Behaviour Research and Therapy 120 (2019) 103438

Contents lists available at ScienceDirect

Behaviour Research and Therapy

journal homepage: www.elsevier.com/locate/brat

Towards integrating personalized feedback research into clinical practice: Development of the Trier Treatment Navigator (TTN)

Wolfgang Lutz*, Julian A. Rubel, Brian Schwartz, Viola Schilling, Anne-Katharina Deisenhofer

University of Trier, Germany

ARTICLE INFO

Keywords:
Routine outcome monitoring
Patient-focused feedback research
Personalized and precision mental health
Treatment navigation

ABSTRACT

In this study, a computer-based feedback, decision and clinical problem-solving system for clinical practice will be described - the Trier Treatment Navigator (TTN). The paper deals with the underlying research concepts related to personalized pre-treatment recommendations for drop-out risk and optimal treatment strategy selection as well as personalized adaptive recommendations during treatment. The development sample consisted of 1234 patients treated with cognitive behavioral therapy (CBT). Modern statistical machine learning techniques were used to develop personalized recommendations.

Drop-out analyses resulted in seven significant predictors explaining 12.0% of variance. The prediction of optimal treatment strategies resulted in differential prediction models substantially improving effect sizes and reliable improvement rates. The dynamic failure boundary reliably identified patients with a higher risk for no improvement or deterioration and indicated the usage of clinical problem-solving tools in risk areas. The probability to be reliably improved for patients identified as at risk for treatment failure was about half of the probability for other patients (35% vs. 62.15%; $\chi^2_{df=1} = 82.77$, $p < .001$).

Results related to the computer-based feedback system are discussed with regard to the implication for clinical applications as well as clinical training and future research possibilities.

1. Introduction

Tailoring psychological treatments to the individual patient is nothing new. In fact, this is how the majority of clinicians work (Lambert, 2013). Most therapists individualize their approach based on their clinical experience with similar patients. However, this means that in clinical practice, personalisation is mostly conducted unsystematically and based on intuition (Perlis, 2016). Given the fact that past research is dominated by the development and validation of treatment protocols, this personalized practice largely lacks empirical evidence. Therefore, it is no wonder that systematically tailoring psychological treatments to individual patients on the basis of empirical evidence has received burgeoning interest in psychotherapy research (e.g., Cohen & DeRubeis, 2018; Howard, Moras, Brill, Martinovich, & Lutz, 1996; Lambert, 2007; Lutz, de Jong, & Rubel, 2015). This may be due to the fact that despite the existence of highly effective treatments, only 40–70% of patients show substantial improvements, whereas a high percentage of patients (depending on the change criteria used) do not profit and about 5–10% deteriorate during the provision of psychological treatments (Lambert, 2013).

Personalized feedback to clinicians and recommendations for decision-making can be roughly compared to navigation systems in cars, helping drivers to find the optimal path through traffic. Transferred to the context of psychotherapy, this means navigating through psychological treatments to find the optimal path to therapeutic change for individual patients. The basic research behind these systems can be broadly categorized into two research traditions: pre-treatment decisions and adaptive decisions during treatment (Lutz, Zimmermann, Müller, Deisenhofer, & Rubel, 2017). Research on pre-treatment decisions was pushed forward by the most recent debate about the implementation of personalized or precision medicine concepts into mental health (e.g., Cohen & DeRubeis, 2018; Huibers et al., 2015). To date, most applications and studies in this research tradition have tried to generate empirically-based decision rules for differential treatment selection. That is, they have attempted to answer the question which treatment is most promising for a specific patient (e.g., cognitive behaviour therapy vs. antidepressants). In recent applications, prediction models in a machine learning framework have been used to identify the

Abbreviations: Trier Treatment Navigator, TTN; Personalized Advantage Index, PAI; expected treatment response, ETR; not-on-track, NOT; on track, OT; Clinical problem-solving tools, CPST

* Corresponding author. Clinical Psychology and Psychotherapy, Department of Psychology, University of Trier, D-54296, Trier, Germany.
E-mail address: wolfgang.lutz@uni-trier.de (W. Lutz).

https://doi.org/10.1016/j.brat.2019.103438
Received 27 November 2018; Received in revised form 24 June 2019; Accepted 7 July 2019
Available online 08 July 2019
0005-7967/

best sets and weighting of patient characteristics to inform these selection decisions. Different strategies have been suggested to empirically define the most promising psychological treatment for a particular patient (e.g., Cuijpers, Ebert, Acarturk, Andersson, & Cristea, 2016; DeRubeis et al., 2014; Kessler et al., 2017; Lutz et al., 2005; Ng & Weisz, 2016). One early approach based its differential treatment predictions on the progress of already treated similar patients called *nearest neighbors* (Lutz et al., 2006, 2017). Other studies have developed indices like the Personalized Advantage Index (PAI) to predict the best available treatment for an individual patient (Cohen & DeRubeis, 2018; Deisenhofer et al., 2018; DeRubeis et al., 2014). Another concept is based on individual patient data network metaregression (Furukawa et al., 2018). This concept was used to enable personalized treatment recommendations for patients suffering from persistent depressive disorder for either pharmacotherapy or the cognitive-behavioral analysis system of psychotherapy (CBASP).

However, these applications all share two limitations: First, all applications used data to retrospectively decide which treatment protocol would have been the more promising treatment and compare the outcomes of patients who received their optimal treatment with those who received their non-optimal treatment. These retrospective analyses are promising, however there is a clear need for the development of easily applicable recommendation systems in order to test prospective applications in randomized controlled trials (RCT) under real world conditions. Second, which treatment protocol to provide is only one of a number of pre-treatment decisions that may be optimized through research on personalized decision-making. It is therefore important for therapists to contemplate beforehand, whether the specific techniques they are trained to provide will be helpful in the case at hand. It is not only interesting which treatment protocol works best for a given patient, but also which treatment strategy or technique within a treatment protocol may be the most promising given the specific intake characteristics of varying patients. In most clinical contexts, a fine-grained empirically-based decision support system that provides information on the optimal treatment strategy or the potential risk of non-completion (drop-out) for incoming patients seems useful to improve the evidence base of treatment (Fisher & Boswell, 2016; Lutz et al., 2017; Lutz et al., 2018; Rubel, Fisher, Husen, & Lutz, 2018; Zimmermann, Rubel, Page, & Lutz, 2017). In order to make prediction research more relevant to the individual therapist, models and statistics that provide recommendations regarding clinical strategies, techniques or risk estimates for non-completion within one treatment protocol seem necessary (Fisher & Boswell, 2016; Lutz et al., 2017; Lutz et al., 2018; Rubel et al., 2018; Zimmermann et al., 2017).

Irrespective of the quality of pre-treatment selection, a substantial number of patients remain who do not improve as expected during treatment and are thus at risk of negative outcomes or premature termination (Zimmermann et al., 2017). However, as has been repeatedly shown, therapists are not good at recognizing and predicting negative developments in patients' distress during treatment (Hannan et al., 2005). Here, the second broad feedback research tradition comes into play: monitoring patient progress in “real-time” over the course of treatment and feeding this information back to the therapist and/or patient to support adaptive or dynamic decision-making. Two names for this research tradition have emerged and are commonly used in the literature: routine outcome monitoring (ROM) and patient-focused feedback research. Using such psychometric feedback during the course of treatment has been shown to be effective to draw therapists’ attention towards problematic developments and thus allows dynamic adaptation of treatment, especially for patients at risk of treatment failure (e.g., Bar-Kalifa et al., 2016; Castonguay, Barkham, Lutz, & McAleavy, 2013; Kendrick et al., 2016; Shimokawa, Lambert, & Smart, 2010). For example, in a well-conducted RCT with $N = 2233$ patients, Delgadillo et al. (2018) examined the effects of feedback on priming therapists to review patients at risk of treatment failure in supervision. When therapists were provided with feedback, these risk patients were significantly less impaired after treatment and had a lower probability of reliable deterioration. Despite studies finding that continuous feedback can help to prevent treatment failure, its effects seem small to moderate (Delgadillo et al., 2018; Kendrick et al., 2016; Lambert, Whipple, & Kleinstäuber, 2018; Lambert et al., 2018). An essential element of such dynamic psychometric feedback are combinations of decision support rules and clinical support or problem-solving tools to improve treatment for patients at risk of treatment failure.

Decision support rules allow the comparison of expected treatment response (ETR) curves with actual client progress (Howard et al., 1996; Lutz et al., 2005; 2006). The aim is to estimate an expected trajectory of recovery for individual patients based on their outcome-relevant pre-treatment characteristics (Castonguay et al., 2013; Lambert et al., 2007; Rubel et al., 2015). With this approach, actual treatment progress can be compared to the expected course of treatment and warning signals can be provided to therapists if a patient's progress falls below a predefined failure boundary (Hooke, Sng, Cunningham, & Page, 2017). These cases showing such negative developments are known in the literature as “not-on-track” (NOT) or signal cases (Finch, Lambert, & Schaalje, 2001; Lambert et al., 2007).

Clinical problem-solving tools (CPST) build on these warning signals by further detailing in which clinical areas treatment adaptations may be needed in an individual NOT case (e.g., Whipple et al., 2003). So far, these recommendations consist of relatively basic information on each of the different problem areas, making suggestions about what could be done to alleviate the respective problem (Lambert, 2007; Whipple et al., 2003). A recently published meta-analysis found feedback effects (mean differences, SMD) for NOT patients of .33 *without* CPSTs and of 0.49 *with* CPSTs (Lambert et al., 2018). Furthermore, feedback nearly doubled clinically significant/reliable change rates in patients at risk of treatment failure, when CPSTs for signal patients were included (OR = 2.40 versus OR = 1.89 without CPSTs, Lambert, et al., 2018). Different problem areas have been discussed in the literature on CPSTs. Lambert (2007), for example, suggested four different psychological domains before referring to psychopharmacological treatment: therapeutic alliance, motivation, social support and life events. White et al. (2015) found that of 107 predicted at-risk patients, only 58% scored high enough to set off an alarm in at least one of these areas. Hence, it was possible to identify an obstacle to a positive treatment outcome for 58% of the cases. Nonetheless, for 42% of the cases, no target area could be identified, which is why adding an additional problem area could enhance previous versions. A fruitful additional area may be emotion regulation, as it has been shown to be relevant in the development and maintenance of various clinical disorders (e.g. Hofmann & Kashdan, 2010).

The recent debate about precision and personalized medicine is reflected in the varying purposes of prediction and adaptation systems. Pre-treatment decision recommendations attempt to forecast drop-out or the most promising treatment approach or strategy given specific patient characteristics; adaptive decision recommendations tailor ongoing treatments to the patient's individual treatment response in “real-time". So far, both approaches have been applied and investigated separately. This paper will present the research behind the development of the Trier Treatment Navigator (TTN), a comprehensive feedback system supporting the following four steps of personalized decision-making:

1. Personalized pre-treatment recommendations:
   1. Prediction of drop-out risk.
   2. Prediction of optimal treatment strategy.
2. Personalized adaptive recommendations during treatment:
   1. Dynamic risk index (failure boundary) to identify patients at risk of treatment failure.
   2. Clinical problem-solving tools (CPST) for personalized treatment adaptation.

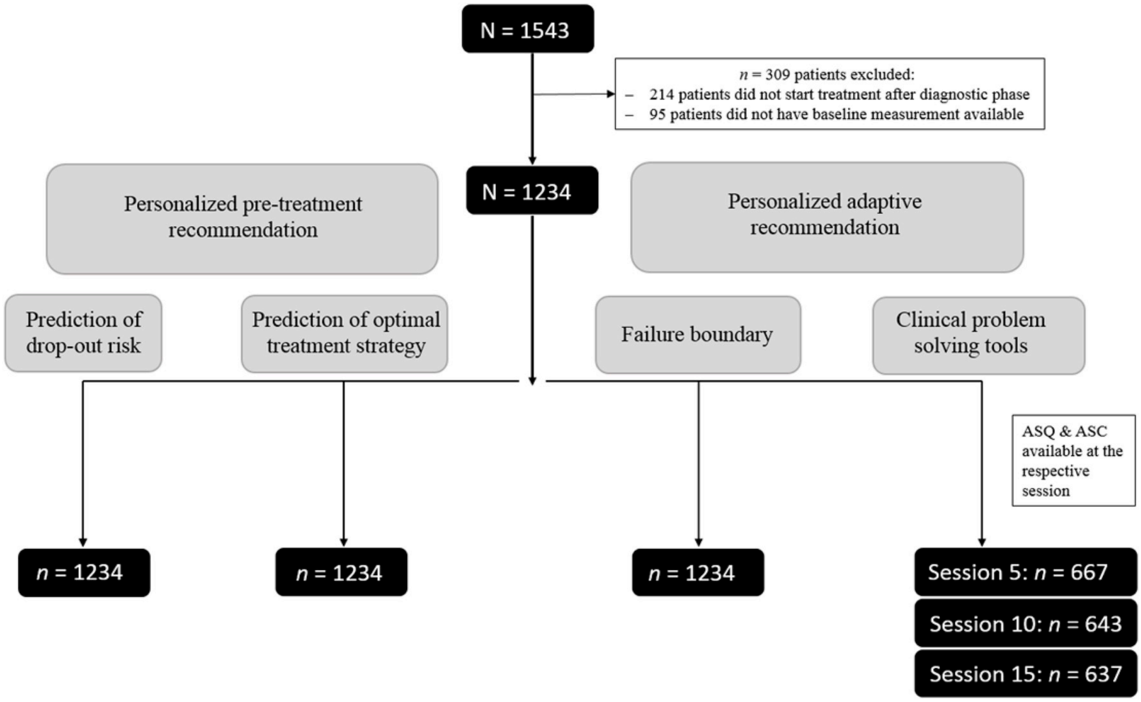

Fig. 1. Patient flow chart.

2. Method

2.1. Patients and treatment

2.1.1. Patients

The development of the TTN was based on a sample of $N = 1543$ patients who applied for individual psychotherapy at a large university outpatient clinic between 2009 and 2016. A flowchart regarding the four research questions is depicted in Fig. 1. Patients included in the analysis ($n = 1234$) had begun individual psychotherapy after the diagnostic phase and either completed treatment (consensual decision to end therapy) or dropped out of treatment. An additional inclusion criterion was the availability of baseline measurements. On average, patients received 30.85 treatment sessions ($SD = 20.24$).

Patients included in the analysis were an average of 35.80 years of age ($SD = 12.61$, minimum = 14, maximum = 76) and the majority was female (61.9%). In the sample, 37.1% of the patients were single and 59.2% were married or in a long-term relationship. In addition, 42.7% had completed university entrance qualifications and 13.3% had at least a bachelor's degree. Of the 1234 patients, 257 (20.8%) were incapable of work. The *Structured Clinical Interview for Axis I DSM-IV Disorders-Patient Edition* (SCID-I; First, Spitzer, Gibbon, & Williams, 1995) and *International Diagnostic Checklist for Personality Disorders (IDCL-P;* Bronisch, Hiller, Mombour, & Zaudig, 1996) were used to make diagnoses at intake. Interviews were conducted by intensively trained independent clinicians. These interviews were videotaped, interviews and diagnoses were discussed in expert consensus teams that included four senior clinicians and final diagnoses were determined by consensual agreement of at least 75% of the team members. Most patients suffered from affective disorders as their primary diagnosis (54.2%), followed by anxiety disorders (15.8%). Additional diagnoses were adjustment disorder (8.0%), PTSD (5.7%), somatoform disorders (3.9%), eating disorders (2.5%), and obsessive compulsive disorder (1.6%). Of the total sample, 21.2% ($n = 262$) fulfilled the criteria for a personality disorder.

According to the patients' pretreatment assessments, the mean scores of the short-form of the Outcome Questionnaire (OQ-30; Ellsworth, Lambert, & Johnson, 2006) and the Brief Symptom Inventory (BSI; Franke, 2000; German translation of Derogatis, 1977) were 1.88 ($SD = 0.55$) and 1.28 ($SD = 0.71$) respectively. These mean scores indicate moderate to severe symptoms in general.

2.1.2. Therapists

Patients were treated by 128 therapists (107 women and 21 men) in clinical training (cognitive-behavioral therapy). All therapists had at least one year of clinical experience in inpatient settings before beginning to see patients at the outpatient clinic. Therapists treated an average of 9.63 patients, while 14 therapists treated only one patient and one therapist treated 24 patients. Each therapist received 1 h of individual or group supervision per case every four sessions. All therapy sessions were videotaped for use in supervision and research. Supervisors were senior clinicians. Individual psychotherapy consisted of weekly sessions.

2.2. Instruments

For an overview of instruments and assessment points over the course of treatment, see Table 1 as well as supplemental material 1.

Table 1
Overview of assessments over the course of treatment.

Pre treatment	Every Session	Every fifth session	Post treatment
• SCID-I • IDCL-P • Sociodemographic and psychometric assessment (see supplemental material)	• HSCL-11	• OQ-30 • ASC • ASQ	• OQ-30 • ASC • ASQ

Note. Structured Clinical Interview for Axis I DSM-IV Disorders-Patient Edition (SCID-I), International Diagnostic Checklist for Personality Disorders (IDCL-P), Hopkins Symptom Checklist – Short Form (HSCL-11), Outcome Questionnaire-30 (OQ-30), Assessment for Signal Clients (ASC), Affective Style Questionnaire (ASQ).

2.2.1. Drop-out

Drop-out was assessed via the therapist's evaluation at the end of each treatment. If termination was planned and consensual, treatment was considered completed (cf. Zimmermann et al., 2017). However, if the patient terminated treatment contrary to the therapist's recommendation, treatment was considered a drop-out. Examples of drop-out are when the patient stopped showing up for appointments and was unable to be contacted or when the patient told the therapist that he or she was no longer interested in therapy despite the therapist advising continuation.

2.2.2. Hopkins Symptom Checklist – short form (HSCL-11)

The HSCL-11 (Lutz, Tholen, Schürch, & Berking, 2006) is an 11-item self-report inventory for the assessment of symptomatic distress. It was developed based on the HSCL-25 (Coyne et al., 1987), which is a brief version of the Hopkins Symptom Checklist-90 (Derogatis et al., 1977). Patients were asked to what degree they suffered from the respective symptom in the last seven days. For each of the eleven symptoms, patients answered on a 4-point Likert scale ranging from 1 ("not at all") to 4 ("extremely"). The mean of the 11 items represents the patient's level of global symptomatic distress for the preceding week. The global score is highly correlated with the BSI ($r = 0.91$; Lutz et al., 2006) and has a high internal consistency ($\alpha = 0.92$; Lutz et al., 2006). In the present sample, the HSCL-11 was administered at the beginning of each session. In the analyses, the HSCL-11 was used as a predictor as well as an outcome instrument.

2.2.3. Outcome Questionnaire-30 (OQ-30)

The OQ-30 is a 30-item self-report measure designed to assess patient outcome over the course of therapy. It is a short form of the OQ-45, comprising the 30 items that are most sensitive to patient change, and has demonstrated high levels of congruence with the OQ-45 in measurement of patient outcome (Ellsworth et al., 2006; Vermeersch et al., 2004). The OQ has three primary dimensions: (a) symptomatic distress, (b) interpersonal relationships, and (c) social role performance. All 30 items can be aggregated to create a total score. Total scores range from 0 to 120, with higher scores reflecting poorer psychological functioning. In a previous study, the OQ-30 showed adequately high internal consistency (pretreatment $\alpha = .89$ and posttreatment $\alpha = 95$; Bar-Kalifa et al., 2016). The instrument was administered before and after treatment as well as after every fifth session. The total score as well as the OQ-30's subscales were used as predictors in the analyses. Since the OQ-30 is also administered every fifth session, it was used as an outcome instrument to measure reliable improvement and deterioration for on-track and not-on-track patients.

2.2.4. Questionnaire for the Evaluation of Psychotherapeutic Progress (FEP-2)

The FEP-2 is a measure of therapeutic progress, which can be used for both change and outcome assessment (Lutz, Tholen, et al., 2006). The instrument has 40 items and measures the dimensions well-being, symptoms, interpersonal relationships, and incongruence with respect to approach and avoidance goals. Items can be aggregated to a total score. It is answered on a 5-point Likert (never, seldom, sometimes, often, and always). The instrument has been shown to be change sensitive as well as reliable (total score $\alpha = 0.93$; subscales range from $\alpha = 0.81$ to 0.87) and is available in the public domain (Böhnke & Lutz, 2014; Lutz et al., 2006). The instrument was administered before treatment. Again, the total score as well as the subscales were tested as predictors in the analyses.

2.2.5. Brief Symptom Inventory (BSI)

General symptom severity was measured using the BSI (Derogatis, 1977; Franke, 2000; German translation of), which is a 53-item self-report inventory that inquiries about physical and psychological symptoms during the past week. It is the brief form of the Derogatis' Symptom Check-List-90 Revised (SCL-90-R; Derogatis, 1977), which assesses 9 subscales with the following dimensions: somatization, obsessive-compulsive, interpersonal sensitivity, depression, anxiety, hostility, phobic anxiety, paranoid ideation, and psychoticism. The items are scored on a 5-point Likert scale ranging from 0 ("not at all") to 4 ("extremely"). The internal consistency of the BSI has been found to be $\alpha = 0.92$ and the retest-reliability $rtt = 0.90$ (Franke, 2000). In this study, the BSI was administered at the beginning of treatment. The Global Severity Index (GSI), which was calculated by averaging all items, as well as the subscales were tested as potential predictors.

2.2.6. Treatment expectations

Two items assessing treatment expectations were used as predictors in the analyses of this study (Lutz et al., 2006). At the beginning of therapy, patients rated the item: "How confident are you, that psychotherapy will be successful in helping you with your problems?" on a 4-point Likert scale ranging from 1 ("not at all confident") to 4 ("very confident"). The second item asked the therapist the same question regarding his/her expectation of patient progress: "How much improvement can realistically be expected if the patient continues treatment?" on a 4-point Likert scale ranging from 1 ("no more improvement") to 4 ("substantially more improvement"). Treatment expectations assessed at the beginning of treatment were included as predictors in the analyses.

2.2.7. Personality style (PSSI–K)

The personality style was assessed using the PSSI-K Kuhl and Kazén (1997). This instrument is consisting of 54 items, belonging to 14 dimensions each measured with 4 items. It assesses personality styles based on DSM-IV personality disorder descriptions and high values in any dimension suggest the existence of a personality disorder (Kuhl & Kazén, 1997). Statements are self-rated on a scale ranging from 0 ("not at all") to 3 ("totally agree"). All but one dimension showed good internal consistencies between $\alpha = 0.70$ and $\alpha = 0.77$ except the narcissism scale with $\alpha = 0.58$. All subscales administered at the beginning of treatment were entered into the analyses as potential predictors.

2.2.8. Assessment for Signal Clients

The Assessment for Signal Clients (ASC; Lambert et al., 2007) is a 40-item self-report measure, which is composed of four subscales (therapeutic alliance, social support, motivation, and life events). Questions are answered on a 5-point Likert-scale ranging from "strongly disagree" to "strongly agree". Sum scores are calculated for each of the subscales, however, there is no total score. Probst, Lambert, Dahlbender, Loew, and Tritt (2014) found satisfactory Cronbach's α coefficients, namely: TA:$\alpha = 0.89$; SS:$\alpha = 0.76$; MO: $\alpha = 0.78$; and LE: $\alpha = 0.71$. Risk signals and CPST were based on previously reported cutoff scores for each subscale (alliance ≤ 39, social support ≤ 23, motivation ≤ 32, critical life events ≤ 23; Lambert et al., 2007). The ASC was later introduced into the routine assessment procedure of the clinic. Therefore, the datasets were smaller for analyses including this instrument. The ASC was administered after every fifth session and used to identify patients at risk for treatment failure.

2.2.9. Affective Style Questionnaire

The Affective Style Questionnaire (Graser et al., 2012; Hofmann & Kashdan, 2010) is a 20-item instrument that assesses emotion regulation based on three broad regulation styles: suppression, acceptance, and reappraisal. Questions are answered on a 5-point Likert-scale ranging from "not true of me at all" to "extremely true of me". Internal consistencies of the subscales are acceptable (suppression: $\alpha = 0.84$, acceptance: $\alpha = 0.72$ and reappraisal: $\alpha = 0.75$; Graser et al., 2012). It is part of the process of identifying signal patients who have serious emotion regulation problems. As soon as a patient reaches the cutoff score on one scale, the CPST focusing on emotions, is activated. For each subscale, the cutoff score is calculated by using the mean of an archive dataset and subtracting 1*SD* (suppression = 3.01–0.74, acceptance = 2.97–0.68 and reappraisal = 4.54–0.75). The ASQ was later introduced into the routine assessment procedure of the clinic. Therefore, the datasets were smaller for analyses including this instrument. Like the ASC, the ASQ was administered every five sessions.

2.2.10. Predictors

Potential predictors for all analyses were grouped into demographic and clinical factors. A total of 78 possible predictor variables (supplemental material 1), routinely collected at intake, were used in the first step of predictor selection. Demographics included for example age, gender, ethnicity and employment status. Clinical factors characterized the history and severity of the patient's condition, including family history of psychiatric problems and chronicity of mental health problems (in years and months), but also variables such as outcome expectancy, dysfunctional attitudes and interpersonal problems (see above and supplemental material 1 for a detailed list of predictors).

2.3. Data analytic strategy

Analyses were directed toward the development of an empirical basis for visualizations supporting the four types of personalized decision-making in the TTN. The TTN was developed to be used until session 25 in accordance with research on dose-response relations, showing that change happens within that time frame, and previous studies on such decision support tools that had a similar treatment length (Lambert, 2013).

All analyses were conducted with the free software environment R version 3.1.1 (R Core Team, 2014). Missing value imputation was performed using the R package missForest v1.4 (Stekhoven, 2013) for all variables with < 20% missings. For variables that were suitable for imputation, between 1.30% and 15.47% of the data were missing (M = 9.42; SD = 2.34). Visualizations in the feedback system were displayed using the R package ggplot2 v2.2.0 (Wickham, 2016). All variable selection models were estimated using the R package SparseLearner v1.0-2 (Guo & Hao, 2015).

2.4. Pre-treatment decision recommendations

2.4.1. Predictor selection strategy for drop-out risk

To select potential predictors of drop-out, a two-step procedure was applied. First, bivariate correlations were used as a screening and second, a more conservative machine learning approach (LASSO, least absolute shrinkage and selection operator; Guo et al., 2015) was applied in order to protect against overfitting.

Accordingly, first the whole set of variables (k = 78, see supplemental material 1) collected before the first session was correlated with drop-out using a point-biserial correlation (n = 1234). Variables with significant correlations were then included in a LASSO logistic regression model.[1]

To obtain regression weights for the prediction of drop-out in new cases, the significant predictors indicated by the LASSO model were set as predictors in a logistic regression using a generalized linear model:

$$\log\left(\frac{dropout}{1-dropout}\right)_i = b_0 + b_1 * X_{1i} + b_2 * X_{2i} + \ldots + b_k * X_{ki} + e_i$$

where $\log(\frac{dropout}{1-dropout})_i$ is the logarithmized odds of dropping out of treatment for patient i, is the regression intercept, b_1 to b_k are the regression coefficients for predictors 1 to k, X_{1i} to X_{ki} are the values of predictors 1 to k for patient i, and e_i is the residual. The model was evaluated using Nagelkerke's R^2 that, as a pseudo-R^2, approximates the explained variance of dichotomous dependent variables. Additionally, a brier score as well as the area under the curve (AUC) of the receiver operating characteristic (ROC) curve were calculated to confirm the result, because R^2 estimates might be treated with caution for dichotomous outcomes (Peng, Lee, & Ingersoll, 2002).[2] The aim of the drop-out prediction was not to classify patients categorically, but to provide therapists with an individual probability to enhance attention for patients at risk. Therefore, the positive predictive value (PPV) as a measure of identifiability of patients at risk for drop-out was evaluated for different thresholds.

2.4.2. Personalized predictions of the optimal treatment strategy

The development of the optimal treatment strategy prediction was also based on the above described dataset of n = 1234 patients that were treated with CBT. Treatment strategies were defined based on therapist ratings of the predominantly applied intervention strategy during the first 10 sessions of treatment. Often in clinical practice, some patients seem not to engage in treatment and lack the necessary compliance for protocol-specified problem-solving interventions. In order to overcome this barrier, Schulte and Eifert (2002), for example, describe two tasks in their dual model that therapists have to deal with especially early in treatment. First, the application of treatment techniques/problem-solving procedures and second, the motivation of patients for treatment by clarifying treatment goals and strengthening the therapeutic relationship, if necessary. According to the dual model, therapists attend and react to therapeutic process problems by switching between a method-oriented treatment strategy and a process-oriented strategy. The authors suggest the development of empirically based criteria to inform therapists what to look for, when to react, and which intervention strategy is preferable to engage the patient early in treatment. Technical details on the definition of treatment strategies can be found in supplemental material 2.

For the recommendation of a treatment strategy, again the above described two-step variable selection procedure was applied. First, the set of 78 potential predictors of pre to post improvement on the HSCL-11, controlled for initial impairment, was examined. All significantly correlating variables were likewise included in a LASSO selection algorithm.[3]

Based on the selected variables, we identified the most similar patients (nearest neighbors) per treatment strategy (problem-solving, mixed, motivation-oriented) for each patient. The nearest neighbor approach is an early machine learning technique, which has an optimal relative human-to-machine decision-making ration, as it remains clinically comprehensible and communicable (Beam & Kohane, 2018). Details on the nearest neighbor procedure are provided in supplemental material 3.

To evaluate the prognostic validity of the nearest neighbor procedure, two thresholds were tested. To be recommended, the predicted effect size[4] for one strategy had to be at least 0.1 (threshold 1) or 0.2 (threshold 2) larger than for the two other treatment strategies in the first ten sessions. Based on these thresholds, patients were categorized into either having been treated with their optimal or their non-optimal treatment strategy. These groups were then compared regarding effect size as well as change rates (reliable improvement/deterioration;

[1] A bootstrap ranking procedure with a 10-fold cross-validated LASSO logistic regression model with 100 bootstraps for binary outcomes was used.

[2] Nagelkerke's R^2 represents the corrected ratio of the likelihood functions of the null model and the model with explanatory variables. The brier score is calculated as the mean squared difference between predicted probabilities and actual outcomes, therefore ranging from 0 to 1, with 0 being a perfect prediction. The ROC curve plots the true positive rate against the false positive rate at various threshold settings, with the AUC as a measure of aggregated classification performance.

[3] A bagging prediction for linear regression models with a 10-fold cross-validation and 50 base-level models for continuous outcome was used.

[4] Calculated as the standardized (using the pre SD) mean difference between the pre-assessment and the assessment before session ten.

Jacobson & Truax, 1991)[5] on the HSCL-11 and OQ-30 until session ten.

2.5. Adaptive decision recommendations during treatment

2.5.1. Dynamic failure boundary to identify patients at risk for treatment failure

The development of the dynamic failure boundary to identify patients at risk for treatment failure was also based on the dataset above described ($n = 1234$). The prediction analyses also relied on nearest neighbors selected as described above, but extended to model a dynamic failure boundary over the course of treatment for a reference group with a positive (i.e., successful) course of treatment.

To model the course of treatment for each patient based on their nearest neighbors, impairment measured by the HSCL-11 was regressed on the logarithmized session variable, the number of sessions as well as an interaction effect between them. In accordance with prior research on change trajectories in psychotherapy, a log-linear (base 10) transformation of the time scale (sessions) was used for these analyses[6]:

$$\text{HSCL-11}_{tj} = \gamma_{00} + \gamma_{10}{}^{*}\ \log(\text{session})_{tj} + \gamma_{01}{}^{*}\#\text{sessions}_j + \gamma_{11}{}^{*}\#\text{sessions}_j{}^{*}\log(\text{session})_{tj}.\ e_{0j} + e_{1j} * \log(\text{session})_{tj} + r_{tj}$$

Where HSCL-11_{tj} is impairment measured by the HSCL-11 at session t for patient j, γ_{00} is the overall intercept, γ_{10} and γ_{01} are the main effects of the time variable (i.e. the logarithmized session number) and the total number of sessions per patient (i.e. #sessions), respectively. Furthermore, a cross-level interaction between the number of sessions and the time variable is modeled, represented by the γ_{11} coefficient. Person-specific deviations from the average mean and slope are realized by adding the random terms e_{0j} and e_{1j} and residuals are modeled by adding the level-1 error term r_{tj}. By centering the variable number of sessions at the overall average session number ($M = 35.85$), these predictions pertain to an assumed average treatment length. This approach accounts for the repeated finding in growth models of naturalistic studies, that the change rate depends on the overall number of sessions (i.e., patients with fewer sessions change faster; e.g., Baldwin, Berkeljon, Atkins, Olsen, & Nielsen, 2009).

The failure boundary for being on- or not-on-track was then calculated based on the upper limit of the 90% confidence interval (e.g., Whipple et al., 2003). In order to dynamically adapt the failure boundary, the prediction was updated each session by additionally taking into account the amount of change from intake up to that session (for nearest neighbor patients with positive change).

Each time, the patient's observed impairment exceeded the boundary, the patient was signaled as not-on-track. To get back on-track, impairment had to fall below the boundary, which was dynamically recalculated each session. Additionally, in order to minimize the effects of measurement error, if a patient crossed the failure boundary, improvement on the HSCL-11 had to be at least reliable (RCI) in relation to the level of impairment the first time the boundary was crossed. As soon as the failure boundary is crossed, the system provides the therapist with a warning signal to alert the clinician that symptom change is not as expected and further clinical adaptation is needed. To evaluate the failure boundary for patients at risk for negative treatment outcome, the relative frequencies of reliable improvement and deterioration from pre-treatment to session 25 for on-track and not-on-track patients were calculated.

2.5.2. Clinical problem-solving tools (CPST) for personalized treatment adaptation

Within the TTN, therapists are supported by different clinical problem-solving tools as soon as a failure boundary is exceeded, that is, symptom progress assessed by the HSCL-11 is not as positive as expected. The CPST are composed of the following problem areas (1) *risk/suicidality*, (2) *motivation/therapy goals,* (3) *therapeutic alliance,* (4) *social* support *and critical life events as well as* (5) *emotion regulation/self-regulation*. Each problem area or domain is assessed every fifth session by either the ASC, ASQ or single items from the OQ-30 and HSCL-11.

The tools have a common structure. First, a thematic overview is presented. Second, critical items of the associated scale are displayed. Third, in-depth questions prompt therapists to make a reference to their own patients. Fourth, the tools provide recommendations for interventions that can help the therapist resolve the problem. Further, details on the development of the tools can be found in supplemental material 4.

The association between crossing the failure boundary based on the HSCL-11, i.e., being not-on-track, at session t + 1 and the problem domains of the CPST at session t was evaluated for the first three measurements of ASC and ASQ (t = 5, t = 10 and t = 15). In logistic regression analyses, being not-on-track was regressed on *risk/suicidality* measured by single items of the OQ-30 and HSCL-11, the ASC subscales *motivation, therapeutic alliance, social* support, and *critical life events* as well as *emotion regulation* measured by the ASQ. Variance explained by the models was again estimated using Nagelkerke's pseudo-R^2.

3. Results

3.1. Pre-treatment decision recommendations

3.1.1. Prediction of drop-out risk

The two-step variable selection procedure initially identified 30 of 78 variables, which significantly correlated with drop-out ($p \leq .05$). Seven of these variables were selected by the LASSO logistic regression model (FEP-2 total score, HSCL-11 total score, OQ-30 subscale interpersonal relations, PSSI–K subscales histrionic and compulsive, therapist-rated treatment expectation, and education). To determine regression weights for patient-specific predictions, these variables were entered into a generalized linear model (see Table 2). Drop-out probability was predicted to be higher for patients with higher impairment on the FEP-2 ($b = -0.697$, $p = .001$), lower impairment on the HSCL-11 ($b = 0.609$, $p = .001$), a more histrionic personality style ($b = 0.359$, $p = .001$), higher impairment of interpersonal relationships ($b = 0.530$, $p < .001$), a less obsessive personality style ($b = -0.320$, $p = .004$), a lower therapist treatment expectation ($b = -0513$, $p < .001$) and a lack of university entrance qualification ($b = -0.610$, $p < .001$) .[7]

Nagelkerke's pseudo-R^2 for the final model was $R^2 = 0.120$, thus explaining 12.0% of variance. The brier score was 0.161 and the ROC curve showed an AUC of 0.667. For a threshold of drop-out probability $\geqq 0.3$, which is just above the average drop-out rate of 22.6%, the positive predictive value (PPV, the probability that a patient drops out, if they are above the risk threshold for drop-out) was = 0.394 and

[5] The reliable change index (RCI) was RCI = 0.38 for the HSCL-11 and RCI = 0.28 for the OQ-30. Reliable improvement refers to a symptom reduction of at least an RCI, reliable deterioration to a symptom increase of at least an RCI.

[6] Dose-effectiveness research has shown a consistent pattern in most analyses of rapid response early in therapy (e.g., Lambert, 2007). This consistent curvilinear pattern is parsimoniously approximated by a log-linear transformation of session number, as in the present case, and widely used in this area of research (see also Stulz, Lutz, Kopta, Minami, & Saunders, 2013).

[7] These are results of the GLM. The resulting coefficients of the BRLasso as well as the GLM are unstandardized regression coefficients. They can be interpreted as the extent to which the logarithmized odds of dropping out of treatment change, when the predictor variable increases by one. An important difference between Lasso and GLM coefficients is the shrinkage in Lasso models. Coefficients of the Lasso model are shrunk to avoid overfitting and to select the variables that are most likely to be predictive, also in new data. Therefore, the Lasso coefficients are smaller than the GLM coefficients with the exception of one variable.

Table 2
Final prediction model for drop-out.

	coefficients		p-value
	BRLasso	GLM	
Intercept		0.739	.210
FEP-2	−0.230	−0.697**	.001
HSCL-11	0.261	0.609**	.001
PSSI–K – subscale histrionic	0.322	0.359**	.001
OQ-30 – subscale interpersonal relationships	0.411	0.530***	< .001
PSSI–K – subscale obsessive-compulsive	−0.416	−0.320**	.004
Therapist's treatment expectation	−0.509	−0.513***	< .001
University entrance qualification	−0.586	−0.610***	< .001

Note. coefficients = unstandardized regression coefficients; BRLasso = Bootstrap Ranking Least absolute shrinkage and selection operator; GLM = Generalized linear model; FEP-2 = Questionnaire for the Evaluation of Psychotherapeutic Progress-2; HSCL-11 = Hopkins Symptom Checklist – Short Form; OQ-30 = Outcome Questionnaire-30; PSSI–K = Personality style and disorder inventory; ** = $p < .01$; *** = $p < .001$.

the negative predictive value (NPV) was 0.822. The accuracy of the prediction was .728 [0.702; 0.753] with a sensitivity of 0.382, a specificity of .829 and a Kappa of .213. Each patient-specific drop-out probability is displayed in the TTN in comparison to the average drop-out rate at the outpatient clinic for treatment planning (e.g., awareness of motivation problems; see Fig. 2 a).

3.1.2. Personalized predictions of the optimal treatment strategy

The two-step variable selection procedure first identified five variables (of 78) that correlated significantly with HSCL-11 scores ($p \leq .05$) over the course of treatment. Second, using the more conservative bagging LASSO procedure reduced the number to four predictors of HSCL-11 score improvement (higher intake HSCL-11 score, higher patient-rated treatment expectation, more previous treatments,[8] and higher chronicity[9] were associated with more improvement on the HSCL-11).

As described, the effect size on the HSCL-11 from pre-treatment to session 10 was calculated for the three strategies based on nearest neighbors. For 310 patients (25.12% of n = 1234), the predicted effect size for one strategy was at least 0.2 larger than for both other strategies. Among these patients, those who were treated with the recommended strategy (optimal) had a significantly higher effect size than patients treated with a non-recommended strategy (non-optimal) on the HSCL-11 ($M_{\text{optimal}} = 0.832$, $M_{\text{non-optimal}} = 0.620$; $t_{201.01} = -1.99$, $p = .048$) as well as on the OQ-30 ($M_{\text{optimal}} = 0.636$, $M_{\text{non-optimal}} = 0.413$; $t_{201.01} = -2.19$, $p = .029$). Additionally, the probability of optimal cases to be reliably improved until session 10 was significantly higher than for non-optimal cases on the HSCL-11 (63.37% vs. 50.20%; $\chi^2_{df=1} = 4.22$, $p = .040$) as well as the OQ-30 (63.37% vs. 49.76%; $\chi^2_{df=1} = 4.54$, $p = .033$) (Table 3).

For 614 patients (49.76% of n = 1234), the predicted effect size for one strategy was at least 0.1 larger than for both other strategies. Among these patients, those who were treated with the optimal strategy showed a higher effect size than patients treated with a non-optimal strategy on the HSCL-11 ($M_{\text{optimal}} = 0.696$, $M_{\text{non-optimal}} = 0.543$; $t_{201.01} = -1.81$, $p = .072$) as well as the OQ-30 ($M_{\text{optimal}} = 0.535$, $M_{\text{non-optimal}} = 0.409$; $t_{201.01} = -1.70$, $p = .090$). Although these differences in effect size were not statistically significant at the 5%-level, the probability for optimal cases to be reliably improved until session 10 was significantly higher than for non-optimal cases on the HSCL-11 (55.93% vs. 46.22%; $\chi^2_{df=1} = 4.37$, $p = .037$) as well as the OQ-30 (54.80% vs. 45.54%; $\chi^2_{df=1} = 3.97$, $p = .046$) (Table 3). For each new patient, the average effect size for each treatment strategy (based on already treated nearest neighbor patients) is visualized in the TTN (see Fig. 2b). If a strategy shows a higher effect size, this strategy was more successful for patients similar to the new patient and is recommended. Accordingly, in the example screenshot (see Fig. 2b) a motivation-oriented strategy was recommended.

3.2. Adaptive decision recommendations during treatment

3.2.1. A dynamic failure boundary to identify patients at risk for treatment failure

Of the 1234 patients used for the validation of the dynamic failure boundary, 498 (40.36%) were not-on-track at least once during the first 25 sessions and 57 (4,6%) had no sufficiently similar positively changing cases and no prediction was calculated. NOT cases had significantly higher HSCL-11 scores after 25 sessions than on track (OT) cases ($M_{OT} = 1.62$, $SD_{OT} = 0.53$; $M_{NOT} = 2.01$, $SD_{NOT} = 0.67$; $t_{df=912.37} = -10.67$, $p < .001$). This corresponds to an effect size for OT patients of $d_{OT} = 1.01$ [0.85 1.17] and an effect size of $d_{NOT} = 0.25$ [0.07 0.42] for NOT patients. Table 4 shows the rates of patients in both groups (OT and NOT), whose scores in the HSCL-11 reliably improved or deteriorated from pre to session 25. The probability for NOT patients to be reliably improved was about half that of OT patients (35% vs. 62.15%; $\chi^2_{df=1} = 82.77$, $p < .001$). Moreover, NOT patients showed a substantially higher probability for reliable deterioration (17% vs. 1.6%; $\chi^2_{df=1} = 92.49$, $p < .001$).

Similarly, NOT cases had significantly higher OQ-30 scores than OT cases ($M_{OT} = 1.41$, $SD_{OT} = 0.65$; $M_{NOT} = 1.68$, $SD_{NOT} = 0.61$; $t_{df=1107} = -7.38$, $p < .001$). This corresponds to an effect size of $d_{OT} = 0.76$ [0.65 0.87] for OT patients and an effect size of $d_{NOT} = 0.42$ [0.30 0.55] for NOT patients. Table 4 shows the rates of patients in both groups (OT and NOT), whose scores in the OQ-30 reliably improved or deteriorated. The probability for NOT patients to be reliably improved was significantly lower than that of OT patients (43.57% vs. 55.96%; $\chi^2_{df=1} = 17.65$, $p < .001$). Moreover, NOT patients showed a substantially higher probability for reliable deterioration (12.45% vs. 3.24%; $\chi^2_{df=1} = 35.28$, $p < .001$).

Furthermore, treatments of NOT cases were significantly longer than those of OT cases ($M_{OT} = 25.82$, $M_{NOT} = 38.47$, $t_{df=1098.7} = -11.28$, $p < .001$).

3.2.2. Monitoring and clinical problem-solving tools (CPST) for personalized treatment adaptation

To evaluate the association between crossing the failure boundary and the problem domains of the CPST, logistic regression analyses were established. At session 5, data on the HSCL-11, OQ-30, ASC, and ASQ was available for $n = 667$ patients[10] Of the six potential predictors, risk/suicidality ($b = 0.266$, $p < .001$) and life events ($b = -0.515$, $p = .003$) were significant predictors of the probability to be not-on-track at session 6. The model explained about 18% of variance in being not-on-track ($R^2_{\text{Nagelkerke}} = 0.180$). At session 10 ($n = 643$), again risk/suicidality ($b = 0.248$, $p < .001$) and life events ($b = -0.734$, $p < .001$) were significant predictors of the probability to be not-on-track at session 11, with an explained variance of $R^2_{\text{Nagelkerke}} = 0.192$. At session 15 ($n = 637$), higher risk/suicidality ($b = 0.112$, $p = .028$), less emotion regulation ($b = -0.320$, $p = .028$), less social support ($b = -0.390$, $p = .041$) and less life events ($b = -0.927$, $p < .001$) were significantly associated with being not-on-track at session 16. The model had a pseudo-R^2 of $R^2_{\text{Nagelkerke}} = 0.218$, thus explaining about

[8] “How much psychotherapy have you had in the past?”

[9] “How long has the problem for which you are presently seeking treatment been a concern to you?”

[10] As described in the methods section, the ASC and ASQ were later introduced into the routine assessment procedure of the clinic. Therefore, the datasets were smaller for these analyses. We compared sociodemographic characteristics, BSI, OQ-30 and HSCL-11 scores of the different datasets with the complete sample (n = 1234) and found no significant differences.

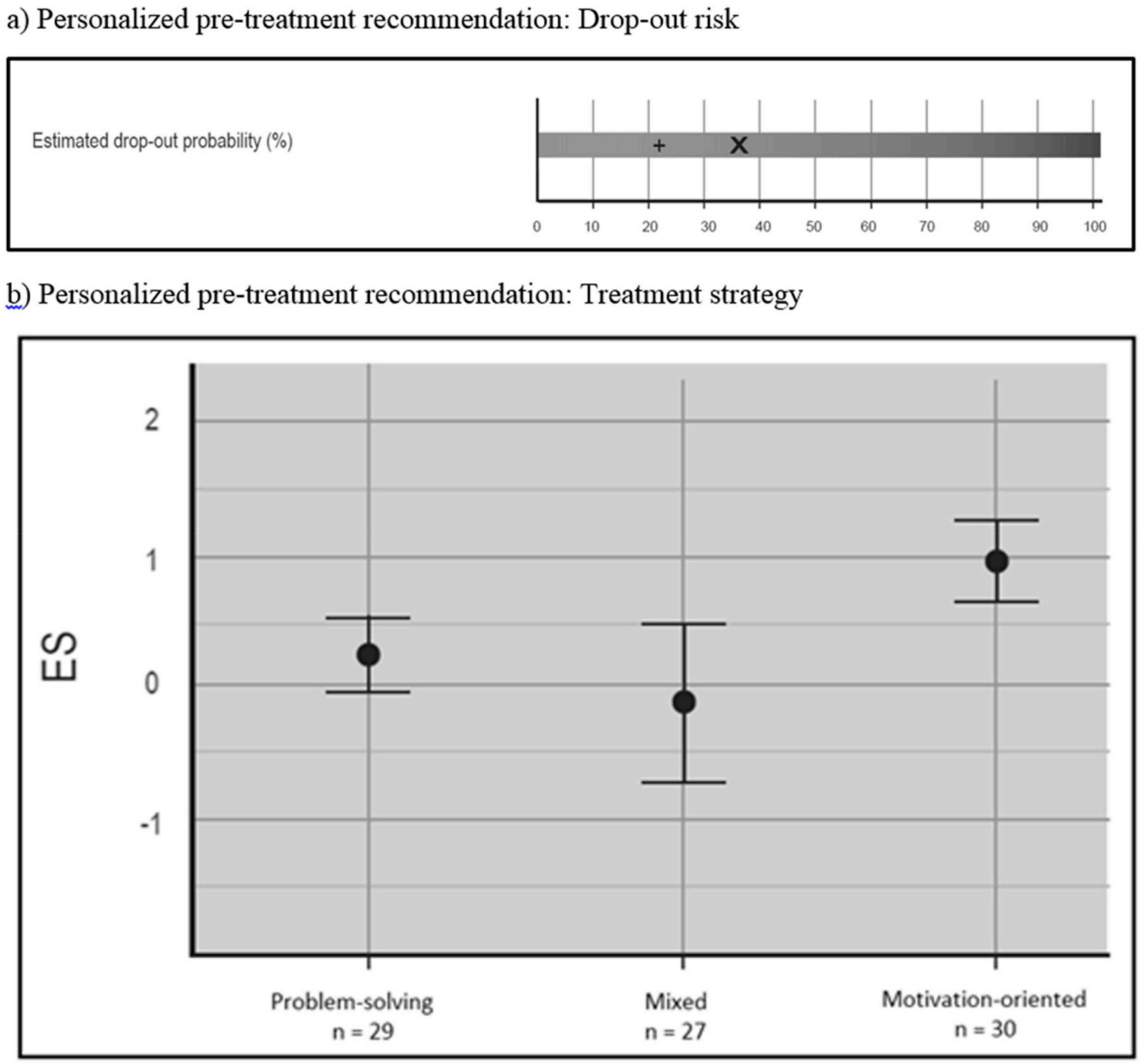

Fig. 2. Screenshot of a patient-specific pre-treatment recommendation as it is displayed to therapists within the Trier Treatment Navigator (TTN).
Note. a) Drop-out risk (in %): X = patient-specific drop-out probability; + = average drop-out rate at the outpatient clinic. b) treatment strategy: ES = effect size; For a maximum of 30 most similar cases already treated, the average effect size (treatment effectiveness) is determined for the three treatment strategies: problem-solving, mixed and motivation-oriented. Motivation-oriented shows the highest effect size, meaning this strategy was more successful in the cases most similar to this specific patient and is preferable to the other two.

Table 3
Effect sizes and relative frequencies of reliable improvement and deterioration from pre-treatment to session 10 for patients treated with the recommended (optimal)/not recommended (non-optimal) treatment strategy.

Diff.	strategy	n (%)	HSCL-11			OQ-30		
			ES	Rel. imp. (% per row)	Rel det. (% per row)	ES	Rel. imp. (% per row)	Rel det. (% per row)
> = 0.2	non-optimal	209 (16.94)	0.620	105 (50.20)	12 (5.74)	0.413	104 (49.76)	25 (11.96)
	optimal	101 (8.18)	0.832	64 (63.37)	7 (6.93)	0.636	64 (63.37)	8 (7.92)
> = 0.1	non-optimal	437 (35.41)	0.543	202 (46.22)	41 (9.38)	0.409	199 (45.54)	68 (15.56)
	optimal	177 (14.34)	0.696	99 (55.93)	17 (9.60)	0.535	97 (54.80)	23 (12.99)

Note. HSCL-11 = Hopkins Symptom Checklist – Short Form; OQ-30 = Outcome Questionnaire-30; Diff. = Critical difference in predicted effect sizes for strategy recommendation, meaning that the predicted effect size for one strategy was at least 0.1 or 0.2 larger than for both other strategies; n = sample size; ES = effect size (Cohen's d); Rel. imp. = reliable improvement; Rel. det. = reliable deterioration.

21.8% of variance in being not-on-track. The full model for session 15 is presented exemplarily in Table 5.

Within the TTN, the therapist receives a figure showing the patient-specific symptomatic progress measured with the HSCL-11 (A, Fig. 3). As soon as the failure boundary of the expected recovery curve is crossed (B & C, Fig. 3), the therapist receives a warning signal within the TTN (D, see Fig. 3), which activates clinical problem-solving tools that can be directly used in clinical practice (CPST, E, Fig. 3).

4. Discussion

The present paper describes the development of the Trier Treatment

Table 4
Relative frequencies of reliable improvement and deterioration from pre-treatment to session 25 for on-track and not-on-track patients.

	n	HSCL-11		OQ-30	
		Rel. imp. (% per row)	Rel. det. (% per row)	Rel. imp. (% per row)	Rel. det. (% per row)
All patients	1177[a]	597 (50.72)	98 (8.33)	602 (51.15)	84 (7.14)
On-track	679	422 (62.15)	11 (1.62)	380 (55.96)	22 (3.24)
Not-on-track	498	175 (35.14)	87 (17.47)	217 (43.57)	62 (12.45)

Note. Rel. imp = reliable improvement; Rel. det. = reliable deterioration. HSCL-11 = Hopkins Symptom Checklist – Short Form; OQ-30 = Outcome Questionnaire-30. [a] 57 of the n = 1234 patients had no sufficiently similar positively changing cases.

Table 5
Logistic regression for the relationship of the probability being not-on-track (session 16) and the problem domains of the clinical problem-solving tools (CPST, session 15).

	coefficient	p-value
Intercept	4.320	.009
Risk/suicidality	0.112*	.028
Motivation	−0.373	.277
Alliance	0.371	.330
Emotion regulation	−0.320*	.028
Social support	−0.390*	.041
Life events	−0.927***	< .001

Note. Logistic regression with the dummy coded variable 1 = not-on-track and −1 = on-track as the dependent variable. CPST are assessed by either the Assessment for Signal Clients (ASC; higher scores indicating more functioning), the Affective Style Questionnaire (ASQ; higher scores indicating more functioning) or single items from the Outcome Questionnaire-30 (OQ-30; higher scores indicating more problems) and the Hopkins Symptom Checklist 11 (HSCL 11; higher scores indicating more problems). The five domains are: risk/suicidality (based on two items from the OQ-30 and one item from the HSCL 11), motivation/therapy goals (based on the ASC motivation scale), therapeutic alliance (based on the ASC alliance scale), emotion regulation/self-regulation (based on the ASQ) and social support/critical life events (based on the ASC scales social support and life events).
* = $p < .05$; ** = $p < .01$; *** = $p < .001$.

Navigator (TTN). The TTN is a comprehensive feedback system that supports therapists in their everyday decision-making. As such, the TTN stands in the tradition of research focusing on the implementation of change measures and feedback as part of clinical practice and training. Advances in new technologies and software allowing the easy implementation of outcome assessments make this area of research an exciting new development. The increased use of change measures and feedback can substantially change the way we think about psychological interventions. The integration of empirical data into the decision-making process has the potential to lead to a major advancement on the path towards “blended” treatment concepts. Therapists personal skills are augmented by tools from E-mental health and facilitate not only new forms of research, but also enable clinicians to adapt treatment decisions with respect to treatment strategies, modalities or treatment length based on patient characteristics and needs. Furthermore, navigation systems and tools can be applied “in real time” to improve treatment, especially for those at risk for treatment failure. In order to facilitate research and the use of the TTN as well as its further development, the software is freely available at github (https://github.com/Psykli/Trier_Treatment_Navigator) and accompanied by a sample dataset of 200 cases and open source videos.

In the present study on $n = 1234$ patients, we identified seven significant predictors of drop-out, which were therefore included in the personalized graphical visualizations of the TTN to forecast attrition. Such a graphical visualization can increase therapists’ awareness of potential treatment drop-out and enable the clinician to better tailor treatment (for instance by focusing on strategies aimed at increasing patient motivation). The nearest neighbor predictions successfully selected optimal treatment strategies for individual patients (problem-solving or motivation-oriented strategies or mixed) for the beginning of treatment (first 10 sessions). We tested two thresholds. Although the more conservative validation strategy yielded better results overall, we were able to show that with both evaluation strategies patients treated with the recommended treatment strategy achieved better effect sizes and improvement rates.

Throughout treatment, the TTN reliably identified patients with a higher risk for no improvement or deterioration within the first 25 sessions. The probability to be reliably improved for patients identified as at risk for treatment failure (35.00%) was about half of the probability for other patients (62.15%). Furthermore, patients identified as NOT by the dynamic failure boundary had longer treatments on average than those who were OT. The results further provide support for the social support/life event tool, the risk tool as well as the emotion regulation/self-regulation tool. Results stress the importance of integrating these components into the TTN. Therapists may deal with such deficits by strengthening the social network or applying social skills training, activating patients’ competences and resources, working out a risk profile or facilitating adaptive emotion regulation strategies. Therapeutic alliance and motivation were not found to be significant predictors of unexpected symptom progress. This may be due to the skewed distribution of these variables. In therapeutic alliance and motivation scales, ceiling effects are common. For instance in our data, the cut-off of the ASC motivation item “*I wonder what I am doing in therapy; actually I find it boring*” was exceeded in less than 1% of the cases until session 25. However, if it is exceeded, it is crucial that the therapist acts on this information in order to prevent drop-out. In terms of therapeutic alliance and motivation, it is still useful to have such tools in the navigation system, however it is more informative to look at the outliers.

The TTN combines precision mental health predictions with outcome monitoring and feedback tools in routine outpatient psychotherapy. It is currently being tested in a prospective randomized controlled trial to test its effectiveness as well as moderating and mediating factors (Lutz et al., 2017).

It should be noted that the TTN is not the only prediction system that is currently being tested in prospective studies. Two other similar initiatives should be mentioned at this point, which both make use of the IAPT (Improving Access to Psychological Therapies) system in the UK. Based on archival data from the IAPT service system, Delgadillo and colleagues (Delgadillo, Huey, Bennet, & McMillan, 2017) developed an index, which indicates whether a patient will benefit from a low intensity intervention or need more intensive treatment. This risk index has been tested in independent samples and is currently being subjected to a prospective test in which patients are assigned to low or high intensity interventions based on their risk score (Delgadillo, Huey, Bennett, & McMillan, 2017;http://www.isrctn.com/ISRCTN11106183). In another attempt to optimize predictions for low and high intensity treatments, Cohen, DeRubeis and colleagues initiated an international “tournament” to which thirteen groups were invited. Each group was provided with the same dataset of patients who had received low or high intensity treatment in the IAPT system. Based on this dataset, each group was asked to develop an algorithm that optimally assigns patients to one of these treatment alternatives. A combination of the algorithms performing best in an independent holdout sample will be subsequently tested in a prospective trial. For more information on the tournament and the resulting trial, see https://osf.io/wxgzu/.

Another advantage of such navigation systems is the potential to easily and dynamically update the system based on new research. For example, compared to previous versions of the TTN, the presented version now comprises dynamic failure boundaries that adaptively

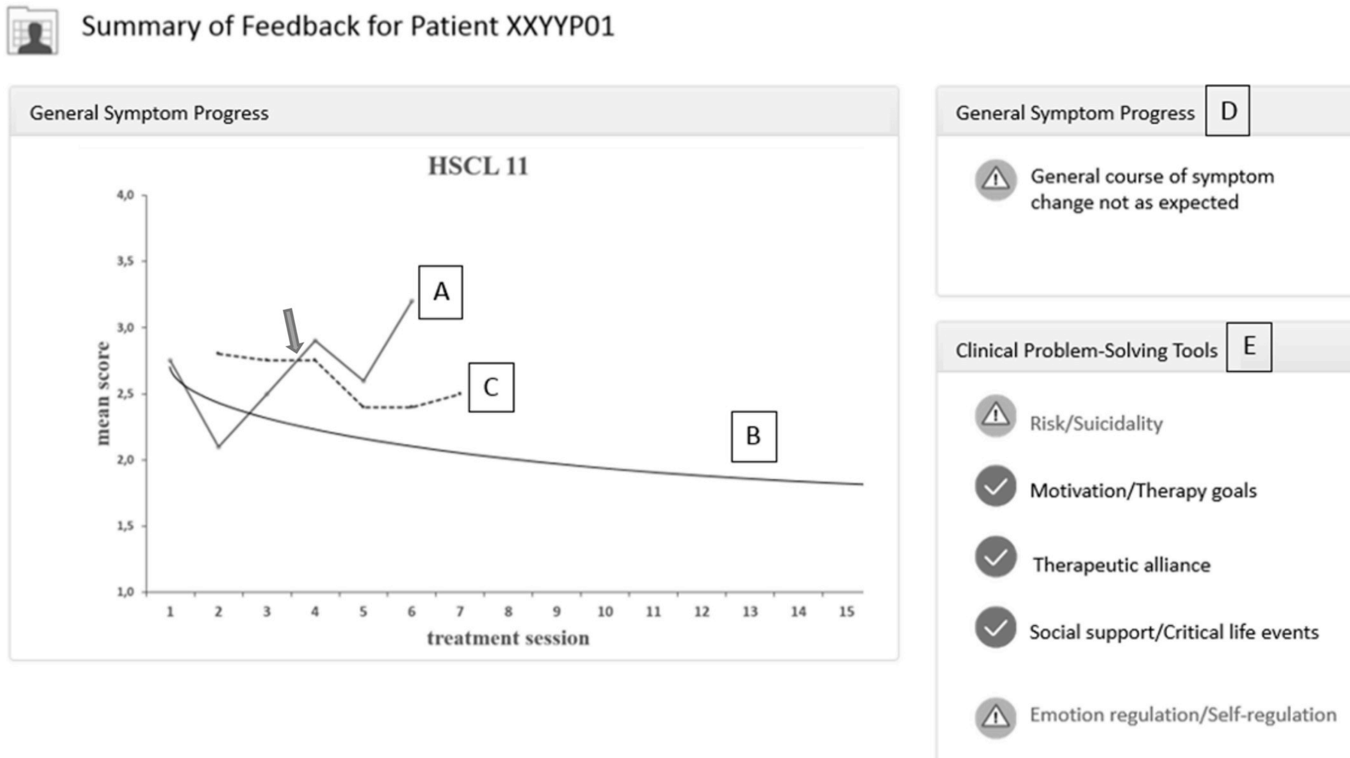

Fig. 3. Screenshot of a patient-specific adaptive recommendation as it is displayed to therapists within the Trier Treatment Navigator (TTN).
Note. The example screenshot shows the symptomatic progress of a patient measured with the Hopkins-Symptom-Checklist-11 (HSCL-11). A: Individual measurement points for the patient measured at the beginning of each session. B: Expected recovery curve. C: Failure boundary. D: As soon as the patient's score exceeds the failure boundary on the HSCL-11 (symbolized in the graph via an arrow), the therapist receives this warning signal, which is defined in more detail in the clinical problem-solving tools (CPST) below. E: CPST are divided into five domains. The exclamation mark indicates those domains where the patient has specific problems. The therapist is able to click on this icon to get access to the activated tools. The check hook signals that the patient has few or no problems in this area.

change during the course of treatment. The TTN has a modular character, with two broad areas (pre-treatment recommendations and adaptive recommendations during treatment), each currently containing two steps of personalized decision-making (drop-out prediction and optimal treatment strategy as well as dynamic risk index and clinical problem-solving tools). This opens up two possibilities for future changes: First, the existing recommendations can be improved by novel measures and/or analytical approaches. Second, additional recommendations could be implemented in the system. Only recently, expected treatment response models have been criticized for not being dynamic enough (Langkaas, Hoffart, & Wampold, 2018). Using only one prediction at the start of treatment places too much weight on the intake measurement and disregards additional information collected at each session. The concept presented here includes a dynamic, adaptive approach that continuously updates the predictions session by session. Furthermore, the more patients and therapists use the system, the larger and better the reference sample becomes and even patients with extreme or rare scores and disorders will have enough nearest neighbors to generate predictive reference sample.

It is important to note that the use of large databases of already treated patients is only one viable approach to personalizing psychological treatments. An example of another area that has lately received burgeoning interest is idiographic symptom dynamics research based on intensive longitudinal patients assessments (e.g., Fisher & Boswell, 2016; Hayes et al., 2018). Patients are asked to answer questions regarding their cognitions, emotions, and behaviors several times a day for a period of time (e.g., two weeks), resulting in a large, patient-specific database. These data can be leveraged, for example, by network models, which visualize symptom dynamics in an easy-to-understand way (e.g., Fisher, Reeves, Lawyer, Medaglia, & Rubel, 2017; Hofmann & Curtiss, 2018; Rubel et al., 2018). Initial pilot studies using these network models as an additional source of information for therapists are currently underway (Kroeze et al., 2017). Others have used intensive longitudinal data collected before therapy to improve predictions of treatment response (e.g., Husen, Rafaeli, Rubel, Bar-Kalifa, & Lutz, 2016) and treatment drop-out (Lutz et al., 2018). All this could easily be implemented into the TTN. As such, the TTN should not be seen as a static product, but rather as a system that enables new research to be implemented into clinical practice and the evaluation of the effects of these implementation efforts.

4.1. Limitations, future directions, and summary

Developing a system that can be applied in routine clinical practice must implement scientific research in a useable way. Several limitations arise from this challenge, which are worth noting. The described concepts and ideas are only early implementations and new territory. Especially critical is how such systems are used by clinicians, whether they are accepted in clinical practice and useful in the long run. There are several aspects that need to be considered with regard to the implementation of such navigation systems. First, technical and software equipment and support are necessary to be able to routinely collect data at every session and further process these data to produce easy-to-understand figures and recommendations.

Another important point is clinical training. Therapists need to be trained to use the navigation system so that it can unfold its potential. This is why such a system and its research and psychometric background need to be part of clinical training, including specific workshops that help therapists understand and deal with the feedback and recommendations they receive. For the current implementation, we have prepared several video clips for therapists that explain the navigation system step-by-step. Furthermore, a contact person within the outpatient clinic was established to answer questions and help with technical problems. Additionally, clinicians are provided with a workshop that conveys the clinical usage of the computer-based system.

Regarding drop-out prediction, it should be noted that specificity is more important than sensitivity when using our system. Therefore, we did not built the prediction model to signal many cases, but to provide therapists with a signal, when there is a higher probability of drop-out compared to the average dropout rate in the outpatient center. Therefore, therapists are not provided with a categorical prediction (i.e., drop-out yes vs. no), but rather the estimated probability that a patient will drop-out based on their intake characteristics. This probability should be interpreted with caution and can be compared with the overall base rate of drop-outs in the clinic. Therapists should be aware of drop-out in all cases. However, if the system additionally tags a patient as at risk of drop-out, this is an important signal to the therapist, because the prediction is highly specific. Patients showing a drop-out probability of $\geqq .3$ (slightly above the average drop-out probability of .22) will end treatment prematurely in 39.4% of cases, while patients with a predicted probability of $\geqq 0.5$ will do so in 64.5% of cases. With an increasing predicted probability, the discrimination between patients who drop out of treatment and those who end treatment regularly improves. On the whole, we are still facing some problems with the identification of drop-out and predictions have to be optimized in further developments of the TTN. Some recent work has been done regarding novel measurement and network analysis techniques (e.g., Lutz et al., 2018).

Although it is a strength of the current feedback system that its predictions are updated after each session, which takes information that is collected throughout treatment into account, this very feature also has the drawback that the failure boundary is less transparent for therapists. As displayed in Fig. 3 C, the failure boundary can make rather large jumps from one session to the next, especially once a patient goes off-track for the first time. However, it is important to note that these jumps help to prevent patients from alternating between OT and NOT from session to session.

Furthermore, although a number of different instruments were applied to inform the prediction of mental health and drop-out, their range is still limited. A number of other clinically relevant aspects, such as social connectedness/support/loneliness, medical issues, etc. are not considered, but are probably important predictors. The selection of predictors is always a crucial issue of prediction models and needs to take both comprehensiveness and sparsity into account. Therefore, it is unlikely that one study can capture all potentially important predictors. Rather, predictor selection is part of a continuous process, where each study helps to inform future studies about what predictors to include.

Along these lines, we do not think that the TTN as described in this report is the final version of a comprehensive treatment navigation system. Instead, in our opinion, it merely stands at the beginning of a development that can change the way psychotherapy is conducted with the aim of integrating research into daily practice and using information from daily practice to improve research and increase the number of patients that profit from psychological treatments.

Acknowledgement

This work was supported by grants from the German Research Foundation (LU 660/10-1, LU 660/8-1). We are thankful to Kaitlyn Boyle for proofreading the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.brat.2019.103438.

References

Baldwin, S. A., Berkeljon, A., Atkins, D. C., Olsen, J. A., & Nielsen, S. L. (2009). Rates of change in naturalistic psychotherapy: Contrasting dose–effect and good-enough level models of change. *Journal of Consulting and Clinical Psychology, 77*(2), 203–211.

Bar-Kalifa, E., Atzil-Slonim, D., Rafaeli, E., Peri, T., Rubel, J., & Lutz, W. (2016). Therapist–client agreement in assessments of clients' functioning. *Journal of Consulting and Clinical Psychology, 84*(12)https://doi.org/10.1037/ccp0000157.

Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *Journal of the American Medical Association, 319*(13), 1317–1318.

Böhnke, J. R., & Lutz, W. (2014). Short versions of existing instruments: Using item information to optimize targeted assessments. *Assessment, 21*(6), 679–693. https://doi.org/10.1177/1073191114529152.

Bronisch, T., Hiller, W., Mombour, W., & Zaudig, M. (1996). *International diagnostic checklists for personality disorders according to ICD-10 and DSM–IV-IDCL-P.* Seattle, WA: Hogrefe and Huber Publishers.

Castonguay, L. G., Barkham, M., Lutz, W., & McAleavy, A. (2013). Practice oriented research: Approaches and applications. In M. J. Lambert (Ed.). *Bergin and Garfield's handbook of psychotherapy and behavior change* (pp. 85–133). (6th ed.). Hoboken, N.J.: Wiley.

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

Coyne, J. C., Kessler, R. C., Tal, M., Turnbull, J., Wortman, C. B., & Greden, J. F. (1987). Living with a depressed person. *Journal of Consulting and Clinical Psychology, 55*(3), 347–352.

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

Deisenhofer, A. K., Delgadillo, J., Rubel, J. A., Böhnke, J. R., Zimmermann, D., Schwartz, B., et al. (2018). Individual treatment selection for patients with posttraumatic stress disorder. *Depression and Anxiety, 35*(6), 541–550.

Delgadillo, J., de Jong, K., Lucock, M., Lutz, W., Rubel, J., Gilbody, S., ... O'Hayon, H. (2018). Feedback-informed treatment versus usual psychological treatment for depression and anxiety: A multisite, open-label, cluster randomised controlled trial. *The Lancet Psychiatry, 5*(7), 564–572.

Delgadillo, J., Huey, D., Bennett, H., & McMillan, D. (2017). Case complexity as a guide for psychological treatment selection. *Journal of Consulting and Clinical Psychology, 85*(9), 835–853.

Derogatis, L. R. (1977). *SCL-90-R: Administration, scoring and procedures manual I.* Baltimore: Clinical psychometric research.

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.

Ellsworth, J. R., Lambert, M. J., & Johnson, J. (2006). A comparison of the Outcome Questionnaire-45and Outcome Questionnaire-30 in classification and prediction of treatment outcome. *Clinical Psychology & Psychotherapy: An International Journal Of Theory & Practice, 13*(6), 380–391.

Finch, A. E., Lambert, M. J., & Schaalje, B. G. (2001). Psychotherapy quality control: The statistical generation of expected recovery curves for integration into an early warning system. *Clinical Psychology & Psychotherapy: An International Journal Of Theory & Practice, 8*(4), 231–242.

First, M. B., Spitzer, R. L., Gibbon, M., & Williams, J. B. W. (1995). *Structured clinical interview for axis I DSM–IV disorders-patient ed. (SCID-I/P)*. New York, NY: Biometrics research department, NY State psychiatric institute.

Fisher, A. J., & Boswell, J. F. (2016). Enhancing the personalization of psychotherapy with dynamic assessment and modeling. *Assessment, 23*(4), 496–506.

Fisher, A. J., Reeves, J. W., Lawyer, G., Medaglia, J. D., & Rubel, J. A. (2017). Exploring the idiographic dynamics of mood and anxiety via network analysis. *Journal of Abnormal Psychology, 126*(8), 1044–1056.

Franke, G. (2000). *BSI: Brief Symptom Inventory von L.R. Derogatis (Kurzform der SCL-90-R) – deutsche Version.* Beltz Test GmbH.

Furukawa, T. A., Efthimiou, O., Weitz, E. S., Cipriani, A., Keller, M. B., Kocsis, J. H., ... Schramm, E. (2018). Cognitive-behavioral analysis system of psychotherapy, drug, or their combination for persistent depressive disorder: Personalizing the treatment choice using individual participant data network metaregression. *Psychotherapy and Psychosomatics, 87*(3), 140–153.

Graser, J., Bohn, C., Kelava, A., Schreiber, F., Hofmann, S. G., & Stangier, u. (2012). Der „Affective Style Questionnaire (ASQ)": Deutsche Adaption und Validitäten. *Diagnostica, 58*(2), 100–111. https://doi.org/10.1026/0012-1924/a000056.

Guo, P., & Hao, Y. (2015). *SparseLearner: Sparse learning algorithms using a LASSO-type penalty for coefficient estimation and model prediction. R package version 1.0-2.* http://CRAN.R-project.org/package=SparseLearner.

Guo, P., Zeng, F., Hu, X., Zhang, D., Zhu, S., Deng, Y., et al. (2015). Improved variable selection algorithm using a LASSO-type penalty, with an application to assessing hepatitis B infection relevant factors in community residents. *PLoS One, 10*(7), e0134151. https://doi.org/10.1371/journal.pone.0134151.

Hannan, C., Lambert, M. J., Harmon, C., Nielsen, S. L., Smart, D. W., Shimokawa, K., et al. (2005). A lab test and algorithms for identifying clients at risk for treatment failure. *Journal of Clinical Psychology, 61*(2), 155–163.

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

Hofmann, S. G., & Kashdan, T. B. (2010). The affective style Questionnaire: Development and psychometric properties. *Journal of Psychopathology and Behavioral Assessment,* *32*(2), 255–263. https://doi.org/10.1007/s10862-009-9142-4.

Hooke, G. R., Sng, A. A., Cunningham, N. K., & Page, A. C. (2017). Methods of delivering progress feedback to optimise patient outcomes: The value of expected treatment trajectories. *Cognitive Therapy and Research, 42*(2), 204–211. https://doi.org/10.1007/s10608-017-9851-z.

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

Huibers, M. J. H., Cohen, Z. D., Lemmens, L. H. J. M., Arntz, A., Peeters, F. P. M. L., Cuijpers, P., et al. (2015). Predicting optimal outcomes in cognitive therapy or interpersonal psychotherapy for depressed individuals using the personalized advantage index approach. *PLoS One, 10*(11), e0140771. https://doi.org/10.1371/journal.pone.0140771.

Husen, K., Rafaeli, E., Rubel, J. A., Bar-Kalifa, E., & Lutz, W. (2016). Daily affect dynamics predict early response in CBT: Feasibility and predictive validity of EMA for outpatient psychotherapy. *Journal of Affective Disorders, 206*, 305–314.

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*, 12–19.

Kendrick, T., El-Gohary, M., Stuart, B., Gilbody, S., Churchill, R., Aiken, L., ... Moore, M. (2016). Routine use of patient reported outcome measures (PROMs) for improving treatment of common mental health disorders in adults. *Cochrane Database of Systematic Reviews,* (7), 1–103.

Kessler, R. C., Van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Ebert, D. D., et al. (2017). Using patient self-reports to study heterogeneity of treatment effects in major depressive disorder. *Epidemiology and Psychiatric Sciences, 26*(1), 22–36.

Kroeze, R., Van Veen, D. C., Servaas, M. N., Bastiaansen, J. A., Oude Voshaar, R., Borsboom, D., et al. (2017). Personalized feedback on symptom dynamics of psychopathology: A proof-of-principle study. *Journal Of Person-Oriented Research, 3*(1), 1–10.

Kuhl, J., & Kazén, M. (1997). *Persönlichkeits-Stil-und Störungs-Inventar: (PSSI). [Personality-Style and Disorder Inventory]. Göttingen: Hogrefe*.

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

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

Lambert, M. J., Bailey, R., Kimball, K., Shimokawa, K., Harmon, S. C., & Slade, K. (2007). *Clinical support tool manual-brief version-40.* Salt Lake City: OQ Measures.

Lambert, M. J., Whipple, J. L., & Kleinstäuber, M. (2018). Collecting and delivering progress feedback. A meta-analysis of routine outcome monitoring. *Psychotherapy, 55*(4), 520–537.

Langkaas, T. F., Wampold, B. E., & Hoffart, A. (2018). Five types of clinical difference to monitor in practice. *Psychotherapy, 55*(3), 241–254.

Lutz, W., de Jong, K., & Rubel, J. (2015). Patient-focused and feedback research in psychotherapy: Where are we and where do we want to go? *Psychotherapy Research, 25*(6), 625–632. https://doi.org/10.1080/10503307.2015.1079661.

Lutz, W., 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*(5), 904–913.

Lutz, W., Saunders, S. M., Leon, S. C., Martinovich, Z., Kosfelder, J., Schulte, D., et al. (2006). Empirical and clinical useful decision making in psychotherapy: Differential predictions with treatment response models. *Psychological Assessment, 18*, 133–144.

Lutz, W., Schwartz, B., Hofmann, S. G., Fisher, A. J., Husen, K., & Rubel, J. A. (2018). Using network analysis for the prediction of treatment dropout in patients with mood and anxiety disorders. A methodological proof-of-concept study. *Scientific Reports, 8*(1), https://doi.org/10.1038/s41598-018-25953-0https://doi.org/10.1038/s41598-018-25953-0.

Lutz, W., Tholen, S., Schürch, E., & Berking, M. (2006). Die Entwicklung, Validität und Reliabilität von Kurzformen gängiger psychometrischer Instrumente zur Evaluation des therapeutischen Fortschrittes in Psychotherapie und Psychiatrie. *Diagnostica, 52*, 11–25.

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)https://doi.org/10.1186/s12888-017-1464-2.

Ng, M. Y., & Weisz, J. R. (2016). Annual research review: Building a science of personalized intervention for youth mental health. *Journal of Child Psychology and Psychiatry, 57*(3), 216–236.

Perlis, R. H. (2016). Abandoning personalization to get to precision in the pharmacotherapy of depression. *World Psychiatry : Official Journal of the World Psychiatric Association (WPA), 15*(3), 228–235. https://doi.org/10.1002/wps.20345.

Probst, T., Lambert, M. J., Dahlbender, R. W., Loew, T. H., & Tritt, K. (2014). Providing patient progress feedback and clinical support tools to therapists: Is the therapeutic process of patients on-track to recovery enhanced in psychosomatic in-patient therapy under the conditions of routine practice? *Journal of Psychosomatic Research, 76*(6), 477–484.

R Core Team (2014). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

Rubel, J. A., Fisher, A. J., Husen, K., & Lutz, W. (2018). Translating person-specific network models into personalized treatments: Development and demonstration of the dynamic assessment treatment algorithm for individual networks (DATA-IN). *Psychotherapy and Psychosomatics, 87*, 249–251. https://doi.org/10.1159/000487769.

Rubel, J., Lutz, W., Kopta, S. M., Köck, K., Minami, T., Zimmermann, D., et al. (2015). Defining early positive response to psychotherapy: An empirical comparison between clinically significant change criteria and growth mixture modeling. *Psychological Assessment, 27*(2), 478–488.

Schulte, D., & Eifert, G. H. (2002). What to do when manuals fail? The dual model of psychotherapy. *Clinical Psychology-Science and Practice, 9*, 312–328.

Shimokawa, K., Lambert, M. J., & Smart, D. W. (2010). Enhancing treatment outcome of patients at risk of treatment failure. Meta-analytic and mega-analytic review of a psychotherapy quality assurance system. *Journal of Consulting and Clinical Psychology, 78*(3), 298–311. https://doi.org/10.1037/a0019247.

Stekhoven, D. J. (2013). *missForest: Nonparametric missing value imputation using random forest. R package version 1.4*.

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

Vermeersch, D. A., Whipple, J. L., Lambert, M. J., Hawkins, E. J., Burchfield, C. M., & Okiishi, J. C. (2004). Outcome Questionnaire: Is it sensitive to changes in counseling center clients? *Journal of Counseling Psychology, 51*(1), 38–49.

Whipple, J. L., Lambert, M. J., Vermeersch, D. A., Smart, D. W., Nielsen, S. L., & Hawkins, E. J. (2003). Improving the effects of psychotherapy. The use of early identification of treatment and problem-solving strategies in routine practice. *Journal of Counseling Psychology, 50*(1), 59–68. https://doi.org/10.1037/0022-0167.50.1.59.

White, M. M., Lambert, M. J., Ogles, B. M., Mclaughlin, S. B., Bailey, R. J., & Tingey, K. M. (2015). Using the Assessment for Signal Clients as a feedback tool for reducing treatment failure. *Psychotherapy Research, 25*(6), 724–734. https://doi.org/10.1080/10503307.2015.1009862.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis.* New York: Springer.

Zimmermann, D., Rubel, J., Page, A. C., & Lutz, W. (2017). Therapist effects on and predictors of non-consensual dropout in psychotherapy. *Clinical Psychology & Psychotherapy, 24*(2), 312–321.