the extreme, the "outcome" for patients who dropped out prior to attending a

single session does not reflect response to CBT or PDT. Additionally, we were less confident in our ability to impute valid week-16 outcomes for these early dropouts. We decided to remove from our analyses patients who attended 3 or fewer therapy sessions. This reduced our sample from 195 to 167. Thus, the final sample comprised 167 patients: 75 in the CBT and 92 in the PDT condition (see Supplemental Figure S1 for a Patient Flow Chart). Baseline sample demographic characteristics for the final sample are presented in Supplemental Table S1.

Interventions. Both PDT and CBT encompassed 16 individual 45-minute sessions within 22 weeks and were conducted according to a published treatment manual (de Jonghe, 2005; Molenaar et al., 2009). CBT was based on the principles described by Beck (1979) and included behavioral activation and cognitive restructuring according to a session-by-session protocol with homework assignments. Short psychodynamic supportive psychotherapy (de Jonghe et al., 2013) represented the psychodynamic intervention. This modality involved an open patient-therapist dialogue that used supportive and insight-facilitating techniques to address the emotional background of the depressive symptoms by discussing current relationships, internalized past relationships, and intrapersonal patterns.

Measures. HAM-D scores were used as the outcome measure for this study. Trained research assistants (master-level graduate students in clinical psychology) assessed the HAM-D according to the Dutch scoring manual (de Jonghe, 1994). Assessors were not blind to treatment condition. Assessors engaged in one-hour peer supervision sessions bi-weekly, in which audiotaped interviews were discussed. The average intraclass correlation coefficient over 46 audiotaped assessments scored by

multiple assessors was .97. Supplemental Table S2 lists the 49 patient characteristics considered during variable selection, all of which were assessed at pre-treatment.

Building the Personalized Advantage Index model. All analyses were performed in R (Team, 2000). Pre-processing and random forest-based imputation of missing data (Stekhoven & Buhlmann, 2012) was performed on baseline and outcome data prior to variable selection. Future efforts in larger samples should perform these steps (especially imputation) separately for the training and test samples to avoid this form of double-dipping. Categorical variables (e.g., relationship status) were turned into binary variables and binary variables without sufficient variability (variables whose smallest category made up < 20% of the sample) were excluded (Kuhn & Johnson, 2013). Outliers for continuous variables were winsorized, and some variables with skewed distributions were log-transformed (see Supplemental Table S2 for more details).

Variable Selection. To select the patient characteristics associated with treatment outcome, we applied a multi-phase selection procedure that combines four different variable selection methods, each of which has been used in recently published treatment selection efforts (Bleich et al., 2014; Iniesta et al., 2016; Keefe et al., 2018; Zilcha-Mano et al., 2016). The first step was to apply three approaches to identify predictors of (differential) treatment response: 1. Random Forest (mobForest package in R; (Garge et al., 2013), 2. Elastic Net Regularization (glmnet package; (Friedman et al., 2010), and 3. Bayesian Additive Regression Trees (BART; bartMachine package; (Kapelner & Bleich, 2016). The second step was to reduce the variables consistently identified by the these three approaches using the stepwise AIC-penalized bootstrapped (Austin & Tu, 2004)

approach (BootStepAIC package; (Rizopoulos, 2009). We will now describe each of these methods in more detail, and discuss their relative strengths and limitations.

Random Forest is a recursive partitioning approach that can accommodate large numbers of predictor variables as well as complex relationships including non-linear and higher order interactions (Kapelner & Bleich, 2016). RF builds upon recursive partitioning approaches like classification and regression trees and model-based recursive partitioning. It addresses model instability by randomly selecting features and creating many "tree models", the predictions of which are aggregated to generate stable predictions (Austin & Tu, 2004). RF also allows for information from weaker predictors to be incorporated in situations where they might otherwise be dominated by stronger predictors, such as in bagging (Garge et al., 2013). The model function of RF can be specified as "y ~ tx," which forces the approach to select splits that maximize the difference in the treatment condition coefficient between subgroups, thus focusing on identifying moderators of the treatment effect. When RF is used for variable selection, the permuted variable importance is generated by comparing the mean square error (MSE) of the predictions in the held-out (out of bag) samples when the real values are used to the MSE when permuted values for a given predictor are used. The extent to which the MSE increases when permuted values are used indicated how "important" that variable is (Garge et al., 2013). Variables that surpass the recommended threshold (which is set based on the largest observed "noise" variable, for which the permuted data improves the MSE, relative to the real data) are selected.

Elastic Net Regularization can provide a hybrid of the Lasso and Ridge regression approaches, combining the L1 and L2 penalizations to allow for the selection of a

parsimonious set of variables that predict outcome (Hastie et al., 2009). We used the R package glmnet (Friedman et al., 2010) to implement ENR variable selection, and used Zou and Hastie's (2005) recommended default value for the alpha parameter (alpha = 0.5). Uses of ENR in the literature of variable selection and treatment selection have only investigated prognostic models in which a single treatment is modeled (Chekroud et al., 2017; Chekroud et al., 2016; Iniesta et al., 2016). Current implementations of ENR in R do not accommodate variable selection for models in which moderators are of primary interest. In order to adapt ENR for the purpose of identifying moderators, we split the training sample into each of the two treatment groups, and then constructed prognostic models within each group. Variables that were retained as predictors in only one condition, or that were selected in both but specified with differing coefficient values, were identified as potential moderators of treatment effects. We refer the reader to Cohen and DeRubeis' (2018) review (specifically, their Figure 1) for a more in-depth discussion of why variables with these relationships are candidate moderators. As one example, consider a variable that is selected by ENR in one treatment condition and specified with a positive coefficient, and that is selected in the other treatment condition but specified with a negative coefficient. This information could suggest that a disordinal relationship exists between that variable and treatment, such that individuals with higher levels on that variable do worse relative to individuals with lower levels in one treatment, whereas individuals with lower levels on that variable do worse relative to individuals with higher levels in the other treatment.

Bayesian Additive Regression Trees builds on ensemble-of-tree methods such as RF by incorporating an underlying Bayesian probability model (Chipman et al., 2010).

BART and RF have similar strengths insofar as they both can handle large numbers of predictors, and can accommodate non-linear and higher order interactions. The inclusion of the Bayesian prior improves upon other tree-ensemble approaches by introducing regularization, which reduces the likelihood that the ensemble will become dominated by any single tree (Genuer et al., 2010). Kapelner and Bleich adapted the bartMachine R-package (2016) to help focus model building on moderators. To achieve this aim, they introduced a parameter that forces the search for variable splits to focus more on treatment than other variables, thus introducing more interactions between treatment and other variables. This is conceptually similar to when researchers only consider interactions between treatment and baseline variables (and not interactions between baseline variables themselves), or to how RF can specify the splitting criteria to evaluate the difference in the treatment coefficient for the model y ~ tx. Bleich and colleagues (2014) adapted BART to extract informed prior information about variable importance, and provide an interaction plot feature that can be used to identify potential 3-way interactions. The ICEbox package in R (Goldstein et al., 2015) allows for the visualization of predictive relationships in BART models, including non-linear and higher-order interactions between variables and treatment. The N most important interactions identified by BART are retained, where N was decided based on the number of variables selected by Random Forest (which uses a permutation test to determine an importance threshold cutoff)[9]. See the Supplemental Variable Selection, as well as Garge et al. (2013) for more details on how the threshold is determined.

[9] We decided to select the number of variables identified by RF, and not to use BART's built-in permutation test for thresholding variable importance because BART's test was created for use in contexts where the variable search was not biased to focus on treatment interactions.

We decided to reduce the variables consistently selected by the above three approaches using a specific fourth approach for the following reason: If the model that was used to generate predictions relied on linear or logistic regression, then the variables selected by RF or BART could lead to a model with poor fit, if, for example, these variables relied on non-linear relationships or higher order interactions. The BootStepAIC package (Rizopoulos, 2009) performs variable selection using a stepwise AIC-penalized bootstrapped approach (Austin & Tu, 2004). By only including the moderator relationships identified in the other three approaches (and their corresponding main effects), this search generated a model emphasizing the prediction of differential treatment response, while reducing the chance that predictors that require unspecified linear or higher-order interactions were included. 10,000 bootstrapped training samples were drawn, and within each training sample backwards elimination was used to select variables that independently contribute to predicting outcome. Austin and Tu (2004) recommend selecting variables that are retained in at least 60% of bootstrapped samples, but this recommendation is specific to prognostic variables (main effects only). As we were interested in interactions, we relied on the consistency of the direction of the coefficients across the 10,000 bootstrapped samples. By using a threshold of 95% consistency in sign of the moderator coefficient, variables with smaller effects that were consistent in the direction with which they predict differential response across treatments could be included. The primary goal of this step was to ensure that the variables selected will function properly and consistently, and increase the likelihood that the final model will replicate in future samples drawn from the same population.

Generating PAIs. Based on the set of variables selected, outcome predictions were generated for each study participant in both of the treatments. To avoid the risk of overconfidence that could occur when evaluating model-performance on individuals whose data were used to set model-weights, these predictions were generated using ten-fold cross validation (CV). 10-fold CV is recommended based on its good bias and variance properties in small samples (Kuhn & Johnson, 2013). For each of the 10 folds, individuals in that fold were held out, and data from the patients in the other 9 folds were used to generate a linear regression model in which end-of-treatment HAM-D score was predicted by the set of selected predictor variables (main effects for each variable and terms representing their interactions with treatment). The data from the patients in the held-out fold were then used to generate predictions for those patients in each treatment. For each individual, the difference in the predicted HAM-D score in CBT and PDT is their PAI. Individuals with a lower predicted HAM-D score in CBT (and thus a better predicted outcome in CBT) were then classified as "CBT-indicated" and individuals who had a better predicted outcome in PDT were labeled "PDT-indicated". The size of the PAI is taken to be an indication of the strength of the treatment recommendation (DeRubeis et al., 2014).

Despite our use of cross-validation during the weight-setting stage, our use of the full sample during variable selection and imputation could lead to model overfitting and inflated relationships (Fiedler, 2011), and as noted by Hastie et al. (2009), represents a form of double-dipping that can increase risk of overconfidence. This fact, along with our small sample size, contributes to our strong recommendation that the model and variables

presented here should not be used to guide treatment decisions unless (or until) they are validated in an external sample.

Evaluating PAIs. To characterize the expected utility of the PAIs for guiding treatment selection, we compared the average end-of-treatment HAMD scores of individuals who got their indicated treatment (based on their PAI scores) against that of participants who received their non-indicated treatment. Next, we looked within the subgroup indicated to need CBT, and compared HAMD scores for those who received their indicated treatment (CBT) to those who received their non-indicated treatment (PDT). We then performed the analogous comparison for those identified as "PDT-indicated." In order to investigate the importance of the strength of these recommendations following earlier PAI efforts (DeRubeis et al., 2014; Huibers et al., 2015), we then evaluated the above comparisons within the strongest 60% of PAIs (the 60% of the largest absolute value PAIs). The entire 10-fold cross-validation procedure and evaluation was repeated 1000 times to account for the influence of the selection of the 10 folds on the results (Kuhn & Johnson, 2013). The findings presented below summarize the results from these 1000 runs.

Results

Variable selection. Table 1 summarizes the results of our new variable selection approach at each stage (See Supplemental for a more detailed discussion of the results from each approach).

Table 1. Summary of variable selection results

	<------	Step 1	------>	Step 2	Result
Variable	Random Forest	Elastic Net	BART	Included in BootStep AIC	Selected by BootStep AIC
Baseline HAM-D	Yes	Yes	Yes	Yes	**Yes**
Age	Yes	No	No	No	N/A
Anxiety Sensitivity (ASI)	No	Yes	Yes	Yes	**Yes**
BAI	Yes	No	No	No	N/A
(BSI 2) Cognitive Problems	Yes	Yes	Yes	Yes	No
(BSI 3) Interpersonal Sensitivities	Yes	Yes	No	Yes	No
(BSI 4) Depressed Mood	Yes	No	Yes	Yes	**Yes**
(BSI 5) Fear	Yes	Yes	Yes	Yes	No
(BSI 7) Phobic Fears	Yes	Yes	No	Yes	No
(BSI 8) Paranoid Thoughts	Yes	Yes	Yes	Yes	**Yes***
Contacted Physician	No	Yes	No	No	N/A
Dysthymia	No	Yes	No	No	N/A
Employed	Yes	Yes	Yes	Yes	No
Episode Duration	No	Yes	No	No	N/A
Inventory of Depressive Symptomatology (IDS)	Yes	Yes	Yes	Yes	No
LEIDS Acceptance	No	Yes	No	No	N/A
LEIDS Hopelessness	No	No	Yes	No	N/A
Mobility	No	Yes	No	No	N/A
NEO Extraversion	No	Yes	Yes	Yes	**Yes**
NEO Neuroticism	Yes	Yes	Yes	Yes	No
NVM Extraversion	Yes	Yes	Yes	Yes	No
NVM Somatization	Yes	Yes	Yes	Yes	No

Pain (VAS)	Yes	Yes	Yes	Yes	No
Psychological Needs (PRF)	Yes	No	Yes	Yes	**Yes**
NEO Neuroticism x Married	N/A	N/A	Yes	Yes	No

Table 1. Summary of variable selection results for all variables selected by at least one approach. Three different variable selection approaches based on Random Forest, Elastic Net Regularization, and Bayesian Additive Regression Trees (BART) were applied to the full set of 49 potential baseline predictors. The 16 potential moderators that were selected by at least two of these three approaches were then submitted, along with one three-way interaction identified by BART (NEO Neuroticism x Married x Treatment), to a final variable selection stage with BootStepAIC. Bold text indicates the variables selected by BootStepAIC based on a criteria of at least 95% consistency of the coefficient sign for the interaction with treatment across 10,000 bootstrapped samples. *although BSI 8 was selected by BootStepAIC, its p-value in the final model built in the full sample was .43, and so, following the recommendation of Kuhn and Johnson (2013) to favor simpler models, it was not included in the final model.

The final model including the variables selected by BootStepAIC was:

Y = tx * (HAM-D Baseline + ASI + Depressed Mood (BSI 4) + NEO Extraversion + Psychological Needs)

Thus, five variables were selected as predictors of differential treatment response: HAM-D score, Brief Symptom Inventory (BSI; (De Beurs & Zitman, 2005) Depressed Mood subscale, Anxiety Sensitivity Index total score (ASI; (Reiss et al., 1986), NEO Five Factor Inventory (NEO-FFI; (Hoekstra et al., 2003) Extraversion subscale, and Patient Request Form Psychological Needs subscale (Veeninga & Hafkenscheid, 2004). We note that although both the depressed mood subscale of the BSI and the HAM-D measure the construct of depression, the correlation between the two scales was low (r=.23). One explanation for this might be that the HAM-D measures a more broad set of symptoms (e.g., sleep, libido, appetite, psychomotor retardation, etc.) than the 6 items

that are captured by the BSI's depressed mood subscale: suicidal thoughts, loneliness, sad/depressed mood, lack of interest, hopelessness, and worthlessness. There were no significant differences between the treatment groups for any of the variables selected for the final model (see Supplemental Table S3). Table 2 presents the final model with weights set using the full sample.

Table 2. Final regression model specified using the full sample.

Variable	*B*	*SE*	*p* value
(Intercept)	13.14	0.53	0.00**
Treatment	–0.50	1.05	0.63
ASI	0.40	0.64	0.53
Depressed Mood (BSI 4)	–0.16	0.63	0.80
HAM-D Baseline	3.30	0.58	0.00**
NEO Extraversion	–1.41	0.58	0.02*
Psychological Needs	0.11	0.56	0.85
Treatment x ASI	–3.22	1.28	0.01*
Treatment x Depressed Mood (BSI 4)	2.97	1.27	0.02*
Treatment x HAM-D Baseline	1.19	1.16	0.31
Treatment x NEO Extraversion	3.10	1.15	0.01**
Treatment x Psychological Needs	1.65	1.46	0.15

*$p < .05$, **$p < .01$

Baseline HAM-D score was included as both a main effect and as an interaction with treatment, but it did not appear to have a significant moderator relationship in the context of the final model. The moderator relationships included in the final model are visualized in Figure 1. We refer the reader to a recent review on treatment selection by

Cohen and DeRubeis (2018) for a detailed discussion of how to approach interpretation of moderator relationships in treatment selection.

Figure 1. Visualization of the moderator relationships.

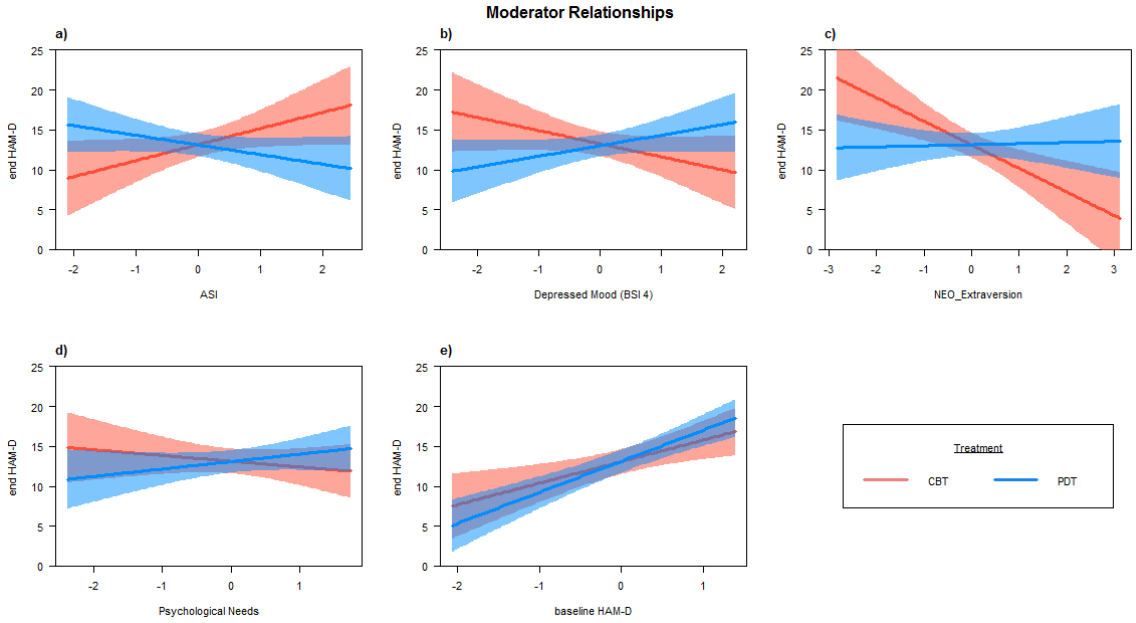


Figure 1. Conditional plots with confidence bands for the conditional mean generated using R package visreg from the final model estimated in the complete sample. Conditioning for each plotted variable uses the mean value for all other variables. The X-axes represent the standardized/centered scores that were used during analysis.

PAI Results. Individuals who received their model-indicated treatment had better outcomes than those who received their non-indicated treatment (see Figure 2). The mean end of treatment HAM-D scores, averaged across the 1,000 CVs for individuals who received their PAI-indicated treatment, was 12.3 (SD=7.6); the average for those receiving their non-indicated treatment was 13.9 (SD=7.9). This reflected, on average, a 1.6 points advantage for those receiving their indicated treatment (95% CI=0.5 to 2.8; Cohen's *d*=0.21, 95% CI=0.07 to 0.37). When we restricted our evaluation to the largest

60% of PAIs (absolute value), we found that the effect of treatment selection grew to 2.6 points (95% CI=1.4 to 3.7; average Cohen's *d*=0.37, 95% CI=0.19 to 0.54).

Figure 2. Comparison of end-of-treatment HAM-D scores for patients randomized to their PAI-indicated treatment with those who were randomized to their non-indicated treatment.

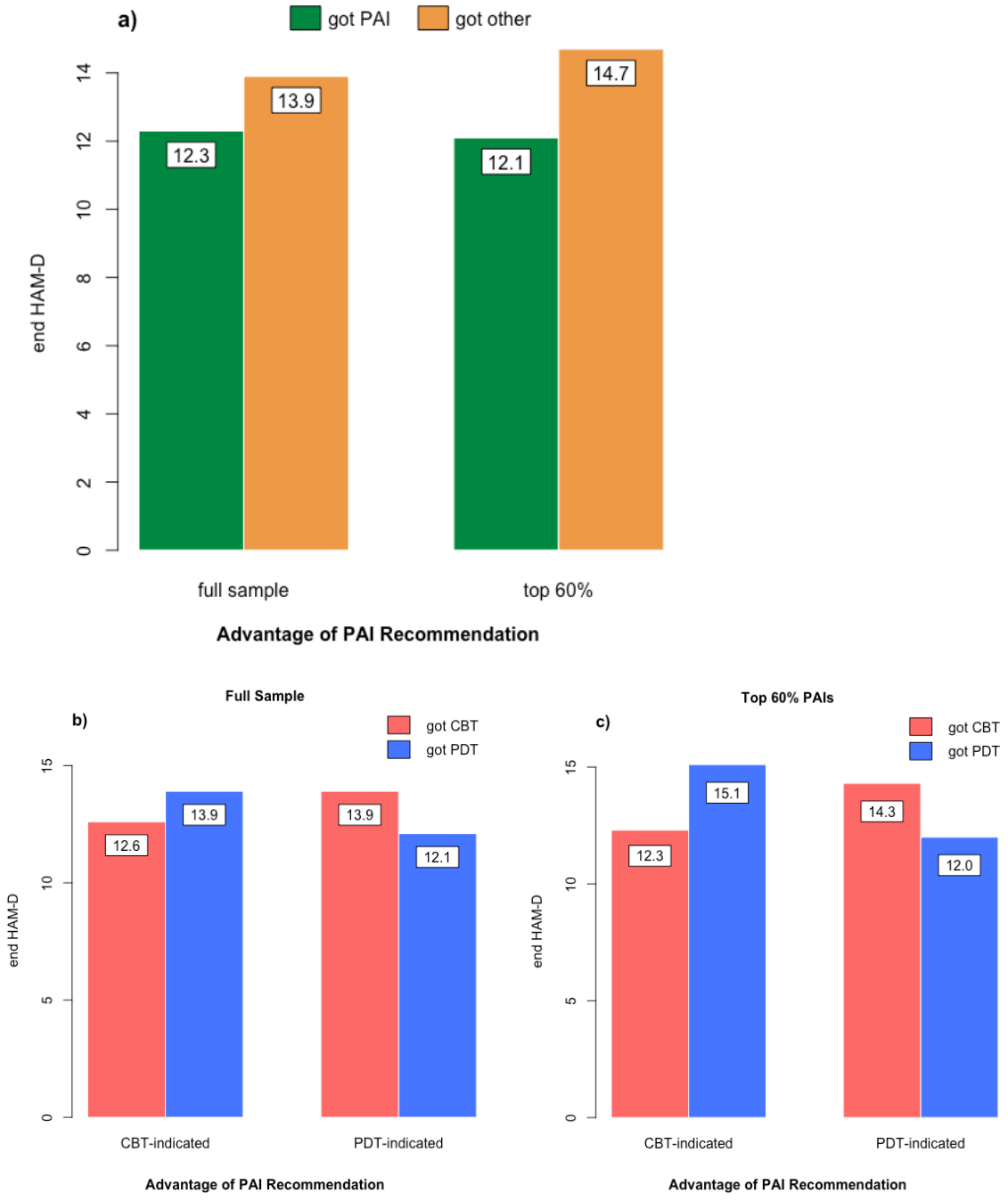


Figure 2. Panel 2a shows this comparison with treatment conditions collapsed for the full sample (left set of bars), and for the 60% of patients with larger PAIs (right set of bars). Figure 2b decomposes the comparison by treatment for the full sample, with those indicated to need CBT represented by the left two bars, and those indicated to need PDT by the right two bars. Figure 2c presents the same breakdown as in figure 2b, but for the 60% of patients with larger PAIs.

Discussion

Helping service-users and clinicians make better-informed treatment decisions is one of the core goals of precision medicine in mental health. In depression, treatment selection models can improve the ability to identify the best treatment among available options (Cohen & DeRubeis, 2018). Here, we have described a treatment selection model based on patient characteristics that could be used to decide between cognitive-behavioral and psychodynamic therapy for those with mild-to-moderate depression not taking antidepressants.

The differential prediction described here relied on four factors: anxiety sensitivity, depression symptom severity, extraversion, and psychological treatment needs. Although in this investigation the aim was to develop and provide a first test of a multivariable model that could inform treatment selection, it nonetheless is important to attempt an understanding of the basis for each variable's contribution to the model. In the following we provide tentative, speculative interpretations of the findings, taking into account the directions of the relationships observed.

The ASI reflects a person's beliefs that anxiety experiences have negative somatic, psychological or social consequences. Higher scores on this measure were associated with superior outcomes in PDT relative to CBT, and vice versa for lower scores. Patients with higher baseline depressed mood, as measured by the BSI, tended to

improve more in CBT, whereas those with lower BSI scores tended to respond better to PDT. Those with higher scores on a measure of extraversion (on the NEO) fared better in CBT than in PDT. The reverse was true for those low on extraversion. The Patient Request Form Psychological Needs scale assesses a patient's needs for a psychological treatment. Higher scores on this measure predicted better response to PDT, relative to CBT, and the reverse prediction was obtained for those with lower scores on this measure. Thus, in contrast to anxious, introverted patients, patients who were relatively more extraverted and who had low psychological treatments needs were better matched to CBT. We would speculate that these patients typically express themselves more and have already talked about their feelings and problems with others without much hesitation. They might be more in need of the structured approach of CBT, directed strongly at adapting behavior and changing cognitions through practical exercises. It may be that PDT is more efficacious for patients who search for a psychological solution to their depressive symptoms. Anxious and introverted patients, who have a tendency to avoid focusing on their problems despite a need to do so, may find that the supportive milieu of PDT fostered explorations of their feelings and problems in a way that was appropriate for their individual needs and capacities.

Limitations

This study has a number of limitations. Treatment selection is likely to be most effective when the interventions under consideration differ substantively and substantially in their mechanisms and targets (Cohen & DeRubeis, 2018). Relative to comparisons between, for example, medications and psychotherapy, the similarity of these two psychotherapies likely resulted in a decreased potential to identify individuals

who are strongly indicated to need one treatment over another. Nevertheless, among the 60% of patients with the strongest PAIs, a *d*-type effect size of 0.37 was observed for receiving the indicated versus contraindicated treatment.

Patients seeking treatment for depression have many options, including other psychotherapies and medication; this model cannot inform the decision of whether or not to pursue treatments other than CBT and PDT. This model would at best be valid for use in similar populations. It cannot be known how the model would perform if it were applied to patients whose values on predictors were outside the observed range on the predictor measures, or if used in the context of a population of those with severe depression or in patients who are also taking antidepressant medications.

Kessler and colleagues (2017) leveled a valid criticism of the early publications on treatment selection in depression (DeRubeis et al., 2014; Huibers et al., 2015), arguing that, because variable selection was performed within the sample on which the model was evaluated, "any attempt to use the coefficients in these models to predict differential treatment response in a new sample of patients would almost certainly yield less positive effects than those suggested by the results of studies" (p. 6). Kriegeskorte and colleagues (2009) also note that this approach can lead to distorted descriptive statistics and invalid statistical inference. The size of the sample used in these analyses, although drawn from the largest RCT comparing these two treatments to date, did not allow for a true hold-out. As described by Hastie et al. (2009), the CV scheme applied here has given an unfair advantage to the predictors as they were chosen on the basis of the full sample, thus we do not approximate an evaluation of our models in a completely independent test set. Thus, we stress that this model would need to be validated in an independent sample

before being considered for use as a clinical decision tool. Despite these shortcomings, we believe this work represents an important step towards the types of studies that would address Kessler's criticisms (e.g., prospective tests). These efforts propose potential moderators and present a specific model that could be investigated in future studies, and identify an important issue that merits increased attention by those interested in precision medicine: methodological heterogeneity in variable selection.

Future Directions

The treatment selection subfield of precision medicine in mental health is still in its developmental stage, and the statistical methods described in this and other similar efforts are constantly evolving (Zilcha-Mano, 2018). Although replication and external validation are essential steps that should precede the implementation of any specific treatment selection model, the publication and discussion of candidate predictors, models and statistical approaches are equally important, as they set the foundation for future efforts.

A wide variety of feature selection techniques have been employed in recent efforts to construct treatment selection models in mental health, and no clear guidance exists as to which approach is best. We have presented an example of a new variable selection approach that incorporates several of the leading techniques in order to identify reliable predictors of differential response to treatment. We propose this specific combination as a starting point and suggest that future efforts should explore different permutations, such as adding other methods (e.g., Support Vector Machines), reducing the number of approaches used, using different combinations, or adjusting the settings within each of these techniques. Examples of the latter include adjusting the thresholds

for the inclusion of variables and specifying different tuning parameters (Kuhn & Johnson, 2013).

Conclusion

We refined the Personalized Advantage Index approach to generate individual treatment recommendations for adults with mild to moderate depression not taking antidepressants who are deciding between CBT versus PDT. Our novel approach synthesized the results of four different variable selection techniques by selecting the patient characteristics that are consistently identified as associated with (differential) treatment outcome. Although no significant efficacy differences were found between CBT and PDT across the total sample, the resulting treatment recommendations suggested that for the majority of the individual patients, one of the treatments could be predicted to more efficacious than the other based on a model including four pre-treatment patient characteristics (anxiety sensitivity, depression symptom severity, extraversion, and psychological treatment needs). The small sample and lack of a separate validation sample indicate the need for prospective tests (Lutz et al., 2017) before using this model for treatment selection, but these findings add to a growing literature on the potential for model-guided treatment recommendations to improve patient outcomes for depression.

Abbreviations used in manuscript: AIC = Akaike Information Criterion, ASI = Anxiety Sensitivity Index, BART = Bayesian Additive Regression Trees, BSI = Brief Symptom Inventory, CBT = Cognitive Behavioral Therapy, DSM-IV = Diagnostic and Statistical Manual of Mental Disorders 4th edition, ENR = Elastic Net Regularization, HAM-D = Hamilton Rating Scale for Depression, IDS = Inventory of Depressive Symptomatology, LEIDS = Leiden Index of Depression Sensitivity, MSE = Mean Squared Error, NEO = Neuroticism-Extraversion-Openness Personality Inventory, NVM = Shortened Dutch Adaptation of the Minnesota Multiphasic

Personality Inventory, PAI = Personalized Advantage Index, PDT = Psychodynamic Therapy, PRF = Patient Request Form, RF = Random Forests, VAS = Visual Analog Scale for Pain

Acknowledgments

We would like to thank Robert DeRubeis for his insightful comments on this manuscript. ZDC was supported by the MQ Foundation under Grant MQ: Transforming mental health MQ14PM_27. The trial was financed by an unrestricted research grant by Wyeth Pharmaceuticals, The Netherlands. Arkin Mental Health Care, The Netherlands, financially supported research logistics and the contributions of ED, HLV, and JJMD. Vrije Universiteit Amsterdam, Faculty of Behavioral and Movement Sciences, Section Clinical Psychology, The Netherlands, financially supported ED's contributions to the study. None of the sponsors had a role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; nor in the preparation, review, or approval of the manuscript. The opinions and assertions contained in this article should not be construed as reflecting the views of the sponsors.

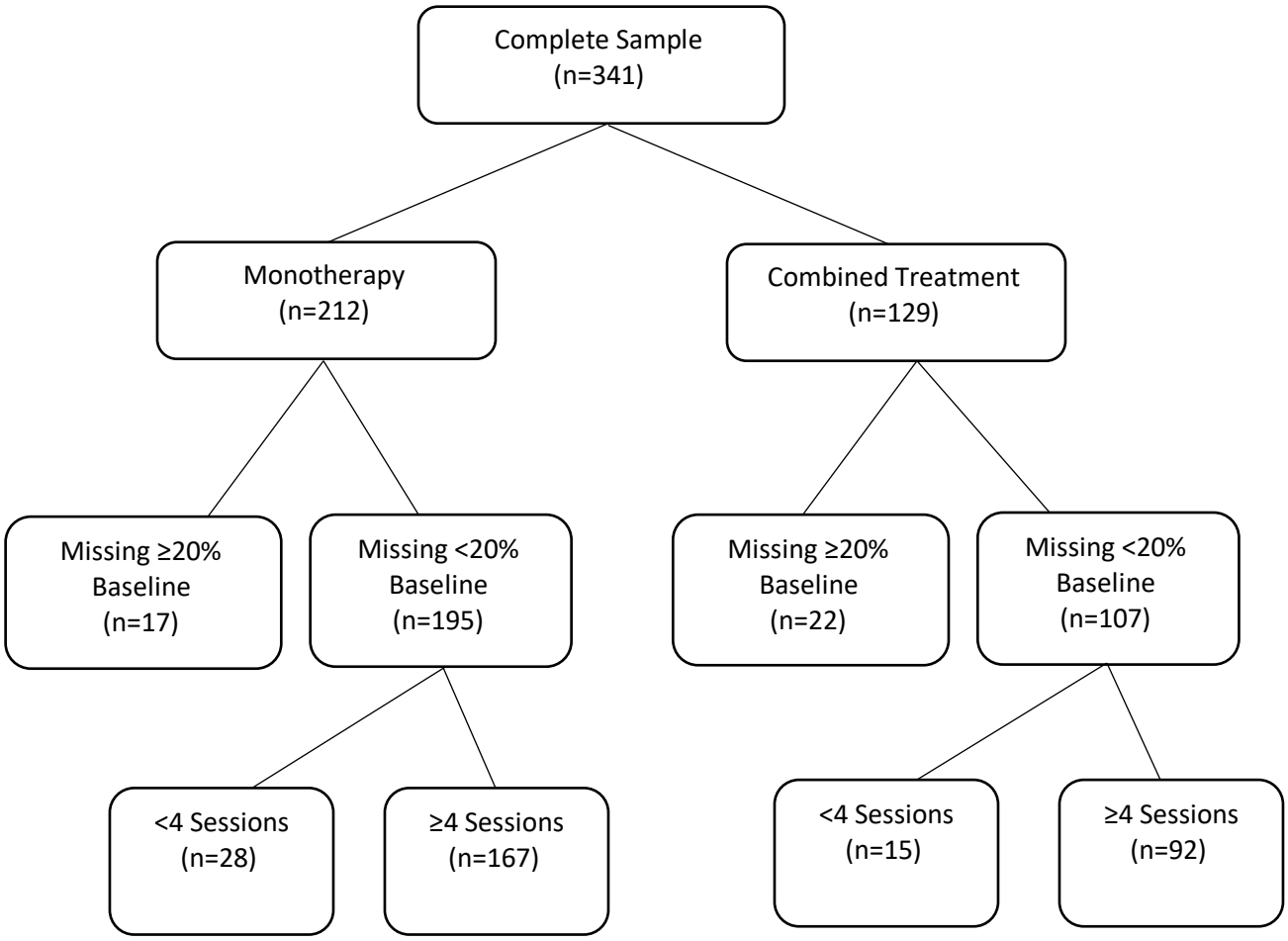
Supplemental Material: Participants

Participants in the trial (n=341) were referred by their general practitioner to one of three outpatient mental health clinics in Amsterdam, The Netherlands. Inclusion criteria were: 1) presence of a depressive episode according to DSM-IV criteria as assessed with the *MINI-International Neuropsychiatric Interview – Plus* (MINI-Plus; (Sheehan et al., 1998), 2) *Hamilton Depression Rating Scale* (HAM-D; (Hamilton, 1960) scores ≥ 14, 3) age between 18 and 65 years, and 4) written informed consent after explanation of the study procedures. Exclusion criteria included presence of psychotic

symptoms or bipolar disorder, severe suicidality warranting immediate intensive treatment or hospitalization, substance misuse/abuse in the last six months, pregnancy, inability to meet trial demands, and use of psychopharmacology or other medications that might influence mental functions. Participants were not compensated for their participation in the study.

As our goal was to build a model to answer how individuals who *received* CBT or PDT fared, we felt that individuals who dropped out very early in treatment would not be meaningful to our models. In the extreme, the "outcome" for patients who dropped out prior to attending a single session does not reflect response to CBT or PDT. We decided to remove from our analyses patients who attended 3 or fewer therapy sessions. This reduced our sample from 195 to 167. Figure S1 presents the patient flow chart for the sample used in this study. Baseline sample demographic characteristics for the final sample are presented in Table S1. Patients in CBT (compared to PDT) were more likely to be married (32% vs. 17%, p = .03), had shorter average episode durations (2.29 vs. 2.84 years, $p < .005$), and fewer serious life events (4.88 vs. 5.73, $p = .03$).

Supplemental Figure 1. Patient flow chart


Complete Sample
(n=341)
Monotherapy
(n=212)
Combined Treatment
(n=129)
Missing ≥20%
Baseline
(n=17)
Missing <20%
Baseline
(n=195)
Missing ≥20%
Baseline
(n=22)
Missing <20%
Baseline
(n=107)
<4 Sessions
(n=28)
≥4 Sessions
(n=167)
<4 Sessions
(n=15)
≥4 Sessions
(n=92)

Supplemental Table 1. Baseline sample characteristics

Demographic characteristics	**CBT (*n* = 75)**	**PDT (*n* = 92)**	Mean difference (95% CI) X^2 (*df*)	*p* Value
Age (years)			2.70 (–0.61-6.01)	.11
Mean (*sd*)	38.13 (10.93)	40.83 (10.65)		
Range	23-64	23-63		
Female (%)	55 (73)	63 (68)	0.47 (1)	.49
Northwest European (%)	50 (67)	58 (63)	0.24 (1)	.63
Married (%)	24 (32)	16 (17)	4.84 (1)	.03*
Employed (%)	31 (41)	44 (48)	0.70 (1)	.40
Prior Medication (%)	33 (44)	39 (42)	0.04 (1)	.83
Episode Duration			0.55 (0.18-0.92)	.003**
Mean (*sd*)	2.29 (1.15)	2.84 (1.22)		
Range	1-4	1-4		
Prior Treatment (%)	24 (32)	31 (34)	0.05 (1)	.82
Serious Life Events			0.85 (0.10-1.60)	.03*
Mean (*sd*)	4.88 (2.47)	5.73 (2.42)		
Range	0-11	1-11		
Education Level			2.44 (2)	.30
Low (%)	13 (17)	20 (22)		
Intermediate (%)	27 (36)	40 (43)		
High (%)	35 (47)	32 (35)		

Table S1. * = $p < .05$, ** = $p < .01$.

Supplemental Material: Data Pre-Processing

Prior to variable selection, we removed baseline variables with greater than 20% missingness. Categorical variables were transformed into binary variables. Furthermore, binary variables with to little variance (whose smallest category made up less than 20% of the sample) were removed; a low representation could lead to coefficient instability during cross-validation. Redundant variables were removed. Some variables were made ordinal, based on distributional or theoretical reasons. After data pre-processing, our new dataset dropped from 195 baseline predictor variables (see Table S2) to 49. Our variables were then mean-centered and standardized to improve numerical stability in future calculations (Kraemer & Blasey, 2004; Kuhn & Johnson, 2013).

Supplemental Table 2. **All baseline predictors**

Variable Label	Explanation	Included yes/no	Reason excluded	Data Pre-Processing
AFHNKL	Dependent personality disorder (PD)	no	too little variance	
ANTISL	Antisocial PD	no	too little variance	
ASI	Anxiety Sensitivity Index	yes		
BAI	Beck Anxiety Inventory	yes		
HDRS_baseline	Hamilton Depression Rating Scale (baseline)	yes		
BRDRLL	Borderline PD	no	too little variance	
BSI_1	Brief Symptom Inventory (BSI) Somatic complaints	yes		
BSI_2	BSI Cognitive problems	yes		
BSI_3	BSI Interpersonal sensitivities	yes		
BSI_4	BSI Depressed mood	yes		
BSI_5	BSI Fear	yes		
BSI_6	BSI Amount Of Hostility	yes		Winsorized 1 high outlier.
BSI_7	BSI Phobic fears	yes		Winsorized 2 high outliers.
BSI_8	BSI Paranoid thoughts	yes		
BSI_9	BSI Psychoticism	yes		Winsorized 1 high outlier.
bsi_tot	BSI total score	no	included as subscales	
clusa	Cluster A Personality	no	too much missingness	
clusb	Cluster B Personality	no	too much missingness	
clusc	Cluster C Personality	no	too much missingness	
das	Dysfunctional Attitudes Scale	no	too much missingness	
dem1	Nationality	no	similar to “Ethnicity”	

Ethnicity	Ethnic/cultural group	yes		Recode to binary as Northwest European (Northwest European) vs. not Northwest European (all other categories)
Married	Marital status	yes		Recode to binary as married (married) vs. not married (all other categories)
dem4	Living Situation	no	similar to “Married”	
Religion	Religion	yes		Recode to binary as religious (all categories besides atheist) vs. not religious (atheist)
dem6	Highest training	no	similar to “Education”	
Education	Educational level (completed)	yes		
dem7	Graduated?	no	similar to “Education”	
dem8	Current job	no	similar to “Employment”	
Employment	Work Situation	yes		Recode to binary as employed (job, student) vs. not employed (sickness, allowance, disability, other).
Main_Earner	Main Earner?	yes		Recode to binary as main earner (main earner, dual earner) vs. not main earner (partner is main earner, parents are main earner, other).
dem11	Job main earner	no	similar to “Main_Earner”	
dem12	Main income source main earner	no	similar to “Main_Earner”	
dem13	Height main income earner	no	similar to “Income”	

Income	Recoded income level	yes		Recode to binary as below poverty line (<= 1273 gross per month) vs. above poverty line (>1273 gross per month)
DEPRSL	Depressive PD	no	similar to "LEIDS_RiskAversion", "ONTWKL", and "dimtot"	
dimtot	Self-report questionnaire for personality disorders	no	too much missingness	
Mobility	EuroQol (EQ) Item 1: Mobility	yes		Recode to binary as no problems with mobility (no problems) vs. problems with mobility (some problems, bedridden)
eq2	EuroQol (EQ) Item 2: Self Care	no	too little variance	Similar scheme as "Mobility"
eq3	EuroQol (EQ) Item 3: Daily activities	no	similar to "HealthStatus"	Similar scheme as "Mobility"
eq4	EuroQol (EQ) Item 4: Pain/complaints	no	similar to "HealthStatus"	Similar scheme as "Mobility"
eq5	EuroQol (EQ) Item 5: General Mood	no	similar to "HealthStatus"	
eq6	EuroQol (EQ) Item 6: Current health status vs. last year health status	no	similar to "HealthStatus"	
HealthStatus	Health Status	yes		Winsorized 3 low outliers.
gaf_ft	GAF score according to pharmacotherapist (for severely depressed patients with combined treatment only)	no	too much missingness	
IDS	Inventory of Depressive Symptomatology, Self-Report	yes		Winsorized 3 high outliers and 6 low outliers.
lasinte_m	Social Integration	no	too much missingness	

lasloss_m	Loss	no	too much missingness	
lasnorm_m	Values and Norms	no	too much missingness	
lasskill_m	Skills	no	too much missingness	
lastrad_m	Traditions	no	too much missingness	
Age	Age in years	yes		
LEIDS_Acceptance	Leiden Index of Depression Sensitivity (LEIDS) acceptance/coping	yes		Log transformed (add 1 to account for values of 0)
LEIDS_Aggression	LEIDS Aggression	yes		Winsorized 1 high outlier.
LEIDS_Perfectionism	LEIDS Control/perfectionism	yes		
LEIDS_Hopelessness	LEIDS Hopelessness/suicidality	yes		
LEIDS_RiskAversion	LEIDS Risk aversion	yes		
LEIDS_Rumination	LEIDS Rumination	yes		Winsorized 1 low outlier.
leids_tot	LEIDS total score	no	included as subscales	
SeriousLifeEvents	Total number of serious life events	yes		
EpisodeDuration	How long have you had these complaints for?	yes		
PriorTx	Did you receive prior treatment for the current depressive episode?	yes		
mide29	How many previous periods in your life did you feel depressed and had these symptoms?	no	similar to "EpisodeDuration"	
Dysthymia	Comorbid dysthymia	yes		
neo_altrui	NEO Altruism	no	too much missingness	
neo_cons	NEO Conscientiousness	no	too much missingness	
NEO_Extraversion	NEO Extraversion	yes		Winsorized 2 high outliers and 2 low outliers.
NEO_Neuroticism	NEO Neuroticism	yes		Winsorized 1 low outlier.
neo_open	NEO Openness	no	no values	
NRCSTL	Narcissistic PD	no	too little variance	

NVM_Extraversion	NVM Extraversion	yes		
NVM_Negativism	NVM Negativism	yes		Winsorize 1 high outlier and 3 low outliers.
NVM_Psychopathology	NVM Serious pathology	yes		
NVM_Somatization	NVM Somatization	yes		
NVM_Embarrassment	NVM Embarrassment	yes		
OBSESL	Obsessive Compulsive PD	no	too much missingness	
ONTWKL	Avoidant PD	no	similar to “DEPRSL”, “persstr”, “clusc”, “dimtot”	
oq_ernst	OQ (from Outcome Questionnaire-45) Ernst	no	similar to “BAI”, “BSI_2”, “BSI_3”, “BSI_4”, “BSI_5”, “BSI_7”, “BSI_9”	
oq_interp	OQ Interpersonal	no	too much missingness	
oq_maatsch	OQ Social	no	too much missingness	
oq_tot	OQ total score	no	total score	
PARANL	Paranoid PD	no	similar to “ps”, “persstr”, “clusa”, “dimtot”	
persstr	Number of personality disorders	no	too much missingness	
pm	Psychological mindedness scale	no	too much missingness	
PRF_PsychNeeds	Patient Request Form: Psychological needs	yes		Winsorize 1 low outlier.
PRF_MedicalNeeds	Patient Request Form: Medical and passive needs	yes		Winsorize 2 high outliers.
ps	Personality disorder	no	too much missingness	
PSAGRL	Passive Aggressive PD	no	too much missingness	
SCHZOL	Schizoid PD	no	too little variance	
SCHZTL	Schizotypal PD	no	too little variance	
Gender	Gender	yes		
THEATL	Histrionic PD	no	too little variance	

Contact_Physician	Number of times contacted GP in the past four weeks	yes		
tic2	Number of times contacted care provider in mental health care institute	no	similar to "Contact_Physician"	
tic3	Number of times contacted psychologist or psychotherapist in private practice	no	too little variance	
tic4	Number of times contacted psychologist or psychotherapist at psychiatry department hospital	no	too little variance	
tic5	Kind of hospital	no	too much missingness	
Contact_Doctor	Number of times contact company doctor	yes		
tic7	Number of times contact medical specialist	no	similar to "Contact_Doctor"	
tic8	Type of medical specialist	no	too much missingness	
tic9	Number of times contact physiotherapist	no	too little variance	
tic10	Number of times contact social work	no	too little variance	
tic11	Number of times contact CAD	no	only 1 value	
tic12	Used home care	no	too little variance	
tic13	Number of times contact alternative healer	no	too little variance	
tic14	Type of alternative healer	no	too much missingness	
tic15	Day or part-time treatment	no	only 1 value	
tic16	Type of institution day/share	no	too much missingness	
tic17	Inpatient treatment in health care institute	no	too little variance	
tic18	Type of (general) health care institute	no	too much missingness	
tic19	Attended a self-help group	no	too little variance	
PriorMed	Any prior medications?	yes		
tic22	Highest education level	no	too much missingness	
tic23	Chronic disease in the past or current year	no	similar to "EpisodeDuration"	

UCL_ActiveAddress	Utrecht Coping List (UCL) Active addressing	yes		Winsorized 2 high outliers and 1 low outlier.
UCL_Avoidance	UCL Avoidance	yes		Winsorized 1 high outlier.
VAS_Pain	Somscore Visual Analog Scale for Pain	yes		

Supplemental Material: Missing Data Imputation

Imputation: To impute missing baseline data, we used a random forest-based imputation strategy (missForest package in R; (Stekhoven & Buhlmann, 2012), which generates a single imputed dataset by averaging over multiple regression trees, thus giving the benefit of multiple imputation without needing to run the primary analyses across multiple imputed datasets (as with multivariate imputation by chained equations). It has been found to outperform other methods of imputation, especially when complex and non-linear interactions are present, and can handle different types of variables (Shah et al., 2014). We imputed missing values using all 49 baseline predictors, as well as other data including longitudinal outcome data, and information about treatment (e.g., number of sessions, etc.) with random forest. The treatment variable was not included during imputation. Integer-type variables with missing values that were imputed were rounded. For imputing missing baseline data, using outcomes for imputation generates coefficients closer to "true" coefficients, compared to not using outcomes for imputation (which produces biased [underestimated] coefficients; (Moons et al., 2006). Following imputation, we removed some of the baseline predictors that were included to improve the quality of our imputation but that were not appropriate to use as predictors in the context of the PAI. Specifically, we removed therapist variables (e.g., therapist age) because these variables do not reflect patient characteristics.

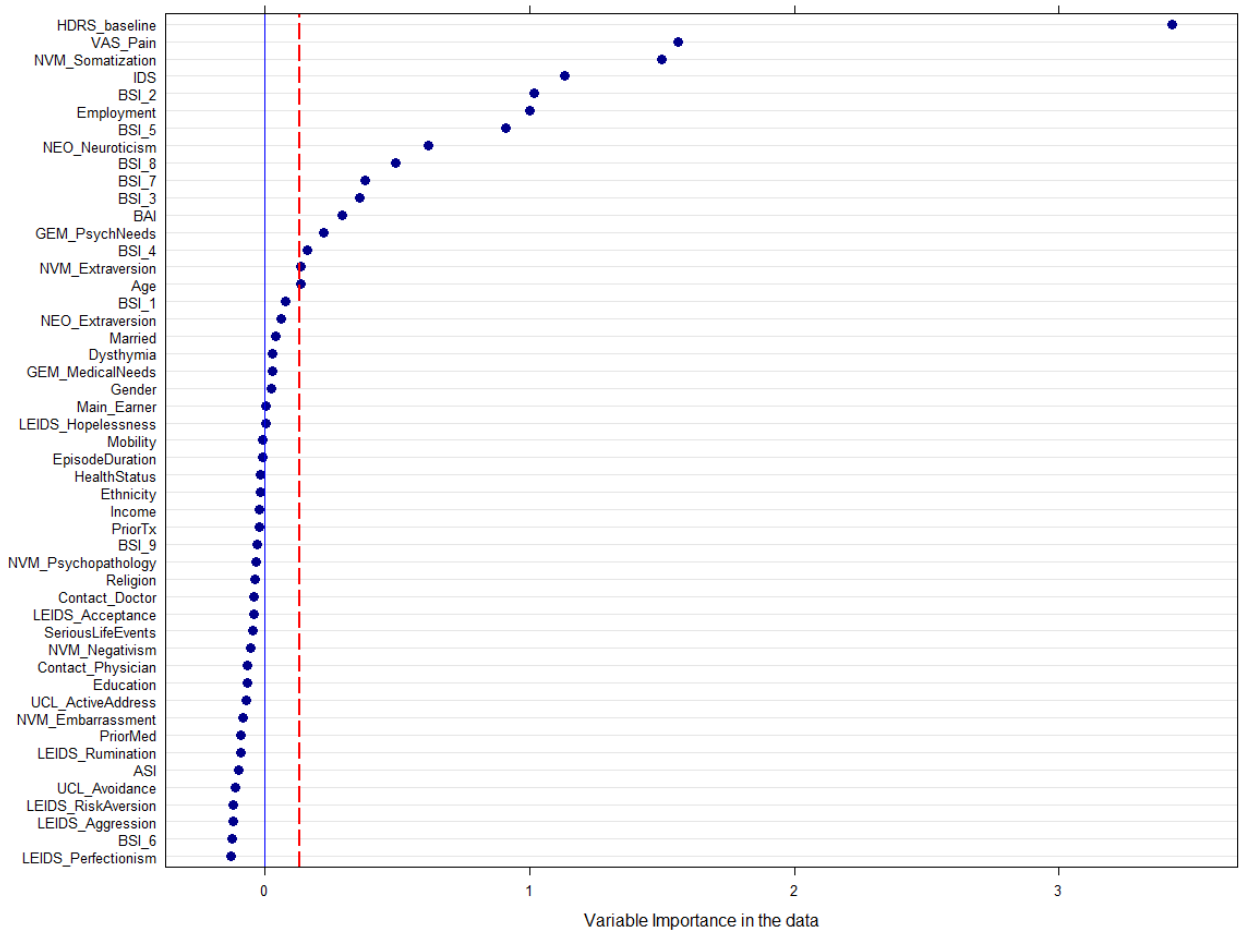
Supplemental Material: Variable Selection

Random Forest

For Random Forest (RF) variable selection, we used the mobForest (Garge et al., 2013) package in R. RF can be made to focus on predictor by treatment group

interactions during the model building process by forcing the approach to select "splits" that maximize the difference in the treatment condition coefficient between subgroups (by specifying the model function of RF as "y ~ treatment"). We set the mobForest mtry criteria, which determines how many variables random forest evaluates at each node for a single tree, to 16 based on the recommendation that mtry equal the number of predictors divided by three (Kuhn & Johnson, 2013). The number of trees was set to 10,000 to stabilize the results. Figure S2 presents the output from the varimplot() command, which provides a visualization of each variable's importance (evaluated by random forest). The importance value for each variable is determined through permutation tests. Variables to the right of the dotted red line (the permutation threshold) are retained.

Supplemental Figure 2. Random forest variable importance plot with permutation test



Elastic Net Regularization

Elastic Net Regularization (ENR) can provide a hybrid of the Lasso and Ridge regression approaches, combining the L1 and L2 penalizations to allow for the selection of a parsimonious set of variables that predict outcome (Hastie et al., 2009). The degree to which ENR approximates these penalties is determined by the alpha setting, which can range between 0 (representing the ridge penalty) and 1 (representing the lasso). Zou and

Hastie (2005) recommend ENR (e.g., using alpha = 0.5) over lasso when dealing with highly correlated predictors, and so we set alpha = 0.5. We used the R package glmnet (Friedman et al., 2010) to implement ENR variable selection. ENR can handle high numbers of potential predictors and can overcome issues of high correlations between baseline variables. Current implementations of ENR in R do not accommodate variable selection for models in which moderators are of primary interest (because they do not link the main effects of variables to their interactions and thus can result in models wherein the interactions, but not their main effects, are included). Uses of ENR in the literature of variable selection and treatment selection have only investigated prognostic models in which a single treatment is modeled. In order to adapt ENR for the purpose of identifying moderators, we split the training sample into each of the two treatment groups, and then constructed prognostic models within each group. Variables that were retained in only one condition, or that were selected but specified with differing coefficients, were retained as potential moderators of treatment effects.

Bayesian Additive Regression Trees

Bayesian Additive Regression Trees (BART) builds on ensemble-of-tree methods such as RF by incorporating an underlying Bayesian probability model (Chipman et al., 2010). We used the bartMachine package (Kapelner & Bleich, 2016) in R for variable selection with BART. Bleich and Kapelner developed bartMachine to extract information about variable importance, and provide an interaction plot feature that can be used to identify potential 3-way interactions, and with Goldstein and colleagues added the ICEbox package in R that allows for the visualization of higher-order and non-linear relationships from BART models (Bleich et al., 2014; Goldstein et al., 2015; Kapelner &

Bleich, 2016). Kapelner and Bleich adapted the bartMachine R package for the purposes of variable selection for the PAI to help focus model building on variables that predict differential treatment response. Kapelner and Bleich adapted the bartMachine R-package (2016) to help focus model building on moderators. To achieve this aim, they introduced a parameter that forces the search for variable splits to focus more on treatment than other variables, thus introducing more interactions between treatment and other variables. This is conceptually similar to when researchers only consider interactions between treatment and baseline variables (and not interactions between baseline variables themselves), or to how RF can specify the splitting criteria to evaluate the difference in the treatment coefficient for the model y ~ tx. For each BART model, the bartMachine package can output variable importance values that represent the proportion of times each variable is chosen as a splitting rule, as well as interaction importance values representing the same information. Two variables are identified as having interacted in a given tree if they appear together in a contiguous downward path from the root node to the terminal node. The developers have not yet developed a significance test for the importance of interactions (Kapelner & Bleich, 2016), and so we decided to take the N most important interactions, where N was the number of moderators selected by RF (which uses a permutation test to determine an importance threshold cutoff). Given instability in the internal tree structures that generate predictions in BART models, we created 5 separate BART models and only retained the interactions that were selected in the majority of the 5 models.

Stepwise AIC-penalized bootstrapped approach

The next step was to reduce the variables consistently identified by at least two of the these three approaches (RF, ENR, BART) using the stepwise AIC-penalized bootstrapped approach (Austin & Tu, 2004) instantiated in the BootStepAIC package in R (Rizopoulos, 2009). By only including the moderator relationships (and not prognostic variables that didn't interact with treatment) identified by the other approaches (and their corresponding main effects), this search generated a model emphasizing the prediction of differential treatment response. Using the BootStepAIC approach as the final step, instead of RF or BART, was aimed to reduce the chance that predictors that require unspecified linear or higher-order interactions were be included. 10,000 bootstrapped training samples were drawn, and within each training sample backwards elimination was used to select variables that independently contribute to predicting outcome. We relied primarily on the consistency of the direction of the coefficients across the 10,000 bootstrapped samples. By using a threshold of 95% consistency in sign of the moderator coefficient, variables with smaller effects that were consistent in the direction with which they predict differential response across treatments can be included. The primary goal of this step was to ensure that the variables selected will function properly and consistently, and increase the likelihood that the final model will replicate in future samples drawn from the same population.

Supplemental Table 3. BootStepAIC variable selection moderator sign consistency output

Coefficient for Treatment Interaction Term	+ (%)	– (%)	Selected by BootStepAIC?
Baseline HAM-D	98	2	**Yes**
Anxiety Sensitivity (ASI)	1	99	**Yes**
(BSI 2) Cognitive Problems	59	41	No
(BSI 3) Interpersonal Sensitivities	44	56	No
(BSI 4) Depressed Mood	97	3	**Yes**
(BSI 5) Fear	85	15	No
(BSI 7) Phobic Fears	83	17	No
(BSI 8) Paranoid Thoughts	**1**	**99**	**Yes***
Employed	56	44	No
Inventory of Depressive Symptomatology (IDS)	87	13	No
NEO Extraversion	97	3	**Yes**
NEO Neuroticism	43	57	No
NVM Extraversion	87	13	No
NVM Somatization	6	94	No
Pain (VAS)	74	26	No
Psychological Needs (PRF)	95	5	**Yes**
NEO Neuroticism x Married**	8	92	No

Table S3. The 16 potential moderators that were selected by at least two of the three approaches (RF, ENR, BART), along with one three-way interaction identified by BART (NEO Neuroticism x Married x Treatment). These were submitted to a final variable selection stage with BootStepAIC. Bold text indicates the variables selected by BootStepAIC based on a criteria of at least 95% consistency of the coefficient sign for the interaction with treatment (presented in the middle columns) across 10,000

bootstrapped samples. *although BSI 8 was selected by BootStepAIC, its p-value in the final model built in the full sample was .43, and so, following the recommendation of Kuhn and Johnson (2013) to favor simpler models, it was not included in the final model. ** NEO Neuroticism x Married represents the 3-way interaction between these two predictors and treatment.

Supplemental Material: Supplemental Results.

Table S4 compares the distributions across the two treatment groups for the five moderator variables included in the final model. Statistical tests revealed no significant group differences in any of the variables in the final model

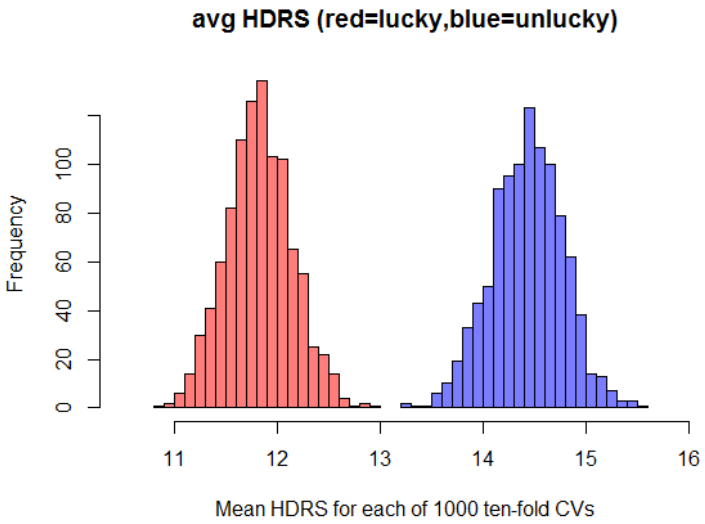
Supplemental Table 4. Predictor variables included in the final model.

Predictor	**CBT (*n* = 75)**	**PDT (*n* = 92)**	Mean difference (95% CI)	*p* Value
HAM-D Baseline				
Mean (*sd*)	19.95 (2.59)	20.00 (3.13)	0.05 (–0.84-0.94)	.91
Range	15-24	14-24		
Anxiety Sensitivity (ASI)				
Mean (*sd*)	34.82 (11.57)	33.71 (12.83)	–1.11 (–4.88-2.66)	.56
Range	17-73	16-70		
NEO Extraversion				
Mean (*sd*)	32.74 (5.97)	32.44 (7.04)	–0.30 (–2.32-1.71)	.77
Range	16-45	14-53		
Depressed Mood (BSI 4)				
Mean (*sd*)	2.28 (0.80)	2.23 (0.80)	–0.05 (–0.30-0.20)	.69
Range	0.17-4	0.33-3.67		
Psychological Needs				
Mean (*sd*)	3.76 (0.72)	3.70 (0.75)	–0.06 (–0.29-0.17)	.60
Range	1.78-5	1.78-5		

1000 CV PAI Runs

The results presented in the primary manuscript present summary findings from the 1000 ten-fold crossvalidations (CV). Figure S3 shows, for each of the 1000 runs of the 10-fold CV, the average HAM-D for the subgroup who got their indicated treatment (lucky, shown in red) and the average HAM-D for the corresponding subgroup who got their non-indicated treatment (unlucky, shown in blue). What is powerful about this result is that there wasn't a single run in which the average score for the lucky group overlapped with the average score for the unlucky group in any of the other runs. This means that even the "worst" set of predictions generated across the 1000 runs resulted in a better outcome for the lucky than the unlucky.

Supplemental Figure 3. Mean HAM-D for each of 1000 ten-fold CVs.



Supplemental Figure 4. Mean HAM-D for the largest 60% PAIs for each of 1000 ten-fold CVs

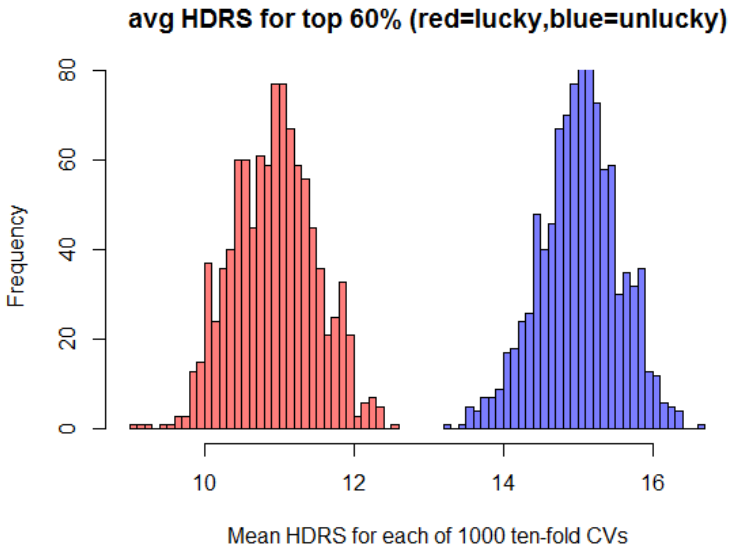


Figure S4 illustrates the above comparison for the largest 60% of PAIs. The same pattern can be observed, with an even larger separation between the two distributions.

CHAPTER 3: Improving Treatment Decisions for Patients with PTSD: A demonstration of model-based treatment selection using the Personalized Advantage Index approach

This work is in preparation as:

Cohen, Z. D., Wiltsey Stirman, S., DeRubeis, R. J., Keefe, J. R., Wiley, J. F., Smith, B. N., & Resick, P. A. (*in prep*). Improving Treatment Decisions for Patients with PTSD: A demonstration of model-based treatment selection using the Personalized Advantage Index approach.

Abstract

Objective: Individuals seeking treatment for Post-Traumatic Stress Disorder (PTSD) choose between numerous evidence-based treatments, including Prolonged Exposure (PE), and Cognitive Processing Therapy (CPT). As these treatments are, on average, equally effective, the decision of which treatment to pursue is complex. A new treatment selection method that uses predictors of differential treatment response could be used to inform this decision process and improve outcomes.

Method: The Personalized Advantage Index (PAI) treatment selection approach was applied to data (N=159) from a randomized comparison of two treatments (CPT vs. PE) for female rape-trauma PTSD. Data-driven variable selection was used to create linear models predicting end-of-treatment PTSD Symptom Scale (end-PSS) scores. Using each patient's predicted outcomes in CPT and PE, the indicated treatment was identified. Outcomes for patients who received their indicated versus non-indicated treatment were compared.

Results: The final model included five moderators: dissociation, childhood sexual abuse, trait angry-temperament, number of separate crime occasions, and daytime sleep-dysfunction. The average end-PSS for individuals receiving their PAI-indicated treatment was 13.68, which is below diagnostic threshold for PTSD, and 18.56 for those receiving their non-indicated treatment, which is above threshold. This corresponds to an advantage of 4.88 points on the PSS (Cohen's *d*=0.42) for receiving the PAI-indicated treatment. Among the 48.8% of the sample whose PAIs exceeded the reliable change index threshold, the mean end-PSS was 12.57 for got-indicated and 19.89 for got-non-indicated. This advantage of 7.31 (Cohen's *d*=0.64) exceeded the reliable change threshold.

Conclusion: Formal decision support tools drawing on information from moderators could improve treatment outcomes. Treatment selection approaches are in developmental stages. If validated in external samples or prospective trials, these approaches could provide clinicians with clear, actionable recommendations for individual patients.

Keywords: PTSD, precision medicine, prolonged exposure, cognitive processing therapy, treatment selection

Significance Statement

Two front-line treatments for PTSD, Prolonged Exposure and Cognitive Processing Therapy appear to be equally effective, but not all patients respond. Clinicians and researchers seek to minimize the likelihood of poor response, but have struggled to understand whether some patients might be better suited for one treatment over another. This paper demonstrates an approach for identifying optimal evidence-based treatments

for individuals with PTSD. It improves upon the consideration of individual predictors of treatment outcome by combining the predictors and estimating the end-of-treatment PTSD symptom scores for each potential treatment.

Introduction

Post-traumatic stress disorder (PTSD) is a chronic, debilitating condition with a lifetime prevalence of 8% (Steinert et al., 2015). Although evidence suggests that patients who respond to psychological interventions for PTSD are likely to sustain their positive response (Resick, Williams, Suvak, Monson, & Gradus, 2012), rates of spontaneous remission are low, and those who drop out of treatment before responding are unlikely to improve on their own (Bradley et al., 2005; Foa et al., 2005; Morina et al., 2014; Perkonigg et al., 2005; Resick et al., 2012; Steinert et al., 2015). Furthermore, a negative experience with treatment could lower the likelihood that patients will seek or fully engage in additional treatment.

Individuals seeking treatment for post-traumatic stress disorder (PTSD) are often confronted with a choice between multiple treatment options (Lancaster et al., 2016; Watts et al., 2013), including several empirically-supported treatments (ESTs) that are now included in treatment guidelines for PTSD (e.g., American Psychological Association, 2017). Unfortunately, little systematic, practical guidance is available to identify the intervention that is most likely to provide the greatest benefit for a given patient (Cohen & DeRubeis, 2018). Prolonged Exposure (PE; (Foa et al., 2007) and Cognitive Processing Therapy (CPT; (Resick & Schnicke, 1993), two trauma focused evidence-based psychotherapies (TF-EBPs) with substantial support for their efficacy and effectiveness for the treatment of PTSD (Bisson & Andrew, 2007; Watts et al., 2013), have been included in treatment guidelines. Both PE and CPT have been shown to be superior to placebos and to other active controls but evidence suggests that they are, on average, equally effective (Bisson & Andrew, 2007; Watts et al., 2013). While their

respective treatment manuals specify some theorized, general contraindications for TF-EBPs, they provide no guidance as to how to choose between TF-EBPs.

However, treatment allocation in PTSD is not random as clinicians make efforts to determine which treatment is most appropriate for their individual patients. Rosen and colleagues (2017) examined over 6,000 patients in VA PTSD teams, 23% of whom initiated CPT and 14% of whom initiated PE, and identified several patient factors that were associated with differential likelihood of receiving TF-EBPs, including factors such as recent hospitalizations and comorbidities. Other studies have shown similar patterns of findings (Mott et al., 2014; Sripada et al., 2017; Sripada et al., *under review*). Qualitative studies from the VA indicate that clinicians believe that certain patients are "not ready" for TF-EBPs due to factors such as psychiatric instability, complicating comorbidities (substance use, personality disorders), and lack of motivation (Cook et al., 2014; Rosen et al., 2016). Whether and why they judge the appropriateness of one TF-EBP over another are questions that fewer studies have investigated.

Raza and Holohan (2015) surveyed clinicians who were VA-trained in both PE and CPT regarding patient factors that they believed might influence them to recommend PE, CPT, either, or neither. They found that clinicians were more likely to recommend PE over CPT for patients with low literacy, low cognitive functioning, and moderate/severe Traumatic Brain Injury, whereas they were more likely to recommend CPT over PE for patients with strong guilt, strong shame, acts of perpetration, and dissociation history. Cook and colleagues (2017) also sought to understand how VA therapists determine which TF-EBP to provide to their patients. Therapists indicated that

they factored cognitive abilities, distress tolerance, and emotional stability into their decisions about which treatment would be most appropriate, but had difficulty operationalizing specific rules for making that determination. Furthermore, findings from qualitative studies have suggested that clinicians' ways of making determinations did not appear to be consistent with guidance in TF-EBP treatment manuals about contraindications for treatment (Cook et al., 2017; Osei-Bonsu et al., 2017). Thus, while treatment selection is occurring in routine practice, it does not appear to be systematic or based on research findings, and does not clearly follow guidance from treatment manuals. Instead, clinicians are left to make their best guesses based on sometimes limited clinical experience (in the case of clinicians who are newly trained in the EBPs), and findings from a confusing literature on individual moderators of treatment outcome (Cohen & DeRubeis, 2018).

Researchers have investigated patient factors that may predict differential outcomes in TF-EBPs in the context of clinical trials of treatments for PTSD. For example, anger has been shown to relate to poorer prognosis in some studies (Foa et al., 2005; Pitman et al., 1991), although after controlling for pretreatment PTSD severity, a subsequent study did not replicate this finding (Cahill et al., 2003; Cahill et al., 2004). In the sample of women treated with PE or CPT that will be used in this study, Rizvi and colleagues (2009) investigated a restricted set of variables and found several prognostic indicators: younger age, lower intelligence, and less education were associated with higher treatment dropout, whereas higher depression and guilt at pretreatment were associated with greater PTSD symptom improvement. In addition, severity was found to be a prognostic predictor, with greater severity predicting greater symptom change. They

also identified two prescriptive factors, age and anger: older women in PE and younger women in CPT had the best overall outcomes, and women with higher baseline levels of anger were more likely to drop out of PE. When these factors were examined in combination along with additional measures from the study, childhood physical abuse, current relationship conflict, trait anger, and racial minority status were associated with a higher likelihood of dropout in PE than CPT (Keefe et al., 2018).

As Cloitre (2011) observed, despite there being over 20 studies examining patient-specific characteristics of treatment outcome, results between studies have been inconsistent and have left clinicians with little useful guidance regarding how to determine the best fitting treatment for their patients. Some of the reasons for the inconsistencies are likely due to studies that were not powered to reliably detect treatment moderators, as well as heterogeneity across the treatment samples (Cloitre et al., 2015). As noted by Cloitre and colleagues (Cloitre et al., 2015), and discussed by Cohen and DeRubeis (2018) in their recent review on treatment selection, this inconsistency likely has several sources. Some inconsistencies may suggest that these factors are not good candidates for supporting treatment selection (e.g., they are associated with weak or "noisy" effects in statistical models), while others might simply require a standardization in methodological approaches (e.g., poor comparability due to inconsistent modeling decisions). Unfortunately, information on differential patient response from randomized studies is rarely integrated and applied clinically (Simon & Perlis, 2010). This may be due, in part, to a lack of a systematic method for translating this information into meaningful/actionable treatment recommendations.

One aim of precision medicine in mental health is to understand the heterogeneity of responses hidden within the average-treatment effect in order to help match patients to their optimal treatment (Cohen & DeRubeis, 2018). For example, Cloitre (2015) suggested that it may be possible to identify a profile comprising moderators that capture key patient-level historical factors, history, diagnoses, and behaviors or circumstances that, taken together, can predict outcome. In this vein, DeRubeis et al. (2014) described a method that integrates pre-treatment information as well as outcome data from patients who have been randomized to one of two (or more) treatments. Since its initial introduction, the approach has been used and adapted by several groups (Cohen et al., *under review;* Deisenhofer et al., 2018; Huibers et al., 2015; Keefe et al., 2018; Vittengl et al., 2017; Webb et al., 2018; Zilcha-Mano et al., 2016). In the simplest version, a multivariable model is constructed that yields, for each patient, a Personalized Advantage Index (PAI). The PAI is an index that is computed by taking the difference between the estimated end-of-treatment PTSD symptom score for two or more treatments, as predicted by a model that combines multiple predictive and prognostic indicators of treatment outcomes. The sign and magnitude of the PAI, derived from these estimates of the expected outcomes for the patient in each treatment, is used to indicate the preferred treatment. If the PAI indicates that a meaningful difference is expected for a given patient, it could be used to guide clinicians and patients toward the selection of an optimal treatment. In what follows, we will provide an example of how model-based treatment recommendations could improve outcomes in the treatment of PTSD.

Recent reviews have argued that in the field of mental health is that it is unlikely that any single variable, in isolation, will have large clinical utility in guiding mental

health treatment selection (Cohen & DeRubeis, 2018; Gillan & Whelan, 2017; Kessler, 2018; Simon & Perlis, 2010). How useful any single variable would be for helping inform an actual treatment recommendation would depend on a variety of factors that have been discussed elsewhere (e.g., (Janes et al., 2011). Several recent empirical efforts have demonstrated that multivariable prediction models can improve upon (and outperform) clinical prediction (e.g., (Kautzky, Baldinger-Melich, et al., 2017; Kautzky, Dold, et al., 2017). Thus, although single-variable models are explored below for illustrative purposes, the primary focus for this paper is on recommendations that were generated by a multivariable model, whose predictors were selected through a data-driven variable selection process (Cohen et al., *under review*).

Methods

Participants. Study participants were women who met DSM-IV criteria for PTSD assessed using the Clinician Administered PTSD Scale (CAPS; Weathers et al., 1999), a standardized, reliable trauma interview. Participants were randomized to PE, CPT, or a six-week waitlist condition, after which these patients were also randomly assigned to either CPT or PE. Here, we combine the participants from the waitlist with patients from the treatment condition to which they were randomized. Further details on trial methodology and patient sample can be found in the primary outcome publication (Resick et al., 2002). Inclusion criteria included having experienced a completed rape in childhood or adulthood, being at least 3 months post-trauma, and, if on medication, being on a stabilized dose (N=48; 30.1%) by client self-report. 79.9% of the sample reported multiple trauma victimizations (mean=6.22, *sd*=4.75). Exclusion criteria included current psychosis, substance dependence, illiteracy, instability of psychiatric medication dosages,

current self-injurious behavior, suicidal intent, and ongoing trauma (stalking or abusive relationship). Although 171 patients entered the study, the 11 patients who dropped out during the waitlist period (prior to being informed of their assignment to treatment) were excluded from these analyses, as their data could not inform a model on differential response to treatment. One additional patient for whom PTSD symptom data were not available at baseline, during, or post-treatment was also excluded. Demographics for the final sample (N=159) are presented in Table 1. (Descriptive statistics and group difference tests for the full set of baseline measures can be found in Supplemental Table S1.)

Table 1: Demographic and clinical characteristics of patient sample

	Cognitive Processing Therapy (N = 78) mean (*sd*)	Prolonged Exposure (N = 81) mean (*sd*)
Age	31.21 (9.53)	32.28 (9.74)
Race (% White)	57 (73.1%)	57 (70.4%)
Years of Education	14.60 (2.02)	14.18 (2.33)
IQ (Quick Test)	98.21 (8.54)	98.52 (9.90)
Years Since Index Rape	8.34 (8.81)	8.38 (7.90)
CAPS	74.91 (18.30)	74.64 (19.23)
PSS	29.54 (8.5)	29.26 (8.84)
BDI	23.38 (10.24)	23.38 (8.34)

Table 1. No significant differences between treatment conditions were found for any variables at baseline. BDI = Beck Depression Inventory; CAPS = Clinician Administered PTSD Scale; PSS = PTSD Symptom Scale.

Procedure. ***Cognitive Processing Therapy***. CPT is a primarily cognitive therapy delivered over 12 one-hour sessions. The original manual, used in this study (Resick &

Schnicke, 1993), includes psychoeducation, a statement of the impact that the trauma has had on the patient's life and beliefs, differentiation between thoughts and emotions, two assigned written accounts of the traumatic event that are reviewed in the subsequent session and then read daily between sessions, and cognitive restructuring of beliefs about the meaning of the trauma and its implications. The second half of the treatment includes modules that focus on disruptions in beliefs about safety, trust, power/control, esteem, and intimacy that may have resulted from the traumatic exposure.

Prolonged Exposure. PE is an exposure-based treatment (Foa & Rothbaum, 1998) that is based upon Emotional Processing Theory, which suggests that PTSD symptomatology is maintained by avoidance of trauma cues, and by negative cognitions about the self, the world, and one's reaction to the trauma. The PE protocol used in this study was nine 90-minute sessions and included psychoeducation and explanation of rationale for PE, breathing retraining, behavioral exposures, and imaginal exposures (Foa et al., 2007). The majority of the sessions involve imaginal exposure of the index event for 45–60 minutes of the session.

Measures. The primary outcome for the trial was the CAPS. The CAPS was only measured at baseline and post-treatment, and there was significant missingness for the post-treatment assessments due to both assessment and treatment drop-out. To have more assessment points for symptom change and minimize the impact of missing CAPS data, we decided to use the self-reported PTSD Symptom Scale (PSS; (Foa et al., 1993) as the outcome, which was assessed at baseline and at weeks 2, 4, 6, 8, 9, 10, and 12 (post-treatment). In light of findings that PTSD patients who drop out of treatment prior to experiencing symptom response are not expected to recover, we decided to use last-

observation-carried-forward (LOCF), a conservative imputation strategy, to model missing outcomes. The use of LOCF symptom scores as outcomes when examining predictors of response to PTSD treatments has been discussed and demonstrated in other prior efforts (Hagenaars et al., 2010). The mean post-treatment PSS score on the LOCF PSS (referred to hereafter as end-PSS) was 14.40 (sd=10.79) for the CPT group and 17.94 (sd=12.67) for the PE group. Diagnostic cutoff scores of 14 (Coffey et al., 2006) and 15 (Wohlfarth et al., 2003) on the PSS have been proposed for PTSD. Mean end-PSS were 3.54 points lower in CPT than PE (pooled sd=11.79; Cohen's d=0.30; t-statistic[10]=1.90; *p*=.059). This difference mirrors the findings reported in the primary report of the full intent-to-treat sample[11], in which post-treatment PSS and CAPS scores were non-significantly lower in CPT than PE (PSS of 13.66 vs. 17.99 and CAPS of 39.08 vs. 44.89, respectively (Resick et al., 2002).

Detailed explanations of all baseline measures are provided in the supplemental methods section, along with descriptive statistics and tests for group differences, which are presented in supplemental Table 1.

Data Analysis. Data preprocessing and missing data are described in the supplemental methods. Missing baseline data were imputed using a single-dataset

[10] Unless otherwise noted, all t-statistics and associated *p*-values are from 2-tailed Welch's two sample t-tests, equal variance not assumed.

[11] Note that the means reported by Resick et al. (2002) from the original ITT sample analyzed separately for the three initially randomized conditions (CPT, PE, and waitlist). Thus, means presented here for CPT and PE were each derived from samples of N=62.

random forest imputation strategy with the *missForest* package in R (Stekhoven & Buhlmann, 2012). More details regarding imputation are provided in the supplemental.

Variable Selection. To leverage the benefits of data-driven variable selection procedures, we ran the treatment selection analyses based on a model in which the variables were selected using a multi-method approach recently introduced by Cohen and colleagues (Cohen et al., *under review*). This approach, which has since been used or adapted in other efforts (Schweizer et al., *submitted;* Webb et al., 2018; Wiltsey Stirman et al., *submitted*), was developed in response to the observed heterogeneity of variable selection approaches used in precision medicine (Cohen & DeRubeis, 2018). It combines four commonly used variable selection approaches in order to identify predictors that are selected consistently across multiple approaches: Random Forests using the *mobForest* package (Garge et al., 2013) in R, Elastic Net Regularization using the *glmnet* package (Friedman et al., 2009), Bayesian Additive Regression Trees using the *bartMachine* package (Bleich et al., 2014; Kapelner & Bleich, 2016), and stepwise AIC-penalized bootstrapped variable selection (Austin & Tu, 2004), using the *bootStepAIC* package (Rizopoulos, 2009). All of the variable selection techniques that comprise the approach utilize cross-validation and/or bootstrapping, thus increasing the stability and generalizability of the identified variables (Hastie et al., 2009). In Step-1, all variables are submitted to each of three variable selection approaches: Random Forest, Elastic Net, and BART. Then, the variables that are selected by at least 2 out of 3 approaches are submitted to the final variable selection approach, bootStepAIC. Following Austin and Tu's (2004) recommendation, variables whose interactions with treatment are retained in at least 60% of 10,000 bootstrapped replicates are included in the final model. The multi-

method approach harnesses benefits of machine learning while preserving the interpretability of classic parametric regression. See the supplemental methods and Cohen et al. (*under review*) for more details on the variable selection approach.

Generation of PAI scores and Evaluation of the Results. The treatment selection model was defined using a linear regression model with the following general form (Tx=treatment, factors=baseline variables selected as potential moderators. The outcome variable, "Y", was post-treatment score on the PSS and the covariate was the pre-treatment score on the PSS):

Y = covariate + Tx + factors + Tx * (factors)

For each patient, predictions are generated describing their expected outcomes in CPT or PE. The prediction that corresponds to the treatment any given patient actually received is called the factual prediction, and the prediction of how he or she would have done in the other treatment is called the counterfactual prediction. The PAI for each patient is the difference in the predicted outcomes between treatment-A (CPT) and treatment-B (PE). This index is signed, and the direction of the index indicates which treatment is indicated by the model to be preferred over the other.

Outcomes of individuals classified as "CPT-indicated" who received CPT were compared to outcomes for "CPT-indicated" individuals who instead received their non-indicated treatment (PE). We next performed the analogous comparison for those identified as "PE-indicated." Then, we collapsed these two comparisons to examine whether, on average, individuals who received their indicated treatment had superior outcomes compared to individuals who did not receive their indicated treatment. More

detailed discussions of the creation and evaluation of PAIs can be found in DeRubeis et al. (2014), Cohen and DeRubeis (2018) and Cohen et al. (*under review*).

10-fold cross-validation (CV). 10-fold CV was used when generating PAIs to protect against over-fitting during the model weight-setting process (Kuhn & Johnson, 2013). The data were split into 10 folds, 9/10 of the data were used to set a model's weight, and then this model generated predictions for the held out 1/10 of the sample. Thus, each patient's predictions were generated from a model in which the weights were set without her data. The 10-fold CV procedure was repeated multiple times (N=1000) to account for variability related to the selection of the 10 folds, and the findings presented below summarize results from the 1000 runs. As noted by Hastie et al. (2009), although this approach attempts to approximate the protection of a true hold out sample, it does not fully protect against the risk of double-dipping because of our utilization of the full sample during the variable selection phase. However, given the small sample size, performing a split-halves analysis or a "complete 10-fold CV" whereby variable selection would be performed within each of the training samples was deemed impractical. Thus, the findings presented below will require replication in an independent sample.

Clinically Significant PAIs. Generally, larger PAIs indicate stronger recommendations for the indicated treatment over the non-indicated treatment. We hypothesized that those patients for whom stronger PAIs were produced would be especially likely to show greater benefits of treatment selection. DeRubeis et al. (2014) divided the sample's PAIs using a cutoff that was suggested by clinical guidelines. Here, we subset larger PAIs based on an index of reliable change of 6.15 points on the PSS that has been reported in the literature (Foa et al., 2002; Larsen et al., 2016).

Results

Data-driven Variable Selection. Of the 57 variables under consideration, in Step-1 14 were selected as potential moderators, as well as one potential 3-way interaction that was indicated by BART (see Table 2). (An additional 13 variables were selected by only one approach; these were not submitted to Step-2, bootStepAIC). Based on the criterion of being retained by bootStepAIC in at least 60% of the 10,000 bootstrapped replicates, five interactions were selected for the final model: trait angry temperament (subscale of the State Trait Anger Expression Inventory [STAXI]; (Spielberger & Sydeman, 1994), number of separate crime occasions, childhood sexual abuse (assessed with the Sexual Abuse Exposure Questionnaire [SAEQ]; (Rowan et al., 1994), dissociation (subscale of the Trauma Symptom Inventory [TSI]; (Briere et al., 1995), and daytime sleep dysfunction (subscale of the Pittsburgh Sleep Quality Index [PSQI]; (Buysse et al., 1989). The main effect for baseline PSS was included in the final model as a covariate, as planned, although it was not selected by bootStepAIC.

Table 2. Variable selection	<------	Step 1	------>	Step 2	Result
Variable	BART	Elastic Net	Random Forest	Included in Bootstep AIC?	Selected by Bootstep AIC?
Baseline PSS	✔	✔	✔	Yes	No*
Depression (BDI)	✔	✔	✔	Yes	No
Trait Angry Temperament (STAXI)	✔	✔	✔	Yes	**Yes**
# Separate Crime Occasions	✔	✔	✔	Yes	**Yes**
Dissociative Experiences (DES)	✔	✔	✔	Yes	No
Parental Violence (AE-III-PP)		✔	✔	Yes	No

Childhood Sexual Abuse (SAEQ)		✔	✔	Yes	**Yes**
Intimacy (PBRS)		✔	✔	Yes	No
Dissociation (TSI)	✔	✔		Yes	**Yes**
Dysfunctional Sexual Behavior (TSI)	✔	✔		Yes	No
Intrusive Experiences (TSI)		✔	✔	Yes	No
Sleep Meds (PSQI)	✔	✔		Yes	No
Daytime Sleep Dysfunction (PSQI)	✔	✔		Yes	**Yes**
Race White	✔	✔		Yes	No
Sleep Meds (PSQI) x Baseline PSS	✔			Yes	No
IQ (Quick Test)		✔		1/3	
Global Guilt (TRGI)		✔		1/3	
Distress (TRGI)			✔	1/3	
Wrongdoing (TRGI)		✔		1/3	
Trait Anger (STAXI)	✔			1/3	
Anger Out (STAXI)		✔		1/3	
Anger Control (STAXI)			✔	1/3	
Esteem (PBRS)		✔		1/3	
Hopelessness (BHS)		✔		1/3	
Sleep Quality (PSQI)		✔		1/3	
Sleep Latency (PSQI)		✔		1/3	
Sleep Efficiency (PSQI)		✔		1/3	
Married		✔		1/3	

Table 2. Summary of variable selection results for all variables selected by at least one approach. Columns under "Step 1" present the results of the three different variable selection approaches based on Random Forest, Elastic Net Regularization, and BART.

The variables that were selected by at least two of these three approaches were then submitted, along with one three-way interaction identified by BART (Sleep Meds [PSQI] x Baseline PSS x Treatment), to a final variable selection stage with BootStepAIC. Bold text indicates the variables that were retained as interactions by BootStepAIC in at least 60% of the 10,000 bootstrapped replicates, and thus were included in the final model. *Although baseline PSS was not selected as a moderator, its main effect was included as a covariate to control for its relationship to outcome. AE-III-PP = Assessing Environments-III-Physical Punishment Scale; BART = Bayesian Additive Regression Trees; BDI = Beck Depression Inventory; BHS = Beck Hopelessness Scale; DES = Dissociative Experiences Scale; PBRS = Personal Beliefs and Reactions Scale; PSS = PTSD Symptom Scale; PSQI = Pittsburgh Sleep Quality Index; SAEQ = Sexual Abuse Exposure Questionnaire; STAXI = State Trait Anger Expression Inventory; TRGI = Trauma Related Guilt Inventory; TSI = Trauma Symptom Inventory.

Final Model. Weights for the models used to generate the PAIs were set using the 10-fold CV scheme described above. Table 3 presents the coefficients from a model where weights were set using the full sample. The interpretation of moderator effects based on the coefficients of regression models can be a complicated endeavor (Kraemer & Blasey, 2004). A recent review on treatment selection contains a more detailed discussion of how to interpret moderator relationships for treatment recommendations (Cohen & DeRubeis, 2018). To facilitate this interpretation, the moderator relationships included in the final model are presented visually in Supplemental Figure S1.

Table 3. Final linear model predicting end-PSS generated using the full sample.

coefficient	**estimate**	**std. error**	**t value**	***p* value**
intercept	16.11	0.82	19.67	<.001***
Treatment (Tx)	3.69	1.64	2.25	.026*
Baseline PSS	4.98	0.96	5.18	<.001***
Trait Angry Temperament (STAXI)	1.05	0.83	1.26	.210
Daytime Sleep Dysfunction (PSQI)	–2.11	0.88	–2.39	.018*
# Separate Crime Occasions	1.48	0.89	1.66	.099†
Dissociation (TSI)	–1.76	0.94	–1.87	.064
Childhood Sexual Abuse (SAEQ)	0.70	0.89	0.79	.434
Tx * Trait Angry Temperament (STAXI)	4.33	1.67	2.59	.011*
Tx * Daytime Sleep Dysfunction (PSQI)	3.83	1.68	2.28	.024*
Tx * # Separate Crime Occasions	4.62	1.80	2.56	.011*
Tx * Dissociation (TSI)	–2.99	1.81	–1.65	.100†
Tx * Childhood Sexual Abuse (SAEQ)	–3.64	1.78	–2.04	.043*

Table 3. Treatment was coded ±0.5 and all predictors were centered/standardized. † = $p <$ .10, * = $p < .05$, *** = $p < .001$. PSS = PTSD Symptom Scale; PSQI = Pittsburgh Sleep Quality Index; SAEQ = Sexual Abuse Exposure Questionnaire; STAXI = State Trait Anger Expression Inventory; TSI = Trauma Symptom Inventory.

Personalized Advantage Index (PAI) Scores.[12] The mean absolute value PAI in the full sample was 7.06, (*sd*=5.31). For CPT-indicated individuals (68.6% of the sample) the mean was 7.87 (*sd*=5.54), and for PE-indicated individuals (31.4%) it was 5.31 (*sd*=4.29).

Approximately half (48.8%) of the sample had PAIs that were large enough (absolute value greater than 6.15) to be considered reliable as describe previously. Breaking down this group by which treatment was indicated, 37.8% of the total sample were predicted to have a reliable advantage of CPT over PE, versus 11.0% for PE over CPT.

Estimated Utility of PAI. The benefit of treatment selection (see Figure 2) was estimated by comparing the observed outcomes of individuals who were randomized to their PAI-indicated treatment (mean end-PSS=13.68; *sd*=10.64) to those who were not (mean end-PSS=18.56; *sd*=12.53). This reflected an average advantage of 4.88 PSS points (Cohen's *d*=0.42; t-statistic=2.64; *p*=.015) for those receiving their indicated treatment. The difference in outcomes for those who received their indicated versus non-indicated treatment (See Figure 3) was larger for CPT-indicated individuals (mean=6.13, Cohen's *d*=0.53, t-statistic=2.73, *p*=.009) than for PE-indicated individuals (2.15, Cohen's *d*=0.19, t-statistic=0.65, *p*=.526). (See Table 4 for full results and all comparisons.)

[12] As noted in the methods section, these results summarize the findings from the 1000 10-fold CVs: unless otherwise indicated, all Ns, means, SDs, 95% CIs, Cohen's *d*s, t-statistics, and p-values have been averaged across 1000 runs.

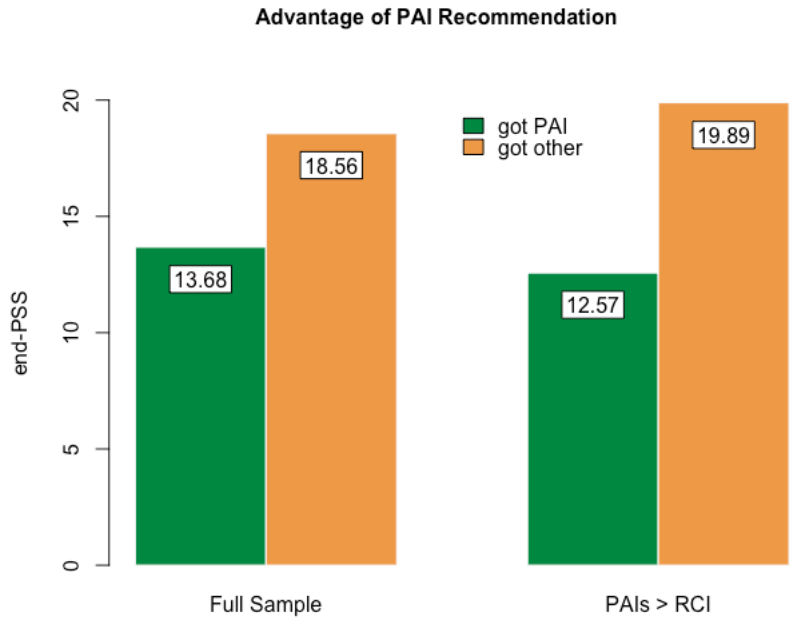


Figure 1. Comparison of end-PSS scores for patients who received their PAI-indicated ("got PAI") treatment versus those who received their non-indicated ("got other") treatment for the full sample (left bars) and for the subset of patients (right bars) with larger PAIs that exceeded the reliable change index (RCI).

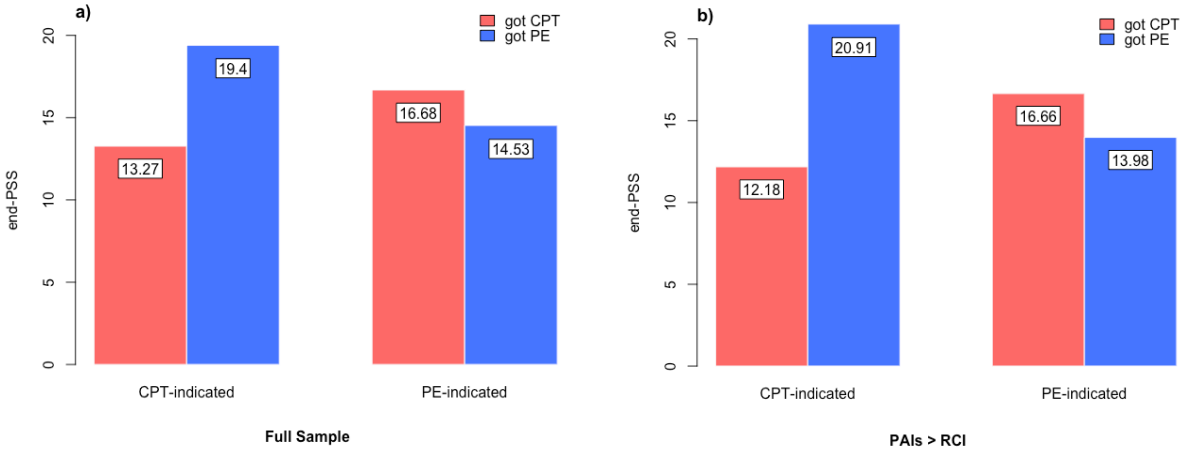


Figure 2. Panel a) Comparison of end-PSS scores for patients who received their PAI-indicated treatment versus those who received their non-indicated broken down by those who were CPT-indicated (left bars) versus PE-indicated (right bars). **Panel b)** The same

comparisons presented in Figure 3a performed in the subset of patients with larger PAIs that exceeded the reliable change index (RCI).

When this evaluation was restricted to the larger PAIs (48.8% of sample with PAIs above the reliability threshold), the observed effect of treatment selection (Figure 2) grew to 7.31 points (Cohen's *d*=0.64, t-statistic=2.80, *p*=.011), which in addition to being statistically significant, surpassed the reliable change threshold. Here, as in the full sample, there was a larger advantage for CPT-indicated individuals (mean difference=8.73, Cohen's *d*=0.75, t-statistic=2.90, *p*=.008) than for PE-indicated individuals (mean difference=2.68, Cohen's *d*=0.26, t-statistic=0.53, *p*=.592). The average end-PSS scores among those who received their PAI-indicated treatments were below the diagnostic cutoffs for probable PTSD in the collapsed sample, the CPT-indicated subgroup, and the PE-indicated subgroup, and above diagnostic threshold for those who received their non-indicated treatment.

Table 4. Observed mean end-PSS scores for patients who received their indicated or non-indicated treatment with group difference tests and effect sizes

PAIs	Random allocation	end-PSS mean (*sd*)	95% CI [lower, upper]	Difference 95% CI [lower, upper]	Cohen's d	t-statistic (*p*-value)
all PAIs	got indicated	13.68 (*10.64*)	[11.51, 15.84]	4.88 [1.22, 8.54]	0.42	2.64 (*p*=.015)
	got non-indicated	18.56 (*12.53*)	[16.25, 20.86]			
CPT-indicated	got indicated	13.27 (*10.64*)	[10.94, 15.61]	6.13 [1.67, 10.59]	0.53	2.73 (*p*=.009)
	got non-indicated	19.40 (*12.53*)	[16.83, 21.97]			
PE-indicated	got indicated	14.53 (*10.28*)	[11.53, 17.53]	2.15 [−4.44, 8.74]	0.19	0.65 (*p*=.526)
	got non-indicated	16.68 (*12.93*)	[13.71, 19.65]			
all PAIs > RCI	got indicated	12.57 (*10.35*)	[10.02, 15.13]	7.31 [2.10, 12.52]	0.64	2.80 (*p*=.011)
	got non-indicated	19.89 (*12.42*)	[17.15, 22.62]			
CPT-indicated > RCI	got indicated	12.18 (*10.37*)	[9.46, 14.90]	8.73 [2.69, 14.76]	0.75	2.90 (*p*=.008)
	got non-indicated	20.91 (*12.75*)	[17.94, 23.87]			
PE-indicated > RCI	got indicated	13.98 (*10.31*)	[10.23, 17.74]	2.68 [−8.6, 13.96]	0.26	0.53 (*p*=.592)
	got non-indicated	16.66 (*11.25*)	[12.93, 20.40]			

Discussion

Although many studies have identified prognostic predictors and moderators of treatment response in PTSD (Cloitre, 2015), they have resulted in little consistent guidance about how to select, from treatments that appear to be equally effective, which treatment is most likely to benefit an individual patient. Meehl (1954) was an early advocate for the superiority of what he described as "actuarial decision-making" over purely clinical decision-making. Clinicians with expertise and experience in a given field are generally skilled at selecting and coding the information that is needed for treatment selection, but often fail the difficult task of integrating information simultaneously from multiple (and sometimes conflicting) sources (Dawes, 1979; Dawes et al., 1989; Grove et al., 2000). In this paper we have described a method for generating treatment recommendations that simultaneously takes into account multiple factors, and yields an estimate of the outcomes that can be expected based on this combination of factors.

We employed a multi-stage variable selection approach that incorporated advanced statistical methods designed to identify robust predictors of treatment response. PAI-based treatment recommendations were generated for each patient in the sample using cross-validation. We found a significant advantage in outcomes for patients who received their model-indicated treatment relative to those who did not. The mean end-PSS of patients who received their indicated treatment was below diagnostic cutoffs for PTSD, whereas the mean end-PSS of those who received their non-indicated treatment exceeded them. These findings, if replicated in an external sample or validated in a prospective study, would suggest that outcomes for patients deciding between CPT or PE could be improved through treatment selection.

When we attended to the strength of the recommendations by looking within the roughly half of the sample with larger PAI recommendations, the statistically significant advantage of receiving the indicated treatment exceeded the reliable change index. This result, which replicates findings from other efforts regarding the importance of attending to the strength of recommendations (Cohen et al., *under review;* DeRubeis et al., 2014; Huibers et al., 2015; Keefe et al., 2018; Webb et al., 2018), likely reflects two features: 1) There are some patients for whom the two treatments are predicted to be more or less equally beneficial, and thus either treatment could be recommend. In these patients, very small PAIs might indicate a slight advantage of one treatment over the other, but this advantage might not be large enough to generate a meaningful effect of treatment selection; or 2) The predictions generated by these models contain noise, and for patients who have PAIs that are close to zero, the assignment of an "indicated treatment" might be too unstable to generate treatment recommendations that are reliable or useful.

In the full sample, as well as in the subset of individuals with larger PAIs, the advantage of receiving the model indicated treatment was larger among the CPT-indicated individuals, compared to those for whom PE was recommended. This was expected based on differences in the magnitudes of the PAIs, and might have been related to the (trend-level) superiority of CPT versus PE outcomes on the PSS in this sample.

The data-driven variable selection approach resulted in a final treatment selection model comprising five variables, all of which have been the focus of prior work on treatment response in PTSD: dissociation, childhood sexual abuse, trait angry temperament, number of separate crime occasions, and daytime sleep dysfunction.

Some of the relationships included in the final model replicated prior findings from the literature. For example, we found that higher anger (trait angry temperament subscale of the STAXI) was associated with worse outcomes in PE (relative both to low trait angry temperament in PE and to high angry temperament in CPT), which aligned with prognostic findings from the literature associating higher baseline anger with worse outcomes in PE (Foa et al., 1995; Pitman et al., 1991). Although Rizvi and colleagues' (2009) previous analyses of these data did not investigate trait angry temperament, they did investigate trait anger, for which they failed to find a significant main or moderating effect on treatment response. In their discussion of the variability of findings regarding the association between anger and outcomes in PE, Rizvi and colleagues (2009) note that different groups have used different measures of anger that were designed to capture different aspects of anger (e.g., state vs. trait, anger at self vs. others, anger at the index trauma vs. situationally reactive anger), as well as different outcome measures, and that this might account for some of the inconsistencies in the literature. Rizvi et al. (2009) did, however, report that higher trait anger was associated with higher dropout for PE than CPT. Unlike our analyses, in which LOCF was used to allow patients who dropped out during treatment to be retained in the sample, Rizvi et al. (2009) decided to restrict their analyses to a completers-only sample. Insofar as there is an association between anger and dropout, analyses of the relationship between anger and symptom response that are performed in a completers only sample risk systematically missing individuals with higher baseline anger. Additionally, because trait anger was a moderator of dropout in this sample, it is possible that Rizvi and colleagues' (2009) use of a completers-only

sample masked the moderating effect of baseline anger on response that was discovered in our analyses.

Dissociation (as measured by the TSI) was included as a moderator in our final model, in which there was a significant negative association between baseline dissociation and end-PSS for those in PE, and no relationship between baseline dissociation and outcome for those in CPT. Hagenaars et al. (2010) examined the relationship between three dissociative phenomena - dissociation (Dissociative Experiences Scale; DES), depersonalization (the mean score of the three dissociation items from the CAPS), and numbing (the mean score of the three numbing items from the PSS-avoidance subscale) - and PE treatment response and found no association between dissociation and response. They examined this regression in both a completers-only and an ITT sample (using LOCF to impute missing PSS scores) and obtained similar findings. However, Hagenaars and colleagues findings came from a different measure of dissociation, the dissociative experiences scale (DES). As a post-hoc exploratory analysis aimed to better approximate their approach, we constructed two regression models using the subset of the sample treated with PE (with and without controlling for baseline PSS) predicting end-PSS using DES scores. These analyses replicated their results, as we found that although higher levels of baseline dissociation (DES) were associated with worse outcomes in the model in which baseline PSS was not included as a covariate, no relationship between dissociation and outcome was found after controlling for baseline PSS. Taken together, these findings highlight the importance of attending to differences in analytic approaches (include variables, outcomes, samples, and statistical methodologies) when interpreting findings on predictors reported in the literature.

Many of the issues described above also complicated efforts to compare our results from the other four moderators with findings from the literature. For example, previous research in PE found that prior trauma in childhood was associated with poorer treatment outcomes (Hembree et al., 2004) We did not replicate this finding: in our final model, higher scores on a continuous variable representing levels of childhood sexual abuse (measured using the SAEQ) was associated with worse outcomes in CPT, but not PE. However, Hembree et al. (2004) used a different assessment tool (the Standardized Assault Interview) that assessed for a childhood history of not only sexual abuse, but also physical abuse, and witnessing extreme family violence, and they included this information as a binary variable that was yes if prior exposure to any of the three criteria was endorsed. In this study, childhood trauma that was non-sexual in nature was assessed with a different tool and included as a separate variable. Additionally, there were sample differences that might be especially important when comparing the relationship between history of childhood sexual assault and outcomes: only 68.5% of their sample had PTSD related to a sexual assault (rape or attempted rape), as compared to 100% of our sample.

These examples demonstrate one reason why we advocate using data-driven treatment selection models to inform treatment recommendations. A recent review by Cohen and DeRubeis (2018) provides a more in-depth discussion of the potential issues that can arise when trying to abstract clinical recommendations from the literature on predictors. Given these barriers, it is unsurprising that little progress has been made to date in the area of precision medicine approaches to mental health treatment selection. Approaches to model building and evaluation are still in the developmental stages. The variable selection approach used here is merely a proposed starting point, many details of

which should be tested and empirically derived and refined in future work. Future efforts could explore different permutations of the composite variable selection methods, such as adding different approaches (e.g., Support Vector Machines), reducing the number of approaches used, or adjusting relevant settings within each of the techniques. Examples of the latter include adjusting the thresholds for the inclusion of variables and specifying different tuning parameters (Kuhn & Johnson, 2013).

Despite these complications, and in part because of them, a growing body of literature describing data-driven approaches to treatment selection in depression (Chekroud et al., 2016; Delgadillo et al., 2017; DeRubeis et al., 2014; Iniesta et al., 2016; Kessler et al., 2017; Perlis, 2013; Petkova et al., 2017; Saunders et al., 2016; Smagula et al., 2016; Vittengl et al., 2017; Wallace et al., 2013; Webb et al., 2018; Zilcha-Mano et al., 2016), anxiety (Niles, Loerinc, et al., 2017; Niles, Wolitzky-Taylor, et al., 2017), psychosis (Koutsouleris et al., 2016), and PTSD (Cloitre et al., 2016; Deisenhofer et al., 2018; Keefe et al., 2018) has emerged, laying the foundation necessary to make research-informed treatment recommendations feasible.

The data-driven variable selection in this study utilized the same sample from which the training samples used for model estimation were drawn. This could lead to model overfitting and inflated relationships (Fiedler, 2011), and as noted by Hastie et al. (2009), represents a form of double-dipping that can increase risk of overconfidence. However, unlike the vast majority of moderator research in the psychiatric literature, we employed a multi-method, multi-step variable selection process that used bootstrapping and cross-validation, incorporating out-of-bag predictions and permutation tests to select variables that are more likely to generalize to a new sample. Moreover, to limit the risk of

bias in the model coefficients, the predictive utility of our model was estimated using repeated 10-fold cross-validation for weight setting. Nevertheless, in light of these concerns and our small sample size, the model and variables presented here should not be used to guide treatment decisions until they have been validated in an external sample. Even then, it is unclear (and perhaps unlikely; Nigatu et al., 2016) that this model would generalize to a population that is different from the one in which it was built (i.e., civilian females with rape-trauma PTSD). At this time, this is the only completed trial comparing CPT and PE for sexual trauma PTSD on which the present model could be tested. A large-scale trial comparing CPT and PE for PTSD in a veteran population (N = 900) may provide the sample size necessary to both develop and prospectively test a single treatment selection model, building off of the model developed in the present study (Schnurr et al., 2015).

Beyond concerns of reliability and generalizability, issues related to ethical and practical barriers to treatment selection will need to be addressed. As noted by Cloitre (2015), "The consideration of treatment matching to patient needs extends beyond symptom acuity and complexity… Treatment matching to patients also requires consideration of access to care and logistical barriers" (p. 500). However, building powerful, reliable models of treatment response is a necessary first step on the road to precision mental health. Once models being proposed in the literature are successfully validated in external samples or different populations, clinicians and patients can begin to use the clear information such approaches provide about expected outcomes in different treatments to improve the shared decision-making process by which they plan and initiate treatment.

Acknowledgments

Z.D.C and R.J.D were supported by the MQ Foundation Psy-IMPACT grant MQ14PM_27. The RCT was supported by Grant NIH-1 R01-MH51509 from the National Institute of Mental Health, awarded to P.A.R.

Supplemental Material: Methods

Participants

Descriptive statistics for the sample on the baseline variables and tests for group differences are presented in Supplemental Table S1.

Supplemental Table 1. Descriptive statistics of baseline variables

	Cognitive Processing Therapy (N = 79) mean (*sd*)	Prolonged Exposure (N = 81) mean (*sd*)	Continuous: Mean difference (95% CI) Categorical: X^2 (*df*)	*p* value	% missing
Age	31.21 (9.53)	32.28 (9.74)	–1.08 (–4.1, 1.94)	.481	0.0%
Race (% White)	57 (73.1%)	57 (70.4%)	0.04 (1)	.839	0.0%
Years of Education	14.60 (2.02)	14.18 (2.33)	0.42 (–0.26, 1.11)	.222	1.9%
IQ (Quick Test)	98.21 (8.54)	98.52 (9.90)	–0.31 (–3.21, 2.59)	.834	6.3%

Years Since Index Rape	8.34 (8.81)	8.38 (7.90)	–0.04 (–2.66, 2.58)	.978	1.3%
Total Sex Crime Exposures	2.38 (2.42)	2.26 (2.65)	0.12 (–0.68, 0.91)	.770	1.3%
Clinician Administered PTSD Scale (CAPS)[13]	74.91 (18.30)	74.64 (19.23)	0.27 (–5.62, 6.15)	.928	0.0%
PTSD Symptom Scale (PSS)	29.54 (8.5)	29.26 (8.84)	0.28 (–2.44, 3.00)	.840	0.0%
Beck Depression Inventory (BDI)	23.38 (10.24)	23.38 (8.34)	0.00 (–2.92, 2.92)	.998	1.9%
Beck Hopelessness Scale (BHS)	9.54 (5.45)	9.67 (5.39)	–0.12 (–1.82, 1.58)	.887	7.5%
Dissociative Experiences Scales (DES)	19.59 (13.06)	22.69 (15.00)	–3.10 (–7.51, 1.32)	.168	6.9%
Childhood Sexual Abuse (SAEQ)	1.11 (1.74)	1.38 (1.93)	–0.27 (–0.84, 0.31)	.360	0.6%
Childhood Physical Abuse (AE–III–PPS)	3.59 (1.98)	4.15 (2.49)	–0.56 (–1.27, 0.15)	.119	1.9%
Current Partner Conflict[14] (CTS)	11 (14.7%)	11 (14.1%)	0.00 (1)	1	3.8%

[13] Due to its overlap with the PSS, the CAPS was not included as a potential predictor

[14] Current Partner Conflict (yes/no) had too little variability to be included (10.8%) and thus was excluded from the variable selection process and missing values were not imputed.

Previous Partner Conflict (CTS)	1.79 (1.36)	1.50 (1.41)	0.29 (–0.15, 0.72)	.194	1.9%
MDD (current)	34 (43.6%)	31 (38.3%)	0.27 (1)	.603	0.6%
Panic Disorder[15] (current)	8 (10.4%)	9 (11.3%)	0.00 (1)	1	1.3%
Alcohol (lifetime)	34 (43.6%)	39 (48.1%)	0.17 (1)	.676	8.2%
Married	19 (24.4%)	20 (24.7%)	0.00 (1)	1	1.3%
Global Guilt (TRGI)	2.35 (1.09)	2.48 (1.11)	–0.13 (–0.48, 0.21)	.457	1.3%
Distress (TRGI)	3.19 (0.55)	3.20 (0.61)	–0.01 (–0.19, 0.17)	.912	1.9%
Guilt Cognitions (TRGI)	1.83 (0.84)	2.09 (0.88)	–0.26 (–0.53, 0.01)	.056†	3.1%
Hindsight Bias (TRGI)	1.88 (1.03)	2.14 (1.15)	–0.26 (–0.61, 0.08)	.131	3.8%
Wrongdoing (TRGI)	1.63 (0.97)	1.97 (0.92)	–0.34 (–0.63, –0.04)	.026*	5.7%
Lack of Justification (TRGI)	2.54 (1.03)	2.68 (0.92)	–0.14 (–0.45, 0.16)	.355	5.7%
State Anger (STAXI)	17.48 (7.33)	17.95 (7.63)	–0.46 (–2.81, 1.88)	.696	4.4%
Trait Anger (STAXI)	20.69 (5.86)	21.43 (5.53)	–0.74 (–2.53, 1.04)	.413	4.4%
Trait Angry Temperament (STAXI)	7.23 (2.80)	7.79 (2.68)	–0.55 (–1.41, 0.31)	.206	4.4%

[15] Panic disorder (yes/no) had too little variability to be included (14.4%) and thus was excluded from the variable selection process and missing values were not imputed.

Trait Angry Reaction (STAXI)	9.97 (3.08)	10.22 (2.84)	–0.26 (–1.18, 0.67)	.585	4.4%
Anger In (STAXI)	19.97 (4.63)	20.73 (4.14)	–0.75 (–2.13, 0.62)	.281	4.4%
Anger Out (STAXI)	15.33 (4.22)	15.63 (4.18)	–0.31 (–1.62, 1.01)	.646	4.4%
Anger Control (STAXI)	20.76 (5.70)	21.81 (5.19)	–1.05 (–2.75, 0.66)	.228	4.4%
Anger Expression (STAXI)	30.66 (10.28)	30.60 (8.95)	0.06 (–2.96, 3.08)	.969	4.4%
Number of Separate Crime Occasions	6.33 (4.89)	6.12 (4.65)	0.21 (–1.28, 1.71)	.780	4.4%
Negative Rape Belief (PBRS)	5.40 (0.68)	5.26 (0.85)	0.15 (–0.10, 0.39)	.236	10.7%
Undoing (PBRS)	2.56 (1.84)	2.13 (1.63)	0.44 (–0.11, 0.98)	.117	6.9%
Self–Blame (PBRS)	3.17 (1.52)	2.95 (1.71)	0.21 (–0.29, 0.72)	.407	10.7%
Safety (PBRS)	2.50 (1.46)	2.53 (1.31)	–0.03 (–0.47, 0.40)	.889	10.7%
Trust (PBRS)	2.73 (0.98)	2.77 (1.13)	–0.04 (–0.37, 0.30)	.828	10.7%
Competence and Power (PBRS)	3.12 (0.92)	3.01 (1.02)	0.11 (–0.19, 0.41)	.477	10.7%
Esteem (PBRS)	3.17 (0.83)	3.20 (0.86)	–0.02 (–0.29, 0.24)	.854	10.7%
Intimacy (PBRS)	3.01 (1.06)	2.83 (1.18)	0.17 (–0.18, 0.53)	.330	10.7%
Anger/Irritability (TSI)	13.17 (5.33)	13.74 (6.33)	–0.57 (–2.41, 1.27)	.541	9.4%
Anxious Arousal (TSI)	14.64 (4.58)	14.77 (4.86)	–0.13 (–1.61, 1.35)	.864	8.8%
Defensive Avoidance (TSI)	16.16 (4.69)	16.94 (4.57)	–0.79 (–2.24, 0.66)	.286	8.8%
Depression (TSI)	14.04 (5.72)	13.39 (5.31)	0.65 (–1.08, 2.38)	.459	8.8%
Dissociation (TSI)	12.07 (4.80)	12.35 (5.53)	–0.29 (–1.91, 1.34)	.727	9.4%
Dysfunctional Sexual Behavior (TSI)	1.52 (0.98)	1.53 (1.00)	–0.01 (–0.32, 0.30)	.966	4.4%
Impaired Self–Reference (TSI)	13.69 (5.94)	13.44 (5.69)	0.25 (–1.57, 2.07)	.787	10.1%
Intrusive Experiences (TSI)	13.84 (4.90)	14.31 (5.81)	–0.47 (–2.16, 1.22)	.581	5.0%
Sexual Concerns (TSI)	11.62 (6.64)	11.68 (6.86)	–0.05 (–2.17, 2.06)	.960	8.8%

Tension–Reduction Behavior (TSI)	5.87 (4.41)	6.17 (4.02)	–0.30 (–1.62, 1.02)	.656	4.4%
Subjective Sleep Quality (PSQI)	1.94 (0.84)	1.88 (0.80)	0.06 (–0.20, 0.32)	.646	0.6%
Sleep Latency (PSQI)	1.94 (1.05)	2.00 (0.92)	–0.07 (–0.38, 0.24)	.672	1.3%
Sleep Duration (PSQI)	1.90 (1.01)	1.74 (0.97)	0.16 (–0.15, 0.47)	.312	2.5%
Habitual Sleep Efficiency (PSQI)	1.24 (1.11)	1.23 (1.17)	0.01 (–0.35, 0.36)	.968	3.1%
Sleep Disturbance (PSQI)	1.72 (0.62)	1.94 (0.66)	–0.22 (–0.42, –0.02)	.034*	2.5%
Use of Sleeping Medication (PSQI)	21 (26.9%)	35 (43.2%)	3.93 (1)	.047*	0.6%
Daytime Sleep Dysfunction (PSQI)	1.70 (0.81)	1.75 (0.78)	–0.05 (–0.30, 0.20)	.687	1.9%
Global Sleep Score (PSQI)	10.76 (3.80)	11.39 (4.26)	–0.64 (–1.90, 0.63)	.321	3.8%

Supplemental Table S1. Baseline predictors. † = $p < .10$, * = $p < .05$, missing values imputed; AE-III-PP = Assessing Environments-III-Physical Punishment Scale; BART = Bayesian Additive Regression Trees; BDI = Beck Depression Inventory; BHS = Beck Hopelessness Scale; CAPS = Clinician Administered PTSD Scale; CTS = Conflict Tactic Scale; DES = Dissociative Experiences Scale; PBRS = Personal Beliefs and Reactions Scale; PSS = PTSD Symptom Scale; PSQI = Pittsburgh Sleep Quality Index; SAEQ = Sexual Abuse Exposure Questionnaire; STAXI = State Trait Anger Expression Inventory; TRGI = Trauma Related Guilt Inventory; TSI = Trauma Symptom Inventory.

Measures

The **PTSD Symptom Scale** (PSS) is a 17-item inventory scale that assesses PTSD symptom severity (Foa et al., 1993). The PSS has demonstrated high internal consistency and strong concurrent and convergent validity with other measures of psychological distress and PTSD symptomatology (Foa et al., 1993). The current study calculated the frequency score for each subscale: reexperiencing, avoidance, and arousal. The alpha coefficient from the current study was .84.

The **Beck Depression Inventory** (BDI) includes 21 self-report questions that assess depression symptomatology (Beck et al., 1996). The BDI allows respondents to evaluate their mental state in the past week based on a four descriptions that differ by severity level. Beck et al. (1988) found satisfactory internal consistency (mean Cronbach's alpha of .86), strong convergent and concurrent validity with other measures of depression. In the current study, the alpha coefficient was .92.

The 44-item **State Trait Anger Expression Inventory** (STAXI) was designed to measure expression, experience and control of anger (Spielberger & Sydeman, 1994). Respondent's scores from the STAXI range from 10-40, with higher scores signifying greater levels of anger. Spielberger and colleagues (1983) found good internal consistency as well as convergent validity for the questionnaire with similar anger measures. To assess participant anger, the current study used the trait anger (frequency of feeling angry), trait angry temperament (reacting with anger without provocation), trait angry reaction (frequency of feeling angry in frustrating situations), anger expression-in (frequency of feeling angry and expressing the anger physically or verbally), anger

expression-out (frequency of feeling angry without expression), and anger control (frequency of controlling expression of anger) subscales. The alpha coefficient from this sample was .86.

The 80-item **Conflict Tactic Scale-2** (CTS) quantifies the severity and frequency of conflict and violence within intimate, interpersonal relationships (Straus et al., 1996); items were added to assess sexual coercion. Respondents are asked to rate accuracy of certain statements across the following areas: negotiation (e.g., "My partner showed care for me even though we disagreed"), physical assault ("my partner chocked me"), sexual coercion ("My partner made my have sex without a condom"), injuries ("I went to a doctor because of a fight with my partner"), and psychological aggression ("My partner called me fat or ugly"). Among these domains, the current study employed the current and previous partner conflict subscales, which represent a composite of the frequency and severity of reported experiences. Past research has shown CTS to have reliability ranging from .79 to .95 and acceptable construct and convergent validity (Straus et al., 1996).

The **Personal Beliefs and Reactions Scale** (PBRS) includes a 55-item inventory that measures rape-victims' maladaptive beliefs about the traumatic event and internal characteristics (Resick et al., 1991; Mechanic & Resick, 1999). Resick and colleagues (1991) found strong test-retest reliability among rape victims ($\alpha = .81$). The current study included each of the eight subscales from the PBRS: rape beliefs, self-blame, undoing, safety, trust, competence and power, esteem, and intimacy.

The **Trauma Symptom Inventory** (TSI) includes 100-items on a four point Likert scale that assess a large range of symptoms related to trauma (Briere et al., 1995). The current study used each of the ten clinical scales: anger/irritability, anxious arousal,

defensive avoidance, depression, dissociation, dysfunctional sexual behavior, impaired self-reference, intrusive experiences, sexual concerns, and tension-reduction behavior. This self-report measure has been reported to have strong internal consistency (α = .74-.90) and good convergent validity with similar trauma measures (e.g., CAPS; (Briere et al., 1995; McDevitt-Murphy et al., 2005).

The **Pittsburgh Sleep Quality Index** (PSQI) is a 19-item inventory that measures individuals' sleep quality in the past month (Buysse et al., 1989). The PSQI has acceptable internal consistency ranging from .67 to 83 (Buysse et al., 1989; Cook et al., 2013) as well as strong validity (Carpenter & Andrykowski, 1998). The current study used the subjective sleep quality, sleep latency, sleep duration, habitual sleep efficiency, sleep disturbances, use of sleeping medication, daytime dysfunction, and the global PSQI score to assess sleep quality of participants.

The **Trauma Related Guilt Inventory** (TRGI) presents 32-items about guilt in relation to hindsight bias, distress, violation of personal standards, and lack of justification (Kubany et al., 1996). Past research revealed high internal consistency and strong correlation with other guilt-related measures (Kubany et al., 1996). The current study analyzed the three scales: distress, global guilt, and guilt cognitions along with the three subscales: hindsight bias, lack of justification, and wrongdoing. The scales and subscales had acceptable internal consistency in the current study ranging from .73 to .92.

The **Beck Hopelessness Scale** (BHS) includes 20 true-false questions that assess future beliefs and expectations as well as loss of motivation (Beck et al., 1974). The BHS has been found to have high internal consistency (α = .93) and good convergent validity with measures of hopelessness (Beck et al., 1974).

The **Dissociative Experiences Scale** (DES) includes 28 self-report items on a 11-point scale from "never" to "always" that assess the frequency (in percent) of experiences related to amnesia, depersonalization-derealization and absorption (Bernstein & Putnam, 1986; Carlson & Putnam, 1993). This measure (total range: 0–100) was designed to measure general dissociative disorders among individuals. Prior research on the DES has demonstrated strong reliability with the mean Cronbach's alpha of .93 (Van IJzendoorn & Schuengel, 1996). There is some evidence that the DES is more related to general dissociative tendencies than to state dissociation (Bremner et al., 1998).

IQ was measured using the **Quick Test** includes 50 items that measure general intelligence (Ammons & Ammons, 1962). The respondent does not need to read, write or speak; instead, the individual uses drawings to explain the meaning of different words. Past research indicates that the QT correlates well with other intelligence measures (i.e., WAIS; (Maloney et al., 1973).

The **Assessing Environment-III Scale** (AE-III) assesses childhood experience of punishment, family atmosphere, marital distress, and rejection from parents (Berger & Knutson, 1998; (Berger et al., 1988). Respondents are asked a series of true-false questions which range in severity levels. Prior research indicates high internal consistency ($\alpha = .85$; (Bluestone, 2005). The "Parent Violence" variable represents the total score from the AE-III parent violence scale.

The **Sexual Abuse Exposure Questionnaire** (SAEQ) assesses the nature (e.g., age of onset, duration, frequency) of sexual abuse experiences in 10 categories of increasingly invasive events ranging from "exposure of genital area" to intercourse

(Rowan et al., 1994). The overall exposure score, which is ranges from 0 to 10 based on the number of categories that are endorsed as having been experienced, has been found to have good test-retest reliability (.73 to .93). In these analyses "Childhood Sexual Abuse" variable represents the SAEQ total overall exposure score.

Data Pre-Processing and Missing Data Imputation

Variables with greater than 20% missingness or less than 20% membership in smallest category were excluded. The highest missingness for variables that remained was 10.7% (see supplemental Table S1). Categorical variables were made binary where appropriate. For example, the categorical variable describing marital status was made binary (married or cohabitating vs. other). Imputation of missing outcome and baseline data was performed using the missForest package in *R* (Stekhoven & Buhlmann, 2012), which implements a non-parametric random forest-based imputation strategy, generating a single imputed dataset by averaging over multiple regression trees. It has been found to outperform other methods of imputation, especially when complex and non-linear interactions are present, and can handle different types of variables (Shah et al., 2014; Waljee et al., 2013). To improve imputation, all available longitudinal symptom measures from the CAPS and PSS were included during the imputation process, after which all except baseline PSS were removed. The treatment variable was not included during imputation. To reduce risk of leverage points, boxplots were used to identify outliers, which were winsorized by setting their values to the closest non-outlier value. The follow variables had outliers: Age (one high), Years Education (one high), TRGI Distress (three low), STAXI Trait Anger (one high), STAXI Trait Angry Temperament (three high), STAXI Anger Out (three high), Total Sex Crime Exposures (nine high),

PBRS Negative Rape Beliefs (five low), TSI Defensive Avoidance (one low), TSI Dissociation (six high), TSI Tension-Reduction Behaviors (three high), and Dissociative Experiences Scale (three high). Due to skewed distributions, one continuous variable (PSQI Sleep Meds) was made binary (yes/no) and two were log transformed to achieve normal distributions (TSI Dysfunctional Sexual Behavior and Previous Partner Conflict from the CTS).

Variable Selection

Following the approach introduced by Cohen and colleagues (*under review*), all potential baseline predictors (see Table S1) were entered simultaneously into each of the first three approaches: Random Forest (RF), Elastic Net Regularization (ENR), and Bayesian Additive Regression Trees (BART). The variables that were consistently identified as having important interactions with treatment in at least 2 of the 3 methods were then subjected to the fourth and final variable selection approach: stepwise AIC-penalized bootstrapped variable selection (bootStepAIC). The variables that were selected by bootStepAIC comprised those that were used in the final model. Additional details on each method are provided below.

Random Forest (RF). RF is a non-parametric recursive partitioning approach to modeling that can accommodate large numbers of predictor variables as well as complex relationships (e.g., higher order interactions and non-linear associations). To perform RF variable selection, we used the *mobForest* package (Garge et al., 2013) in R, which can be made to focus its search on moderators (as opposed to main effects or prognostic variables). The mtry criteria, which determines how many variables random forest evaluates at each node for a single tree, was set to 19 (# predictors divided by 3). The number of trees was set to 10,000 to stabilize the results. Variables that surpass a permutation-based importance threshold were selected.

Elastic Net Regularization (ENR). ENR combines the Ridge (L1 penalization) and Lasso (L2 penalization) approaches, allowing for the selection of a parsimonious set of variables related to outcome (Hastie et al., 2009). ENR can accommodate large numbers of variables and is robust to high predictor covariance (Friedman et al., 2010; Zou & Hastie, 2005). ENR was performed using the *glmnet* package (Friedman et al., 2009). The *glmnet* alpha parameter was set to 0.5. The *glmnet* package is unable to accommodate variable selection in the context of moderators. Thus, we took the approach of building two prognostic models: the sample was split into two groups, one for each treatment, and variable selection was performed separately for each sample. To stabilize our results, we ran ENR 5 times (for each group), and the variables that were retained consistently throughout all 5 runs were considered. Variables that were selected in only one condition, or that were selected in both but specified with differing coefficients, were selected as potential moderators.

Bayesian Additive Regression Trees (BART). Variable selection with BART was performed using the *bartMachine* package (Kapelner & Bleich, 2016). BART, which builds on ensemble-of-tree methods (e.g., RF) by incorporating an underlying Bayesian probability model, has been adapted to extract information about variable importance (Bleich et al., 2014; Goldstein et al., 2015). The variable selection routine can be focused on identifying moderators by forcing the variable splitting search to focus more (here we set this parameter to 10-times more) on the treatment variable than other variables, thus introducing more interactions between treatment and other variables. This can be thought of similar to what researchers do when they only consider interactions between treatment group and baseline variables (and not interactions between baseline variables themselves). The N most important interactions identified by BART were retained, where N was determined based on the number of variables selected by RF based on its permutation test importance threshold cutoff. To account for variability in internal model structure, BART was run five times, and only variables that were among the N most important for all five runs were retained.

Variables selected by at least two of the three approaches were submitted to **Stepwise AIC-penalized bootstrapped variable selection** (Austin & Tu, 2004), which was performed using the *bootStepAIC* package in R (Rizopoulos, 2009). *bootStepAIC* was chosen as the final approach due to our use of linear regression for our final model, which made it essential that all included variables function in the context of a linear regression. A moderator selected by one of our two ensemble-of-tree-based approaches RF or BART could lead to a model with poor fit, if, for example, that variable relied on an unspecified non-linear relationships or higher order interaction. Within each of 10,000

bootstrapped training samples, backwards elimination was performed. The bootstrapped replicates were examined for internal consistency (Austin & Tu, 2004), and moderators that were retained in at least 60% of the bootstrapped samples were included in the final regression model.

Supplemental Material: Results

Supplemental Figure 1. Visualization of moderator relationships.

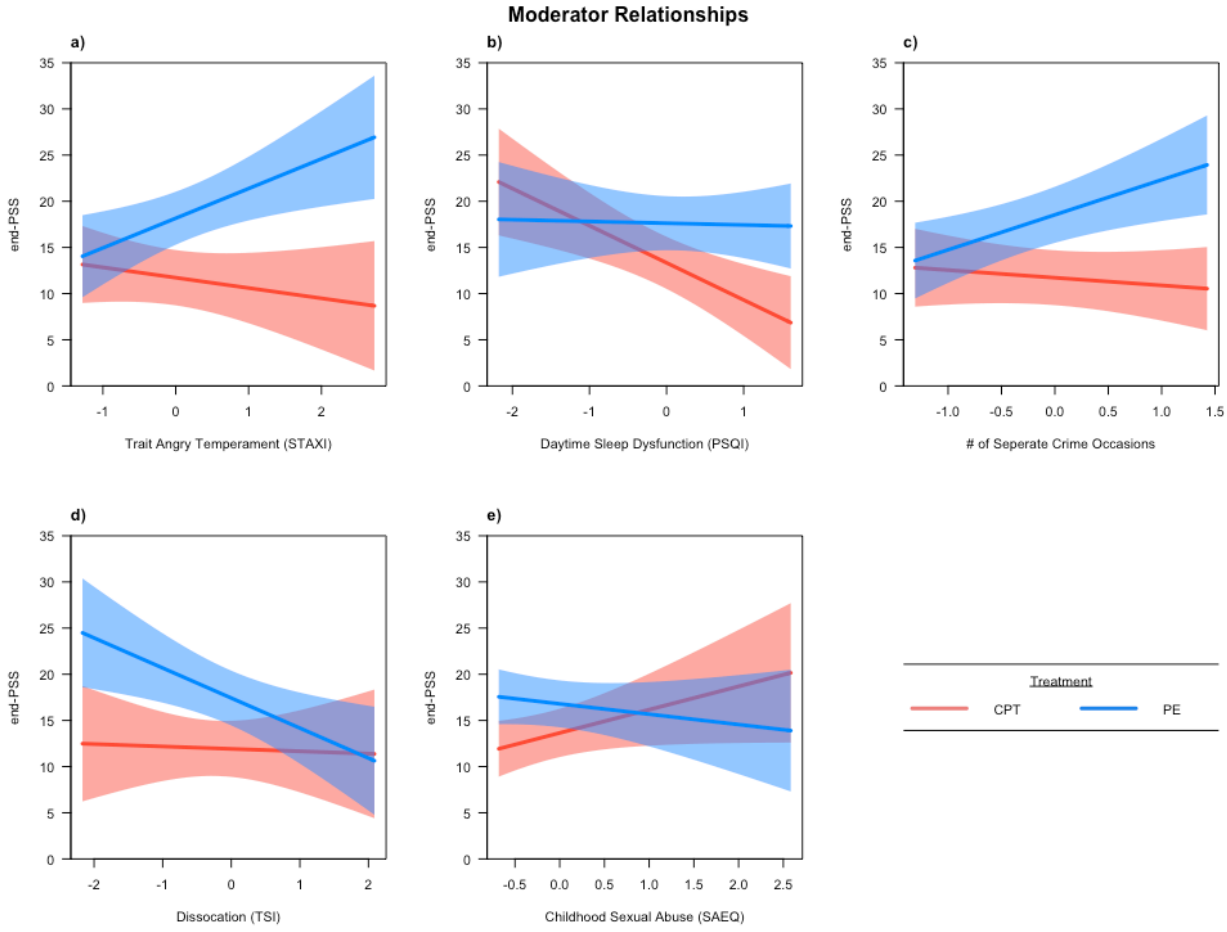


Figure S1. Moderator relationships from the final model visualized using the R package *visreg* (Breheny & Burchett, 2013). Conditional plots with confidence bands for the conditional mean from the final model estimated in the complete sample. Conditioning for each plotted variable uses the mean value for all other variables. The Y-axis represents the predicted end-of-treatment score on the PSS, and the X-axis represents the standardized/centered score for each variable that was used during analysis. PSS = PTSD Symptom Scale; STAXI = State Trait Anger Expression Inventory; TSI = Trauma Symptom Inventory; PSQI = Pittsburgh Sleep Quality Index; SAEQ = Sexual Abuse Exposure Questionnaire.

BIBLIOGRAPHY

Chapter 1 References

Allen, J., Mattson, M., Miller, W., Tonigan, J., Connors, G., Rychtarik, R., . . . Litt, M. (1997). Matching alcoholism treatments to client heterogeneity. *Journal of studies on alcohol, 58*(1), 7-29.

American Psychiatric Association. (2010). *Practice Guideline for the Treatment of Patients with Major Depressive Disorder*. Retrieved from Arlington, VA:

Amsterdam, J. D., Lorenzo-Luaces, L., & DeRubeis, R. J. (2016). Step-wise loss of antidepressant effectiveness with repeated antidepressant trials in bipolar II depression. *Bipolar Disorders, 18*(7), 563-570. doi:10.1111/bdi.12442

Amsterdam, J. D., & Shults, J. (2009). Does tachyphylaxis occur after repeated antidepressant exposure in patients with Bipolar II major depressive episode? *Journal of affective disorders, 115*(1-2), 234-240. doi:10.1016/j.jad.2008.07.007

Amsterdam, J. D., Wang, C.-H., Shwarz, M., & Shults, J. (2009). Venlafaxine versus lithium monotherapy of rapid and non-rapid cycling patients with bipolar II major depressive episode: a randomized, parallel group, open-label trial. *Journal of affective disorders, 112*(1), 219-230.

Ashar, Y. K., Chang, L. J., & Wager, T. D. (2017). Brain Mechanisms of the Placebo Effect: An Affective Appraisal Account. *Annual Review of Clinical Psychology, 13*, 73-98.

Austin, P. C., & Tu, J. V. (2004). Bootstrap methods for developing predictive models. *The American Statistician, 58*(2), 131-137.

Bagby, R. M., Ryder, A. G., & Cristi, C. (2002). Psychosocial and clinical predictorsof response to pharmacotherapy for depression. *Journal of Psychiatry and Neuroscience, 27*(4), 250.

Barber, J. P., & Muenz, L. R. (1996). The role of avoidance and obsessiveness in matching patients to cognitive and interpersonal psychotherapy: Empirical findings from the Treatment for Depression Collaborative Research Program. *Journal of Consulting and Clinical Psychology, 64*(5), 951.

Barbui, C., Cipriani, A., Patel, V., Ayuso-Mateos, J. L., & van Ommeren, M. (2011). Efficacy of antidepressants and benzodiazepines in minor depression: systematic review and meta-analysis. *The British Journal of Psychiatry, 198*(1), 11-16.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J Pers Soc Psychol, 51*(6), 1173-1182.

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

Bennabi, D., Aouizerate, B., El-Hage, W., Doumy, O., Moliere, F., Courtet, P., . . . Vaiva, G. (2015). Risk factors for treatment resistance in unipolar depression: a systematic review. *Journal of affective disorders, 171*, 137-141.

Berg, J. M., Kennedy, J. C., Dunlop, B. W., Ramirez, C. L., Stewart, L. M., Nemeroff, C. B., . . . Craighead, W. E. (2017). The structure of personality disorders within a depressed sample: Implications for personalizing treatment. *Personalized Medicine in Psychiatry, 1*, 59-64.

Bernecker, S. L., Coyne, A. E., Constantino, M. J., & Ravitz, P. (2017). For whom does interpersonal psychotherapy work? A systematic review. *Clinical psychology review*. doi:https://doi.org/10.1016/j.cpr.2017.07.001

Berwian, I. M., Walter, H., Seifritz, E., & Huys, Q. J. (2016). Predicting relapse after antidepressant withdrawal - a systematic review. *Psychological medicine, 47*(3), 426-437. doi:10.1017/S0033291716002580

Beutler, L. E., & Clarkin, J. F. (1990). *Systematic treatment selection: Toward targeted therapeutic interventions*: Routledge.

Beutler, L. E., Engle, D., Mohr, D., Daldrup, R. J., Bergan, J., Meredith, K., & Merry, W. (1991). Predictors of differential response to cognitive, experiential, and self-directed psychotherapeutic procedures. *J Consult Clin Psychol, 59*(2), 333-340.

Beutler, L. E., & Harwood, T. M. (2000). *Prescriptive psychotherapy: A practical guide to systematic treatment selection*: Oxford University Press on Demand.

Beutler, L. E., Someah, K., Kimpara, S., & Miller, K. (2016). Selecting the most appropriate treatment for each patient. *International Journal of Clinical and Health Psychology, 16*(1), 99-108.

Biernacka, J., Sangkuhl, K., Jenkins, G., Whaley, R., Barman, P., Batzler, A., . . . Chen, C. (2015). The International SSRI Pharmacogenomics Consortium (ISPC): a genome-wide association study of antidepressant treatment response. *Translational psychiatry, 5*(4), e553.

Bleich, J., Kapelner, A., George, E. I., & Jensen, S. T. (2014). Variable Selection for Bart: An Application to Gene Regulation. *Annals of Applied Statistics, 8*(3), 1750-1781. doi:10.1214/14-Aoas755

Bossuyt, P. M., & Parvin, T. (2015). Evaluating biomarkers for guiding treatment decisions. *EJIFCC, 26*(1), 63.

Brunoni, A. R., Sampaio-Junior, B., Moffa, A. H., Borrione, L., Nogueira, B. S., Aparício, L. V. M., . . . Tavares, D. (2015). The Escitalopram versus Electric Current Therapy for Treating Depression Clinical Study (ELECT-TDCS): rationale and study design of a non-inferiority, triple-arm, placebo-controlled clinical trial. *São Paulo Medical Journal, 133*(3), 252-263.

Bursac, Z., Gauss, C. H., Williams, D. K., & Hosmer, D. W. (2008). Purposeful selection of variables in logistic regression. *Source code for biology and medicine, 3*(1), 17.

Byar, D. P. (1985). Assessing apparent treatment—covariate interactions in randomized clinical trials. *Statistics in Medicine, 4*(3), 255-263.

Byar, D. P., & Corle, D. K. (1977). Selecting optimal treatment in clinical trials using covariate information. *Journal of chronic diseases, 30*(7), 445-459.

Byrne, S. E., & Rothschild, A. J. (1998). Loss of antidepressant efficacy during maintenance therapy: possible mechanisms and treatments. *The Journal of clinical psychiatry*.

Chakraborty, B., & Moodie, E. (2013). *Statistical methods for dynamic treatment regimes*: Springer.

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66*(1), 7.

Cheavens, J. S., Strunk, D. R., Lazarus, S. A., & Goldstein, L. A. (2012). The compensation and capitalization models: a test of two approaches to individualizing the treatment of depression. *Behaviour research and therapy, 50*(11), 699-706.

Chekroud, A. M., Gueorguieva, R., Krumholz, H. M., Trivedi, M. H., Krystal, J. H., & McCarthy, G. (2017). Reevaluating the efficacy and predictability of antidepressant treatments: a symptom clustering approach. *JAMA psychiatry, 74*(4), 370-378.

Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., & Trivedi, M. H. (2016). Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry, 3*. doi:10.1016/s2215-0366(15)00471-x

Chi, K. F., Korgaonkar, M., & Grieve, S. M. (2015). Imaging predictors of remission to anti-depressant medications in major depressive disorder. *Journal of affective disorders, 186*, 134-144.

Christensen, H., Griffiths, K. M., & Farrer, L. (2009). Adherence in internet interventions for anxiety and depression: systematic review. *Journal of medical Internet research, 11*(2).

Cloitre, M., Petkova, E., Su, Z., & Weiss, B. (2016). Patient characteristics as a moderator of post-traumatic stress disorder treatment outcome: combining symptom burden and strengths. *BJPsych Open, 2*(2), 101-106. doi:10.1192/bjpo.bp.115.000745

Cohen, Z. D., Kim, T., Van, H. L., Dekker, J. J., & Driessen, E. (*under review*). Recommending cognitive-behavioral versus psychodynamic therapy for mild to moderate adult depression. *psyArXiv preprint at https://psyarxiv.com/njus6*. doi:DOI 10.17605/OSF.IO/6QXVE

Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716.

Craske, M. G., Meuret, A. E., Ritz, T., Treanor, M., & Dour, H. J. (2016). Treatment for anhedonia: a neuroscience driven approach. *Depression and anxiety, 33*(10), 927-938.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American psychologist, 12*(11), 671.

Cuijpers, P., Ebert, D. D., Acarturk, C., Andersson, G., & Cristea, I. A. (2016). Personalized psychotherapy for adult depression: A meta-analytic review. *Behavior therapy, 47*(6), 966-980.

Cuijpers, P., Huibers, M. J., & Furukawa, T. A. (2017). The Need for Research on Treatments of Chronic Depression. *JAMA psychiatry, 74*(3), 242-243.

Cuijpers, P., Reynolds, C. F., Donker, T., Li, J., Andersson, G., & Beekman, A. (2012). Personalized treatment of adult depression: medication, psychotherapy, or both? A systematic review. *Depression and anxiety, 29*(10), 855-864.

Cuijpers, P., Van Straten, A., Warmerdam, L., & Smits, N. (2008). Characteristics of effective psychological treatments of depression: a metaregression analysis. *Psychotherapy Research, 18*(2), 225-236.

d'Agostino, R. B. (1998). Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med, 17*(19), 2265-2281.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American psychologist, 34*(7), 571.

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.

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

Deisenhofer, A. K., Delgadillo, J., Rubel, J. A., Böhnke, J. R., Zimmermann, D., Schwartz, B., & Lutz, W. (2018). Individual treatment selection for patients with posttraumatic stress disorder. *Depression and anxiety, 35*(6), 541-550.

Delgadillo, J., Huey, D., Bennett, H., & McMillan, D. (2017). Case complexity as a guide for psychological treatment selection. *Journal of Consulting and Clinical Psychology, 85*(9), 835.

Delgadillo, J., Moreea, O., & Lutz, W. (2016). Different people respond differently to therapy: A demonstration using patient profiling and risk stratification. *Behaviour research and therapy, 79*, 15-22.

DeRubeis, R. J., Cohen, Z. D., Forand, N. R., Fournier, J. C., Gelfand, L. A., & Lorenzo-Luaces, L. (2014). The Personalized Advantage Index: translating research on prediction into individualized treatment recommendations. A demonstration. *PLoS One, 9*(1), e83875. doi:10.1371/journal.pone.0083875

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

DeRubeis, R. J., Hollon, S. D., Amsterdam, J. D., Shelton, R. C., Young, P. R., Salomon, R. M., . . . Brown, L. L. (2005). Cognitive therapy vs medications in the treatment of moderate to severe depression. *Archives of general psychiatry, 62*(4), 409-416.

Dichter, G. S., Gibbs, D., & Smoski, M. J. (2015). A systematic review of relations between resting-state functional-MRI and treatment response in major depressive disorder. *Journal of affective disorders, 172*, 8-17.

Dodd, S., & Berk, M. (2004). Predictors of antidepressant response: a selective review. *International journal of psychiatry in clinical practice, 8*(2), 91-100.

Doove, L. L., Dusseldorp, E., Van Deun, K., & Van Mechelen, I. (2014). A comparison of five recursive partitioning methods to find person subgroups involved in meaningful treatment-subgroup interactions. *Adv. Data Analysis and Classification, 8*(4), 403-425.

Driessen, E., Cuijpers, P., Hollon, S. D., & Dekker, J. J. (2010). Does pretreatment severity moderate the efficacy of psychological treatment of adult outpatient depression? A meta-

analysis. *Journal of Consulting and Clinical Psychology, 78*(5), 668-680. doi:10.1037/a0020570

Drysdale, A. T., Grosenick, L., Downar, J., Dunlop, K., Mansouri, F., Meng, Y., . . . Etkin, A. (2017). Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature medicine, 23*(1), 28-38.

Dunlop, B. W., Kelley, M. E., Aponte-Rivera, V., Mletzko-Crowe, T., Kinkead, B., Ritchie, J. C., . . . Team, P. R. (2017). Effects of Patient Preferences on Outcomes in the Predictors of Remission in Depression to Individual and Combined Treatments (PReDICT) Study. *Am J Psychiatry*, appiajp201616050517. doi:10.1176/appi.ajp.2016.16050517

Dunlop, B. W., Kelley, M. E., Mletzko, T. C., Velasquez, C. M., Craighead, W. E., & Mayberg, H. S. (2012). Depression beliefs, treatment preference, and outcomes in a randomized trial for major depressive disorder. *J Psychiatr Res, 46*(3), 375-381. doi:10.1016/j.jpsychires.2011.11.003

El-Mallakh, R. S., Roberts, R. J., El-Mallakh, P. L., Findlay, L. J., & Reynolds, K. K. (2016). Pharmacogenomics in psychiatric practice. *Clinics in laboratory medicine, 36*(3), 507-523.

Elkin, I., Shea, M. T., Watkins, J. T., Imber, S. D., Sotsky, S. M., Collins, J. F., . . . Docherty, J. P. (1989). National Institute of Mental Health treatment of depression collaborative research program: General effectiveness of treatments. *Archives of general psychiatry, 46*(11), 971-982.

Fabbri, C., Hosak, L., Mossner, R., Giegling, I., Mandelli, L., Bellivier, F., . . . Serretti, A. (2017). Consensus paper of the WFSBP Task Force on Genetics: Genetics, epigenetics and gene expression markers of major depressive disorder and antidepressant response. *World J Biol Psychiatry, 18*(1), 5-28. doi:10.1080/15622975.2016.1208843

Fernandez, K. C., Fisher, A. J., & Chi, C. (2017). Development and initial implementation of the Dynamic Assessment Treatment Algorithm (DATA). *PLoS One, 12*(6), e0178806.

Fineberg, N. A., Brown, A., Reghunandanan, S., & Pampaloni, I. (2012). Evidence-based pharmacotherapy of obsessive-compulsive disorder. *International Journal of Neuropsychopharmacology, 15*(8), 1173-1191.

Fisher, A. J., & Boswell, J. F. (2016). Enhancing the Personalization of Psychotherapy With Dynamic Assessment and Modeling. *Assessment, 23*(4), 496-506. doi:10.1177/1073191116638735

Forand, N. R., Huibers, M. J., & DeRubeis, R. J. (2017). Prognosis Moderates the Engagement–Outcome Relationship in Unguided cCBT for Depression: A Proof of Concept for the Prognosis Moderation Hypothesis. *Journal of Consulting and Clinical Psychology*. doi:http://dx.doi.org/10.1037/ccp0000182

Fournier, J. C., DeRubeis, R. J., Hollon, S. D., Dimidjian, S., Amsterdam, J. D., Shelton, R. C., & Fawcett, J. (2010). Antidepressant drug effects and depression severity: a patient-level meta-analysis. *Jama, 303*(1), 47-53.

Fournier, J. C., DeRubeis, R. J., Shelton, R. C., Gallop, R., Amsterdam, J. D., & Hollon, S. D. (2008). Antidepressant medications v. cognitive therapy in people with depression with or without personality disorder. *Br J Psychiatry, 192*(2), 124-129. doi:10.1192/bjp.bp.107.037234

Fournier, J. C., DeRubeis, R. J., Shelton, R. C., Hollon, S. D., Amsterdam, J. D., & Gallop, R. (2009). Prediction of response to medication and cognitive therapy in the treatment of moderate to severe depression. *Journal of Consulting and Clinical Psychology, 77*(4), 775.

Gabrieli, J. D., Ghosh, S. S., & Whitfield-Gabrieli, S. (2015). Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. *Neuron, 85*(1), 11-26.

Gail, M., & Simon, R. (1985). Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, 361-372.

Garge, N. R., Bobashev, G., & Eggleston, B. (2013). Random forest methodology for model-based recursive partitioning: the mobForest package for R. *BMC Bioinformatics, 14*, 125. doi:10.1186/1471-2105-14-125

Gillan, C. M., & Daw, N. D. (2016). Taking Psychiatry Research Online. *Neuron, 91*(1), 19-23. doi:10.1016/j.neuron.2016.06.002

Gillan, C. M., & Whelan, R. (2017). What big data can do for treatment in psychiatry. *CURRENT OPINION IN BEHAVIORAL SCIENCES, 18*, 34-42.

Gordon, E., Rush, A. J., Palmer, D. M., Braund, T. A., & Rekshan, W. (2015). Toward an online cognitive and emotional battery to predict treatment remission in depression. *Neuropsychiatr Dis Treat, 11*, 517-531. doi:10.2147/NDT.S75975

Green, K. C., & Armstrong, J. S. (2015). Simple versus complex forecasting: The evidence. *Journal of Business Research, 68*(8), 1678-1685.

Group, P. C. R. (2008). Patients' preferences within randomised trials: systematic review and patient level meta-analysis. *The BMJ, 337*.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law, 2*(2), 293-323. doi:10.1037//1076-8971.2.2.293

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: a meta-analysis. *Psychological assessment, 12*(1), 19.

Gunn, J., Wachtler, C., Fletcher, S., Davidson, S., Mihalopoulos, C., Palmer, V., . . . Dowrick, C. (2017). Target-D: a stratified individually randomized controlled trial of the diamond clinical prediction tool to triage and target treatment for depressive symptoms in general practice: study protocol for a randomized controlled trial. *Trials, 18*(1), 342.

Gunter, L., Zhu, J., & Murphy, S. (2011). Variable selection for qualitative interactions. *Statistical methodology, 8*(1), 42-55.

Gunter, L., Zhu, J., & Murphy, S. (2011). Variable selection for qualitative interactions in personalized medicine while controlling the family-wise error rate. *Journal of biopharmaceutical statistics, 21*(6), 1063-1078.

Haby, M. M., Donnelly, M., Corry, J., & Vos, T. (2006). Cognitive behavioural therapy for depression, panic disorder and generalized anxiety disorder: a meta-regression of factors that may predict outcome. *Australian and New Zealand Journal of Psychiatry, 40*(1), 9-19.

Hamburg, M. A., & Collins, F. S. (2010). The path to personalized medicine. *N Engl J Med, 2010*(363), 301-304.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). New York: Springer.

Hingorani, A. D., van der Windt, D. A., Riley, R. D., Abrams, K., Moons, K. G., Steyerberg, E. W., . . . Hemingway, H. (2013). Prognosis research strategy (PROGRESS) 4: stratified medicine research. *Bmj, 346*, e5793.

Hirschfeld, R. M. A. (2000). Psychosocial predictors of outcome in depression. In B. FE & K. DJ (Eds.), *Psychopharmacology: The Fourth Generation of Progress* (pp. 1113-1121). New York, NY: Raven Press.

Hollon, S. D., Areán, P. A., Craske, M. G., Crawford, K. A., Kivlahan, D. R., Magnavita, J. J., . . . Bufka, L. F. (2014). Development of clinical practice guidelines. *Annual Review of Clinical Psychology, 10*, 213-241.

Hollon, S. D., Thase, M. E., & Markowitz, J. C. (2002). Treatment and prevention of depression. *Psychological Science in the public interest, 3*(2), 39-77.

Holmes, E. A., Craske, M. G., & Graybiel, A. M. (2014). A call for mental-health science. *Nature, 511*(7509), 287.

Howland, R. H. (2014). Pharmacogenetic testing in psychiatry: not (quite) ready for primetime. *J Psychosoc Nurs Ment Health Serv, 52*(11), 13-16. doi:10.3928/02793695-20141021-09

Huang, Y., Gilbert, P. B., & Janes, H. (2012). Assessing Treatment-Selection Markers using a Potential Outcomes Framework. *Biometrics, 68*(3), 687-696.

Huang, Y., Laber, E. B., & Janes, H. (2015). Characterizing expected benefits of biomarkers in treatment selection. *Biostatistics*, kxu039.

Huibers, M. J., Cohen, Z. D., Lemmens, L. H., Arntz, A., Peeters, F. P., Cuijpers, P., & DeRubeis, R. J. (2015). Predicting Optimal Outcomes in Cognitive Therapy or Interpersonal Psychotherapy for Depressed Individuals Using the Personalized Advantage Index Approach. *PLoS One, 10*(11), e0140771.

Hunter, A. M., Cook, I. A., Abrams, M., & Leuchter, A. F. (2013). Neurophysiologic effects of repeated exposure to antidepressant medication: are brain functional changes during antidepressant administration influenced by learning processes? *Medical hypotheses, 81*(6), 1004-1011.

Hunter, A. M., Cook, I. A., Greenwald, S., Tran, M. L., Miyamoto, K. N., & Leuchter, A. F. (2011). The Antidepressant Treatment Response (ATR) index and treatment outcomes in a placebo-controlled trial of fluoxetine. *Journal of Clinical Neurophysiology, 28*(5), 478-482.

Imai, K., & Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics, 7*(1), 443-470.

Iniesta, R., Malki, K., Maier, W., Rietschel, M., Mors, O., Hauser, J., . . . Uher, R. (2016). Combining clinical variables to optimize prediction of antidepressant treatment outcomes. *J Psychiatr Res, 78*, 94-102. doi:10.1016/j.jpsychires.2016.03.016

Iniesta, R., Stahl, D., & McGuffin, P. (2016). Machine learning, statistical learning and the future of biological research in psychiatry. *Psychol Med, 46*(12), 2455-2465. doi:10.1017/S0033291716001367

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

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112): Springer.

Jamshidian, M., & Jalal, S. (2010). Tests of homoscedasticity, normality, and missing completely at random for incomplete multivariate data. *Psychometrika, 75*(4), 649-674.

Janes, H., Pepe, M. S., Bossuyt, P. M., & Barlow, W. E. (2011). Measuring the performance of markers for guiding treatment decisions. *Annals of internal medicine, 154*(4), 253-259.

Jappe, L. M., Klimes-Dougan, B., & Cullen, K. R. (2013). Brain imaging and the prediction of treatment outcomes in mood and anxiety disorders *Functional brain mapping and the endeavor to understand the working brain*: InTech.

Jollans, L., & Whelan, R. (2016). The clinical added value of imaging: A perspective from outcome prediction. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 1*(5), 423-432.

Joyce, P. R., & Paykel, E. S. (1989). Predictors of drug response in depression. *Archives of general psychiatry, 46*(1), 89-99.

Kapelner, A., & Bleich, J. (2016). bartMachine: Machine Learning with Bayesian Additive Regression Trees. *Journal of Statistical software, 70*(4), 1-40. doi:10.18637/jss.v070.i04

Kapelner, A., Bleich, J., Cohen, Z. D., DeRubeis, R. J., & Berk, R. (*under review*). Inference for treatment regime models in personalized medicine. *arXiv preprint arXiv:1404.7844*.

Katsnelson, A. (2013). Momentum grows to make 'personalized' medicine more 'precise'. *Nat Med, 19*(3), 249. doi:10.1038/nm0313-249

Keefe, J. R., Wiltsey Stirman, S., Cohen, Z. D., DeRubeis, R. J., Smith, B. N., & Resick, P. A. (2018). In rape trauma PTSD, patient characteristics indicate which trauma-focused treatment they are most likely to complete. *Depression and anxiety, 35*(4), 330-338.

Kemp, A. H., Brunoni, A. R., & Machado-Vieira, R. (2015). Predictors of treatment response in major depressive disorder *Treatment-Resistant Mood Disorders* (pp. 53-60): Oxford University Press, Oxford.

Kemp, A. H., Gordon, E., Rush, A. J., & Williams, L. M. (2008). Improving the prediction of treatment response in depression: integration of clinical, cognitive, psychophysiological, neuroimaging, and genetic measures. *CNS Spectr, 13*(12), 1066-1086; quiz 1087-1068.

Kessler, R. C. (2018). The potential of predictive analytics to provide clinical decision support in depression treatment planning. *Current opinion in psychiatry, 31*(1), 32-39.

Kessler, R. C., van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Cai, T., . . . de Jonge, P. (2016). Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports. *Molecular psychiatry, 21*(10), 1366.

Kessler, R. C., van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Ebert, D. D., . . . Zaslavsky, A. M. (2017). Using patient self-reports to study heterogeneity of treatment effects in major depressive disorder. *Epidemiol Psychiatr Sci, 26*(1), 22-36. doi:10.1017/S2045796016000020

Khan, A., Leventhal, R. M., Khan, S. R., & Brown, W. A. (2002). Severity of depression and response to antidepressants and placebo: an analysis of the Food and Drug Administration database. *Journal of clinical psychopharmacology, 22*(1), 40-45.

Kim, T., Dufour, S., Cohen, Z. D., Sylvia, L. G., Deckersbach, T., & Nierenberg, A. (*submitted*). Applying machine-learning techniques for treatment selection in bipolar disorder for patients deciding between lithium or quetiapine. *PsyArXiv preprint at https://psyarxiv.com/xk3s9/*. doi:DOI 10.17605/OSF.IO/HDQY2

King, M., Walker, C., Levy, G., Bottomley, C., Royston, P., & Weich, S. (2008). Development and validation of an international risk prediction algorithm for episodes of major depression in general practice attendees: the PredictD study. *Arch Gen Psychiatry, 65*. doi:10.1001/archpsyc.65.12.1368

Kingslake, J., Dias, R., Dawson, G. R., Simon, J., Goodwin, G. M., Harmer, C. J., . . . Dourish, C. T. (2017). The effects of using the PReDicT Test to guide the antidepressant treatment of depressed patients: study protocol for a randomised controlled trial. *Trials, 18*(1), 558.

Kirsch, I., Deacon, B. J., Huedo-Medina, T. B., Scoboria, A., Moore, T. J., & Johnson, B. T. (2008). Initial severity and antidepressant benefits: a meta-analysis of data submitted to the Food and Drug Administration. *PLoS medicine, 5*(2), e45.

Klerman, G. L., & Weissman, M. M. (1994). *Interpersonal psychotherapy of depression: A brief, focused, specific strategy*: Jason Aronson, Incorporated.

Kocsis, J. H., Leon, A. C., Markowitz, J. C., Manber, R., Arnow, B., Klein, D. N., & Thase, M. E. (2009). Patient preference as a moderator of outcome for chronic forms of major depressive disorder treated with nefazodone, cognitive behavioral analysis system of psychotherapy, or their combination. *The Journal of clinical psychiatry*.

Kool, S., Schoevers, R., de Maat, S., Van, R., Molenaar, P., Vink, A., & Dekker, J. (2005). Efficacy of pharmacotherapy in depressed patients with and without personality disorders: a systematic review and meta-analysis. *Journal of affective disorders, 88*(3), 269-278.

Koutsouleris, N., Kahn, R. S., Chekroud, A. M., Leucht, S., Falkai, P., Wobrock, T., . . . Hasan, A. (2016). Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. *The Lancet Psychiatry, 3*(10), 935-946.

Kraemer, H. C. (2013). Discovering, comparing, and combining moderators of treatment on outcome after randomized clinical trials: a parametric approach. *Stat Med, 32*(11), 1964-1973. doi:10.1002/sim.5734

Kraemer, H. C., & Blasey, C. M. (2004). Centring in regression analyses: a strategy to prevent errors in statistical inference. *Int J Methods Psychiatr Res, 13*(3), 141-151.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*: Springer Science & Business Media.

Lam, R. W., Milev, R., Rotzinger, S., Andreazza, A. C., Blier, P., Brenner, C., . . . Evans, K. R. (2016). Discovering biomarkers for antidepressant response: protocol from the Canadian biomarker integration network in depression (CAN-BIND) and clinical characteristics of the first patient cohort. *BMC psychiatry, 16*(1), 105.

Layard, R., Clark, D., Knapp, M., & Mayraz, G. (2007). Cost-benefit analysis of psychological therapy. *National Institute Economic Review, 202*(1), 90-98.

Lener, M. S., & Iosifescu, D. V. (2015). In pursuit of neuroimaging biomarkers to guide treatment selection in major depressive disorder: a review of the literature. *Annals of the New York Academy of Sciences, 1344*(1), 50-65.

Leuchter, A. F., Cook, I. A., Gilmer, W. S., Marangell, L. B., Burgoyne, K. S., Howland, R. H., . . . Fava, M. (2009). Effectiveness of a quantitative electroencephalographic biomarker for predicting differential response or remission with escitalopram and bupropion in major depressive disorder. *Psychiatry research, 169*(2), 132-138.

Leykin, Y., Amsterdam, J. D., DeRubeis, R. J., Gallop, R., Shelton, R. C., & Hollon, S. D. (2007). Progressive resistance to a selective serotonin reuptake inhibitor but not to cognitive therapy in the treatment of major depression. *Journal of Consulting and Clinical Psychology, 75*(2), 267.

Leykin, Y., DeRubeis, R. J., Gallop, R., Amsterdam, J. D., Shelton, R. C., & Hollon, S. D. (2007). The relation of patients' treatment preferences to outcome in a randomized clinical trial. *Behavior therapy, 38*(3), 209-217.

Lisoway, A., Zai, C., Tiwari, A., & Kennedy, J. (2017). DNA Methylation and Clinical Response to Antidepressant Medication in Major Depressive Disorder: A Review and Recommendations. *Neuroscience letters*.

Lo, A., Chernoff, H., Zheng, T., & Lo, S.-H. (2015). Why significant variables aren't automatically good predictors. *Proceedings of the National Academy of Sciences, 112*(45), 13892-13897.

Lorenzo-Luaces, L., DeRubeis, R. J., & Bennett, I. M. (2015). Primary care physicians' selection of low-intensity treatments for patients with depression. *Family medicine, 47*, 511-516.

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

Luedtke, A. R., & van der Laan, M. J. (2016). Super-learning of an optimal dynamic treatment rule. *The international journal of biostatistics, 12*(1), 305-332.

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

Lutz, W., Saunders, S. M., Leon, S. C., Martinovich, Z., Kosfelder, J., Schulte, D., . . . Tholen, S. (2006). Empirically and clinically useful decision making in psychotherapy: differential predictions with treatment response models. *Psychol Assess, 18*(2), 133-141. doi:10.1037/1040-3590.18.2.133

Lutz, W., Zimmermann, D., Müller, V. N., Deisenhofer, A.-K., & Rubel, J. A. (2017). Randomized controlled trial to evaluate the effects of personalized prediction and adaptation tools on treatment outcome in outpatient psychotherapy: study protocol. *BMC psychiatry, 17*(1), 306.

Ma, J., Stingo, F. C., & Hobbs, B. P. (2016). Bayesian predictive modeling for genomic based personalized treatment selection. *Biometrics, 72*(2), 575-583.

MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annu. Rev. Psychol., 58*, 593-614.

Mayberg, H. S., Lozano, A. M., Voon, V., McNeely, H. E., Seminowicz, D., Hamani, C., . . . Kennedy, S. H. (2005). Deep brain stimulation for treatment-resistant depression. *Neuron, 45*(5), 651-660.

McCullough Jr, J. P. (2003). *Treatment for chronic depression: Cognitive behavioral analysis system of psychotherapy (CBASP)* (Vol. 13): Educational Publishing Foundation.

McGirr, A., Berlim, M., Bond, D., Fleck, M., Yatham, L., & Lam, R. (2015). A systematic review and meta-analysis of randomized, double-blind, placebo-controlled trials of ketamine in the rapid treatment of major depressive episodes. *Psychological medicine, 45*(4), 693-704.

McGrath, C. L., Kelley, M. E., Holtzheimer, P. E., Dunlop, B. W., Craighead, W. E., Franco, A. R., . . . Mayberg, H. S. (2013). Toward a neuroimaging treatment selection biomarker for major depressive disorder. *JAMA psychiatry, 70*(8), 821-829.

McHugh, R. K., Whitton, S. W., Peckham, A. D., Welge, J. A., & Otto, M. W. (2013). Patient preference for psychological vs pharmacologic treatment of psychiatric disorders: a meta-analytic review. *J Clin Psychiatry, 74*(6), 595-602. doi:10.4088/JCP.12r07757

Meehl, P. E. (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. doi:http://dx.doi.org/10.1037/11281-000

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*(4), 806.

Mergl, R., Henkel, V., Allgaier, A. K., Kramer, D., Hautzinger, M., Kohnen, R., . . . Hegerl, U. (2011). Are treatment preferences relevant in response to serotonergic antidepressants and cognitive-behavioral therapy in depressed primary care patients? Results from a randomized controlled trial including a patients' choice arm. *Psychother Psychosom, 80*(1), 39-47. doi:10.1159/000318772

Mick, R., & Ratain, M. J. (1994). Bootstrap validation of pharmacodynamic models defined via stepwise linear regression. *Clinical Pharmacology & Therapeutics, 56*(2), 217-222.

Mickey, R. M., & Greenland, S. (1989). The impact of confounder selection criteria on effect estimation. *American journal of epidemiology, 129*(1), 125-137.

National Health Service, D. (2016). *Psychological Therapies, Annual Report on the use of IAPT services: England 2015-16*. Retrieved from London, UK:

National Institute for Health and Clinical Excellence. (2009). *Depression: Treatment and Management of Depression in Adults*. Retrieved from London, UK:

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

Naudet, F., Maria, A. S., & Falissard, B. (2011). Antidepressant response in major depressive disorder: a meta-regression comparison of randomized controlled trials and observational studies. *PLoS One, 6*(6), e20811.

Nelson, J. C., Delucchi, K. L., & Schneider, L. S. (2013). Moderators of outcome in late-life depression: a patient-level meta-analysis. *American Journal of Psychiatry, 170*(6), 651-659.

Nemeroff, C. B., Heim, C. M., Thase, M. E., Klein, D. N., Rush, A. J., Schatzberg, A. F., . . . Dunner, D. L. (2003). Differential responses to psychotherapy versus pharmacotherapy in patients with chronic forms of major depression and childhood trauma. *Proceedings of the National Academy of Sciences, 100*(24), 14293-14296.

Newton-Howes, G., Tyrer, P., Johnson, T., Mulder, R., Kool, S., Dekker, J., & Schoevers, R. (2013). Influence of personality on the outcome of treatment in depression: systematic review and meta-analysis. *Journal of Personality Disorders, 28*(4), 577-593.

Nigatu, Y. T., Liu, Y., & Wang, J. (2016). External validation of the international risk prediction algorithm for major depressive episode in the US general population: the PredictD-US study. *BMC psychiatry, 16*, 256. doi:10.1186/s12888-016-0971-x

Niles, A. N., Loerinc, A. G., Krull, J. L., Roy-Byrne, P., Sullivan, G., Sherbourne, C. D., . . . Craske, M. G. (2017). Advancing Personalized Medicine: Application of a Novel Statistical Method to Identify Treatment Moderators in the Coordinated Anxiety Learning and Management Study. *Behavior therapy, 48*(4), 490-500.

Niles, A. N., Wolitzky-Taylor, K. B., Arch, J. J., & Craske, M. G. (2017). Applying a novel statistical method to advance the personalized treatment of anxiety disorders: A composite moderator of comparative drop-out from CBT and ACT. *Behav Res Ther, 91*, 13-23. doi:10.1016/j.brat.2017.01.001

Nuzzo, R. (2014). Statistical errors. *Nature, 506*(7487), 150.

Olbrich, S., & Arns, M. (2013). EEG biomarkers in major depressive disorder: discriminative power and prediction of treatment response. *International Review of Psychiatry, 25*(5), 604-618. doi:10.3109/09540261.2013.816269

Olbrich, S., van Dinteren, R., & Arns, M. (2015). Personalized medicine: review and perspectives of promising baseline EEG biomarkers in major depressive disorder and attention deficit hyperactivity disorder. *Neuropsychobiology, 72*(3-4), 229-240.

Paez, J. G., Jänne, P. A., Lee, J. C., Tracy, S., Greulich, H., Gabriel, S., . . . Boggon, T. J. (2004). EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science, 304*(5676), 1497-1500.

Pao, W., & Miller, V. A. (2005). Epidermal growth factor receptor mutations, small-molecule kinase inhibitors, and non–small-cell lung cancer: current knowledge and future directions. *Journal of Clinical Oncology, 23*(11), 2556-2568.

Papakostas, G. I., & Fava, M. (2010). *Pharmacotherapy for depression and treatment-resistant depression*: World Scientific.

Parmigiani, G. (2002). *Modeling in medical decision making: a Bayesian approach*: Wiley.

Passos, I. C., Mwangi, B., & Kapczinski, F. (2016). Big data analytics and machine learning: 2015 and beyond. *The Lancet Psychiatry, 3*(1), 13-15.

Pauker, S. G., & Kassirer, J. P. (1980). The threshold approach to clinical decision making. *N Engl J Med, 302*(20), 1109-1117. doi:10.1056/NEJM198005153022003

Paul, G. L. (1967). Strategy of outcome research in psychotherapy. *J Consult Psychol, 31*(2), 109-118.

Perlis, R., Iosifescu, D., Castro, V., Murphy, S., Gainer, V., Minnier, J., . . . Gallagher, P. (2012). Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. *Psychological medicine, 42*(1), 41-50.

Perlis, R. H. (2013). A clinical risk stratification tool for predicting treatment resistance in major depressive disorder. *Biol Psychiatry, 74*(1), 7-14. doi:10.1016/j.biopsych.2012.12.007

Perlis, R. H. (2014). Pharmacogenomic testing and personalized treatment of depression. *Clinical chemistry, 60*(1), 53-59. doi:10.1373/clinchem.2013.204446

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

Perlis, R. H., Fijal, B., Dharia, S., Heinloth, A. N., & Houston, J. P. (2010). Failure to replicate genetic associations with antidepressant treatment response in duloxetine-treated patients. *Biological psychiatry, 67*(11), 1110-1113.

Perlis, R. H., Patrick, A., Smoller, J. W., & Wang, P. S. (2009). When is pharmacogenetic testing for antidepressant response ready for the clinic? A cost-effectiveness analysis based on data from the STAR* D study. *Neuropsychopharmacology, 34*(10), 2227-2236.

Petkova, E., Ogden, R. T., Tarpey, T., Ciarleglio, A., Jiang, B., Su, Z., . . . Grannemann, B. D. (2017). Statistical analysis plan for stage 1 EMBARC (Establishing Moderators and Biosignatures of Antidepressant Response for Clinical Care) study. *Contemporary Clinical Trials Communications, 6*, 22-30.

Phillips, M. L., Chase, H. W., Sheline, Y. I., Etkin, A., Almeida, J. R., Deckersbach, T., & Trivedi, M. H. (2015). Identifying predictors, moderators, and mediators of antidepressant response in major depressive disorder: neuroimaging approaches. *American Journal of Psychiatry, 172*(2), 124-138. doi:10.1176/appi.ajp.2014.14010076

Pizzagalli, D. A. (2011). Frontocingulate dysfunction in depression: toward biomarkers of treatment response. *Neuropsychopharmacology, 36*(1), 183-206.

Polyakova, M., Stuke, K., Schuemberg, K., Mueller, K., Schoenknecht, P., & Schroeter, M. L. (2015). BDNF as a biomarker for successful treatment of mood disorders: a systematic & quantitative meta-analysis. *Journal of affective disorders, 174*, 432-440.

Pompili, M., Venturini, P., Palermo, M., Stefani, H., Seretti, M. E., Lamis, D. A., . . . Girardi, P. (2013). Mood disorders medications: predictors of nonadherence–review of the current literature. *Expert review of neurotherapeutics, 13*(7), 809-825.

Raza, G. T., & Holohan, D. R. (2015). Clinical treatment selection for posttraumatic stress disorder: Suggestions for researchers and clinical trainers. *Psychol Trauma, 7*(6), 547-554. doi:10.1037/tra0000059

Renjilian, D. A., Perri, M. G., Nezu, A. M., McKelvey, W. F., Shermer, R. L., & Anton, S. D. (2001). Individual versus group therapy for obesity: effects of matching participants to their treatment preferences. *Journal of Consulting and Clinical Psychology, 69*(4), 717.

Rivero-Santana, A., Perestelo-Perez, L., Pérez-Ramos, J., Serrano-Aguilar, P., & De las Cuevas, C. (2013). Sociodemographic and clinical predictors of compliance with antidepressants for depressive disorders: systematic review of observational studies. *Patient preference and adherence, 7*, 151.

Rosell, R., Carcereny, E., Gervais, R., Vergnenegre, A., Massuti, B., Felip, E., . . . Sanchez, J. M. (2012). Erlotinib versus standard chemotherapy as first-line treatment for European patients with advanced EGFR mutation-positive non-small-cell lung cancer (EURTAC): a multicentre, open-label, randomised phase 3 trial. *The lancet oncology, 13*(3), 239-246.

Rubenstein, L., Rayburn, N., Keeler, E., Ford, D., Rost, K., & Sherbourne, C. (2007). Predicting outcomes of primary care patients with major depression: Development of a depression prognosis index. *Psychiatr Serv, 58*. doi:10.1176/ps.2007.58.8.1049

Rush, A. J., Trivedi, M. H., Stewart, J. W., Nierenberg, A. A., Fava, M., Kurian, B. T., . . . Husain, M. M. (2011). Combining medications to enhance depression outcomes (CO-MED): acute and long-term outcomes of a single-blind randomized study. *American Journal of Psychiatry, 168*(7), 689-701.

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.

Schleidgen, S., Klingler, C., Bertram, T., Rogowski, W. H., & Marckmann, G. (2013). What is personalized medicine: sharpening a vague term based on a systematic literature review. *BMC Med Ethics, 14*, 55. doi:10.1186/1472-6939-14-55

Schneider, R. L., Arch, J. J., & Wolitzky-Taylor, K. B. (2015). The state of personalized treatment for anxiety disorders: A systematic review of treatment moderators. *Clin Psychol Rev, 38*, 39-54. doi:10.1016/j.cpr.2015.02.004

Schuch, F. B., Dunn, A. L., Kanitz, A. C., Delevatti, R. S., & Fleck, M. P. (2016). Moderators of response in exercise treatment for depression: A systematic review. *J Affect Disord, 195*, 40-49. doi:10.1016/j.jad.2016.01.014

Schwaederle, M., Zhao, M., Lee, J. J., Eggermont, A. M., Schilsky, R. L., Mendelsohn, J., . . . Kurzrock, R. (2015). Impact of precision medicine in diverse cancers: a meta-analysis of phase II clinical trials. *Journal of Clinical Oncology, 33*(32), 3817-3825.

Schweizer, S., Cohen, Z., Hayes, R., DeRubeis, R., Crane, C., Kuyken, W., & Dalgleish, T. (*submitted*). Relapse Prevention for Antidepressant Medication (ADM) Responders with Recurrent Depression: Using the Personalized Advantage Index to Decide Between Maintenance ADM and Mindfulness-Based Cognitive Therapy.

Serretti, A., Chiesa, A., Calati, R., Perna, G., Bellodi, L., & De Ronchi, D. (2009). Common genetic, clinical, demographic and psychosocial predictors of response to pharmacotherapy in mood and anxiety disorders. *International clinical psychopharmacology, 24*(1), 1-18.

Serretti, A., Gibiino, S., & Drago, A. (2011). Specificity profile of paroxetine in major depressive disorder: Meta-regression of double-blind, randomized clinical trials. *Journal of affective disorders, 132*(1), 14-25.

Silveira, H., Moraes, H., Oliveira, N., Coutinho, E. S. F., Laks, J., & Deslandes, A. (2013). Physical exercise and clinically depressed patients: a systematic review and meta-analysis. *Neuropsychobiology, 67*(2), 61-68.

Simon, G. E., & Perlis, R. H. (2010). Personalized medicine for depression: can we match patients with treatments? *American Journal of Psychiatry, 167*(12), 1445-1455.

Smagula, S. F., Wallace, M. L., Anderson, S. J., Karp, J. F., Lenze, E. J., Mulsant, B. H., . . . Lotrich, F. E. (2016). Combining moderators to identify clinical profiles of patients who will, and will not, benefit from aripiprazole augmentation for treatment resistant late-life major depressive disorder. *Journal of psychiatric research, 81*, 112-118.

Smith, D. F. (2013). Quest for biomarkers of treatment-resistant depression: shifting the paradigm toward risk. *Frontiers in psychiatry, 4*, 57. doi:10.3389/fpsyt.2013.00057

Souslova, T., Marple, T. C., Spiekerman, A. M., & Mohammad, A. A. (2013). Personalized medicine in Alzheimer's disease and depression. *Contemporary clinical trials, 36*(2), 616-623.

Stephan, K. E., Schlagenhauf, F., Huys, Q. J., Raman, S., Aponte, E. A., Brodersen, K. H., . . . Dolan, R. (2017). Computational neuroimaging strategies for single patient predictions. *Neuroimage, 145*, 180-199.

Steyerberg, E. (2008). *Clinical prediction models: a practical approach to development, validation, and updating*: Springer Science & Business Media.

Strawbridge, R., Arnone, D., Danese, A., Papadopoulos, A., Vives, A. H., & Cleare, A. (2015). Inflammation and clinical response to treatment in depression: a meta-analysis. *European Neuropsychopharmacology, 25*(10), 1532-1543.

Swift, J. K., & Callahan, J. L. (2009). The impact of client treatment preferences on outcome: a meta-analysis. *Journal of clinical psychology, 65*(4), 368-381.

Swift, J. K., Callahan, J. L., & Vollmer, B. M. (2011). Preferences. *Journal of clinical psychology, 67*(2), 155-165.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

Tiemens, B., Bocker, K., & Kloos, M. (2016). Prediction of treatment outcome in daily generalized mental health care practice: first steps towards personalized treatment by clinical decision support. *European Journal for Person Centered Healthcare, 4*(1), 24-32.

Trivedi, M. H., McGrath, P. J., Fava, M., Parsey, R. V., Kurian, B. T., Phillips, M. L., . . . Toups, M. (2016). Establishing moderators and biosignatures of antidepressant response in clinical care (EMBARC): Rationale and design. *Journal of psychiatric research, 78*, 11-23.

Trivedi, M. H., Rush, A. J., Wisniewski, S. R., Nierenberg, A. A., Warden, D., Ritz, L., . . . McGrath, P. J. (2006). Evaluation of outcomes with citalopram for depression using measurement-based care in STAR* D: implications for clinical practice. *American Journal of Psychiatry, 163*(1), 28-40.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review, 90*(4), 293.

Uher, R., Huezo-Diaz, P., Perroud, N., Smith, R., Rietschel, M., Mors, O., . . . Henigsberg, N. (2009). Genetic predictors of response to antidepressants in the GENDEP project. *The pharmacogenomics journal, 9*(4), 225-233.

Uher, R., Perlis, R., Henigsberg, N., Zobel, A., Rietschel, M., Mors, O., . . . Bajs, M. (2012). Depression symptom dimensions as predictors of antidepressant treatment outcome: replicable evidence for interest-activity symptoms. *Psychological medicine, 42*(5), 967-980.

Uher, R., Tansey, K. E., Dew, T., Maier, W., Mors, O., Hauser, J., . . . McGuffin, P. (2014). An inflammatory biomarker as a differential predictor of outcome of depression treatment with escitalopram and nortriptyline. *Am J Psychiatry, 171*(12), 1278-1286. doi:10.1176/appi.ajp.2014.14010094

van Straten, A., Tiemens, B., Hakkaart, L., Nolen, W., & Donker, M. (2006). Stepped care vs. matched care for mood and anxiety disorders: a randomized trial in routine practice. *Acta Psychiatrica Scandinavica, 113*(6), 468-476.

Vittengl, J. R., Clark, L. A., Thase, M. E., & Jarrett, R. B. (2017). Initial Steps to inform selection of continuation cognitive therapy or fluoxetine for higher risk responders to cognitive therapy for recurrent major depressive disorder. *Psychiatry research, 253*, 174-181.

Vittengl, J. R., Jarrett, R. B., Weitz, E., Hollon, S. D., Twisk, J., Cristea, I., . . . Dunlop, B. W. (2016). Divergent outcomes in cognitive-behavioral therapy and pharmacotherapy for adult depression. *American Journal of Psychiatry, 173*(5), 481-490.

Wallace, M. L., Frank, E., & Kraemer, H. C. (2013). A novel approach for developing and interpreting treatment moderator profiles in randomized clinical trials. *JAMA psychiatry, 70*(11), 1241-1247.

Wang, R., & Ware, J. H. (2013). Detecting moderator effects using subgroup analyses. *Prevention Science*, 1-10.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *Am Stat, 70*(2), 129-133.

Watkins, E., Newbold, A., Tester-Jones, M., Javaid, M., Cadman, J., Collins, L. M., . . . Mostazir, M. (2016). Implementing multifactorial psychotherapy research in online virtual environments (IMPROVE-2): study protocol for a phase III trial of the MOST randomized component selection method for internet cognitive-behavioural therapy for depression. *BMC psychiatry, 16*(1), 345.

Webb, C. A., Trivedi, M. H., Cohen, Z. D., Dillon, D. G., Fournier, J. C., Goer, F., . . . Parsey, R. (2018). Personalized prediction of antidepressant v. placebo response: evidence from the EMBARC study. *Psychological medicine*, 1-10.

Weisz, J. R., Krumholz, L. S., Santucci, L., Thomassin, K., & Ng, M. Y. (2015). Shrinking the gap between research and practice: Tailoring and testing youth psychotherapies in clinical care contexts. *Annual Review of Clinical Psychology, 11*, 139-163.

Weitz, E. S., Hollon, S. D., Twisk, J., van Straten, A., Huibers, M. J., David, D., . . . Cristea, I. A. (2015). Baseline depression severity as moderator of depression outcomes between cognitive behavioral therapy vs pharmacotherapy: an individual patient data meta-analysis. *JAMA psychiatry, 72*(11), 1102-1109.

Wellek, S. (1997). Testing for absence of qualitative interactions between risk factors and treatment effects. *Biometrical journal, 39*(7), 809-821.

Westover, A. N., Kashner, T. M., Winhusen, T. M., Golden, R. M., Nakonezny, P. A., Adinoff, B., & Henley, S. S. (2015). A systematic approach to subgroup analyses in a smoking cessation trial. *The American journal of drug and alcohol abuse, 41*(6), 498-507.

Widaman, K. F., Helm, J. L., Castro-Schilo, L., Pluess, M., Stallings, M. C., & Belsky, J. (2012). Distinguishing ordinal and disordinal interactions. *Psychol Methods, 17*(4), 615-622. doi:10.1037/a0030003

Williams, L. M., Rush, A. J., Koslow, S. H., Wisniewski, S. R., Cooper, N. J., Nemeroff, C. B., . . . Gordon, E. (2011). International Study to Predict Optimized Treatment for Depression (iSPOT-D), a randomized clinical trial: rationale and protocol. *Trials, 12*(1), 4.

Winter, S. E., & Barber, J. P. (2013). Should treatment for depression be based more on patient preference? *Patient preference and adherence, 7*, 1047.

World Health Organization. (2017). *Depression and Other Common Mental Disorders: Global Health Estimates* (CC BY-NC-SA 3.0 IGO). Retrieved from Geneva:

Yakovlev, A. Y., Goot, R. E., & Osipova, T. T. (1994). The choice of cancer treatment based on covariate information. *Statistics in Medicine, 13*(15), 1575-1581.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12*(6), 1100-1122.

Zilcha-Mano, S., Keefe, J. R., Chui, H., Rubin, A., Barrett, M. S., & Barber, J. P. (2016). Reducing Dropout in Treatment for Depression: Translating Dropout Predictors Into Individualized Treatment Recommendations. *The Journal of clinical psychiatry, 77*(12), e1584-e1590.

Zimmerman, M., Clark, H. L., Multach, M. D., Walsh, E., Rosenstein, L. K., & Gazarian, D. (2015). *Have treatment studies of depression become even less generalizable? A review of the inclusion and exclusion criteria used in placebo-controlled antidepressant efficacy trials published during the past 20 years*. Paper presented at the Mayo Clinic Proceedings.

Zimmerman, M., Clark, H. L., Multach, M. D., Walsh, E., Rosenstein, L. K., & Gazarian, D. (2016). Symptom Severity and the Generalizability of Antidepressant Efficacy Trials: Changes During the Past 20 Years. *Journal of clinical psychopharmacology, 36*(2), 153-156.

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

Chapter 2 References

Austin, P. C., & Tu, J. V. (2004). Bootstrap methods for developing predictive models. *The American Statistician, 58*(2), 131-137.

Barber, J. P., & Muenz, L. R. (1996). The role of avoidance and obsessiveness in matching patients to cognitive and interpersonal psychotherapy: Empirical findings from the Treatment for Depression Collaborative Research Program. *Journal of Consulting and Clinical Psychology, 64*(5), 951.

Barth, J., Munder, T., Gerger, H., Nüesch, E., Trelle, S., Znoj, H., . . . Cuijpers, P. (2013). Comparative efficacy of seven psychotherapeutic interventions for patients with depression: a network meta-analysis. *PLoS medicine, 10*(5), e1001454.

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

Bleich, J., Kapelner, A., George, E. I., & Jensen, S. T. (2014). Variable Selection for Bart: An Application to Gene Regulation. *Annals of Applied Statistics, 8*(3), 1750-1781. doi:10.1214/14-Aoas755

Chekroud, A. M., Gueorguieva, R., Krumholz, H. M., Trivedi, M. H., Krystal, J. H., & McCarthy, G. (2017). Reevaluating the efficacy and predictability of antidepressant treatments: a symptom clustering approach. *JAMA psychiatry, 74*(4), 370-378.

Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., & Trivedi, M. H. (2016). Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry, 3*. doi:10.1016/s2215-0366(15)00471-x

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics, 4*(1), 266-298.

Cohen, Z. D., & DeRubeis, R. J. (2018). Treatment Selection in Depression. *Annual Review of Clinical Psychology, 14*.

De Beurs, E., & Zitman, F. (2005). De Brief Symptom Inventory (BSI). *De betrouwbaarheid en validiteit van een handzaam alternatief voor de SCL-90. Leiden: Leids universitair medisch centrum*.

de Jonghe, F. (1994). *Leidraad voor het scoren van de Hamilton Depression Rating Scale [Hamilton Depression Rating Scale scoring manual]*. Amsterdam: Benecke.

de Jonghe, F. (2005). *Kort en Krachtig: Kortdurende Psychoanalytische Steungevende Psychotherapie [Short and Snappy: Short-term Psychoanalytic Supportive Psychotherapy]*. Amsterdam: Benecke NI.

de Jonghe, F., de Maat, S., Van, R., Hendriksen, M., Kool, S., van Aalst, G., & Dekker, J. (2013). Short-term psychoanalytic supportive psychotherapy for depressed patients. *Psychoanalytic Inquiry, 33*(6), 614-625.

Deisenhofer, A. K., Delgadillo, J., Rubel, J. A., Böhnke, J. R., Zimmermann, D., Schwartz, B., & Lutz, W. (2018). Individual treatment selection for patients with posttraumatic stress disorder. *Depression and anxiety, 35*(6), 541-550.

DeRubeis, R. J., Cohen, Z. D., Forand, N. R., Fournier, J. C., Gelfand, L. A., & Lorenzo-Luaces, L. (2014). The Personalized Advantage Index: translating research on prediction into individualized treatment recommendations. A demonstration. *PLoS One, 9*(1), e83875. doi:10.1371/journal.pone.0083875

Driessen, E., Smits, N., Dekker, J., Peen, J., Don, F., Kool, S., . . . Van, H. (2016). Differential efficacy of cognitive behavioral therapy and psychodynamic therapy for major depression: a study of prescriptive factors. *Psychological medicine, 46*(4), 731-744.

Driessen, E., Van, H. L., Don, F. J., Peen, J., Kool, S., Westra, D., . . . Twisk, J. W. (2013). The efficacy of cognitive-behavioral therapy and psychodynamic therapy in the outpatient treatment of major depression: a randomized clinical trial. *American Journal of Psychiatry, 170*(9), 1041-1050.

Driessen, E., Van, H. L., Peen, J., Don, F. J., Kool, S., Westra, D., . . . Dekker, J. J. (2015). Therapist-rated outcomes in a randomized clinical trial comparing cognitive behavioral therapy and psychodynamic therapy for major depression. *Journal of affective disorders, 170*, 112-118.

Driessen, E., Van, H. L., Peen, J., Don, F. J., Twisk, J. W., Cuijpers, P., & Dekker, J. J. (2017). Cognitive-behavioral versus psychodynamic therapy for major depression: Secondary outcomes of a randomized clinical trial. *Journal of Consulting and Clinical Psychology, 85*(7), 653.

Driessen, E., Van, H. L., Schoevers, R. A., Cuijpers, P., Van Aalst, G., Don, F. J., . . . Peen, J. (2007). Cognitive Behavioral Therapy versus Short Psychodynamic Supportive Psychotherapy in the outpatient treatment of depression: a randomized controlled trial. *BMC psychiatry, 7*(1), 58.

Fiedler, K. (2011). Voodoo correlations are everywhere—not only in neuroscience. *Perspectives on Psychological Science, 6*(2), 163-171.

Fournier, J. C., DeRubeis, R. J., Shelton, R. C., Hollon, S. D., Amsterdam, J. D., & Gallop, R. (2009). Prediction of response to medication and cognitive therapy in the treatment of moderate to severe depression. *Journal of Consulting and Clinical Psychology, 77*(4), 775.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical software, 33*(1), 1.

Garge, N. R., Bobashev, G., & Eggleston, B. (2013). Random forest methodology for model-based recursive partitioning: the mobForest package for R. *BMC Bioinformatics, 14*, 125. doi:10.1186/1471-2105-14-125

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters, 31*(14), 2225-2236.

Gibbons, M. B. C., Gallop, R., Thompson, D., Luther, D., Crits-Christoph, K., Jacobs, J., . . . Crits-Christoph, P. (2016). Comparative effectiveness of cognitive therapy and dynamic psychotherapy for major depressive disorder in a community mental health setting: a randomized clinical noninferiority trial. *JAMA psychiatry, 73*(9), 904-911.

Gillan, C. M., & Whelan, R. (2017). What big data can do for treatment in psychiatry. *CURRENT OPINION IN BEHAVIORAL SCIENCES, 18*, 34-42.

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics, 24*(1), 44-65.

Hamilton, M. (1960). A rating scale for depression. *Journal of neurology, neurosurgery, and psychiatry, 23*(1), 56.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). New York: Springer.

Hoekstra, H., Ormel, J., & De Fruyt, F. (2003). NEO-FFI Big Five persoonlijkheidsvragenlijst [NEO-FFI Big Five personality questionnaire]: Lisse. The Nederlands: Swets & Zeitlinger BV.

Hoffart, A., & Johnson, S. U. (2017). Psychodynamic and Cognitive-Behavioral Therapies Are More Different Than You Think: Conceptualizations of Mental Problems and Consequences for Studying Mechanisms of Change. *Clinical Psychological Science, 5*(6), 1070-1086. doi:10.1177/2167702617727096

Huibers, M. J., Cohen, Z. D., Lemmens, L. H., Arntz, A., Peeters, F. P., Cuijpers, P., & DeRubeis, R. J. (2015). Predicting Optimal Outcomes in Cognitive Therapy or Interpersonal Psychotherapy for Depressed Individuals Using the Personalized Advantage Index Approach. *PLoS One, 10*(11), e0140771.

Iniesta, R., Hodgson, K., Stahl, D., Malki, K., Maier, W., Rietschel, M., . . . Uher, R. (2018). Antidepressant drug-specific prediction of depression treatment outcomes from genetic and clinical variables. *Scientific reports, 8*(1), 5530.

Iniesta, R., Malki, K., Maier, W., Rietschel, M., Mors, O., Hauser, J., . . . Uher, R. (2016). Combining clinical variables to optimize prediction of antidepressant treatment outcomes. *J Psychiatr Res, 78*, 94-102. doi:10.1016/j.jpsychires.2016.03.016

Kapelner, A., & Bleich, J. (2016). bartMachine: Machine Learning with Bayesian Additive Regression Trees. *Journal of Statistical software, 70*(4), 1-40. doi:10.18637/jss.v070.i04

Keefe, J. R., Wiltsey Stirman, S., Cohen, Z. D., DeRubeis, R. J., Smith, B. N., & Resick, P. A. (2018). In rape trauma PTSD, patient characteristics indicate which trauma-focused treatment they are most likely to complete. *Depression and anxiety, 35*(4), 330-338.

Kessler, R. C. (2018). The potential of predictive analytics to provide clinical decision support in depression treatment planning. *Current opinion in psychiatry, 31*(1), 32-39.

Kessler, R. C., van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Ebert, D. D., . . . Zaslavsky, A. M. (2017). Using patient self-reports to study heterogeneity of treatment effects in major depressive disorder. *Epidemiol Psychiatr Sci, 26*(1), 22-36. doi:10.1017/S2045796016000020

Kikkert, M. J., Driessen, E., Peen, J., Barber, J. P., Bockting, C., Schalkwijk, F., . . . Dekker, J. J. (2016). The role of avoidant and obsessive-compulsive personality disorder traits in matching patients with major depression to cognitive behavioral and psychodynamic therapy: A replication study. *Journal of affective disorders, 205*, 400-405.

Kraemer, H. C. (2013). Discovering, comparing, and combining moderators of treatment on outcome after randomized clinical trials: a parametric approach. *Stat Med, 32*(11), 1964-1973. doi:10.1002/sim.5734

Kraemer, H. C., & Blasey, C. M. (2004). Centring in regression analyses: a strategy to prevent errors in statistical inference. *Int J Methods Psychiatr Res, 13*(3), 141-151.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nat Neurosci, 12*(5), 535-540. doi:10.1038/nn.2303

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*: Springer Science & Business Media.

Lutz, W., Zimmermann, D., Müller, V. N., Deisenhofer, A.-K., & Rubel, J. A. (2017). Randomized controlled trial to evaluate the effects of personalized prediction and adaptation tools on treatment outcome in outpatient psychotherapy: study protocol. *BMC psychiatry, 17*(1), 306.

Molenaar, P., Don, F., van den Bout, J., Sterk, F., & Dekker, J. (2009). *Cognitieve gedragstherapie bij depressie [Cognitive behavioral therapy for depression]*. The Netherlands: Bohn Stafleu van Loghum.

Moons, K. G., Donders, R. A., Stijnen, T., & Harrell, F. E., Jr. (2006). Using the outcome for imputation of missing predictor values was preferred. *J Clin Epidemiol, 59*(10), 1092-1101. doi:10.1016/j.jclinepi.2006.01.009

Niles, A. N., Loerinc, A. G., Krull, J. L., Roy-Byrne, P., Sullivan, G., Sherbourne, C. D., . . . Craske, M. G. (2017). Advancing Personalized Medicine: Application of a Novel Statistical Method to Identify Treatment Moderators in the Coordinated Anxiety Learning and Management Study. *Behavior therapy, 48*(4), 490-500.

Niles, A. N., Wolitzky-Taylor, K. B., Arch, J. J., & Craske, M. G. (2017). Applying a novel statistical method to advance the personalized treatment of anxiety disorders: A composite moderator of comparative drop-out from CBT and ACT. *Behav Res Ther, 91*, 13-23. doi:10.1016/j.brat.2017.01.001

Reiss, S., Peterson, R. A., Gursky, D. M., & McNally, R. J. (1986). Anxiety sensitivity, anxiety frequency and the prediction of fearfulness. *Behaviour research and therapy, 24*(1), 1-8.

Rizopoulos, D. (2009). BootStepAIC: bootstrap stepAIC. *R package version, 1*(0).

Schweizer, S., Cohen, Z., Hayes, R., DeRubeis, R., Crane, C., Kuyken, W., & Dalgleish, T. (*submitted*). Relapse Prevention for Antidepressant Medication (ADM) Responders with Recurrent Depression: Using the Personalized Advantage Index to Decide Between Maintenance ADM and Mindfulness-Based Cognitive Therapy.

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *Am J Epidemiol, 179*(6), 764-774. doi:10.1093/aje/kwt312

Sheehan, D., Janavs, J., Baker, R., Harnett-Sheehan, K., Knapp, E., Sheehan, M., . . . Amorim, P. (1998). MINI-Mini International neuropsychiatric interview-english version 5.0. 0-DSM-IV. *Journal of Clinical Psychiatry, 59*, 34-57.

Smagula, S. F., Wallace, M. L., Anderson, S. J., Karp, J. F., Lenze, E. J., Mulsant, B. H., . . . Lotrich, F. E. (2016). Combining moderators to identify clinical profiles of patients who will, and will not, benefit from aripiprazole augmentation for treatment resistant late-life major depressive disorder. *Journal of psychiatric research, 81*, 112-118.

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

Team, R. C. (2000). R language definition. *Vienna, Austria: R foundation for statistical computing*.

Turner, E. H., Matthews, A. M., Linardatos, E., Tell, R. A., & Rosenthal, R. (2008). Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine, 358*(3), 252-260.

Uher, R., Perlis, R., Henigsberg, N., Zobel, A., Rietschel, M., Mors, O., . . . Bajs, M. (2012). Depression symptom dimensions as predictors of antidepressant treatment outcome: replicable evidence for interest-activity symptoms. *Psychological medicine, 42*(5), 967-980.

Veeninga, A., & Hafkenscheid, A. (2004). KORT INSTRUMENTEEL De Patienten Behoeften Vragenlijst (PBV). *Gedragstherapie, 37*, 197-204.

Vittengl, J. R., Clark, L. A., Thase, M. E., & Jarrett, R. B. (2017). Initial Steps to inform selection of continuation cognitive therapy or fluoxetine for higher risk responders to cognitive therapy for recurrent major depressive disorder. *Psychiatry research, 253*, 174-181.

Wallace, M. L., Frank, E., & Kraemer, H. C. (2013). A novel approach for developing and interpreting treatment moderator profiles in randomized clinical trials. *JAMA psychiatry, 70*(11), 1241-1247.

Webb, C. A., Trivedi, M. H., Cohen, Z. D., Dillon, D. G., Fournier, J. C., Goer, F., . . . Parsey, R. (2018). Personalized prediction of antidepressant v. placebo response: evidence from the EMBARC study. *Psychological medicine*, 1-10.

World Health Organization. (2017). *Depression and Other Common Mental Disorders: Global Health Estimates* (CC BY-NC-SA 3.0 IGO). Retrieved from Geneva:

Zilcha-Mano, S. (2018). Major developments in methods addressing for whom psychotherapy may work and why. *Psychotherapy Research*, 1-16.

Zilcha-Mano, S., Keefe, J. R., Chui, H., Rubin, A., Barrett, M. S., & Barber, J. P. (2016). Reducing Dropout in Treatment for Depression: Translating Dropout Predictors Into

Individualized Treatment Recommendations. *The Journal of clinical psychiatry, 77*(12), e1584-e1590.

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

Chapter 3 References

Ammons, R. B., & Ammons, C. (1962). The Quick Test (QT): provisional manual. *Psychological Reports*.

Austin, P. C., & Tu, J. V. (2004). Bootstrap methods for developing predictive models. *The American Statistician, 58*(2), 131-137.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). Beck depression inventory-II. *San Antonio, 78*(2), 490-498.

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

Beck, A. T., Weissman, A., Lester, D., & Trexler, L. (1974). The measurement of pessimism: the hopelessness scale. *Journal of Consulting and Clinical Psychology, 42*(6), 861.

Berger, A. M., Knutson, J. F., Mehm, J. G., & Perkins, K. A. (1988). The self-report of punitive childhood experiences of young adults and adolescents. *Child Abuse & Neglect, 12*(2), 251-262.

Bernstein, E. M., & Putnam, F. W. (1986). Development, reliability, and validity of a dissociation scale. *The Journal of nervous and mental disease, 174*(12), 727-735.

Bisson, J., & Andrew, M. (2007). *Psychological treatment of post-traumatic stress disorder (PTSD)*: Wiley Online Library.

Bleich, J., Kapelner, A., George, E. I., & Jensen, S. T. (2014). Variable Selection for Bart: An Application to Gene Regulation. *Annals of Applied Statistics, 8*(3), 1750-1781. doi:10.1214/14-Aoas755

Bluestone, C. (2005). Personal disciplinary history and views of physical punishment: Implications for training mandated reporters. *Child abuse review, 14*(4), 240-258.

Bradley, R., Greene, J., Russ, E., Dutra, L., & Westen, D. (2005). A multidimensional meta-analysis of psychotherapy for PTSD. *American Journal of Psychiatry, 162*(2), 214-227.

Breheny, P., & Burchett, W. (2013). Visualization of regression models using visreg. *R Package*, 1-15.

Bremner, J. D., Krystal, J. H., Putnam, F. W., Southwick, S. M., Marmar, C., Charney, D. S., & Mazure, C. M. (1998). Measurement of dissociative states with the clinician-administered dissociative states scale (CADSS). *Journal of traumatic stress, 11*(1), 125-136.

Briere, J., Elliott, D. M., Harris, K., & Cotman, A. (1995). Trauma Symptom Inventory: Psychometrics and association with childhood and adult victimization in clinical samples. *Journal of interpersonal violence, 10*(4), 387-401.

Buysse, D. J., Reynolds, C. F., Monk, T. H., Berman, S. R., & Kupfer, D. J. (1989). The Pittsburgh Sleep Quality Index: a new instrument for psychiatric practice and research. *Psychiatry research, 28*(2), 193-213.

Carlson, E. B., & Putnam, F. W. (1993). An update on the dissociative experiences scale. *Dissociation: progress in the dissociative disorders*.

Carpenter, J. S., & Andrykowski, M. A. (1998). Psychometric evaluation of the Pittsburgh sleep quality index. *Journal of psychosomatic research, 45*(1), 5-13.

Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., & Trivedi, M. H. (2016). Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry, 3*. doi:10.1016/s2215-0366(15)00471-x

Cloitre, M., Bryant, R. A., & Schnyder, U. (2015). What Works for Whom? *Evidence Based Treatments for Trauma-Related Psychological Disorders* (pp. 499-512): Springer.

Cloitre, M., Petkova, E., Su, Z., & Weiss, B. (2016). Patient characteristics as a moderator of post-traumatic stress disorder treatment outcome: combining symptom burden and strengths. *BJPsych Open, 2*(2), 101-106. doi:10.1192/bjpo.bp.115.000745

Coffey, S. F., Gudmundsdottir, B., Beck, J. G., Palyo, S. A., & Miller, L. (2006). Screening for PTSD in motor vehicle accident survivors using the PSS-SR and IES. *Journal of traumatic stress, 19*(1), 119-128.

Cohen, Z. D., & DeRubeis, R. J. (2018). Treatment Selection in Depression. *Annual Review of Clinical Psychology, 14*.

Cohen, Z. D., Kim, T., Van, H. L., Dekker, J. J., & Driessen, E. (*under review*). Recommending cognitive-behavioral versus psychodynamic therapy for mild to moderate adult depression. *psyArXiv preprint at https://psyarxiv.com/njus6*. doi:DOI 10.17605/OSF.IO/6QXVE

Cook, J. M., Dinnen, S., Simiola, V., Thompson, R., & Schnurr, P. P. (2014). VA residential provider perceptions of dissuading factors to the use of two evidence-based PTSD treatments. *Professional Psychology: Research and Practice, 45*(2), 136.

Cook, J. M., Simiola, V., Hamblen, J. L., Bernardy, N., & Schnurr, P. P. (2017). The influence of patient readiness on implementation of evidence-based PTSD treatments in Veterans Affairs residential programs. *Psychological Trauma: Theory, Research, Practice, and Policy, 9*(S1), 51.

Cook, J. M., Thompson, R., Harb, G. C., & Ross, R. J. (2013). Cognitive− behavioral treatment for posttraumatic nightmares: An investigation of predictors of dropout and outcome. *Psychological Trauma: Theory, Research, Practice, and Policy, 5*(6), 545.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American psychologist, 34*(7), 571.

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

Deisenhofer, A. K., Delgadillo, J., Rubel, J. A., Böhnke, J. R., Zimmermann, D., Schwartz, B., & Lutz, W. (2018). Individual treatment selection for patients with posttraumatic stress disorder. *Depression and anxiety, 35*(6), 541-550.

Delgadillo, J., Huey, D., Bennett, H., & McMillan, D. (2017). Case complexity as a guide for psychological treatment selection. *Journal of Consulting and Clinical Psychology, 85*(9), 835.

DeRubeis, R. J., Cohen, Z. D., Forand, N. R., Fournier, J. C., Gelfand, L. A., & Lorenzo-Luaces, L. (2014). The Personalized Advantage Index: translating research on prediction into individualized treatment recommendations. A demonstration. *PLoS One, 9*(1), e83875. doi:10.1371/journal.pone.0083875

Fiedler, K. (2011). Voodoo correlations are everywhere—not only in neuroscience. *Perspectives on Psychological Science, 6*(2), 163-171.

Foa, E. B., Hembree, E. A., Cahill, S. P., Rauch, S. A., Riggs, D. S., Feeny, N. C., & Yadin, E. (2005). Randomized trial of prolonged exposure for posttraumatic stress disorder with and without cognitive restructuring: outcome at academic and community clinics. *Journal of Consulting and Clinical Psychology, 73*(5), 953.

Foa, E. B., Hembree, E. A., & Rothbaum, B. O. (2007). Prolonged exposure therapy for PTSD: Emotional processing of traumatic experiences – Therapist guide: New York, NY: Oxford University Press.

Foa, E. B., Riggs, D. S., Dancu, C. V., & Rothbaum, B. O. (1993). Reliability and validity of a brief instrument for assessing post-traumatic stress disorder. *Journal of traumatic stress, 6*(4), 459-473.

Foa, E. B., Riggs, D. S., Massie, E. D., & Yarczower, M. (1995). The impact of fear activation and anger on the efficacy of exposure treatment for posttraumatic stress disorder. *Behavior therapy, 26*(3), 487-499.

Foa, E. B., & Rothbaum, B. O. (1998). Treatment manuals for practitioners: Treating the trauma of rape: Cognitive-behavioral therapy for PTSD. New York: Guilford Press.

Foa, E. B., Zoellner, L. A., Feeny, N. C., Hembree, E. A., & Alvarez-Conrad, J. (2002). Does imaginal exposure exacerbate PTSD symptoms? *Journal of Consulting and Clinical Psychology, 70*(4), 1022.

Friedman, J., Hastie, T., & Tibshirani, R. (2009). glmnet: Lasso and elastic-net regularized generalized linear models. *R package version, 1*(4).

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical software, 33*(1), 1.

Garge, N. R., Bobashev, G., & Eggleston, B. (2013). Random forest methodology for model-based recursive partitioning: the mobForest package for R. *BMC Bioinformatics, 14*, 125. doi:10.1186/1471-2105-14-125

Gillan, C. M., & Whelan, R. (2017). What big data can do for treatment in psychiatry. *CURRENT OPINION IN BEHAVIORAL SCIENCES, 18*, 34-42.

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics, 24*(1), 44-65.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: a meta-analysis. *Psychological assessment, 12*(1), 19.

Hagenaars, M. A., van Minnen, A., & Hoogduin, K. A. (2010). The impact of dissociation and depression on the efficacy of prolonged exposure treatment for PTSD. *Behaviour research and therapy, 48*(1), 19-27.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). New York: Springer.

Hembree, E. A., Street, G. P., Riggs, D. S., & Foa, E. B. (2004). Do assault-related variables predict response to cognitive behavioral treatment for PTSD? *Journal of Consulting and Clinical Psychology, 72*(3), 531.

Huibers, M. J., Cohen, Z. D., Lemmens, L. H., Arntz, A., Peeters, F. P., Cuijpers, P., & DeRubeis, R. J. (2015). Predicting Optimal Outcomes in Cognitive Therapy or Interpersonal Psychotherapy for Depressed Individuals Using the Personalized Advantage Index Approach. *PLoS One, 10*(11), e0140771.

Iniesta, R., Malki, K., Maier, W., Rietschel, M., Mors, O., Hauser, J., . . . Uher, R. (2016). Combining clinical variables to optimize prediction of antidepressant treatment outcomes. *J Psychiatr Res, 78*, 94-102. doi:10.1016/j.jpsychires.2016.03.016

Janes, H., Pepe, M. S., Bossuyt, P. M., & Barlow, W. E. (2011). Measuring the performance of markers for guiding treatment decisions. *Annals of internal medicine, 154*(4), 253-259.

Kapelner, A., & Bleich, J. (2016). bartMachine: Machine Learning with Bayesian Additive Regression Trees. *Journal of Statistical software, 70*(4), 1-40. doi:10.18637/jss.v070.i04

Kautzky, A., Baldinger-Melich, P., Kranz, G. S., Vanicek, T., Souery, D., Montgomery, S., . . . Lanzenberger, R. (2017). A New Prediction Model for Evaluating Treatment-Resistant Depression. *The Journal of clinical psychiatry, 78*(2), 215-222.

Kautzky, A., Dold, M., Bartova, L., Spies, M., Vanicek, T., Souery, D., . . . Fabbri, C. (2017). Refining Prediction in Treatment-Resistant Depression: Results of Machine Learning Analyses in the TRD III Sample. *The Journal of clinical psychiatry, 79*(1).

Keefe, J. R., Wiltsey Stirman, S., Cohen, Z. D., DeRubeis, R. J., Smith, B. N., & Resick, P. A. (2018). In rape trauma PTSD, patient characteristics indicate which trauma-focused treatment they are most likely to complete. *Depression and anxiety, 35*(4), 330-338.

Kessler, R. C. (2018). The potential of predictive analytics to provide clinical decision support in depression treatment planning. *Current opinion in psychiatry, 31*(1), 32-39.

Kessler, R. C., van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Ebert, D. D., . . . Zaslavsky, A. M. (2017). Using patient self-reports to study heterogeneity of treatment effects in major depressive disorder. *Epidemiol Psychiatr Sci, 26*(1), 22-36. doi:10.1017/S2045796016000020

Koutsouleris, N., Kahn, R. S., Chekroud, A. M., Leucht, S., Falkai, P., Wobrock, T., . . . Hasan, A. (2016). Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. *The Lancet Psychiatry, 3*(10), 935-946.

Kraemer, H. C., & Blasey, C. M. (2004). Centring in regression analyses: a strategy to prevent errors in statistical inference. *Int J Methods Psychiatr Res, 13*(3), 141-151.

Kubany, E. S., Haynes, S. N., Abueg, F. R., Manke, F. P., Brennan, J. M., & Stahura, C. (1996). Development and validation of the Trauma-Related Guilt Inventory (TRGI). *Psychological assessment, 8*(4), 428.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*: Springer Science & Business Media.

Lancaster, C. L., Teeters, J. B., Gros, D. F., & Back, S. E. (2016). Posttraumatic stress disorder: overview of evidence-based assessment and treatment. *Journal of clinical medicine, 5*(11), 105.

Larsen, S. E., Stirman, S. W., Smith, B. N., & Resick, P. A. (2016). Symptom exacerbations in trauma-focused treatments: Associations with treatment outcome and non-completion. *Behaviour research and therapy, 77*, 68-77.

Maloney, M. P., Steger, H. G., & Ward, M. P. (1973). The Quick Test as a measure of general intelligence in an urban community psychiatric hospital. *Psychological Reports*.

McDevitt-Murphy, M. E., Weathers, F. W., & Adkins, J. W. (2005). The use of the Trauma Symptom Inventory in the assessment of PTSD symptoms. *Journal of traumatic stress, 18*(1), 63-67.

Meehl, P. E. (1954). Clinical versus statistical prediction: A theoretical analysis and a review of the evidence. doi:http://dx.doi.org/10.1037/11281-000

Morina, N., Wicherts, J. M., Lobbrecht, J., & Priebe, S. (2014). Remission from post-traumatic stress disorder in adults: A systematic review and meta-analysis of long term outcome studies. *Clinical psychology review, 34*(3), 249-255.

Mott, J. M., Stanley, M. A., Street, R. L., Jr., Grady, R. H., & Teng, E. J. (2014). Increasing engagement in evidence-based PTSD treatment through shared decision-making: a pilot study. *Mil Med, 179*(2), 143-149. doi:10.7205/MILMED-D-13-00363

Niles, A. N., Loerinc, A. G., Krull, J. L., Roy-Byrne, P., Sullivan, G., Sherbourne, C. D., . . . Craske, M. G. (2017). Advancing Personalized Medicine: Application of a Novel Statistical Method to Identify Treatment Moderators in the Coordinated Anxiety Learning and Management Study. *Behavior therapy, 48*(4), 490-500.

Niles, A. N., Wolitzky-Taylor, K. B., Arch, J. J., & Craske, M. G. (2017). Applying a novel statistical method to advance the personalized treatment of anxiety disorders: A composite moderator of comparative drop-out from CBT and ACT. *Behav Res Ther, 91*, 13-23. doi:10.1016/j.brat.2017.01.001

Osei-Bonsu, P. E., Bolton, R. E., Stirman, S. W., Eisen, S. V., Herz, L., & Pellowe, M. E. (2017). Mental health providers' decision-making around the implementation of evidence-based treatment for PTSD. *The journal of behavioral health services & research, 44*(2), 213-223.

Perkonigg, A., Pfister, H., Stein, M. B., Höfler, M., Lieb, R., Maercker, A., & Wittchen, H.-U. (2005). Longitudinal course of posttraumatic stress disorder and posttraumatic stress disorder symptoms in a community sample of adolescents and young adults. *American Journal of Psychiatry, 162*(7), 1320-1327.

Perlis, R. H. (2013). A clinical risk stratification tool for predicting treatment resistance in major depressive disorder. *Biol Psychiatry, 74*(1), 7-14. doi:10.1016/j.biopsych.2012.12.007

Petkova, E., Tarpey, T., Su, Z., & Ogden, R. T. (2017). Generated effect modifiers (GEM's) in randomized clinical trials. *Biostatistics, 18*(1), 105-118.

Pitman, R. K., Altman, B., Greenwald, E., Longpre, R. E., Macklin, M., Poire, R., & Steketee, G. (1991). Psychiatric complications during flooding therapy for posttraumatic stress disorder. *Journal of Clinical Psychiatry*.

Raza, G. T., & Holohan, D. R. (2015). Clinical treatment selection for posttraumatic stress disorder: Suggestions for researchers and clinical trainers. *Psychol Trauma, 7*(6), 547-554. doi:10.1037/tra0000059

Resick, P. A., Nishith, P., Weaver, T. L., Astin, M. C., & Feuer, C. A. (2002). A comparison of cognitive-processing therapy with prolonged exposure and a waiting condition for the treatment of chronic posttraumatic stress disorder in female rape victims. *Journal of Consulting and Clinical Psychology, 70*(4), 867.

Resick, P. A., & Schnicke, M. (1993). *Cognitive processing therapy for rape victims: A treatment manual* (Vol. 4): Sage.

Resick, P. A., Williams, L. F., Suvak, M. K., Monson, C. M., & Gradus, J. L. (2012). Long-term outcomes of cognitive–behavioral treatments for posttraumatic stress disorder among female rape survivors. *Journal of Consulting and Clinical Psychology, 80*(2), 201.

Rizopoulos, D. (2009). BootStepAIC: bootstrap stepAIC. *R package version, 1*(0).

Rizvi, S. L., Vogt, D. S., & Resick, P. A. (2009). Cognitive and affective predictors of treatment outcome in cognitive processing therapy and prolonged exposure for posttraumatic stress disorder. *Behaviour research and therapy, 47*(9), 737-743.

Rosen, C. S., Clothier, B., Noorbaloochi, S., Smith, B. N., Orazem, R., & Sayer, N. (2017). *Which veterans receive evidence-based psychotherapy for PTSD?* Paper presented at the 32nd Annual Meeting of the International Society of Traumatic Stress Studies, Chicago, IL.

Rosen, C. S., Matthieu, M., Stirman, S. W., Cook, J., Landes, S., Bernardy, N., . . . Finley, E. (2016). A review of studies on the system-wide implementation of evidence-based psychotherapies for posttraumatic stress disorder in the Veterans Health Administration. *Administration and Policy in Mental Health and Mental Health Services Research, 43*(6), 957-977.

Rowan, A. B., Foy, D. W., Rodriguez, N., & Ryan, S. (1994). Posttraumatic stress disorder in a clinical sample of adults sexually abused as children. *Child Abuse Negl, 18*(1), 51-61.

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.

Schnurr, P. P., Chard, K. M., Ruzek, J. I., Chow, B. K., Shih, M.-C., Resick, P. A., . . . Lu, Y. (2015). Design of VA Cooperative Study #591: CERV-PTSD, Comparative Effectiveness Research in Veterans with PTSD. *Contemporary Clinical Trials, 41*(Supplement C), 75-84. doi:https://doi.org/10.1016/j.cct.2014.11.017

Schweizer, S., Cohen, Z., Hayes, R., DeRubeis, R., Crane, C., Kuyken, W., & Dalgleish, T. (*submitted*). Relapse Prevention for Antidepressant Medication (ADM) Responders with Recurrent Depression: Using the Personalized Advantage Index to Decide Between Maintenance ADM and Mindfulness-Based Cognitive Therapy.

Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O., & Hemingway, H. (2014). Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *Am J Epidemiol, 179*(6), 764-774. doi:10.1093/aje/kwt312

Simon, G. E., & Perlis, R. H. (2010). Personalized medicine for depression: can we match patients with treatments? *American Journal of Psychiatry, 167*(12), 1445-1455.

Smagula, S. F., Wallace, M. L., Anderson, S. J., Karp, J. F., Lenze, E. J., Mulsant, B. H., . . . Lotrich, F. E. (2016). Combining moderators to identify clinical profiles of patients who will, and will not, benefit from aripiprazole augmentation for treatment resistant late-life major depressive disorder. *Journal of psychiatric research, 81*, 112-118.

Spielberger, C. D., Jacobs, G., Russell, S., & Crane, R. S. (1983). Assessment of anger: The state-trait anger scale. *Advances in personality assessment, 2*, 159-187.

Spielberger, C. D., & Sydeman, S. J. (1994). State-Trait Anxiety Inventory and State-Trait Anger Expression Inventory.

Sripada, R. K., Bohnert, K. M., Ganoczy, D., & Pfeiffer, P. N. (2017). Documentation of Evidence-Based Psychotherapy and Care Quality for PTSD in the Department of Veterans Affairs. *Administration and Policy in Mental Health and Mental Health Services Research*, 1-9.

Sripada, R. K., Pfeiffer, P. N., Rauch, S. A. M., Ganoczy, D., & Bohnert, K. M. (*under review*). Who Gets Evidence-Based Treatment? Factors Associated with Receipt of Evidence-Based Psychotherapy for PTSD in VA.

Steinert, C., Hofmann, M., Leichsenring, F., & Kruse, J. (2015). The course of PTSD in naturalistic long-term studies: High variability of outcomes. A systematic review. *Nordic journal of psychiatry, 69*(7), 483-496.

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

Straus, M. A., Hamby, S. L., Boney-McCoy, S., & Sugarman, D. B. (1996). The revised conflict tactics scales (CTS2) development and preliminary psychometric data. *Journal of family issues, 17*(3), 283-316.

Van IJzendoorn, M. H., & Schuengel, C. (1996). The measurement of dissociation in normal and clinical populations: Meta-analytic validation of the Dissociative Experiences Scale (DES). *Clinical psychology review, 16*(5), 365-382.

Vittengl, J. R., Clark, L. A., Thase, M. E., & Jarrett, R. B. (2017). Initial Steps to inform selection of continuation cognitive therapy or fluoxetine for higher risk responders to cognitive therapy for recurrent major depressive disorder. *Psychiatry research, 253*, 174-181.

Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J., Balis, U., . . . Higgins, P. D. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ Open, 3*(8), e002847.

Wallace, M. L., Frank, E., & Kraemer, H. C. (2013). A novel approach for developing and interpreting treatment moderator profiles in randomized clinical trials. *JAMA psychiatry, 70*(11), 1241-1247.

Watts, B. V., Schnurr, P. P., Mayo, L., Young-Xu, Y., Weeks, W. B., & Friedman, M. J. (2013). Meta-analysis of the efficacy of treatments for posttraumatic stress disorder. *The Journal of clinical psychiatry, 74*(6), e541-550.

Weathers, F. W., Ruscio, A. M., & Keane, T. M. (1999). Psychometric properties of nine scoring rules for the Clinician-Administered Posttraumatic Stress Disorder Scale. *Psychological assessment, 11*(2), 124.

Webb, C. A., Trivedi, M. H., Cohen, Z. D., Dillon, D. G., Fournier, J. C., Goer, F., . . . Parsey, R. (2018). Personalized prediction of antidepressant v. placebo response: evidence from the EMBARC study. *Psychological medicine*, 1-10.

Wiltsey Stirman, S., Lunney, C., Cohen, Z., DeRubeis, R., Wiley, J., & Schnurr, P. (*submitted*). A Prognostic Index to Inform Selection of a Trauma-Focused or Non-Trauma-Focused Treatment for PTSD.

Wohlfarth, T. D., van den Brink, W., Winkel, F. W., & ter Smitten, M. (2003). Screening for Posttraumatic Stress Disorder: an evaluation of two self-report scales among crime victims. *Psychological assessment, 15*(1), 101.

Zilcha-Mano, S., Keefe, J. R., Chui, H., Rubin, A., Barrett, M. S., & Barber, J. P. (2016). Reducing Dropout in Treatment for Depression: Translating Dropout Predictors Into Individualized Treatment Recommendations. *The Journal of clinical psychiatry, 77*(12), e1584-e1590.

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