Psychotherapy Research

Routledge
Taylor & Francis Group

ISSN: 1050-3307 (Print) 1468-4381 (Online) Journal homepage: https://www.tandfonline.com/loi/tpsr20

Predicting personalized process-outcome associations in psychotherapy using machine learning approaches—A demonstration

Julian A. Rubel, Sigal Zilcha-Mano, Julia Giesemann, Jessica Prinz & Wolfgang Lutz

To cite this article: Julian A. Rubel, Sigal Zilcha-Mano, Julia Giesemann, Jessica Prinz & Wolfgang Lutz (2019): Predicting personalized process-outcome associations in psychotherapy using machine learning approaches—A demonstration, Psychotherapy Research, DOI: 10.1080/10503307.2019.1597994

To link to this article: https://doi.org/10.1080/10503307.2019.1597994

Published online: 26 Mar 2019.

Psychotherapy Research, 2019
https://doi.org/10.1080/10503307.2019.1597994

Predicting personalized process-outcome associations in psychotherapy using machine learning approaches—A demonstration

JULIAN A. RUBEL[1], SIGAL ZILCHA-MANO[2], JULIA GIESEMANN[1], JESSICA PRINZ[1], & WOLFGANG LUTZ[1]

[1]*Department of Psychology, University of Trier, Trier, Germany &* [2]*Department of Psychology, University of Haifa, Haifa, Israel*

(Received 26 March 2018; revised 12 March 2019; accepted 12 March 2019)

Abstract
Objective: Personalized treatment methods have shown great promise in efficacy studies across many fields of medicine and mental health. Little is known, however, about their utility in process-outcome research. This study is the first to apply personalized treatment methods in the field of process-outcome research, as demonstrated based on the alliance-outcome association. **Method:** Using a sample of 741 patients, individual regressions were fitted to estimate within-patient effects of the alliance-outcome association. The Boruta algorithm was used to identify patient intake characteristics that moderate the within-patient alliance-outcome association. The nearest neighbor approach was used to identify patients whose relevant pretreatment characteristics were similar to those of a target patient. The alliance-outcome associations of the most similar patients were subsequently used to predict the alliance-outcome association of the target patient. **Results:** Irrespective of the number of selected nearest neighbors, the correlation between the observed and predicted alliance-outcome associations was low and insignificant. According to the true error of the prediction, the demonstrated approach was unable to improve predictions made with a simple comparison model. **Conclusion:** The study demonstrated the application of personalized treatment methods in process-outcome research and opens many new paths for future research.

Keywords: personalized mental health; nearest neighbor; alliance-outcome research; within- and between-patients effects; longitudinal data; moderators of alliance-outcome association

Clinical or Methodological Significance of this Article

We propose an approach that uses data of already treated patients to derive predictions for future patients about which process variables might be especially important in this patient's treatment. Providing therapists with this prediction might enable them to better personalize their treatments to the needs of their patients.

Introduction

For decades, psychotherapy research trials comparing the efficacy of distinct treatments have failed to compellingly demonstrate the differential effects of these treatments (e.g., Barth et al., 2013). Recent studies using machine learning methods have shown that although treatments do not differ significantly from one another at the sample level, an individual patient may be more likely to benefit from a particular treatment than from others (e.g., DeRubeis et al., 2014). In all fields of medicine, including mental health, personalized treatment selection acknowledges individual differences between patients, by explicitly taking them into account to derive patient-specific decisions (e.g., Cohen & DeRubeis, 2018; Hamburg & Collins, 2010). However, thus far, personalized decisions were restricted to the selection of whole treatment packages (e.g., CBT vs. IPT) rendering them less suitable for practitioners who are only trained in one theoretical orientation. To address this restriction, the current study demonstrates an alternative approach that

Correspondence concerning this article should be addressed to Julian A. Rubel, Department of Psychology, University of Trier, Trier D-54296, Germany. Email: rubel@uni-trier.de

tries to provide recommendations on a strategy and technique level rather than the treatment level (cf. Goldfried, 1980).

Several approaches have been developed to derive individual-level recommendations for treatment selection. Most empirical methods that identify the best treatment for an individual patient are based on multivariable prediction modeling (e.g., Cohen & DeRubeis, 2018). Usually, patient intake characteristics (e.g., sex, age, symptom severity, chronicity, etc.) are used in multivariable prediction models to generate differential predictions for individuals based on their values on these variables. Two commonly applied methods are the *personalized advantage index* (PAI; e.g., DeRubeis et al., 2014) and the *nearest neighbor* (NN) approach (e.g., Lutz et al., 2005). In both methods, it is assumed that an algorithm-based assignment to treatment alternatives, based on the patient's intake characteristics, can optimize the effects of therapy and reduce the number of patients not benefiting from these treatments.

The PAI predicts the optimal treatment for new patients based on information on patients who have previously been treated (e.g., DeRubeis et al., 2014; Deisenhofer et al., 2018). Based on theoretical models and empirical findings, relevant pre-treatment patient characteristics are identified in advance to differentiate between two or more treatment alternatives. Next, variables identified as significantly interacting with treatment condition (i.e., indicating differential treatment response) are used to predict an outcome score for each treatment alternative for each patient. The differences between the predicted outcome scores for each treatment are represented in the PAI. The PAI quantifies the potential advantage of one treatment over another for an individual patient.

To date, several studies have used the PAI for treatment selection. In the first study, DeRubeis et al. (2014) reanalyzed data from a randomized controlled trial comparing cognitive behavioral therapy (CBT) and antidepressant medication (ADM) for depression. Although on average the outcomes of the 50 patients receiving CBT did not differ from those of the 104 patients receiving ADM, the authors were able to build a predictive model based on patient intake characteristics that indicated a potential differential response at the individual level. For about 60% of the sample, treatment response predictions showed meaningful clinical differences. Dividing this group in those patients who received the optimal treatment indicated by the model and those who did not resulted in a significant advantage for the optimal group ($d = 0.58$). In another study, Zilcha-Mano et al. (2016) applied the PAI to predict which of three depression treatments (supportive expressive therapy (SET), ADM, and placebo pill) would result in lower dropout rates for individual patients. Again, on average, the three groups ($N = 156$) did not differ significantly in dropout rates. Based on patient intake characteristics, the authors build a prediction model that allowed the division of patients into those who received their optimal, second optimal, and least optimal treatment. Despite the non-significant average dropout differences, those who received their optimal treatment had about half the dropout risk (24.4%) of those who received their least optimal treatment (47.4%).

The NN approach originated in avalanche research (Brabec & Meister, 2001), where it has been used to predict the risk of an avalanche occurring on a certain day. Avalanche research uses large databases containing potentially relevant parameters for each day, such as temperature and barometric pressure. For any day of interest, the 30 (or more) most similar days are selected, based on the relevant parameters, and the frequency of avalanches occurring on these days is used to predict the probability of an avalanche occurring on the day of interest. Lutz et al. (2005) adapted this methodology to predict treatment response in a sample of 203 psychotherapy outpatients in the UK. Similar to the avalanche prediction model, the response of a new patient to treatment was predicted based on the treatment responses of previously treated patients most similar to the new one. The similarity of previously treated patients to the new patient was calculated using Euclidean distances between the relevant predictor variables. Applying the NN approach to differential treatment selection, Lutz, Saunders, et al. (2006) also tested the predictive validity and clinical utility of this approach. The authors produced predictions for each patient for different treatment protocols—CBT vs. integrative CBT and interpersonal treatment (IPT), and checked whether a given patient's prediction was better for one of these treatments than for the other. On average, the authors did not identify a significant difference between the two treatment approaches, however, for about one third of patients, it was possible to obtain clinically meaningful differential predictions: one approach was predicted to result in significantly better outcomes than the other. For the remaining two-thirds, there was no difference between the predicted change curves of the two protocols.

Although these approaches have changed how treatment efficacy is being studied and even conceptualized, little use has been made of these methods in process-outcome research to date. Transforming heterogeneity between patients from a confounder into a resource appears to be the next step in process-outcome research (e.g., Zilcha-Mano, 2018). In the field of treatment efficacy, such approaches are limited to comparisons between complete treatment packages (e.g., CBT vs. IPT). Therefore, the potential practical implications of such treatment selection studies are limited. For example, if a practitioner is trained in CBT, but not in IPT and given a recommendation to treat a patient with IPT, the therapist would have no other choice but to refer the patient to an IPT therapist, if one were available. Therefore, it makes sense to extend the focus of personalized psychotherapy research by extending treatment selection models to process-outcome research. This may yield recommendations regarding the interventions, strategies, and processes that are likely to be most effective for a given patient, including therapeutic alliance, therapeutic directiveness, and techniques focusing on cognitions, emotions, and behavioral activation, among others (see also Hayes & Hofmann, 2018; Norcross, 2011). We propose an approach aimed at generating recommendations about which change processes may be especially helpful for individual patients. This approach builds on recent investigations in personalized treatment selection using machine learning methodology (e.g., Cohen & DeRubeis, 2018). The approach involves four general steps (cf. Gillan & Whelan, 2017). In the first step, the process-outcome association (P-OA) of interest is quantified at the within-patient (WP) level, as we are interested in how change over time in a process variable influences symptom change over the course of treatment (e.g., Falkenström, Finkel, Sandell, Rubel, & Holmqvist, 2017). Thus, a separate estimate of the P-OA for each patient is obtained.

In the second step, relevant patient intake characteristics are identified that moderate the individual P-OAs, indicating whether a certain WP P-OA is expected to be higher or lower for a given patient. To make predictions that can be generalized to patients who were not part of the sample upon which prediction model development was based, methods are needed to prevent overfitting. An overfitted model makes good predictions in the sample in which it is developed, but performs badly when making out-of-sample predictions (e.g., James, Witten, Hastie, & Tibshirani, 2013).

In the third step, the validity of the derived model is evaluated by testing its predictions on a new dataset, i.e., not the one used to build the prediction model. For the validation dataset, a prediction is generated for each individual patient using the prediction model. The predicted P-OAs can then be compared with the observed P-OAs to obtain an overall measure of prediction accuracy.

In the fourth step, the usefulness of the prediction model is tested in a randomized controlled trial, in which therapists are provided with person-specific predictions for some of their patients, but not for others. For example, half the patients may be randomized to a condition where therapists receive information about the process variables that are expected to have the greatest influence on outcome, and the other half to a condition where the therapists receive no such information.

Ideally, the proposed approach should be implemented to use several process variables simultaneously. However, because this is the first attempt to implement a personalized treatment approach to process variables, our objective is to introduce the method rather than demonstrate its optimal performance. To simplify our task, we focused on a single variable: the working alliance, defined as the emotional bond between therapist and patient, and the degree of their agreement about the goals of therapy and the tasks that are required to achieve them (Bordin, 1979; Hatcher & Barends, 2006). In their meta-analysis, Flückiger, Del Re, Wampold, and Horvath (2018) reported a mean correlation of the alliance-outcome association of $r = .278$ (95% CI [.26, .30]), irrespective of therapeutic orientation, alliance measures used, perspective of the rater (patient, therapist, or observer), or time of assessment. We chose the alliance because it is the most commonly investigated P-OA in psychotherapy research, and the one most consistently related to treatment outcome. Moreover, it has been shown, that the WP alliance-outcome association is not the same for each patient, but that several patient characteristics can moderate this relationship (e.g., Falkenström, Granström, & Holmqvist, 2013; Lorenzo-Luaces, DeRubeis, & Webb, 2014; Zilcha-Mano & Errázuriz, 2015). Although previous studies have investigated moderators of the WP alliance-outcome association (e.g., Falkenström et al., 2013; Lorenzo-Luaces et al., 2014; Zilcha-Mano & Errázuriz, 2015), as of to date, they have not attempted to use information about patient-level moderators to make prospective statistical predictions about the size of the WP alliance-outcome correlation. Moderators identified at the patient level in studies on the WP alliance-outcome association include treatment length, initial symptom severity, number of prior depressive episodes, and personality problems.

The present study is the first to introduce an approach aimed at translating research on P-OA into individually tailored recommendations for therapists. We tested whether an exemplary method for producing personalized treatment recommendations can be used to make personalized predictions regarding the importance of different process variables based on data from patients who have already been treated. We illustrate this approach using the example of the WP alliance-outcome association. This approach, however, may be used to investigate all P-OAs relevant to psychotherapy. Because the present paper is intended to demonstrate this approach, some of the suggestions are preliminary and require further systematic testing. The present study focuses on the first three steps of the four-step model, although a comprehensive process should include all four.

Method

Participants and Treatments

We used a sample of 792 psychotherapy patients from an outpatient clinic. Patients were treated with integrative CBT (including interpersonal and emotion-focused elements). Fifty-one patients were removed because of a lack of variance in their alliance ratings, leaving 741 patients.[1] Sixty-four percent of patients were female. The mean age of the patients was $M = 36$ ($SD = 12.7$). Patients were treated mainly for affective disorders (57.8%) or anxiety disorders (18.4%), followed by adjustment disorder (8.8%), somatoform disorders (4.1%), PTSD (3.9%), eating disorders (2.6%), and other disorders (4.1%). All patients were treated for a minimum of 10 and a maximum of 113 sessions, mean treatment length being $M = 38$ sessions ($SD = 17.2$). Individual psychotherapy consisted of one weekly session. Treatment was open-ended in length, but restricted by health insurance regulations.

Patients were treated by 115 therapists with an average number of 6.44 patients per therapist (range 1–17). All therapists had at least 1 year of training before beginning to see patients in this outpatient clinic. Each therapist received one hour of individual supervision or group supervision per month. All therapy sessions were videotaped for use in supervision. Supervisors were senior clinicians.

To be able to validate our model on an unseen dataset, we randomly split the data into a training (about two-thirds of the sample, $n = 490$) and a test sample (about one-third of the sample, $n = 251$). The development of the prediction model is based entirely on patients in the training sample. The test sample was used to validate the prediction model by comparing the predicted and observed alliance-outcome associations for patients in this sample.

Instruments

Hopkins-Symptom Checklist-11. At the beginning of each therapy session, the Hopkins Symptom Checklist short form-11 (HSCL-11; Lutz, Tholen, Schürch, & Berking, 2006) was administered. The HSCL-11 is an 11-item self-report inventory used to assess symptomatic impairment, and is a brief version of the SCL-90-R (Derogatis, 1992). Items are rated on a 4-point Likert scale, ranging from 1 = not at all to 4 = extremely. The mean value of the items is an indicator of the symptomatic impairment of the patient in the previous week. In the present sample, the HSCL-11 reliably detected differences in systematic changes in symptom distress over the weeks with a high reliability of the change score ($R_c = .84$; Cranford et al., 2006). The average internal consistency over all sessions was also high for the current sample ($\alpha = .92$).

Berne Post-Session Report. At the end of each therapy session, the Berne Post-Session Report (BPSR; Flückiger, Regli, Zwahlen, Hostettler, & Caspar, 2010) was administered. The BPSR is used to assess processes of change in a given therapy session based on patient (P) or therapist (T) reports. For the purpose of our analysis, only the alliance scale of the patient questionnaire was used (cf. Rubel, Rosenbaum, & Lutz, 2017). In the present sample, the alliance subscale reliably detected differences in systematic changes in alliance over the weeks with a high reliability of the change score ($R_c = .94$; Cranford et al., 2006). The average internal consistency over all sessions was also high in the current sample ($\alpha = .95$).

Determining WP Alliance-Outcome Associations

For each patient in both the training and the test samples, we fitted the following individual regression model:

$$Symptoms_{t+1} = \beta_0 + \beta_1 * Alliance_t + \beta_2 * Symptoms_t + \varepsilon \qquad (1)$$

where $Symptoms_{t+1}$ is the dependent variable, representing a patient's symptom severity in session $t + 1$, β_0 is the intercept, β_1 is the coefficient of interest (i.e., the alliance-outcome association), representing the effect of alliance at time point t on symptoms at time point $t + 1$ ($Symptoms_{t+1}$), controlled for the autoregressive effects of symptoms (i.e., $\beta_2 t$), and ε is the error term. As these individual regression models contain WP variation only, their coefficients represent WP associations. We preferred this individual-level regression approach over the more typical two-level hierarchical linear models (sessions nested within patients) to prevent endogeneity bias

(Falkenström et al., 2017).[2] For each patient, the effect of the alliance represented by the β_1 coefficient in equation (1) was saved and used in further analyses.

Selecting Predictors of the WP Alliance-Outcome Association

The extracted WP alliance-outcome association (β_1) of each patient in the training sample was then used as the dependent variable to identify significant moderators of the WP alliance-outcome association. Significant moderators were identified using the Boruta R package (Kursa & Rudnicki, 2010). Boruta is a machine learning algorithm based on random forests (RF; e.g., Breiman, 2001). In the random forest (RF) approach, several regression trees are grown with different subsets of the whole set of predictor variables. Subsequently, the importance of each predictor is quantified by averaging its contribution over all the different regression trees. Although RF can rank variables based on their importance, it does not provide recommendations about which variables to select as relevant and which not. Boruta extends RF by a procedure in which the real predictor variables are tested against random variables. These random variables are created by shuffling the real predictor variables, after which they are added as new fake variables (*shadow features*) to the dataset. Thus, Boruta includes randomly shuffled copies of each variable to the dataset. Since the values in these variables are randomly assigned to patients, there should be no systematic association between the fake variables and the dependent variable. A random forest model is then estimated with this extended dataset, containing both the real variables and the shuffled fake variables. Boruta identifies as relevant predictors the variables that explain significantly more variation in the dependent variable than the best fake variable (Kursa & Rudnicki, 2010).

A total of 95 possible moderators (Table I), routinely collected at intake, were fed into the Boruta algorithm. All continuous variables were z-standardized, and all categorical variables dummy-coded before fitting the model. Missing data in the potential predictor variables were imputed with the MissForest R package (Stekhoven & Buhlmann, 2012).

Table I. Predictors entered into the *Boruta* algorithm. Predictors were collected routinely at intake. Predictors in **bold** were selected by the *Boruta* algorithm.

Categorical predictors
Diagnosis
Sex
Education
Occupation
Marital status
Daily or occasional medication: yes/no
8 sociodemographic items:
Nationality
Household
Housing situation
Last occupation
Current occupation
Upcoming pension
Work ability in the last 12 months
Children: yes/no
Continuous predictors
Age
OQ total score
OQ Symptom distress (SD) subscale
OQ Social role functioning (SR) subscale
OQ Interpersonal relationship (IR) subscale
Questionnaire for the evaluation of psychotherapy (FEP2) total score
FEP2 Well-being subscale
FEP2 Discomfort subscale
FEP2 Incongruence subscale
FEP2 Interpersonal Problems subscale
Emotionality Inventory (EMI) total score
EMI Anxiety subscale
EMI Depression subscale
EMI Inhibition subscale
EMI Security subscale
EMI Wellbeing subscale
Brief Symptom Inventory (BSI) total score
BSI Somatic problem subscale
BSI Obsessive Compulsive subscale
BSI Uncertainty subscale
BSI Depression subscale
BSI Anxiety subscale
BSI Hostility subscale
BSI Phobia subscale
BSI Paranoid subscale
BSI Psychoticism subscale
BSI Additional
Interpersonal Problems (IIP-32) total score
IIP-32 Autocratic/dominant subscale
IIP-32 Confrontational subscale
IIP-32 Unapproachable subscale
IIP-32 Introverted subscale
IIP-32 Submissive subscale
IIP-32 Exploitable
IIP-32 Caring subscale
IIP-32 Expressive subscale
Incongruence questionnaire (INK-23) total score
INK-23 Approach subscale
INK-23 Avoidance subscale
Dysfunctional attitudes scale—short form (DAS-K) total score
DAS-K Recognition subscale
DAS-K Performance subscale
GAF last week
GAF last 12 month

Generating Patient-Specific Predictions of the Alliance-Outcome Association

To derive predictions for patients in the test sample, we used the NN approach described above. We chose this method because it is highly intuitive and has

  - Inventory of Stressful Life-Events (ILE)—Score for number of events
  - ILE—Score for stress
    - ILE Number of events in patient's life subscale
    - ILE Number of events in life of close relationships subscale
    - ILE Number of events in life of distant relationships subscale
  - Pain sensation scale
  - Personality style and disorder inventory—short-form (PSSI-K)
    - PSSI-K subscale—Antisocial personality style
    - PSSI-K subscale—Paranoid personality style
    - PSSI-K subscale—Schizoid personality style
    - PSSI-K subscale—Avoidant personality style
    - PSSI-K subscale—Compulsive personality style
    - PSSI-K subscale—Schizotypal personality style
    - PSSI-K subscale—Rhapsodic personality style
    - PSSI-K subscale—Narcissistic personality style
    - PSSI-K subscale—Negativistic personality style
    - PSSI-K subscale—Dependent personality style
    - PSSI-K subscale—Borderline personality style
    - PSSI-K subscale—Histrionic personality style
    - **PSSI-K subscale—Depressive personality style**
    - PSSI subscale—Altruistic personality style
- Patient rated well-being items
    - Stress
    - **Psychological Wellbeing**
    - Acute Stress
    - **Emotional and psychological functioning**
    - Recent life satisfaction
    - Current energy level and sense of health
    - Current emotional and psychological functioning
  - Therapy expectations
    - Importance of psychotherapy
    - Difficulties attending psychotherapy
    - Confidence in the helpfulness of psychotherapy in dealing with problems
    - Amount of previous psychotherapy
    - Chronicity of the problem
    - Estimated future coping
  - Therapist rated wellbeing
    - Patient's recent discomfort
    - Current effect of psychotherapy on the patient
    - Expected patient improvement with further psychotherapy

several advantages over more complex methods for the purpose of demonstrating the idea of personalized process-outcome research. To obtain a prediction for each patient in the test sample, we used the alliance-outcome associations (β_1) of patients from the training sample who were most similar to those of the target patient (with regard to the significant moderator variables chosen in the previous stage). For the current demonstration, we tested different numbers of NN (i.e. 1, 5, 10, 20, 30, 40, 50, 100, 200, and 300).

For each individual prediction, the alliance-outcome associations of the NN were aggregated using a weighted average. The individual associations were weighted by the number of therapy sessions of each patient, taking into account that the association for patients with more sessions was based on more data than that for patients with fewer sessions. The weighted mean was calculated as follows:

$$Pred.A - OA_i = \sum_{j=1}^{k} Nr.ofsessions_j * Obs.A - OA_j \bigg/ \sum_{j=1}^{k} Nr.ofsessions_j \qquad (2)$$

where $Pred.\ A - OA_i$ is the predicted alliance-outcome association for patient i, $Nr.ofsession_j$ is the total number of therapy sessions of patient j, and $Obs.A - OA_j$ is the observed alliance-outcome association (β_1) of patient j (Rukhin & Vangel, 1998). The alliance-outcome association of each patient in the test sample was predicted based on the aggregated and weighted alliance-outcome association of the NN in the training sample.

Model Validation

The accuracy of the model is measured using the true error of the prediction (i.e., the average absolute difference between predicted and actual scores across the 251 patients in the test sample) as well as the correlation of the predicted and observed WP alliance-outcome associations for patients in the test dataset.

Results

Boruta identified 11 relevant moderators of the WP alliance-outcome association (bold variables in Table I). These variables were subsequently used to determine similarity between patients. To validate the predictions of this method, the most similar patients in the training sample were identified for each of the 251 patients in the test sample and used to calculate individual predictions. Table I shows the correlation of predicted and observed WP alliance-outcome associations as well as the true error of the predictions for varying numbers of selected nearest neighbors. Irrespective of the number of chosen NN, the correlation between predicted and observed scores was low and insignificant. Correlations ranged from −.07 (300 NN) to .05 (30 NN). The true error ranged from 0.37 (1 NN) to 0.20 (100, 200, and 300 NN). To evaluate the true error scores, we compared them with the error that would have resulted if we had used the average weighted WP alliance score from the complete training sample as a prediction for each patient in the test sample (see last row of Table II). This results in an average true error of 0.19, which can be seen as a possible benchmark. Good predictions need to

Table II. Correlation (r) and average absolute difference (true error) between predicted and observed scores for different numbers of selected similar patients (nearest neighbors).

#of NN	r	True error
1	−.00	0.37
5	−.02	0.26
10	−.00	0.24
20	−.00	0.22
30	.05	0.21
40	.02	0.21
50	.01	0.21
100	−.01	0.20
200	−.00	0.20
300	−.07	0.20
490 (complete training sample)	–	0.19

Note. NN = nearest neighbors.

produce a true error that is significantly lower than this lower boundary. Since all predictions based on the presented approach result in higher errors, we can conclude that these did not improve the prediction compared to the model based on the complete training sample.

Discussion

This paper introduces an approach for translating moderators of WP process-outcome associations into recommendations for individual psychotherapy patients. The approach is illustrated using the example of the WP association between patient-rated therapeutic alliance and symptom impairment. Based on patients' baseline characteristics that moderated the WP alliance-outcome association, the patients most similar to a new patient were used to generate a prediction for the new patient.

In contrast to our expectations, we were not able to cross-validate the predictions made with this approach in an independent test sample. On average, there was no association between the predicted WP alliance-outcome associations and the observed associations, irrespective of the number of selected nearest neighbors. Given the successful application of prediction models in the selection of different treatments in similar or even smaller sized samples (e.g., Cohen & DeRubeis, 2018), it may come as a surprise that we failed to make accurate predictions of WP alliance-outcome associations using the current approach. Below we list several possible reasons for failing to make accurate predictions.

First, in the case at hand, predictions are more difficult to make than in typical treatment selection circumstances. In the case of treatment selection models, accurate prediction of specific scores is not the primary objective, but rather, these models seek to predict the expected difference between two treatment alternatives. Even if the predicted scores do not exactly match the observed scores in the models, they may still accurately differentiate between the treatments (e.g., Cohen & DeRubeis, 2018). By contrast, in the approach presented here, we sought to predict the exact WP process-outcome association. Furthermore, the predicted process-outcome relation is affected by more sources of error than the predicted outcome scores in treatment selection models, where the dependent variable (i.e., generally, the post-treatment symptom score) is biased only by measurement error. In the approach presented here, however, there are three different error sources for the dependent variable: the measurement error of the process variable (i.e., the alliance), the outcome variable (i.e., symptoms), and the error made when estimating the relationship between these two. Therefore, the dependent variable in the approach described here, (i.e., the process-outcome relationship) is likely much noisier than the outcome scores used in typical treatment selection models, and therefore more difficult to assess, and it may require more data points to accurately identify true associations than in treatment selection models.

A second possible explanation for the failure to make accurate predictions in the present study has to do with our sample size. Although our sample was quite large for psychotherapy process-outcome studies (e.g., Crits-Christoph, Connolly Gibbons, & Mukherjee, 2013), it may have been too small for the implementation of machine learning approaches. Small samples might produce a situation in which similar patients (the nearest neighbors) are not similar enough to make valid predictions. A larger sample would increase the probability of having a pool of sufficiently similar patients that would allow more accurate predictions. A third potential explanation concerns predictor weighting. In the current approach, each predictor received the same weight in our model, although some predictors may have a stronger influence than others. This differential influence could be accounted for by assigning a stronger weight to more influential predictors when defining the similarity of the patients. With predictor weighting, the similarity of two patients would depend more on strongly weighted predictors than on less influential ones.

The three potential explanations above concerned technical and methodological reasons (sample size, number of available time points, weights of predictors). However, more fundamental *post hoc* explanations can also be suggested. One such explanation is that pre-treatment predictors are not sufficient to draw inferences about individual dynamics of the alliance over the course of treatment, and their effects on subsequent session outcome. Attention must also be paid to the process of treatment (techniques implemented by the therapists, therapists' level of responsiveness to the patient) and to how it interacts with the baseline predictors. According to this explanation, pre-treatment predictors are not sufficient to predict the WP alliance-outcome association. Although this explanation appears less likely given previous research that has identified moderators of the alliance-outcome association, as reviewed in the introduction section, it cannot be completely ruled out based on the current findings.

Finally, it is possible to argue that the failure of the model to predict the WP alliance-outcome association may be a consequence of ignoring the therapist level in our analyses, especially in view of findings according to which the alliance mainly explains outcome differences between therapists, not within therapists (Baldwin, Wampold, & Imel, 2007). Note, however, that these findings pertain to *between*-patient associations, so that patients with higher early alliance have better outcomes than patients with lower early alliance levels. There is no evidence to date showing that the inclusion of the therapist-level changes *within*-patient effects. On the contrary, based on the methodological literature, it may be expected that the inclusion of such a higher-order effect would not cause a change in the estimations of interest in the present study (Van Landeghem, De Fraine, & Van Damme, 2005, p. 426). To test this notion in the current dataset as a sensitivity analysis, we conducted a three-level multi-level model partitioning the variance in WP alliance-symptom associations into between-patient (i.e., within-therapist) and between-therapists variation. As expected, the between-therapists variance term was negligibly small ($VAR_{between\text{-}therapist} = 7.38^{*}e^{-11}$) and substantially smaller than between-patient variation ($VAR_{between\text{-}patient} = 0.004$). Therefore, it is reasonable to assume that applying models that account for therapist differences would not have resulted in better estimates.

Limitations and Future Directions

The proposed approach is part of a family of methods that may be used in a variety of ways. At each step along the demonstrated approach, several alternative methods can be chosen, all of which have similar aims as the ones demonstrated in the current illustration. Because the present paper did not attempt to come up with the best possible approach for predicting personalized alliance-outcome associations, we made several pragmatic decisions. Below are a few issues that deserve future systematic investigation.

First, the accurate estimation of the alliance-outcome association by individual regression models depends on the length of the treatment (number of sessions). Longer treatments provide a larger database for estimating this relationship than shorter treatments do. It is not possible to discern whether a bad match between predicted and observed alliance-outcome associations is due to a poor prediction or an inadequately measured observed association. It is possible, therefore, that the current testing of this approach by comparing predicted and observed process-outcome associations may not have been the most appropriate. Alternatively, it may be instructive to conduct a study in which therapists are provided with predictions generated with this approach for one group of patients, and no information or alternative information for another group. If outcomes for the group of patients for which this information is provided are better than those of the other group, the method can be said to produce clinically helpful information, irrespective of our ability to validate the described approach as we attempted to here.

Second, there are several different predictor selection algorithms in addition to Boruta, such as LASSO, backward selection, and others (e.g., Cohen, Kim, Van, Dekker, & Driessen, 2019). Future research needs to systematically test these methods on different real and simulated datasets to reach conclusions about which method is best suited for the current approach.

Third, the removal of 51 patients due to no variation in alliance measures over time emphasizes the fact that WP associations can only be investigated for patients who show at least a minimum amount of variation in the process variable of interest. As such, for those patients who do not experience meaningful WP shifts in the alliance during treatment it is not possible to estimate the patient-specific association of the alliance and symptom change. However, that does not necessarily imply that, for example, the alliance is unimportant for these patients and that a stable good alliance cannot exert positive effects. It is a limitation of the presented method that these kinds of effects cannot be appropriately estimated. These effects may only be observed with between-patient comparisons (i.e., comparisons of symptom scores between patients with low and high average alliance levels). However, given the observational nature of these comparisons, it is unclear whether these can be translated into patient-specific recommendations (e.g., Falkenström et al., 2017).

Finally, the value of the predicted coefficient (the predicted alliance-outcome association) may be of limited use to therapists. It may not be helpful to provide therapists with predictions regarding a single variable, such as the alliance-outcome association, but rather more beneficial to provide them with a patient-specific profile of the predicted effect of various processes, strategies, and techniques. This may enable therapists to prioritize the processes that have the highest predicted benefit for each individual patient.

To provide a definitive answer to all the aforementioned open questions is a task for future research. Here, we presented the proposed approach with the aim of encouraging the formulation of hypotheses about the importance of different process variables for individual patients. Further research into moderators of process-outcome associations is needed to provide personalized psychotherapy recommendations, and thereby personalized mental health care. Despite the failure of the presented model to make predictions, we believe that a shift from treatment selection to process/strategy/technique selection models will represent a significant leap forward in the field of personalized psychotherapy (e.g., Lutz, Zimmermann, Müller, Deisenhofer, & Rubel, 2017; Norcross, 2011).

Acknowledgement

This work was supported by grants from the German Research Foundation (LU 660/10-1, LU 660/8-1). We thank Kaitlyn Boyle for proofreading earlier versions of this manuscript.

Funding

This work was supported by grants from the German Research Foundation (LU 660/10-1, LU 660/8-1).

Notes

[1] Note that for patients without variability it is not possible to calculate associations with other variables. There are various options for addressing this problem other than removing the cases. One alternative would be to assume an alliance-outcome association of 0 for patients with no variability in alliance scores. We conducted a sensitivity analysis to test whether results differed if we included these 51 patients with an alliance-outcome association of 0. Results were not substantially different from those reported in the main analyses, and can be obtained from the first author upon request.

[2] Endogeneity bias results when the lagged dependent variable (i.e., the Symptom$_t$ variable) is included as a predictor in a random intercept multi-level model, where the random intercept is intrinsically correlated with the person-time error, which violates one of the basic assumptions of regression analysis (Baltagi, 2013). The result of endogeneity tends to be that the effect of the lagged dependent variable is estimated to be too large, and the effect of other predictors (i.e., the alliance in this application) to be too small (Allison, 2015). As, in our opinion, controlling for the effects of previous symptoms is important, we decided to circumvent the problem of endogeneity by estimating separate person-specific ordinary least squares regression models that do not estimate a random intercept.

References

Allison, P. (2015). Don’t put lagged dependent variables in mixed models. Retrieved from https://statisticalhorizons.com/lagged-dependent-variables

Baldwin, S. A., Wampold, B. E., & Imel, Z. E. (2007). Untangling the alliance-outcome correlation: Exploring the relative importance of therapist and patient variability in the alliance. *Journal of Consulting and Clinical Psychology*, *75*, 842–852. doi:10.1037/002-006X.75.6.842

Baltagi, B. H. (2013). *Economic analysis of panel data*. Chichester, United Kingdom: Wiley.

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. doi:10.1176/appi.focus.140201

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

Brabec, B., & Meister, R. (2001). A nearest-neighbor model for regional avalanche forecasting. *Annals of Glaciology*, *32*, 130–134.

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

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

Cohen, Z., Kim, T. T., Van, H. L., Dekker, J. J. M., & Driessen, E. (2019). A demonstration of a multi-method variable selection approach for treatment selection: Recommending cognitive-behavioral versus psychodynamic therapy for mild to moderate adult depression. *Psychotherapy Research*. Advance Online Publication. doi:10.17605/OSF.IO/6QXVE

Cranford, J. A., Shrout, P. E., Iida, M., Rafaeli, E., Yip, T., Bolger, N. (2006). A procedure for evaluating sensitivity to within-person change: Can mood measures in diary studies detect change reliably. *Personality and Social Psychology Bulletin*, *32* (7), 917–929.

Crits-Christoph, P., Connolly Gibbons, M. B., & Mukherjee, D. (2013). Psychotherapy process-outcome research. In M. J. Lambert (Ed.), *Bergin and Garfield’s handbook of psychotherapy and behavior change* (6th ed., pp. 298–340). New York: John Wiley & Sons.

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

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

DeRubeis, R. J., Cohen, Z. D., Forand, N. R., Fournier, J. C., Gelfand, L. A., Lorenzo-Luaces, L., & Cho, W. C. S. (2014). The personalized advantage index: Translating research on prediction into individualized treatment recommendations. A demonstration. *PLoS ONE*, *9*, e83875. doi:10.1371/journal.pone.0083875

Falkenström, F., Finkel, S., Sandell, R., Rubel, J., & Holmqvist, R. (2017). Dynamic models of individual change in psychotherapy process research. *Journal of Consulting and Clinical Psychology*, *85* (6), 537–549. doi:10.1037/ccp0000203

Falkenström, F., Granström, F., & Holmqvist, R. (2013). Therapeutic alliance predicts symptomatic improvement session by session. *Journal of Counseling Psychology*, *60*(3), 317–328. doi:10.1037/a0032258

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

Flückiger, C., Regli, D., Zwahlen, D., Hostettler, S., & Caspar, F. (2010). Der Berner Patienten- und Therapeutenstundenbogen 2000 [The Berne post-session reports for patients and therapists 2000]. *Zeitschrift für Klinische Psychologie und Psychotherapie*, *39*, 71–79. doi:10.1026/1616-3443/a000015

Gillan, C. M., & Whelan, R. (2017). What big data can do for treatment in psychiatry. *Current Opinion in Behavioral Sciences*, *18*, 34–42. doi:10.1016/j.cobeha.2017.07.003

Goldfried, M. R. (1980). Toward the delineation of therapeutic change principles. *American Psychologist*, *35*(11), 991–999. doi:10.1037/0003-066X.35.11.991

Hamburg, M. A., & Collins, F. S. (2010). The path to personalized medicine. *New England Journal of Medicine*, *363*, 301–304. doi:10.1056/NEJMp1006304

Hatcher, R. L., & Barends, A. W. (2006). How a return to theory could help alliance research. *Psychotherapy: Theory, Research, Practice, Training*, *43*, 292–299. doi:10.1037/0033-3204.43.3.292

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

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

Kursa, M. B., & Rudnicki, W. R. (2010). Feature selection with the *Boruta* package. *Journal of Statistical Software*, *36*(11), 1–13. Retrieved from http://www.jstatsoft.org/v36/i11/

Lorenzo-Luaces, L., DeRubeis, R. J., & Webb, C. A. (2014). Client characteristics as moderators of the relation between the therapeutic alliance and outcome in cognitive therapy for depression. *Journal of Consulting and Clinical Psychology*, *82*, 368–373. doi:10.1037/a003594

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

Lutz, W., Saunders, S. M., Leon, S. C., Martinovich, Z., Kosfelder, J., Schulte, D., … Tholen, S. (2006). Empirical and clinical useful decision making in psychotherapy: Differential Predictions with Treatment Response Models. *Psychological Assessment*, *18*(2), 133–141.

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

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), 306–317. doi:10.1186/s12888-017-1464-2

Norcross, J. C. (2011). *Psychotherapy relationships that work: Evidence-based responsiveness*. New York, NY: Oxford University Press.

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

Rukhin, A. L., & Vangel, M. G. (1998). Estimation of a common mean and weighted means statistics. *Journal of the American Statistical Association*, *93*, 303–308. doi:10.2307/2669626

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

Van Landeghem, G., De Fraine, B., & Van Damme, J. (2005). The consequence of ignoring a level of nesting in multilevel analysis: A comment. *Multivariate Behavioral Research*, *40*(4), 423–434. doi:10.1207/s15327906mbr4004_2

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

Zilcha-Mano, S., & Errázuriz, P. (2015). One size does not fit all: Examining heterogeneity and identifying moderators of the alliance-outcome association. *Journal of Counseling Psychology*, *62* (4), 579–591. doi:10.1037/cou0000103

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. doi:10.4088/JCP.15m10081