Project Title: Predicting Optimal Treatment Outcomes for Patients with Complex Anxiety Disorders

Authors and Affiliations:
M. S. Heimvik
M. A. Helmich
S. U. Johnson
K. M. Pålerud
R. Hagen
A. Hoffart

Univerisity of Oslo, Complexity in treatment Outcome, Psychopathology and Epidemiology (COPE), Modum Bad

Contact details:
margrsh@uio.no

Preregistration of analyses of preexisting data.

1. **What is the hypothesis that will be investigated?**

Anxiety disorders, despite their acknowledged burden, remain among the most prevalent untreated psychiatric conditions (Bandelow & Michaelis, 2015). While group-level effective treatments for anxiety disorders are accessible, a noteworthy challenge persists in clinical practice: variability in individual treatment responses, indicating that many patients show limited improvement. One compelling explanation for this variability is provided by aptitude-by-treatment interactions (ATI), where the effectiveness of various psychological therapy models, components, or techniques is seen as contingent on the specific characteristics of the patients (Nye et al., 2023). Recent studies have employed machine learning (ML) techniques to optimise this personalisation process. In an early study, Lutz et al. (2005) utilised nearest-neighbour modelling to estimate the rate of symptom change and variability across therapy sessions. Later, DeRubeis et al. (2014) developed the Personalised Advantage Index (PAI), an algorithm that predicts individualised treatment outcomes. The PAI aims to predict how patients would react to two different types of treatment based on the interaction effects between baseline variables and treatment conditions. Other research groups have developed similar approaches (Lutz et al., 2019; Schwartz et al., 2021). More recently, tree-based techniques have been applied to improve these methods further, intending to incorporate more variables and generate more accurate predictions (Keefe et al., 2018; Schwartz et al., 2021).

Cognitive behavioural therapy (CBT) and metacognitive therapy (MCT) are established treatments for anxiety disorders, but it is still unclear which therapy is more beneficial for given individuals (Bystritsky et al., 2013; Wells et al., 2020). This study aims to analyse an extensive collection of clinical data to determine the personal effectiveness of CBT and MCT for complex anxiety disorders. By analysing real-world data from patients who received CBT or MCT, we will examine the moderating effects of factors such as patient characteristics and treatment interactions. The goal is to identify the therapy that provides the most significant benefit for each patient, helping clinicians make informed treatment decisions.

Moreover, we aim to investigate whether the implicit preference for assigning more complex clinical presentations (i.e., increased number and type of comorbidities and chronicity of disorders) to MCT is evident in the PAI scores. Since MCT is applied in a transdiagnostic approach and CBT is applied in a disorder-specific approach at the clinic, we anticipate that

MCT will better serve cases with more complex clinical presentations, as shown by higher PAI scores signifying improved outcomes. By examining the relationship between implicit treatment assignment and PAI scores, we seek insights into how clinical presentation complexity may influence treatment allocation decisions. This analysis can help deepen our understanding of how treatment decisions are made in practice and their potential impact on treatment outcomes.

RQ: Can routinely collected baseline measures be utilised to predict treatment outcomes in complex anxiety disorders, using ML to enhance the process of optimal treatment selection?

Primary hypothesis: Patients who received their optimal treatment based on their PAI prediction would have reduced anxiety measures at the end of treatment compared to those who received their non-optimal treatment.

Secondary hypothesis: Patients with more complex clinical presentations will have MCT assigned as their optimal treatment, and their PAI scores will indicate a greater advantage in terms of treatment outcomes in MCT (PAI properties discussed in question six).

2. How will the crucial variables be operationalised?

To operationalise the crucial variables in the hypotheses, the following measurements will be used:

- Treatment assignment: This is a retrospective analysis, meaning treatments have already been conducted according to the treatment assignment procedure at the clinic. Therefore, the sample is quasi-randomised. We will examine the records to identify which patients were assigned to their model-predicted optimal versus non-optimal treatment.
- PAI scores: PAI scores will be calculated as a quantitative measure of the predicted advantage of the optimal treatment over the non-optimal treatment for each individual (see question six for further explanation).
- Complexity of cases: Complexity will be operationalised by using established criteria or assessments commonly used in clinical practice. This includes factors such as the severity and chronicity of the anxiety disorders, as well as the presence and number of comorbid disorders (see question seven for further explanation).

Measures:

This study aims to use quality register data (routinely collected data) for the analysis. Nine questionnaires were administered at admission, and eight were administered post-treatment. The primary outcome measure is symptom severity, which will be assessed by Beck Anxiety Inventory (BAI) ratings. See below for a comprehensive overview of each questionnaire as well as a timeline of data collection.

The measurement timeline is as follows:

- Assessment measurements: These are taken weeks to months before the patient is admitted to the clinic. These measurements are conducted during a five-day assessment, when the patient is referred or applies to the clinic.
- Admission measurements: These measurements are taken directly before the patient begins treatment at the clinic. They are collected when the patient arrives at the clinic.
- Post-treatment measurements: These measurements are taken after the patient has completed their treatment.

Measure	Measured at		
	Assessment	**Admission**	**Post-treatment**
BAI	X	X	X
SCL-90	X	X	X
IIP-64	X	X	X
BDI	X	X	X
PSWQ	X	X	X
ATQ-23	X	X	X
MCQ-30	X	X	X
Therapist registration form	X	X	X
Patient registration form	X		

Note. BAI = Beck Anxiety Inventory (Beck et al., 1988); SCL-90 = Symptom Checklist-90 (Derogatis, 1983); IIP-64 = Inventory of Interpersonal Problems-64-Circumplex (Horowitz et al., 1988); BDI = Beck Depression Inventory (Beck et al., 1996); PSWQ = Penn State Worry Questionnaire (Meyer et al., 1990); ATQ-23 = Anxious Thought Questionnaire-23 (Johnson et al., 2018); MCQ-30 = Meta-Cognitions Questionnaire 30 (Wells & Cartwright-Hatton, 2004), and see question 6 for more information about the contents of the therapist and patient registration form.

3. What is the source of the data included in the analyses?

The data for this study was collected from individuals referred for treatment at the Department of Anxiety Disorders at Modum Bad Psychiatric Centre in Norway between 2016 and 2024. Modum Bad is a specialised hospital with an inpatient program designed for individuals with treatment-resistant anxiety disorders.

4. How will this data be obtained?

Data access was requested by Sverre Urnes Johnson from the data protection officer at Modum Bad. The data is part of the quality register at Modum Bad, and this use of data does not require any specific approval from the Regional Ethical Committee (REK Southeast (99244)). Modum Bad Psychiatric Centre permitted all the researchers to use the dataset for the current study. Subsets of the data have been previously analysed in several articles (Johnson et al., 2017, 2018; Johnson & Hoffart, 2019; Pålerud et al., n.d.).

5. Are there any exclusion criteria for the data?

Patients: Inclusion and exclusion criteria for patients at Modum Bad are detailed in (Johnson et al., 2017; Pålerud et al., n.d.).

Statistical analysis: Generally, we will adopt an intention-to-treat approach, thus including cases in which missing data could be imputed rather than deleting them listwise. We will use person-mean imputation for questionnaires with ≤30% missing answers, using the patient's available observations. Questionnaires with over 30% missing values will be considered missing to maintain the integrity of the data analysis. If a patient is missing admission scores, these will be substituted with their assessment scores. For patients missing fewer than two questionnaire sum scores, we will impute the missing scores using the missForest method (Stekhoven & Bühlmann, 2012). Individuals completely missing more than two questionnaire sum scores will be excluded from the analysis. We will include personal characteristics such as age, gender, and chronicity, if these variables have missing values, they will be left unaltered.

6. What are the planned statistical analyses?

This study aims to replicate previously used methods to analyse the comparative advantage of treatments for a given individual (Keefe et al., 2018; van Bronswijk et al., 2021; Zilcha-Mano et al., 2016). In this case, CBT vs. MCT for a patient with complex anxiety disorder using their baseline measures as predictors for end-of-treatment symptom severity. The analysis will use a two-stage ML method: (1) a tree-based method for variable selection and then (2) the PAI approach to test the comparative advantage of one treatment over another (DeRubeis et al., 2014; Garge et al., 2013). The first analysis stage applies a bootstrap aggregation of model-based recursive partitioning by the random forest algorithm called mobForest (MoB) analysis (Garge et al., 2013). This non-parametric approach is often used in classification problems involving numerous predictor variables (features) and complex interactions. This statistical method aims to construct a reliable predictive model and to discover the importance of different predictor variables in forecasting the primary outcome measure. This primary outcome measure, the end-BAI, is the foundation for subsequent analyses as it represents the central variable of interest. After identifying these variables, the second stage involves building a multivariate statistical model, otherwise known as the personalized advantage index (PAI), involving interaction terms representing prescriptive variables (DeRubeis et al., 2014). Then, the PAI for a specific patient is determined as the difference between their predicted outcomes for two treatments, namely, treatment A (CBT) and treatment B (MCT).

While there are no definitive guidelines on the appropriate dataset size for ML, some suggest that having 300 patients per treatment arm might be adequate (Luedtke et al., 2019). Equally, a widely recognized standard for reliable predictive modelling posits a minimum requirement of 10 events per variable, a criterion that even robust methods like logistic regression endorse (van der Ploeg et al., 2014). Given these considerations, the current sample size may be inadequate for ML analysis, rendering this study more of a preliminary exploration to identify potential avenues and prospects for future research.

Variable input

To reduce the dimensionality of the dataset, the items from each questionnaire will be inputted into the random forest model as follows:

Questionnaire	Dimensions	References
BAI	Sum Score	(Beck & Steer, 1990)
SCL-90	Somatization, Obsessive-Compulsive, Interpersonal Sensitivity, Depression, Anxiety, Hostility, Phobic Anxiety, Paranoid Ideation, and Psychoticism	(Kostaras et al., 2020)

	Global indexes: Global Severity Index (GSI), Positive Symptom Distress Index (PSDI), Positive Symptoms Total (PST)	
IIP-64	Domineering / Controlling (PA), Vindictive / Self-Centred (BC), Cold / Distant (DE), Socially Inhibited (FG), Non-assertive (HI), Overly accommodating/ Exploitable (JK), Self-Sacrificing (LM), Intrusive/Needy (NO)	(Alden et al., 1990; Bush et al., 2012)
BDI-II	Cognitive-Affective (CA), Somatic-Affective (SA)	(Aasen, 2001; Beck et al., 1996)
PSWQ	Core Worries, Uncontrollability of Worry, Worry Engagement	(Pallesen et al., 2006)
ATQ-23	Sum Score	(Johnson & Hoffart, 2013)
MCQ-30	Cognitive confidence, Positive Beliefs about worry, Cognitive Self-Consciousness, Negative beliefs about uncontrollability and danger, Beliefs about need to control thought	(Wells & Cartwright-Hatton, 2004)
Therapist registration form	No alterations: The therapist registration form includes detailed questions to gather information regarding the patient's admission and discharge, assigned treatment group, drop-out details, history of abuse and addiction, medication history, and psychological assessments. The form also encompasses items on the patient's GAF function and symptom levels, ICD-10 diagnosis, MINI and SCID-II interview, and personality disorder diagnoses based on the SCID-II survey.	
Patient registration form	No alterations: The patient register form includes important demographic and background information about the patient such as birth year, marital status, number of children under the age of 18, information on work and income situation (i.e., paid work (full/part-time), sick pay, work settlement allowance, disability benefit, student loans, etc.), recent employment status and duration, sick leave history, previous treatment for mental illness, and duration of symptoms.	

All values will be centred as per Kraemer & Blasey's (2004) recommendations. For continuous independent variables, mean-centring will be done by subtracting the mean value of each variable from all its observed values, aligning the mean of the distribution with zero. The mean for continuous independent variables is calculated as the average value of a variable across all participants included in the analysis. For dichotomous independent variables (including "treatment"), dummy coding will be set as 1/2 and −1/2, ensuring that the mean of these dichotomous independent variables is also centred around zero (Kraemer & Blasey, 2004).

<u>Stage one:</u> Variable selection

As previously stated, variable selection will be done with mobForest. Random forests create multiple bootstrap samples from the original data and build regression tree models on each of these samples. These trees search for binary splits within the data, considering random subsets of predictors at each split, intending to maximize group similarity. The goal is to find splits where one side demonstrates significantly different model parameters related to the outcome compared to the other side. The data is divided based on the variable with the strongest impact as a moderator (Garge et al., 2013). Predictions for new data are obtained by averaging predictions from all trees, thus reducing prediction variance (Gatnar, 2008).

Tuning parameters:

Setting up forest controls: The analysis will start off with default settings to set up the forest (Garge et al., 2013).

Possible adjustments: While earlier studies have indicated that random forests do not tend to overfit as the number of trees increases, more recent research has uncovered a bias in favour of certain types of variables (Breiman, 2000). This bias includes preferences for variables with specific characteristics, such as those with numerous categories, numeric variables, many missing values, and correlated variables (Hothorn et al., 2006; Kim & Loh, 2001; Strobl et al., 2008). Thus, although it is commonly observed that the default value of 'mtry' is typically the best choice for accurate predictions in real-world studies, it has been suggested that when dealing with correlated predictor variables, it is beneficial to explore various values of 'mtry' (Strobl et al., 2008).

Some adjustments that will be considered in the current study:

- Zilcha-Mano et al. (2016), tuning parameters. 20,000 bootstrapped replicates, set the minimum α level for splits at 0.10 and established a minimum of 15 patients as a prerequisite for considering a node for splitting.
- Similarly, Keefe et al. (2018) also conducted 20,000 bootstrapped replicates, ensuring that each split involved a minimum of 10 patients with a significance level of less than 0.10. Furthermore, they selected 10 random variables for each node.
- In contrast, van Bronswijk et al. (2021) generated a total of 10,000 trees, and for each split, they required a minimum significance level of 0.10 and a minimum subgroup size of 15 individuals.

Furthermore, this study plans to use the framework promoted by Hothorn et al. (2006), which enables the use of p-values for variable selection and as a stopping criterion which will eliminate the need to prune trees (Hothorn et al., 2006). These p-values offer the advantage of being comparable across variables of varying types.

Variable importance:

The next part of the analysis includes assessing the predictor variables' importance and selecting relevant variables for further exploration. To achieve this, several tests will be conducted. First, the models will be aggregated; we will test the mean rule and majority voting combination rules for class labels (Gatnar, 2008).

Calculating variable importance scores:

The default "permutation accuracy importance" method calculates variable importance scores using the default method found in the MoB package (Breiman, 2001; Genuer et al., 2010). While this approach is straightforward, it may introduce biases by favouring correlated predictor variables over uncorrelated ones. To address this issue, the conditional importance test, as proposed by Strobl et al. (2009), will be utilised. A comparison of results between the two methods will be made to evaluate the implications of the chosen variable importance calculation method.

Variable selection:

A conservative approach will be applied to select the variables from the results obtained from the variable importance scores; Variables with importance scores that are either negative, zero, or positive but fall within the same range as the negative values are considered for exclusion from further investigation (Strobl et al., 2009). The rationale behind this approach is that variables with such importance scores are more likely to be irrelevant, as their importance fluctuates around zero. Variables with positive importance scores exceeding this range are deemed informative and thus retained for further exploration. We will also test a less conservative approach, selecting variables if their variable importance score exceeds the threshold of the absolute value of the lowest-ranking variable (van Bronswijk et al., 2021).

Visualization:

Acknowledging the problematic statistical properties associated with significance tests in this context, particularly their dependence on the arbitrarily chosen number of trees in the ensemble ('ntree') for importance averaging, we have decided not to report the exact variable importance scores. As highlighted by Strobl et al. (2009), such reporting can potentially lead to misleading interpretations. Hence, in this study, our focus will be on visualising the ranking of the most important features, rather than values themselves. To facilitate a comprehensive understanding of feature interactions and the model's performance, we will use several visualization techniques including plots for correlation matrix, confusion matrix, AUC, partial dependence, calibration, and contribution.

Performance and error estimations:

This will be addressed by assessing the out-of-bag (OOB) accuracy. OOB accuracy provides a reliable measure of predictive accuracy using cases unseen during model training, ensuring a robust evaluation of model performance.

Exploratory variable selection:

Incorporating the identified tuning parameters from the random forest, we aim to explore the applicability of XG Boost, a gradient-boosted regression tree for variable selection (Chen & Guestrin, 2016; Elith et al., 2008). Studies have shown that XG Boost tends to outperform the MoB Forest analysis in terms of predictive accuracy. However, to our knowledge, its application to datasets like ours has yet to be conducted. Given its common use in very large datasets, we recognise the need first to test the tuning parameters on a random forest model to gain insights into the optimal parameter settings before employing the gradient boosting technique on our dataset (Sommer et al., 2019; Sterner et al., 2021). Following the development of the XG Boost model, we will compare the predictive accuracy of both the XG Boost and random forest model to assess their respective performance in our specific dataset.

<u>Stage two:</u> Comparative advantage analysis (PAI)

The following is based on DeRubeis' (2004) PAI analysis. In general, the second stage involves generating predictions using a regression model that incorporates the predictors identified above, using a leave-one-out cross-validation strategy (Efron & Gong, 1983; Harrell Jr et al., 1996). This procedure, also known as jackknife, requires building N models, each with a sample size of (N-1) (Abdi & Williams, 2010). The expected treatment outcome for an unadministered therapy to a patient is derived from the data of individuals in the alternative treatment group. Therefore, each patient receives two distinct predictions: one for MCT and one for CBT. This analysis results in two types of predictions: the "factual prediction", which estimates the end-BAI score for the treatment that each patient received, and the "counterfactual prediction", which is the estimate of the patient's end-BAI score in the treatment they did not receive. Both predictions are generated using the same model, with end-BAI as the dependent variable. Independent variables include main effects for

"treatment", prognostic and prescriptive variables, as well as terms representing the interactions between "treatment" and the prescriptive variables. Prognostic variables predict end-BAI scores irrespective of the treatment, whereas prescriptive variables predict differential outcomes based on whether the treatment was MCT or CBT (DeRubeis et al., 2014). For each of the N patients, the factual prediction is calculated by inputting the patient's observed values for all independent variables into the prediction model.

Calculating counterfactuals:
With values centred as per Kraemer & Blasey's (2004) recommendations, we can then calculate each patient's counterfactual prediction by substituting the value of the other treatment (either 1/2 or -1/2 based on the patient's actual assignment) in the "treatment" main effect term and all terms representing interactions of "treatment" with the prescriptive variables. If a patient received MCT (coded as 1/2), their counterfactual prediction would be calculated by substituting the treatment variable in the model with −1/2, representing the outcome if they had received CBT instead. This process will be applied to both the main effect of treatment and the interaction terms involving treatment and other prescriptive variables in the model. This counterfactual approach allows us to predict what would happen to each patient's outcome if assigned to the alternative treatment, thereby enabling a comparison between the treatment actually received and the potential outcome of the alternative treatment. This method ensures that the predictions are unbiased because each patient's outcome is predicted without using their own data in the model that generates their prediction (Efron & Gong, 1983). Furthermore, the accuracy of this set of predictions reflects what would be expected if the same procedure is used to predict outcomes in another set of patients randomly drawn from the same patient population, assuming they would be assigned to treatments in the same random manner (Abdi & Williams, 2010).

Calculating PAI:
Next, we will evaluate the PAI value for each patient. The PAI for each patient is calculated as the difference between the two predicted scores (counterfactual and factual). Essentially, it represents the numerical advantage (in units of the outcome measure, the BAI score) of one treatment over the other for that specific patient.
PAI properties:
Using these predicted scores, three properties of the predictions will be calculated and assessed:

1. The "true error" of the factual prediction: This is the mean of the absolute difference between the observed outcomes (actual treatment results) and the predicted outcomes (factual predictions) for each patient. It measures the accuracy of the prediction model in estimating the actual treatment outcomes. A lower true error indicates that the model's predictions are close to the actual results.
2. The standard error of the set of predictions: Quantifies the variability or dispersion of the predicted scores around the mean predicted value. It helps to understand the reliability of the prediction; a smaller standard error suggests that the predictions are consistently close to the mean prediction, indicating a more precise model.
3. The magnitude of the predicted difference (PAI) for each patient: This represents the predicted difference in treatment outcomes between receiving the treatment predicted to be more effective (optimal) and the other treatment (non-optimal).

Testing the utility of PAI:
Finally, we will test the utility of PAI by comparing the mean observed difference in end-BAI units of patients randomly assigned to their 'optimal' treatment (based on PAI) versus those

assigned to their 'non-optimal' treatment. This comparison helps validate whether the PAI can effectively predict which treatment will be more beneficial for a specific patient.

7. What are the criteria for confirming and disconfirming the hypotheses?

Primary hypothesis: Patients who received their optimal treatment based on their PAI prediction would have reduced anxiety measures at the end of treatment compared to those who received their non-optimal treatment.

To test this hypothesis, we will analyse the PAI indices in the following:

1. We will assess the true error of the factual predictions to determine the accuracy of the predictions. We define one standard deviation of treatment outcomes as the benchmark.
2. We will examine the standard error of the predictions to evaluate the precision and reliability of the PAI.
3. Additionally, we will use the adjusted Jacobsen-Traux method (reliable change index, RCI) to measure the significance and magnitude of the treatment effect, comparing the improvement in anxiety symptoms between patients receiving their optimal versus those receiving their non-optimal treatment.
    - The RCI, the gold standard for measuring clinically significant change, is defined by a return to normal criterion (Jacobson & Revenstorf, 1988; Jacobson & Truax, 1991). However, this criterion often falls short in practicality, especially in inpatient settings and among patients with comorbid or chronic disorders. Therefore, we will consider adjustments to the calculation of clinical change:
        - While some previous researchers have used or recommended 1 SD as a normative change criterion for clinical significance, the minimally important differences (MID) research by Norman et al. (2003) supports reducing this criterion to .5 SD, considering the specific patient population. Findings show that a .5 standard deviation consistently identifies a reliable change in chronic medical patients, regardless of the disease or measurement tool (Norman et al., 2003).
        - Additionally, different confidence levels could be incorporated into the RCI formula; for example, a cutoff of 1.96 (i.e., 95% confidence level) could be applied to less stringent or non-standardised clinical significance variables, while a cutoff of .84 (i.e., 80% confidence level) might be appropriate for standardised tools like symptom rating scales (Wise, 2003, 2004). Therefore, we will use a cutoff of .84 to assess reliable change for our primary outcome measure, the BAI.

Secondary hypothesis: Patients with more complex clinical presentations will have MCT assigned as their optimal treatment, and their PAI scores will indicate a greater advantage in terms of treatment outcomes in MCT.

To test this hypothesis, we will use two approaches.

Traditional approach: We define complexity as commonly used in clinical practice and conduct a linear regression where PAI scores are regressed on the complexity of variables. This assesses whether higher complexity is associated with higher PAI scores, indicating greater advantage in MCT:

- We define complexity as:
    - Number of Diagnoses: The total count of diagnosed disorders (counts).
    - Comorbidity: Presence of mood disorder (binary; yes/no).

- Additional Complexity: Presence of a personality disorder (binary; yes/no).
- Chronicity: Duration of the problem (continuous; measured in years).

Exploratory data-driven approach: We plan to examine the number of features utilised by the PAI for each individual case. The number of these features considered will define the complexity of each case. Cases that use a larger number of features can be deemed highly complex, as the model requires more variables to accurately predict outcomes. We will then explore whether these cases with higher complexity are associated with higher PAI scores, indicating a greater advantage in MCT.

8. Have the analyses been validated on a subset of the data? If yes, please specify and provide the relevant files.

The analysis has not yet been validated on a subset of the data.

9. What is known about the data that could be relevant for the tested hypotheses?

Given that this subset of the data has yet to be analysed, there is no pre-existing knowledge about the mean values of the variables or any other characteristics of this sample. Therefore, we are starting the analysis without prior insights into the specific metrics or attributes of this data set.

10. Please provide a brief timeline for the different steps in the preregistration.

The analyses are scheduled for summer 2024, with the manuscript anticipated by winter 2024-2025, and submission to a scientific journal planned by spring 2025.

This preregistration template is based on the following article:
Mertens, G., & Krypotos, A. M. (2019). Preregistration of analyses of preexisting data. *Psychologica Belgica*, *59*(1), 338-352. http://doi.org/10.5334/pb.493

References:

Aasen, H. (2001). *An Empirical Investigation of Depression Symptoms: Norms, Psychometric Characteristics and Factor Structure of the Beck Depression Inventory-II* [Master thesis, The University of Bergen]. https://bora.uib.no/bora-xmlui/handle/1956/1773

Abdi, H., & Williams, L. J. (2010). Jackknife. *Encyclopedia of Research Design*, *2*, 655–660.

Alden, L. E., Wiggins, J. S., & Pincus, A. L. (1990). Construction of circumplex scales for the Inventory of Interpersonal Problems. *Journal of Personality Assessment*, *55*(3–4), 521–536. https://doi.org/10.1207/s15327752jpa5503&4_10

Bandelow, B., & Michaelis, S. (2015). Epidemiology of anxiety disorders in the 21st century. *Dialogues in Clinical Neuroscience*, *17*(3), 327–335. https://doi.org/10.31887/DCNS.2015.17.3/bbandelow

Beck, A. T., Epstein, N., Brown, G., & Steer, R. A. (1988). An inventory for measuring clinical anxiety: Psychometric properties. *Journal of Consulting and Clinical Psychology*, *56*(6), 893.

Beck, A. T., & Steer, R. A. (1990). Manual for the Beck anxiety inventory. *San Antonio, TX: Psychological Corporation*.

Beck, A. T., Steer, R. A., & Brown, G. (1996). Beck depression inventory–II. *Psychological Assessment*.

Breiman, L. (2000). Bias, Variance , And Arcing Classifiers. *Technical Report 460, Statistics Department, University of California*.

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

Bush, A. L., Patel, A. B., Allen, J. G., Teal, C., Latini, D. M., Ellis, T. E., Herrera, S., & Frueh, B. C. (2012). Factor Structure and Convergent Validity of the Inventory of Interpersonal Problems in an Inpatient Setting. *Journal of Psychiatric Practice®*, *18*(3), 145. https://doi.org/10.1097/01.pra.0000415072.36121.2d

Bystritsky, A., Khalsa, S. S., Cameron, M. E., & Schiffman, J. (2013). Current Diagnosis and Treatment of Anxiety Disorders. *Pharmacy and Therapeutics*, *38*(1), 30–57.

Centring in regression analyses: A strategy to prevent errors in statistical inference—Kraemer—2004—International Journal of Methods in Psychiatric Research—Wiley Online Library. (n.d.). Retrieved 16 January 2024, from https://onlinelibrary-wiley-com.ezproxy.uio.no/doi/abs/10.1002/mpr.170

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Derogatis, L. R. (1983). SCL-90-R: Administration, scoring and procedures. *Manual II for the R (Evised) Version and Other Instruments of the Psychopathology Rating Scale Series*.

DeRubeis, R. J., Cohen, Z. D., Forand, N. R., Fournier, J. C., Gelfand, L. A., & Lorenzo-Luaces, L. (2014). The Personalized Advantage Index: Translating Research on Prediction into Individualized Treatment Recommendations. A Demonstration. *PLoS ONE*, *9*(1), e83875. https://doi.org/10.1371/journal.pone.0083875

Efron, B., & Gong, G. (1983). A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. *The American Statistician*, *37*(1), 36–48. https://doi.org/10.1080/00031305.1983.10483087

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*(4), 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Garge, N. R., Bobashev, G., & Eggleston, B. (2013). Random forest methodology for model-based recursive partitioning: The mobForest package for R. *BMC Bioinformatics*, *14*(1), 125. https://doi.org/10.1186/1471-2105-14-125

Gatnar, E. (2008). Fusion of Multiple Statistical Classifiers. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, & R. Decker (Eds.), *Data Analysis, Machine Learning and Applications* (pp. 19–27). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-78246-9_3

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*(14), 2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014

Harrell Jr, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, *15*(4), 361–387.

Horowitz, L. M., Rosenberg, S. E., Baer, B. A., Ureño, G., & Villaseñor, V. S. (1988). Inventory of interpersonal problems: Psychometric properties and clinical applications. *Journal of Consulting and Clinical Psychology*, *56*(6), 885.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674.

Jacobson, N. S., & Revenstorf, D. (1988). Statistics for assessing the clinical significance of psychotherapy techniques: Issues, problems, and new developments. *Behavioral Assessment*, *10*(2), 133–145.

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

Johnson, S. U., & Hoffart, A. (2013). Anxious Thought Questionnaire – 23. *Manuscript in Preparation.*

Johnson, S. U., & Hoffart, A. (2019). Moderators and predictors of outcome in metacognitive and cognitive behavioural therapy for co-morbid anxiety disorders. *Clinical Psychology & Psychotherapy*, *26*(4), 399–408. https://doi.org/10.1002/cpp.2361

Johnson, S. U., Hoffart, A., Nordahl, H. M., Ulvenes, P. G., Vrabel, K., & Wampold, B. E. (2018). Metacognition and cognition in inpatient MCT and CBT for comorbid anxiety disorders: A study of within-person effects. *Journal of Counseling Psychology*, *65*(1), 86–97. https://doi.org/10.1037/cou0000226

Johnson, S. U., Hoffart, A., Nordahl, H. M., & Wampold, B. E. (2017). Metacognitive therapy versus disorder-specific CBT for comorbid anxiety disorders: A randomized controlled trial. *Journal of Anxiety Disorders*, *50*, 103–112. https://doi.org/10.1016/j.janxdis.2017.06.004

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. https://doi.org/10.1002/da.22731

Kim, H., & Loh, W.-Y. (2001). Classification Trees With Unbiased Multiway Splits. *Journal of the American Statistical Association*, *96*(454), 589–604. https://doi.org/10.1198/016214501753168271

Kostaras, P., Martinaki, S., Asimopoulos, C., Maltezou, M., & Papageorgiou, C. (2020). The use of the Symptom Checklist 90-R in exploring the factor structure of mental disorders and the neglected fact of comorbidity. *Psychiatry Research*, *294*, 113522. https://doi.org/10.1016/j.psychres.2020.113522

Kraemer, H. C., & Blasey, C. M. (2004). Centring in regression analyses: A strategy to prevent errors in statistical inference. *International Journal of Methods in Psychiatric Research*, *13*(3), 141–151. https://doi.org/10.1002/mpr.170

Luedtke, A., Sadikova, E., & Kessler, R. C. (2019). Sample Size Requirements for Multivariate Models to Predict Between-Patient Differences in Best Treatments of Major Depressive Disorder. *Clinical Psychological Science*, *7*(3), 445–461. https://doi.org/10.1177/2167702618815466

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

Meyer, T. J., Miller, M. L., Metzger, R. L., & Borkovec, T. D. (1990). Development and validation of the penn state worry questionnaire. *Behaviour Research and Therapy*, *28*(6), 487–495. https://doi.org/10.1016/0005-7967(90)90135-6

Norman, G. R., Sloan, J. A., & Wyrwich, K. W. (2003). Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. *Medical Care*, *41*(5), 582–592. https://doi.org/10.1097/01.MLR.0000062554.74615.4C

Nye, A., Delgadillo, J., & Barkham, M. (2023). Efficacy of personalized psychological interventions: A systematic review and meta-analysis. *Journal of Consulting and Clinical Psychology*, *91*(7), 389–397. https://doi.org/10.1037/ccp0000820

Pålerud, K. M., Helmich, M. A., Hoffart, A., Ebrahimi, O. V., Snuggerud, T. R., & Johnson, S. U. (n.d.). *Sudden Gains in Cognitive Behavioral Therapy and Metacognitive Therapy for Complex Anxiety Disorders*.

Pallesen, S., Nordhus, I. H., Carlstedt, B., Thayer, J. F., & Johnsen, T. B. (2006). A Norwegian adaptation of the Penn State Worry Questionnaire: Factor structure, reliability, validity and norms. *Scandinavian Journal of Psychology*, *47*(4), 281–291. https://doi.org/10.1111/j.1467-9450.2006.00518.x

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

Sommer, J., Sarigiannis, D., & Parnell, T. (2019). *Learning to Tune XGBoost with XGBoost* (No. arXiv:1909.07218). arXiv. http://arxiv.org/abs/1909.07218

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

Sterner, P., Goretzko, D., & Pargent, F. (2021). *Everything has its Price: Foundations of Cost-Sensitive Learning and its Application in Psychology*. OSF. https://doi.org/10.31234/osf.io/7asgz

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*, 1–11.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*(4), 323–348. https://doi.org/10.1037/a0016973

van Bronswijk, S. C., DeRubeis, R. J., Lemmens, L. H. J. M., Peeters, F. P. M. L., Keefe, J. R., Cohen, Z. D., & Huibers, M. J. H. (2021). Precision medicine for long-term depression outcomes using the Personalized Advantage Index approach: Cognitive therapy or interpersonal psychotherapy? *Psychol. Med*, *51*(2), 279–289. https://doi.org/10.1017/S0033291719003192

van der Ploeg, T., Austin, P. C., & Steyerberg, E. W. (2014). Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology*, *14*(1), 137. https://doi.org/10.1186/1471-2288-14-137

Wells, A. (2009). *Metacognitive therapy for anxiety and depression*. Guilford Press.

Wells, A., Capobianco, L., Matthews, G., & Nordahl, H. M. (Eds.). (2020). *Metacognitive Therapy: Science and Practice of a Paradigm*. Frontiers Media SA. https://doi.org/10.3389/978-2-88966-244-9

Wells, A., & Cartwright-Hatton, S. (2004). A short form of the metacognitions questionnaire: Properties of the MCQ-30. *Behaviour Research and Therapy*, *42*(4), 385–396.

Wise, E. A. (2003). Psychotherapy Outcome and Satisfaction: Methods Applied to Intensive Outpatient Programming in a Private Practice Setting. *Psychotherapy: Theory, Research, Practice, Training*, *40*(3), 203–214. https://doi.org/10.1037/0033-3204.40.3.203

Wise, E. A. (2004). Methods for Analyzing Psychotherapy Outcomes: A Review of Clinical Significance, Reliable Change, and Recommendations for Future Directions. *Journal of Personality Assessment*, *82*(1), 50–59. https://doi.org/10.1207/s15327752jpa8201_10

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), 17155. https://doi.org/10.4088/JCP.15m10081