Psychotherapy Research, 2024
https://doi.org/10.1080/10503307.2024.2334047

RESEARCH ARTICLE

Bridging the feedback implementation gap: A comparison of empirical and rational decision rules in naturalistic psychotherapy

STEPHAN RAMSPERGER, MICHAEL WITTHÖFT, & ANNE-KATHRIN BRÄSCHER

Department for Clinical Psychology, Psychotherapy and Experimental Psychopathology, Johannes Gutenberg University Mainz, Mainz, Germany

(Received 30 May 2023; revised 7 March 2024; accepted 14 March 2024)

Abstract

Objective: Previous research indicates positive effects of feedback based on rational or empirical decision rules in psychotherapy. The implementation of these usually session-to-session-based feedback systems into clinical practice, however, remains challenging. This study aims to evaluate decision rules based on routine outcome monitoring with reduced assessment frequency. **Method:** Data routinely collected every 5–20 sessions of $N = 3758$ patients treated with CBT in an outpatient clinic ($M_{sessions} = 42.8$, $SD = 15.4$) were used to develop feedback decision rules based on the *expected treatment response* and *nearest neighbors* approach, the *reliable change index,* and method of *percental improvement*. The detection of patients at risk of treatment failure served as primary endpoint. **Results:** Significantly lower reliable improvement, higher reliable deterioration rates, and smaller effect sizes were found for patients identified at risk of treatment failure by all rules. The *nearest neighbors*-based approach showed the highest sensitivity regarding the detection of reliably deteriorated cases. **Conclusion:** Consistent with previous research, the empirical models outperformed the rational rules. Still, the first-time used *percental improvement*-based rule also showed satisfactory results. Overall, the results point to the potential of basic feedback systems that might be easier to implement in practice than session-to-session based systems.

Keywords: cognitive behavioral therapy; feedback; expected treatment response; nearest neighbors; percental improvement

Clinical or methodological significance of this article: This study shows that even on a less frequent routine outcome monitoring structure, empirical, and rational decision rules can be used to identify patients at risk of treatment failure to a satisfactory degree, with the empirical *nearest neighbors* method showing the best results. This indicates the potential of a possibly more easy-to-implement basic feedback system for clinical practice.

Objective

In the last decades, patient-focused psychotherapy research became more and more popular (Lutz, de Jong, et al., 2015), going along with the development of a variety of increasingly sophisticated routine outcome monitoring (ROM) and feedback systems (Delgadillo et al., 2018; Finch et al., 2001; Lutz et al., 2019). The idea is that by routinely collecting (mainly patient-reported) outcome measures over the course of treatment and reporting them back to the therapist, treatment outcomes will improve. In particular, the early detection of

Correspondence concerning this article should be addressed to Stephan Ramsperger, Department for Clinical Psychology, Psychotherapy and Experimental Psychopathology, Johannes Gutenberg University Mainz, Wallstr. 3, Mainz 55112, Germany. Email: stramspe@uni-mainz.de

patients with unfavorable treatment courses and the resulting ability to counteract these deteriorations by adapting interventions represents an essential objective of feedback systems (Boswell, 2020; Lambert et al., 2002). This is because of in spite the extensive evidence that psychotherapy is effective in treating mental disorders, the proportion of patients deteriorating during treatment is a non-negligible 5–10% (Hansen et al., 2002). At the same time, unfortunately, clinicians often fail to recognize these deteriorations, emphasizing the need for continuous monitoring (Hannan et al., 2005; Hatfield et al., 2010). The effectiveness of ROM and feedback, especially for those so-called *not on track* patients (NOT; also *signal cases*), has been demonstrated in multiple meta-analyses and systematic reviews, showing overall improved outcomes (de Jong et al., 2021; Østergård et al., 2020) and reduced deterioration (Lambert et al., 2018; Lambert & Shimokawa, 2011), and dropout rates (de Jong et al., 2021).

In general, two broad classes of decision rules to assess treatment progress and identify NOT cases as the underlying basis of these feedback systems are distinguished (Lutz et al., 2009). First, rationally derived decision rules refer to predefined criteria and expert assessments based on clinical as well as statistical expertise regarding what can be considered as sufficient treatment progress. Arguably the most widespread example of this approach is the concept of clinically significant change by Jacobson and Truax (1991), defining meaningful change on the basis of a twofold criterion. Using the *reliable change index* (RCI), it is first determined whether a change in test scores is statistically significant given the reliability of the respective test instrument. In addition, if a patient crosses the cutoff between a clinical and a healthy (functional) reference sample, it can be considered not only a statistically reliable but also a clinically significant improvement.

Along with the RCI, although less widely used, the method of *percental improvement* (PI) has become established in recent years for evaluating treatment outcomes in psychotherapy research (Hiller et al., 2012). This method, in the tradition of psychopharmacological research, expresses the degree of symptom change in percentage, whereby a change equal or greater to 50% in the pathological range is usually considered significant (e.g., Lecrubier, 2002; Schlagert & Hiller, 2017b). Advocates of the PI method name as advantages among other that differences in initial impairment are considered and that it is easily communicable, especially to patients (Hiller et al., 2012). Even more important, early and dramatic changes in the treatment process defined by the PI criterion proved to be predictive of treatment outcome (Erekson et al., 2018; Schlagert & Hiller, 2017a). As a disadvantage, the criterion of exactly 50% necessary change appears highly arbitrary. To the authors' knowledge, despite its benefits, the PI method has never been evaluated in the context of feedback systems.

From the rational decision rules outlined above, the empirically derived rules can be distinguished. Based on large sets of ROM data from previously treated patients, the expected treatment response (ETR) of a given patient can be modeled using statistical methods. Consequently, the predicted course of treatment of a patient can be compared with the observed one (Lutz et al., 2009). By means of *failure boundaries* around the modeled treatment course (e.g., based on the 90% confidence interval; Lutz et al., 2019), strong deviations can then be registered and the patient can be reported to the respective clinician as NOT (Finch et al., 2001; Lutz et al., 1999).

The ETR concept was first introduced by Lutz et al. (1999) using multilevel growth curve modeling (also known as hierarchical linear modeling) to predict individual treatment trajectories on the basis of seven intake characteristics of previously treated patients. Since then, alternative and more advanced empirical methods for generating ETR curves have been developed (e.g., Finch et al., 2001). A more recent approach represents an early machine learning technique, called *nearest neighbor* method (NN; Lutz et al., 2005, 2019). Within this method, the presented intake characteristics are first (and only) used to identify previously treated patients who are most similar to a given patient starting treatment, hence his or hers *nearest neighbors*. This homogeneous NN subset is then used to predict the treatment course of the index patient. The approach is thus intended to be more in line with how practitioners use their clinical experience (Lutz et al., 2005).

Previous studies using rational decision rules have shown that they can reliably predict treatment outcomes (Lutz et al., 2022). Nevertheless, direct comparisons between rational and the described empirical methods indicate a slight superiority of the latter regarding their overall predictive accuracy (Spielmans et al., 2006) and their ability to correctly identify patients with negative treatment outcomes (Lambert et al., 2002; Lutz et al., 2006). In addition, the NN approach showed to be superior to an alternative ETR approach in predicting client progress (Lutz et al., 2005). However, these major comparison studies are not without limitations with (1) all using the Outcome Questionnaire (OQ-45; for a description, see Lambert, 2015) as outcome measure[1] and (2) comparing the respective empirical method with the same rational decision rule loosely based on an adaptation of clinically significant change concepts (Lambert et al., 2002; Lutz et al.,

2006; Spielmans et al., 2006). The restriction of these previous comparison studies to a rational decision rule specifically conceptualized for the Outcome Questionnaire should be considered when interpreting previous findings. Furthermore, the rather complex rational decision algorithm with several decision matrices grouped according to treatment phases (for examples, see Lambert et al., 2002) poses a challenge for the transfer to other outcome measures and into clinical practice.

Considering the bigger picture, although the ROM approach has not only the ambition but also the potential to narrow the gap between research and practice (Newnham & Page, 2010), it is the implementation of these very monitoring and feedback applications into clinical practice that is proving difficult (Boswell, 2020; Lutz et al., 2022). Surveys from across countries indicate that ROM is rarely used in clinical practice (e.g., Ionita & Fitzpatrick, 2014; Jensen-Doss et al., 2018). Moreover, whereas most current feedback systems rely on session-to-session assessments (de Jong et al., 2021) to enable a prompt intervention for NOT cases, in particular this high frequent application of outcome measures every 1–2 sessions appears to be rare in clinical routine care (5.2%; Jensen-Doss et al., 2018). In our recent survey among psychotherapists working in an outpatient setting in Germany, ROM was mainly not used at all (40.8%) or rarely (28%). Of the psychotherapists using ROM, only 3.3% reported an assessment every 1–2 sessions (Unpublished observations). Moreover, the motivation for such a session-to-session administration also seems to be limited (Jensen-Doss et al., 2018). This is especially relevant, since therapists' attitudes and the commitment to use feedback influence its effects and thereby treatment outcomes (de Jong et al., 2012; Lutz, Rubel, et al., 2015).

One of the most cited reasons hindering the implementation of ROM is the additional administrative and time burden it causes for clinicians (Boswell et al., 2015; Kaiser et al., 2018; Mellor-Clark et al., 2016). Clinicians seem to simply lack time in their everyday practical routine. Some are even concerned that the administrative process may have a negative impact on the therapeutic alliance (Boswell, 2020). This is probably of particular importance for the application of outcome measures on a session-to-session basis. Thus, reducing the monitoring frequency could provide an opportunity to ease the administrative burden.

Regarding the empirical prediction models, there is also the additional challenge of large amounts of patient outcome data being required for their development and use. However, as illustrated earlier, few clinicians or even researchers are likely to have access to such large samples of routinely collected outcome measures (Wise et al., 2016). Furthermore, most empirically based feedback systems are based on rather short treatment durations (for an overview of mean treatment durations in feedback studies, see de Jong et al., 2021). Thus, their generalizability to long-term treatments is still lacking evidence and should be examined. Desirably, we will soon reach a point where outcome data is shared and pooled so that empirical models can be better estimated and subsequently broadly applied in practice. Until then, however, we need practical transitional solutions.

In conclusion, the following questions arise: Can the empirical predictive models be successfully applied to a naturalistic sample with longer treatment durations and less frequent ROM? How does this affect their predictive accuracy? And, based on these conditions, do the more extensive empirically based decision rules continue to outperform the rationally based decision rules, which might be faster and easier to implement into clinical practice?

With this study, we therefore pursue several goals: To address time and administrative burden as one of the most commonly cited barriers to the adoption of ROM, we aim to apply two empirical methods, namely *expected treatment response* (Lutz et al., 1999) and *nearest neighbors* (Lutz et al., 2005), already established for session-to-session based administration and relatively short treatments to a naturalistic outpatient sample with less frequent (every 5–20 sessions) ROM. We will then compare the two empirically based methods with two simpler rational methods (*reliable change index* and *percental improvement*) and evaluate them based on their predictive accuracy and ability to detect deterioration.

Method

Participants

The total sample consisted of $N = 7037$ patients who started psychotherapy at a university outpatient clinic in Germany between 2001 and January of 2021. Patients who did not finish treatment before data extraction ($n = 1316$), with less than three assessments (i.e., less than 15 therapy sessions; $n = 1691$) or incomplete baseline measures ($n = 272$) were excluded (Figure 1). Patients included in the remaining sample ($N = 3758$) were on average 35.81 years old ($SD = 12.96$, age range 16–87) and were predominantly female (66.7%). Of the patients, 49.3% had a degree for university entrance level. At the beginning of therapy, 7.4% of patients were unemployed and 22.7% unable to work. The most often diagnosed primary diagnoses were affective disorders (37.1%), followed by anxiety disorders with 24.1%.

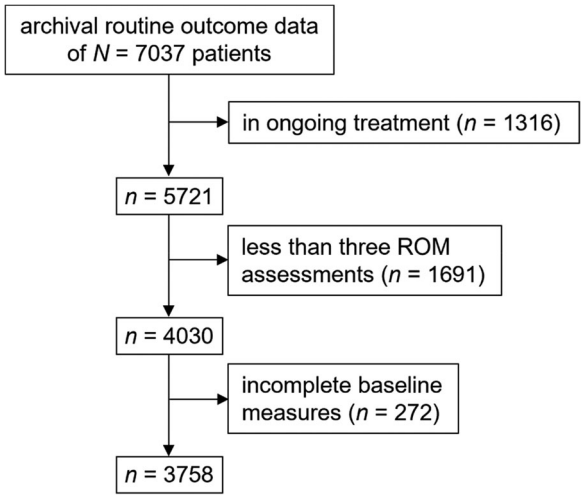

Figure 1. CONSORT flow diagram.

Further primary diagnoses were somatic symptom and related disorders (13.1%), feeding and eating disorders (8.6%), personality disorders (7.5%), trauma- and stressor-related disorders (3.9%), obsessive-compulsive disorders (3.6%), neurodevelopmental disorders (1.7%), schizophrenia spectrum and other psychotic disorders (1.9%), disruptive, impulse-control and, conduct disorders (0.9%) or substance-related and addictive disorders (0.9%). Comorbidity was given in 69.1% cases, with 41.1% diagnosed with two mental disorders, 20.2% with three and 7.7% with four or more diagnoses.

A subsample analysis was conducted for patients with a diagnosed primary or comorbid depressive disorder with the German version of the Beck Depression Inventory (BDI; Hautzinger et al., 1994, 2006) as an external outcome measure. Details on the subsample ($n = 2506$) can be found in Supplemental material 1.

Procedure

For the development and evaluation of the feedback systems, archival routine outcome monitoring data of the outpatient clinic was used. The outpatient clinic entails a distinctive quality management system, certified according to DIN EN ISO 9001 since 2005, securing a standardized intake, diagnostic and progress monitoring procedure (Hiller et al., 2006). Prior to treatment, all patients gave written informed consent regarding the use of their anonymized treatment data for research purposes. Furthermore, patients were diagnosed using the German version of the Structured Clinical Interviews for DSM-IV Axis I Disorders (SCID-I; Wittchen et al., 1997) and personality disorders (SCID-II; Fydrich et al., 1997) at the beginning of treatment. Therapy was conducted by psychologists in advanced psychotherapeutic training (90.0%) and licensed psychotherapists treating the patients with cognitive behavioral therapy. The treatment lasted on average 42.79 sessions ($SD = 15.36$). Ethical approval for the study was granted by the local Ethics Committee (021-JGU-psychEK-S025).

Measures

Brief symptom inventory (BSI). The BSI (Franke, 2000; German version) is a shortened version of the Symptom Checklist-90-Revised (SCL-90-R; Derogatis, 1992), comprising 53 items. It acquires the subjectively perceived impairment due to psychological and physical symptoms in the last seven days based on a Likert-scale from 0 ("not at all") to 4 ("extremely"). The Global Severity Index (GSI) calculated as the average of all items represents the intensity of general psychological distress, higher values indicating more severe distress. In this study, the BSI was administered at baseline and thereafter at sessions 10, 20, 40, 55, 75, 90, and 95 (if applicable) serving as predictor, progress, and outcome measure. The internal consistencies ranging between $\alpha_{min} = .95$ and $\alpha_{max} = .99$ for the present sample can be considered as high.

Predictor variables. A variety of potential treatment progress predictors were routinely collected after registration and during the diagnostic phase in the beginning of treatment, information given by the patient or their therapist. Complementary to the GSI-BSI baseline score (i.e., initial impairment), the baseline score of the BDI (see Supplemental material 2 for details), sociodemographic variables (e.g., gender, age), biographic information, treatment and mental disorder history including previous complications (e.g., previous treatment, medication) as well as intrinsic treatment motivation (rated by the therapist on a 5-point Likert scale from 0 ["nonexistent"] to 4 ["very high"]), and diagnostic information (e.g., primary diagnosis) were considered. See Supplemental material 3 for a full list of included predictors.

Data Analysis

Data analysis was conducted using the free software R version 4.0.4 (R Core Team, 2021). For the development of the empirical approaches the packages glmnet v4.1-3 (Friedman et al., 2010), cluster v2.1.2 (Maechler et al., 2021), lme4 v1.1-27.1 (Bates et al., 2015), and lmerTest v3.1-3

(Kuznetsova et al., 2017) were used in particular. The model evaluation was mainly conducted using epiR v2.0.43 (Stevenson & Sergeant, 2022) and pROC v1.18.0 (Robin et al., 2011).

Operationalization of NOT Cases and Outcome Evaluation

For the evaluation of the feedback approaches, a patient was classified as *not on track* (NOT; i.e., not on the path expected or desired by a curative procedure) case if the respective approach (or failure boundary) generated a NOT signal at least once during the course of treatment. The detection of reliably deteriorated cases (according to Jacobson & Truax, 1991) at the end of therapy was selected as main endpoint. Therefore, sensitivity and specificity rates, positive (PPV) and negative predictive values (NPV), and the area under the curve (AUC) were calculated. Cases were categorized as true positive (TP) if they were identified as NOT whilst being reliably deteriorated at the end of treatment.

For the evaluation and selection of the decision rules, sensitivity was weighted higher than specificity and the PPV. Since feedback tends to have the largest effect for patients at risk of negative treatment outcome (Lambert et al., 2018), the detection of these cases (i.e., sensitivity) seemed crucial to us. Further, we considered the costs of false negatives, i.e. the non-recognition of a potential treatment failure, to be greater than the costs of a false positive case. Consequently, sensitivity (or recall/hit rate) was regarded as more relevant than the PPVs (or precision) of the approaches. Moreover, sensitivity has the additional merit over the PPV that it is a prevalence-independent parameter and is therefore, unlike the PPV (Altman & Bland, 1994; Tharwat, 2021), not influenced by imbalanced data (which in our case can be assumed with deterioration rates of 5–10%; Hansen et al., 2002). Lastly, sensitivity and specificity allow benchmarking with the previous decision rule comparison studies (e.g., Spielmans et al., 2006), as these were regularly reported in contrast to the PPVs and NPVs.

As additional evaluation parameters, differences between OT and NOT cases were examined by comparing relative frequencies of reliable pre to post treatment improvement and deterioration. The classification of treatment outcome into reliable improvement and deterioration was based on the RCI[2] by Jacobson and Truax (1991). For the BSI-GSI, the parameters used to calculate the RCI were extracted from the German manual of the BSI (Franke, 2000). Using the standard deviation of the representative sample of outpatients ($SD_1 = 0.72$) and the test-retest reliability ($r_{xx} = .90$), a critical difference of 0.63 was obtained.

As the two rational decision rules (see below) also rely on the RCI and could therefore have an artificial advantage, effect size differences between OT and NOT cases were used as a further unrelated main evaluation criterion. In addition, a subsample analysis using the BDI as a second outcome measure, distinct from the BSI-GSI forming the feedback basis, was calculated to further investigate the results obtained by the rational approaches (see Supplemental materials 1–2, 4, 8–9 for details). Cutoff and reliable change scores for the BDI are provided in Supplemental material 4.

Rational Decision Rules

RCI approach. Due to its widespread use, the first rational rule was based on the RCI by Jacobson and Truax (1991). A patient was considered not on track if the reliable deterioration criterion was met, using the baseline value as reference point. A change in GSI-BSI scores ≥ 0.63 was considered statistically reliable.

PI approach. A second rational rule (hereafter named PI approach) was developed based on the PI method supplemented by the reliable deterioration criterion by Jacobson and Truax (1991). In the pertinent PI research (e.g., Hiller et al., 2012), solely the pathological range of a psychometric scale is used to calculate the degree of change of a patient. Accordingly, in our study, patients with a baseline score above the clinical cutoff[3] of ≥ 0.56 were defined as NOT if they had an increase of at least 50% between baseline and the current session on the pathological range (0.56–4.00) of the BSI-GSI. To cover the entire sample, we then deviated from the previous conceptualization and used the complete BSI-GSI scale as a reference point for calculating the percentage change for patients with a baseline score in the non-pathological range. As the PI criterion is rather liberal in the upper end of a scale (i.e., a deterioration of 50% is only achieved with large score differences; see Hiller et al., 2012; Schlagert & Hiller, 2017b), the 50% criterion in our PI approach was supplemented by the RCI as a fixed critical difference. Following Schlagert and Hiller's (2017b) conceptualization of deterioration, patients were labeled as NOT, if they fulfilled either the 50% or the RCI criterion in any session.

ROC curve-based PI approach. Supplementing the rational PI approach using the established 50% criterion, we determined a modified "optimal"

percentage change criterion based on the Receiver Operating Characteristic (ROC) curve in a preliminary analysis. The aim was to be able to factor in our postulated higher costs of a false negative compared to a false positive feedback signal as well as the prevalence of reliable deterioration. The optimal PI criterion was determined with the Youden's index (Youden, 1950) modified for prevalence and costs by Perkins and Schisterman (2006), using the percentage change values between baseline and the respective assessment time as predictor for reliable deterioration. The relative costs of a false negative compared to a false positive classification were weighted by the factor five and a deterioration prevalence of 10% (Hansen et al., 2002) in psychotherapy was assumed. Except for the percentage change value, this approach modification was conceptualized identically to the original PI approach described above. By calibrating the optimal change value using our archive data, the ROC curve-based PI approach is, strictly speaking, no longer a genuine rational rule. It can rather be seen as a hybrid approach combining empirical and rational components.

Empirical Decision Rules

Preliminary analysis. Before the development of the empirically based feedback systems, another preliminary analysis to identify the predictors to be used in these models was conducted. The two-step procedure previously employed by Lutz et al. (2019) and Mütze et al. (2022) was followed. First, 41 potential prediction variables of pre to post improvement on the GSI-BSI, controlled for initial impairment, were examined. Next, all significantly correlating variables were included in a Least Absolute Shrinkage and Selection Operator (LASSO) regression model to select the most predictive variables to optimize model interpretation and protect against overfitting (Tibshirani, 1996). The LASSO linear regression model for continuous outcome was fitted using 10-fold-cross validation.

With both empirically based approaches, the original ETR and NN, multilevel growth curve models were used to model individual treatment trajectories with treatment session (level 1) nested within patients (level 2). In accordance with prior research on the negatively accelerating relationship between the number of sessions and treatment outcome in psychotherapy (Howard et al., 1986), a base-10 log-linear transformation on session number was used. The main difference between the two empirical approaches thus lies in the use of predictors, which is described below. Different failure boundaries were calculated according to Lutz et al. (1999). If the observed course of treatment deviated from the expected one at least once during the course of treatment by more than the upper boundary of the two-tailed 50%, 68%, 80%, 90% or 95% failure boundary (see Bone et al., 2021; Delgadillo et al., 2018; Finch et al., 2001; Lutz et al., 1999, 2019), patients were considered at risk for negative treatment outcome and labeled as NOT. Multiple failure boundaries were tested to identify the one with the highest detection rate given our sample.

ETR approach. After the preliminary analysis, the model described by Lutz et al. (1999), adapted for the predictors selected, was calculated. Patients' treatment progress was modeled using a multilevel fixed intercept and random slopes model. Predictors were entered in the level 2 model, initial impairment on the GSI-BSI serving as single intercept predictor. All interval scaled predictors were grand mean centered. Given that baseline GSI-BSI is used as first measurement point, it extracts all reliable variation when entered in the level 2 model predicting intercept (Lutz et al., 1999, 2005). Hence, no random effect was included.

NN approach. For the NN approach, a dissimilarity matrix using Gower's coefficient (Gower, 1971) was calculated to be able to include categorical predictors for the identification of a patient's most similar cases (i.e., NN). This resulted in distances between patients ranging from 0 indicating a complete overlap and 1 as maximal dissimilarity possible. Taking again prior research into consideration, baseline GSI-BSI was weighted double compared to the other predictors[4] because of its assumed high predictive value (Lutz et al., 2019, prioritizing initial impairment in the NN selection process). Subsequently, the 30 patients with the lowest dissimilarities were selected as NN (Mütze et al., 2022). In contrast to the ETR approach, individual treatment courses were then modeled using an unconditional fixed intercept random slopes growth model based solely on the index patient's NN subset. The index patient's slope was calculated as the average slope of its NN (Lutz et al., 2005, 2006).

Results

Preliminary Analyses

Out of 22 variables correlating significantly with pre to post improvement on the GSI-BSI, 5 were selected by the conservative LASSO procedure. Unstandardized regression weights were obtained from a general linear model predicting pre to post

improvement. Initial impairment ($b = 0.59$, $p < .001$) was the strongest predictor. Additionally, previous inpatient treatment ($b = -0.10$, $p < .001$), intrinsic treatment motivation ($b = 0.06$, $p < .001$), a personality disorder ($b = -0{,}17$, $p < .001$), and a patient's overall number of diagnoses ($b = -0.06$, $p < .001$) were selected, with the final model explaining 38 percent of variance (see Supplemental material 5).

In terms of the ROC-based determination of the percentage change criterion, a deterioration of 35.24% was found to be optimal for the detection of treatment failures. Hence, the modified PI approach will be referred as $PI_{35.24}$ approach in the following.

Comparison of Rational and Empirical Decision Rules

Regarding the complete sample, 110 patients (3%) reliably deteriorated during treatment while 1341 patients (36%) showed at least reliable improvement at the end of therapy. This corresponds to an overall effect size of $d_{\text{pre-post}} = 0.75$ [0.71;0.78], with an average improvement from $M_{\text{pre}} = 1.17$ ($SD_{\text{pre}} = 0.68$) to $M_{\text{post}} = 0.70$ ($SD_{\text{post}} = 0.59$), which can be considered as moderate effect (Ellis, 2010).

The prediction accuracy of the reliable deterioration cases by the decision rules is shown in Table I. For better clarity, only the failure boundaries with the highest sensitivity and specificity or PPV values for the empirical methods are displayed. Further details on the ETR model can be found in Supplemental material 6, results for the other failure boundaries in Supplemental material 7. For both the NN and ETR method, the 50% failure boundaries showed the highest sensitivity with regard to the prediction of reliable deterioration, the NN method being slightly more sensitive (.87 vs .81 for the ETR method). Here, 96 out of 110 cases were correctly identified as NOT during treatment, resulting in the highest true positive rate (87.3%) of all decision rules, outperforming the ETR-based boundaries as well as the rational decision rules. As for specificity, the 95% failure boundaries of both empirical methods showed the highest rates. However, the overall highest specificity was shown by the rational RCI approach, with only 130 cases (3.6%) falsely labeled as NOT. Correspondingly, the RCI approach also showed with .30 a significantly higher PPV than all other approaches. The positive (.07-.30) and negative predictive values (.98-.99) of all investigated rules were low and very high, respectively. Of the selected approaches, the $PI_{35.24}$ approach had the highest AUC value, followed by the original rational PI and the empirical NN approach with a 50% failure boundary ($NN_{50\%}$). Differences to the other rules were, however, not significant.

For the rational approaches, the time difference between the first NOT signal given by the approach and the treatment outcome assessment were calculated for the true positive cases. A mean time difference of $M_{n=56} = 28.57$ treatment sessions ($SD = 14.48$) for the RCI approach and $M_{n=78} = 30.00$ sessions ($SD = 15.69$) for the PI approach were found.

Table II shows the reliable improvement and deterioration rates of OT and NOT cases, cases being differentiated by the respective decision rule. The probability for NOT cases to be reliably deteriorated by the end of treatment was significantly higher in every decision rule and failure boundary condition compared to OT cases. Accordingly, the probability for NOT cases to be reliably improved was

Table I. Contingency table indices and associated evaluation parameters for the rational and empirical decision rules (and failure boundaries).

Approach	TP *n* (%)	FN *n* (%)	FP *n* (%)	TN *n* (%)	Sensitivity [CI]	Specificity [CI]	PPV[a] [CI]	NPV [CI]	AUC [CI]
PI									
50%	78 (70.9)	32 (29.1)	619 (17.0)	3029 (83.0)	.71 [.62;.79]	.83 [.82;.84]	.11 [.09;.14]	.99 [.99;.99]	.77 [.73;.81]
35.24%	90 (81.8)	20 (18.2)	727 (19.9)	2921 (80.1)	.82 [.73;.89]	.80 [.79;.81]	.11 [.09;.13]	.99 [.99;.99]	.81 [.77;.85]
RCI	56 (50.9)	54 (49.1)	130 (3.6)	3518 (96.4)	.51 [.42;.61]	.96 [.96;.97]	.30 [.24;.37]	.99 [.98;.99]	.74 [.69;.78]
NN									
50%	96 (87.3)	14 (12.7)	1220 (33.4)	2428 (66.6)	.87 [.80;.93]	.67 [.65;.68]	.07 [.06;.09]	.99 [.99;.99]	.77 [.74;.80]
95%	56 (50.9)	54 (49.1)	308 (8.4)	3340 (91.6)	.51 [.41;.61]	.92 [.91;.92]	.15 [.12;.20]	.98 [.98;.99]	.71 [.67;.76]
ETR									
50%	89 (80.9)	21 (19.1)	1103 (30.2)	2545 (69.8)	.81 [.72;.88]	.70 [.68;.71]	.08 [.06;.09]	.99 [.99;.99]	.75 [.72;.79]
95%	54 (49.1)	56 (50.9)	260 (7.1)	3388 (92.9)	.49 [.39;.59]	.93 [.92;.94]	.17 [.13;.22]	.98 [.98;.99]	.71 [.66;.76]

Note: $N = 3758$. TP = True Positive, patients identified as NOT meeting the criterion of reliable deterioration at the end of treatment. FN = False Negative, FP = False Positive, TN = True Negative. PPV = positive predictive value, NPV = negative predictive value. AUC = Area under the curve. CI = 95% confidence interval. PI = precent improvement approach, RCI = reliable change index approach, NN = nearest neighbors method, ETR = expected treatment response approach. Indented percentages for NN and ETR represent failure boundaries.
[a]TP / (TP + FP).

Table II. Relative frequencies of reliable improvement, deterioration and, effect sizes from pre to post treatment for OT and NOT patients by rational and empirical decision rules (and failure boundaries).

	Reliable improvement			Reliable deterioration			Cohen's *d* [CI][a]	
Approach	OT *n* (%)	NOT *n* (%)	χ^2 (*df*=1)	OT *n* (%)	NOT *n* (%)	χ^2 (*df*=1)	OT	NOT
PI								
50%	1291 (42.2)	50 (7.2)	303.07***	32 (1.0)	78 (11.2)	205.66***	0.94 [0.90;0.98]	−0,11 [−0.18;−0.04]
35.24%	1277 (43.4)	64 (7.8)	352.82***	20 (0.7)	90 (11.0)	240.39***	0.98 [0.94;1.02]	−0,08 [−0.14;−0.01]
RCI	1319 (36.9)	22 (11.8)	48.52***	54 (1.5)	56 (30.1)	508.79***	0.81 [0.78;0.85]	−0,35 [−0.51;−0.20]
NN								
50%	995 (40.7)	346 (26.3)	77.48***	14 (0.6)	96 (7.3)	135.97***	1.03 [0.98;1.08]	0.37 [0.32;0.43]
95%	1274 (37.5)	67 (18.4)	52.42***	54 (1.6)	56 (15.4)	220.13***	0.83 [0.80;0.87]	0.05 [−0.05;0.16]
ETR								
50%	1012 (39.4)	329 (27.6)	49.70***	21 (0.8)	89 (7.5)	126.60***	0.99 [0.95;1.04]	0.40 [0.34;0.46]
95%	1261 (36.6)	80 (25.5)	15.55***	56 (1.6)	54 (17.2)	245.56***	0.85 [0.81; 0.89]	0.14 [0.03;0.25]

Note: N = 3758. OT = on track, NOT = not on track. *d* = effect size pre to post treatment. ***$p < .001$. CI = 95% confidence interval. PI = precent improvement approach, RCI = reliable change index approach, NN = nearest neighbors method, ETR = expected treatment response approach. Indented percentages for NN and ETR represent failure boundaries.
[a]Negative effect sizes indicate higher impairment at the end of treatment.

significantly smaller than for OT patients for all decision rules, with the most substantial difference shown by the $PI_{35.24}$ approach (7.8% vs 43.4%) slightly ahead of the original PI approach. A similar pattern can be seen in terms of the pre to post treatment effect sizes of NOT and OT patients. While the effect sizes of NOT patients can all be considered small, similarly large effects were found for OT patients across all decision rules. The biggest differences in effect sizes between OT and NOT cases were found in the rational approaches with $\Delta d = 1.16$ for the RCI, and $\Delta d = 1.05$ for the 50% and 1.06 for the 35.24% PI approach, respectively. For the NOT cases, effect sizes were differentiated depending on the number of signals received (Table III). The effect sizes become smaller or even negative as the number of signals increases.

Depressive Disorder Subsample Analysis

The results of the BDI subsample analysis are presented at length in Supplemental materials 8 and 9. All approaches were less sensitive when using an external outcome measure, $NN_{50\%}$ continuing to have the highest sensitivity at .79. For the rational rules, differences in sensitivity between the main and BDI sample were .12 for the PI and .18 for the RCI approach. The empirically optimized $PI_{35.24}$ showed a difference of .18 equal to the RCI approach. Regarding the empirical rules, $ETR_{50\%}$ showed with .04 the smallest difference and $NN_{95\%}$ with .19 the biggest. There were only marginal differences in terms of specificity. The biggest drop in PPV is seen in the RCI approach from .30 to .14. The biggest effect size differences between OT and

Table III. Effect sizes from pre to post treatment for NOT cases differentiated by the number of signals received over their course of treatment.

Approach	1 Signal *d* [CI][a]	2 Signals *d* [CI]	3+ Signals *d* [CI]
PI			
50%	0.02 [−0.07;0.10]	−0.28 [−0.43;−0.12]	−0.69 [−0.99;−0.39]
35.24%	0.05 [−0.02;0.12]	−0.24 [−0.37;−0.11]	−0.69 [−0.96;−0.41]
RCI	−0.22 [−0.39; −0.05]	−0.66 [−1.05;−0.26]	−1.34 [−2.14;−0.54]
NN			
50%	0.47 [0.40;0.54]	0.25 [0.15;0.36]	0.21 [0.04;0.38]
95%	0.14 [0.03;0.25]	−0.15 [−0.43;0.14]	−0.71 [−1.43;0.02]
ETR			
50%	0.51 [0.43;0.58]	0.25 [0.14;0.36]	0.2 [−0.01;0.41]
95%	0.18 [0.06;0.30]	0.11 [−0.20;0.43]	−0.32 [−1.03;0.40]

Note: *N* = 3758. NOT = not on track. *d* = Cohen's *d*, effect size pre to post treatment. CI = 95% confidence interval. PI = precent improvement approach, RCI = reliable change index approach, NN = nearest neighbors method, ETR = expected treatment response approach. Indented percentages for NN and ETR represent failure boundaries.
[a]Negative effect sizes indicate higher impairment at the end of treatment.

NOT cases using the BDI as outcome measure are shown by the rational approaches with $\Delta d_{PI} = 0.79$ and $\Delta d_{RCI} = 0.89$.

Discussion

In the present study, two empirical, two rational, and a hybrid decision rule for providing feedback regarding a patient's treatment course were investigated, and compared on the basis of intermittent ROM. Consistent with previous research, the empirical decision rules were found to be superior to the rational rules. With the highest detection rate of NOT cases (87.3%), the best result could be achieved using the NN-based rule with a 50% failure boundary ($NN_{50\%}$). When comparing the rational rules, the PI was found to be superior to the RCI approach with regard to sensitivity. By using the "optimal" percentage change value of 35.24% determined by means of ROC curve analysis instead of the established 50%, the PI approach became more sensitive towards the detection of treatment failures. The hybrid $PI_{35.24}$ approach also had the highest value in terms of the AUC. But with a true positive rate of 70.9% and 81.8% respectively, the hit rates of the PI approaches were still (substantially) lower than with the empirical NN method.

To address the question on how the reduction of the monitoring frequency affects the accuracy of the decision rules, we consulted the previous studies comparing rational and empirical decision rules. Since these studies evaluated well-established and widely used decision rules, they will serve as a benchmark for our findings. The idea is that our results should be comparable to those based on session-to-session based approaches, since a substantially poorer detection rate has no benefit for clinical practice even with a facilitated implementation. Whereas Lambert et al. (2002) reported a substantially higher predictive accuracy of their empirical decision rule with a hit rate (i.e., true positives) of 100% compared to those examined in our study ($NN_{50\%} = 87.3\%$ and $ETR_{50\%} = 80.9\%$), similar hit rates to ours were found in the replication (81.2%; Spielmans et al., 2006). However, in our case, considerably more patients were incorrectly labeled as NOT by the empirical rules (18.0% and 19.1% vs $NN_{50\%} = 33.4\%$ and $ETR_{50\%} = 30.2\%$). At last, in the direct comparison of the NN-based models, our replication on the intermittent ROM structure showed higher sensitivity values ($NN_{50\%} = .87$ vs .68), while maintaining comparable specificities ($NN_{50\%} = .67$ vs .68; Lutz et al., 2006). Corresponding to the empirical rules, the PI approach examined in this study showed once lower (80.6%; Lambert et al., 2002) and once quite similar (68.7%; Spielmans et al., 2006) hit rates in the comparison of the rational rules (PI = 70.9%). However, there was a difference with respect to the false positive rate, which was (considerably) lower for the PI approach in both cases (PI = 17.0% vs 20.8% and 39.6%). Finally, Lutz et al. (2006) reported a lower performance of the rational rule in both accuracy measures (sensitivity: .71 vs .57; specificity: .83 vs .66). In addition, the data-driven reduction of the PI criterion to 35.24% led to increased hit rates, which are within the range of the results of the rational rule reported by Lambert et al. (2002) and the empirical rule of Spielmans et al. (2006).

In summary, the above comparisons indicate that even based on an intermittent ROM structure, deteriorations during treatment can be detected to a comparable extent. This points to the applicability and thus the potential of the methods investigated, especially NN and PI, for routine clinical care.

Strengths and Limitations

As outlined above, we selected sensitivity rather than specificity or the PPV and NPV as main prediction target. However, all highly sensitive rules produced a considerable percentage of false warning signals (17.0–33.4%). Especially considering the therapists' reported fears of negative assessment (de Jong, 2016), a feedback system should not excessively label cases incorrectly as NOT. A worst-case scenario would be that a high number of false warning signals could lead to them being ignored by practitioners (Spielmans et al., 2006). Here in particular, the comprehensibility and communicability of the rational PI approach could be an advantage for both therapists and patients. Even if eventually no deterioration occurs, it would be clear to therapists and patients that a 50% worsening of symptoms during therapy is meaningful and should be addressed.

Looking at the PPVs as a further evaluation parameter, all values can be interpreted as low. The highest PPV was achieved by the rational RCI rule with .30. Nevertheless, even this PPV indicates that of all patients who received a NOT signal, ultimately only 30% show a deterioration. As positive and negative predictive values are sensitive to imbalanced data, the low (or very high) values could be influenced by the low prevalence of deterioration cases in our sample (Tharwat, 2021). At 3%, it is slightly below the frequently reported 5–10% (Hansen et al., 2002). This might be because the therapists in training who provided most of the treatments in the outpatient clinic received mandatory supervision every fourth session, including the regular review of

treatment videos. The supervision could have had a positive effect on the treatment outcomes (Bambling et al., 2006), thus reducing the rate of treatment failures. The application of the presented approaches in samples with higher prevalence rates could be beneficial for the further investigation of the PPVs.

The low precision of the feedback approaches also highlights two important aspects. First, thorough training of therapists and accompanying supervision are required so that therapists can understand the feedback system and correctly interpret warning signals. Second, clinical judgment remains crucial and should be complemented but not replaced by the feedback system (Castonguay et al., 2013). Warning signals and psychometric scores should not be regarded as absolute certainties, but rather be openly discussed as an impulse in therapy (e.g., as a gateway to previously unaddressed topics) and evaluated jointly by the patient and therapist.

In this sense, it might also be worth looking at the relative frequencies of reliable improvement and effect sizes for NOT and OT patients to better assess the magnitude of a signal. The probability to be reliably improved at the end of treatment for patients identified as NOT by all empirical approaches was a little less than half of the probability for other patients. This corresponds to reported rates in previous feedback studies (e.g., Lutz et al., 2019; Spielmans et al., 2006). An even clearer difference can be seen with the rational approaches. NOT cases selected by the PI approach reached the reliable improvement criterion about six times less frequently (42.2% vs 7.2%). Despite its data-based optimization, the $PI_{35.24}$ approach achieved only marginally better results. Looking at the effect sizes, we find similar patterns with considerable differences between NOT and OT patients. While NOT patients with a signal generated by the empirical approaches (esp. $NN_{95\%}$) achieved on average only small effects in treatment, the NOT patients labeled by the rational rules even exhibited negative effect sizes, in the sense of a higher symptom burden at the end of therapy. Thus, all but in particular the rational approaches can reliably differentiate between satisfactory and unfavorable treatment courses. If a patient receives two or even more signals, the treatment effects become increasingly unfavorable and should be addressed with greater urgency.

To our knowledge, the present study was the first to use the PI criterion as a rational rule for generating warning signals. Since, in line with previous research, the 50% deterioration criterion was supplemented by the criterion of reliable deterioration, the PI approach can be seen less as an exclusive new alternative than as an extension of the more often used RCI approach. Interestingly, the PI approach proved to be significantly more sensitive in detecting negative treatment outcomes than the RCI approach alone, showing the potential of this approach. In previous research, the PI criterion was limited to the pathological range of a given psychometric scale (see Schlagert & Hiller, 2017b). Despite our efforts to develop an evidence-based, practice-oriented, and easy to use decision rule, we had to deviate from the previous established conceptualization due to a considerable proportion of patients with a non-pathological baseline value in our sample (20%). Similar proportions have been reported in other samples (e.g., 29% in Rubel et al., 2015), suggesting that our modification is necessary to make the PI approach broadly applicable.

By using ROC curve analysis, accounting for the relative cost of a false negative classification and the deterioration base rate of 10%, an optimal percentage change value of 35.24% was determined. The increase in sensitivity compared to the PI approach using the established 50% criterion indicates the potential of this modified criterion. With regard to the additional evaluation parameters, however, the ROC curve-based calibration of the PI criterion loses some of its added value. The $PI_{35.24}$ approach only differentiated between OT and NOT cases by $\Delta d = 0.01$ better than the original PI approach. Moreover, the sensitivity of the PI approach improved only marginally using the modified percentage value in the subsample analysis. It should also be noted that the selection of the weight of the relative costs for a false negative classification was not empirically grounded. In this regard, Perkins and Schisterman (2006) themselves point out that the costs are often difficult to assess. Overall, the results of the modified $PI_{35.24}$ approach indicate that more research is needed. It remains to be evaluated whether the data-based optimization will prove to be generalizable and thus beneficial, also considering that the striking simplicity of the PI criterion gets lost to some extent. Further optimization might be achieved by adjusting the weighting of the relative costs.

One of the limitations of the present study concerns the use of the reliable deterioration criterion in both rational, and the hybrid approach (as an independent rule and supplement) as well as for the assessment of treatment outcomes. We chose the criteria by Jacobson and Truax (1991) for treatment outcome evaluation as they are widely used in clinical research, in particular for the evaluation of feedback (e.g., Lutz et al., 2006). Given an average of (almost) 30 sessions, there was a considerable time gap, in which a great deal of change can occur between the first signals provided by rational approaches and

outcome assessment. Still, the present results might be biased to the advantage of the rational approaches.

To address this potential bias, we selected effect size differences as an additional RCI-independent evaluation criterion. In this regard, the rational approaches perform particularly well (see above). Furthermore, we carried out the subsample analysis using the BDI as additional outcome measure. In regard of sensitivity, the decrease from the main to the subsample of the rational approaches was comparable to the empirical (RCI and $PI_{35.24}$ approach similar to the 95%, original PI approach similar to the 50% failure boundaries). In contrast, the RCI approach showed a more substantial drop in PPVs compared to the other rules. Although the RCI approach still had the highest PPV, this could indicate a bias. Overall, the inconclusive results of the subsequent analysis as well as the singular nature of our findings limit the generalizability of our results. Accordingly, a replication and validation of the rational rules in different treatment settings using various outcome measures is needed. Balancing costs and benefits of the rational (e.g., immediate implementation possible, less psychometric knowledge and training required vs lower detection rates) and empirical decision rules (e.g., large amounts of data necessary, higher software requirements, advanced statistical-methodological knowledge needed vs higher detection rates), the PI approach combining the PI and RCI method could nevertheless be a useful transitional solution due to its satisfactory results until sufficient amounts of data are available to compute the NN-based empirical rule.

In terms of the empirical approaches, this study also has limitations. Using LASSO regression, five variables were identified to predict treatment outcome: Initial impairment (on the GSI-BSI), previous inpatient treatment, intrinsic treatment motivation, a diagnosed personality disorder, and overall number of diagnoses. The predictors identified are in line with previous research (e.g., Huibers et al., 2015; Lutz et al., 2019; Lutz, Rubel, et al., 2015; Mütze et al., 2022), indicating their generalizability. However, evidence on dynamic prediction models and dynamic failure boundaries indicate that the predictive accuracy can be improved by adding information from subsequent treatment sessions and early change information (Bone et al., 2021; Lutz et al., 2005, 2019). Thus, the restriction to baseline variables could limit the predictive accuracy of the models. Future studies should therefore investigate the effect on prediction accuracy of the basic feedback systems by including progress information. Keeping the goal of a facilitated implementation in mind, the additional benefit should be balanced with the increased complexity of the models. In this sense, potential parameters for a further simplification should also be explored. In line with previous research, we selected 30 patients with the lowest dissimilarities as NN (Lutz et al., 2019; Mütze et al., 2022). Nevertheless, it has been shown that the performance of the NN model varies only marginally with the number of nearest neighbors used (Lutz et al., 2005, comparing numbers of 10–50 neighbors). Accordingly, the applicability of models with fewer NN (e.g., 10; Lutz et al., 2006) to our intermittent ROM structure should be tested. Moreover, in accordance with previous findings, initial impairment (on the GSI-BSI) explained by far the largest amount of variance in treatment outcome in our study. Empirical models based on a large number of predictors could complicate their implementation in routine clinical care, as their standardized assessment requires time. Subsequently, future studies might also include models such as the alternative ETR approach by Finch et al. (2001), which are based solely on the baseline score of the monitoring instrument, on which the feedback system is based.

Analogous to current ETR-based feedback systems, including those we replicated and validated on the intermittent ROM structure, we also modeled the course of therapy according to the dose–response model for psychotherapy (Finch et al., 2001; Lutz et al., 2005, 2019). Although replicated and consistent support can be found for the curvilinear relationship between treatment length and outcome, there is evidence for different response trajectories in subgroups of patients (e.g., with more chronic or severe psychopathology; Robinson et al., 2020). On an individual level, treatment trajectories also appear to be highly variable and heterogeneous. For instance, Dyason et al. (2021) identified constant (no change) and linear as the most common patterns of change whereas only 3% of patients followed the dose–response models' negative accelerating course in their sample. Overall, there is little research available with respect to long-term psychotherapies. So far, findings are rather inconsistent, with indications for a linear (Nordmo et al., 2021), but also for a curvilinear course (Sembill et al., 2019). Thus, to further improve the accuracy of the empirical feedback approaches, especially for long-term psychotherapies, a more individualized or subgroup-dependent modulation (e.g., primary diagnosis dependent) of treatment trajectories could be performed in future studies.

To achieve a broader implementation of monitoring and feedback in clinical practice, we have chosen the approach of reducing the assessment frequency. In our case, the assessment times were adapted to the number of sessions for short- and long-term CBT treatments provided by the health care and insurance

system in Germany. However, this also has its drawbacks for research and practice. In practice, the decision support provided by the feedback system decreases and unfavorable treatment courses may be detected more slowly. A higher monitoring frequency in the first phase of treatment (e.g., an additional assessment at session 5) could bring additional benefit, as early treatment trajectories and changes in particular appear to have a high predictive value regarding treatment outcome (Rubel et al., 2015; Schlagert & Hiller, 2017a). At the same time, data gathered through repeated measurements are less useful for practice-oriented research compared to those based on session-to-session assessments (Lutz et al., 2022). Also, on a theoretical level, a higher monitoring and thus feedback frequency is associated with a higher feedback effectiveness (see Contextualized Feedback Intervention Theory; Sapyta et al., 2005). However, in a more recent meta-analysis by de Jong et al. (2021), no significant differences were found between continuous and intermittent feedback. Feedback frequency thus represents an interesting moderator that should be explored in more detail in future research. To further investigate the influence of a reduced feedback frequency, as it was used in our study, ideally a session-to-session and a less frequent feedback should be directly compared in a randomized controlled trial. Besides the efficacy itself (in terms of reduced deterioration rates or overall improved treatment outcomes), the acceptance of the feedback systems by practitioners and patients as well as the perceived benefit for clinical practice should be considered, too.

Practical Implications

To counterbalance the shortcomings (i.e., low sensitivity and specificity or PPV, respectively) of the investigated approaches and failure boundaries, the use of two distinct warning signals could be useful in practice (see Lambert et al., 2002, also using multiple warning signals). In the context of the empirical decision rules, the 50% failure boundary could be assigned a yellow warning signal. In this way, its high detection rate is used, while at the same time taking its FP rate into account. As a supplement, the 95% failure boundary could then be provided with a red warning signal. Due to the failure boundary's high specificity, an urgent need for intervention can be assumed in this case. Correspondingly, a yellow warning signal could be used for the 50% (or 35.24%) deterioration criterion (high sensitivity) and a red signal for reliable deterioration (RCI approach; highest specificity and PPV) within the scope of the rational rules.

Conclusions

We evaluated and compared two well-established empirical, two genuine, and one hybrid rational decision rule as basis for a scaled-down feedback system on a naturalistic outpatient sample. The results suggest that the empirical decision rules that are commonly based on session-by-session ROM are generalizable to a less-frequent monitoring structure. Benchmarking our results with those of previous research on session-to-session based feedback, satisfactory detection rates and prediction accuracies could be achieved. Thus, the present study demonstrated an opportunity that could enable a broader implementation of feedback systems in clinical practice, reducing the current gap between research and practice. In accordance with previous research, the empirically based decision rules (especially $NN_{50\%}$) outperformed the rational rules. Still, the first time used PI approach, combining PI- and RCI-based decision making, might be a good (transitional) solution for clinical practice due to its also satisfactory results. The use of the empirically optimized PI criterion could lead to further improved results.

Acknowledgements

We thank Kaline Mütze for the open and always helpful exchange of experiences and expertise.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Statement of Ethics

The study has been performed in accordance with the principles stated in the Declaration of Helsinki. Ethical approval for the study was granted by the Ethics Committee of the Psychological Institute of the Johannes Gutenberg University Mainz in Germany (021-JGU-psychEK-S025). All patients gave written informed consent regarding the use of their anonymized treatment data for research purposes.

Data Availability Statement

The archival data for this study are not publicly available due to restrictions that could compromise research participant privacy. Analysis code and data are available upon reasonable request by emailing the corresponding author.

Supplemental Data

Supplemental data for this article can be accessed https://doi.org/10.1080/10503307.2024.2334047

Notes

[1] Lutz et al. (2006) using the shortened OQ-30.

[2] The RCI was calculated using the following formula (Jacobson & Truax, 1991), with SD_1 defined as the standard deviation of the reference population and r_{xx} as the reliability of the questionnaire:

$$\text{RCI} = 1.96 \times \sqrt{2 \times (SD_1 \times \sqrt{1 - r_{xx}})^2}$$

[3] The cutoff between the pathological and non-pathological range was calculated using formula c for unequal variances in the functional and dysfunctional populations by Jacobson and Truax (1991):

$$\text{cutoff} = \frac{SD_0 M_1 + SD_1 M_0}{SD_0 + SD_1}$$

with M_0 and SD_0 standing for the mean and standard deviation in the non-pathological reference sample and M_1 and SD_1 for the pathological sample, respectively. For the BSI-GSI, $M_1 = 1.32$ and $SD_1 = 0.72$ from the representative outpatient sample and $M_0 = 0.31$ and $SD_0 = 0.23$ from the adult sample reported in the manual by Franke (2000) were taken.

[4] The NN approach without double weighting of the BSI-GSI baseline score resulted in only marginal differences in the detection of NOT cases.

ORCID

Stephan Ramsperger http://orcid.org/0009-0002-1825-0473
Michael Witthöft http://orcid.org/0000-0002-4928-4222
Anne-Kathrin Bräscher http://orcid.org/0000-0002-2621-5689

References

Altman, D. G., & Bland, J. M. (1994). Statistics notes: Diagnostic tests 2: Predictive values. *BMJ*, *309*(6947), 102. https://doi.org/10.1136/bmj.309.6947.102

Bambling, M., King, R., Raue, P., Schweitzer, R., & Lambert, W. (2006). Clinical supervision: Its influence on client-rated working alliance and client symptom reduction in the brief treatment of major depression. *Psychotherapy Research*, *16*(3), 317–331. https://doi.org/10.1080/10503300500268524

Bates, D., Maechler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–49. https://doi.org/10.18637/jss.v067.i01

Bone, C., Simmonds-Buckley, M., Thwaites, R., Sandford, D., Merzhvynska, M., Rubel, J., Deisenhofer, A., Lutz, P. W., & Delgadillo, J. (2021). Dynamic prediction of psychological treatment outcomes: Development and validation of a prediction model using routinely collected symptom data. *The Lancet Digital Health*, *3*(4), e231–e240. https://doi.org/10.1016/S2589-7500(21)00018-2

Boswell, J. F. (2020). Monitoring processes and outcomes in routine clinical practice: A promising approach to plugging the holes of the practice-based evidence colander. *Psychotherapy Research*, *30*(7), 829–842. https://doi.org/10.1080/10503307.2019.1686192

Boswell, J. F., Kraus, D. R., Miller, S. D., & Lambert, M. J. (2015). Implementing routine outcome monitoring in clinical practice: Benefits, challenges, and solutions. *Psychotherapy Research*, *25*(1), 6–19. https://doi.org/10.1080/10503307.2013.817696

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

de Jong, K. (2016). Deriving implementation strategies for outcome monitoring feedback from theory, research and practice. *Administration and Policy in Mental Health and Mental Health Services Research*, *43*(3), 292–296. https://doi.org/10.1007/s10488-014-0589-6

de Jong, K., Conijn, J. M., Gallagher, R. A. V., Reshetnikova, A. S., Heij, M., & Lutz, M. C. (2021). Using progress feedback to improve outcomes and reduce drop-out, treatment duration, and deterioration: A multilevel meta-analysis. *Clinical Psychology Review*, *85*(2019), 102002. https://doi.org/10.1016/j.cpr.2021.102002

de Jong, K., van Sluis, P., Nugter, M. A., Heiser, W. J., & Spinhoven, P. (2012). Understanding the differential impact of outcome monitoring: Therapist variables that moderate feedback effects in a randomized clinical trial. *Psychotherapy Research*, *22*(4), 464–474. https://doi.org/10.1080/10503307.2012.673023

Delgadillo, J., de Jong, K., Lucock, M., Lutz, W., Rubel, J., Gilbody, S., Ali, S., Aguirre, E., Appleton, M., Nevin, J., O'Hayon, H., Patel, U., Sainty, A., Spencer, P., & McMillan, D. (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. https://doi.org/10.1016/S2215-0366(18)30162-7

Derogatis, L. R. (1992). *SCL-90-R: Administration, scoring & procedures manual-II, for the R (revised) version and other instruments of the psychopathology rating scale series* (2nd ed.). Clinical Psychometric Research Incorporated.

Dyason, K. M., Low-Choy, S., O'Donovan, A., & Shanley, D. C. (2021). Modeling individualized trajectories of symptom change to improve feedback procedures in psychotherapy. *Journal of Consulting and Clinical Psychology*, *89*(1), 34–48. https://doi.org/10.1037/ccp0000614

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambride University Press. https://doi.org/10.1017/CBO9780511761676

Erekson, D. M., Horner, J., & Lambert, M. J. (2018). Different lens or different picture? Comparing methods of defining dramatic change in psychotherapy. *Psychotherapy Research*, *28*(5), 750–760. https://doi.org/10.1080/10503307.2016.1247217

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*, *8*(4), 231–242. https://doi.org/10.1002/cpp.286

Franke, G. H. (2000). *BSI: Brief symptom inventory von L.R. Derogatis (Kurzform des SCL-90-R) – Deutsche Version – Manual*. Beltz Test GmbH.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent.

Journal of Statistical Software, *33*(1), 1–22. https://doi.org/10.18637/jss.v033.i01

Fydrich, T., Renneberg, B., Schmitz, B., & Wittchen, H.-U. (1997). *SKID-II. Strukturiertes klinisches Interview für DSM-IV. Achse II: Persönlichkeitsstörungen. Interviewheft. Eine deutschsprachige, erweiterte Bearbeitung der amerikanischen Originalversion des SCID-II von Michael B. First, Robert L. Spitzer, Miriam Gibbon*. Hogrefe.

Gower, A. J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, *27*(4), 857–871. https://doi.org/10.2307/2528823

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

Hansen, N. B., Lambert, M. J., & Forman, E. M. (2002). The psychotherapy dose-response effect and its implications for treatment delivery services. *Clinical Psychology: Science and Practice*, *9*(3), 329–343. https://doi.org/10.1093/clipsy.9.3.329

Hatfield, D. R., McCullough, L., Frantz, S. H. B., & Krieger, K. (2010). Do we know when our clients get worse? An investigation of therapists' ability to detect negative client change. *Clinical Psychology & Psychotherapy*, *17*(1), 25–32. https://doi.org/10.1002/cpp.656

Hautzinger, M., Bailer, M., Worall, H., & Keller, F. (1994). *Beck-depressions-inventar (BDI)*. Hans Huber.

Hautzinger, M., Keller, F., & Kühner, C. (2006). *BDI-II – beck depressions-inventar revision – manual*. Harcourt Test Services.

Hiller, W., Bleichhardt, G., Haaf, B., Legenbauer, T., Mauer-Matzen, K., & Rübler, D. (2006). Zertifizierung einer Psychotherapeutischen Hochschulambulanz nach DIN en ISO 9001. *Zeitschrift Fur Klinische Psychologie Und Psychotherapie*, *35*(3), 225–233. https://doi.org/10.1026/1616-3443.35.3.225

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

Howard, K. I., Kopta, S. M., Krause, M. S., & Orlinsky, D. E. (1986). The dose-effect relationship in psychotherapy. *American Psychologist*, *41*(2), 159–164. https://doi.org/10.1037/0003-066X.41.2.159

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

Ionita, G., & Fitzpatrick, M. (2014). Bringing science to clinical practice: A Canadian survey of psychological practice and usage of progress monitoring measures. *Canadian Psychology*, *55*(3), 187–196. https://doi.org/10.1037/a0037355

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

Jensen-Doss, A., Haimes, E. M. B., Smith, A. M., Lyon, A. R., Lewis, C. C., Stanick, C. F., & Hawley, K. M. (2018). Monitoring treatment progress and providing feedback is viewed favorably but rarely used in practice. *Administration and Policy in Mental Health and Mental Health Services Research*, *45*(1), 48–61. https://doi.org/10.1007/s10488-016-0763-0

Kaiser, T., Schmutzhart, L., & Laireiter, A. R. (2018). Attitudes of Austrian psychotherapists towards process and outcome monitoring. *Administration and Policy in Mental Health and Mental Health Services Research*, *45*(5), 765–779. https://doi.org/10.1007/s10488-018-0862-1

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/JSS.V082.I13

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

Lambert, M. J., & Shimokawa, K. (2011). Collecting client feedback. *Psychotherapy*, *48*(1), 72–79. https://doi.org/10.1037/a0022238

Lambert, M. J., Whipple, J. L., Bishop, M. J., Vermeersch, D. A., Gray, G. V., & Finch, A. E. (2002). Comparison of empirically-derived and rationally-derived methods for identifying patients at risk for treatment failure. *Clinical Psychology and Psychotherapy*, *9*(3), 149–164. https://doi.org/10.1002/cpp.333

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. https://doi.org/10.1037/pst0000167

Lecrubier, Y. (2002). How do you define remission? *Acta Psychiatrica Scandinavica*, *106*(s415), 7–11. https://doi.org/10.1034/j.1600-0447.106.s415.2.x

Lutz, W., de Jong, K., & Rubel, J. A. (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., Lambert, M. J., Harmon, S. C., Tschitsaz, A., Schürch, E., & Stulz, N. (2006). The probability of treatment success, failure and duration — what can be learned from empirical data to support decision making in clinical practice? *Clinical Psychology and Psychotherapy*, *13*, 223–232. https://doi.org/10.1002/cpp.496

Lutz, W., Leach, C., Barkham, M., Lucock, M., Stiles, W. B., Evans, C., Noble, R., & 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. https://doi.org/10.1037/0022-006X.73.5.904

Lutz, W., Martinovich, Z., & Howard, K. I. (1999). Patient profiling: An application of random coefficient regression models to depicting the response of a patient to outpatient psychotherapy. *Journal of Consulting and Clinical Psychology*, *67*(4), 571–577. https://doi.org/10.1037/0022-006X.67.4.571

Lutz, W., Rubel, J. A., Schiefele, A. K., Zimmermann, D., Böhnke, J. R., & Wittmann, W. W. (2015). Feedback and therapist effects in the context of treatment outcome and treatment length. *Psychotherapy Research*, *25*(6), 647–660. https://doi.org/10.1080/10503307.2015.1053553

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

Lutz, W., Schwartz, B., & Delgadillo, J. (2022). Measurement-based and data-informed psychological therapy. *Annual Review of Clinical Psychology*, *18*, 71–98. https://doi.org/10.1146/annurev-clinpsy-071720-014821

Lutz, W., Stulz, N., Martinovich, Z., Leon, S., & Saunders, S. M. (2009). Methodological background of decision rules and feedback tools for outcomes management in psychotherapy. *Psychotherapy Research*, *19*(4–5), 502–510. https://doi.org/10.1080/10503300802688486

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., Studer, M., Roudier, P., Gonzalez, J., Kozlowski, K.,

Schubert, E., & Murphy, K. (2021). *Finding groups in data: Cluster analysis extended Rousseeuw et al.. R package version 2.1.2.* https://svn.r-project.org/R-packages/trunk/cluster/

Mellor-Clark, J., Cross, S., Macdonald, J., & Skjulsvik, T. (2016). Leading horses to water: Lessons from a decade of helping psychological therapy services use routine outcome measurement to improve practice. *Administration and Policy in Mental Health and Mental Health Services Research*, *43*(3), 279–285. https://doi.org/10.1007/s10488-014-0587-8

Mütze, K., Witthöft, M., Lutz, W., & Bräscher, A. (2022). Matching research and practice: Prediction of individual patient progress and dropout risk for basic routine outcome monitoring. *Psychotherapy Research*, *32*(3), 358–371. https://doi.org/10.1080/10503307.2021.1930244

Newnham, E. A., & Page, A. C. (2010). Bridging the gap between best evidence and best practice in mental health. *Clinical Psychology Review*, *30*(1), 127–142. https://doi.org/10.1016/j.cpr.2009.10.004

Nordmo, M., Monsen, J. T., Høglend, P. A., & Solbakken, O. A. (2021). Investigating the dose–response effect in open-ended psychotherapy. *Psychotherapy Research*, *31*(7), 859–869. https://doi.org/10.1080/10503307.2020.1861359

Østergård, O. K., Randa, H., & Hougaard, E. (2020). The effect of using the partners for change outcome management system as feedback tool in psychotherapy—a systematic review and meta-analysis. *Psychotherapy Research*, *30*(2), 195–212. https://doi.org/10.1080/10503307.2018.1517949

Perkins, N. J., & Schisterman, E. F. (2006). The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *American Journal of Epidemiology*, *163*(7), 670–675. https://doi.org/10.1093/aje/kwj063

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*, 1–8. https://doi.org/10.1186/1471-2105-12-77

Robinson, L., Delgadillo, J., & Kellett, S. (2020). The dose-response effect in routinely delivered psychological therapies: A systematic review. *Psychotherapy Research*, *30*(1), 79–96. https://doi.org/10.1080/10503307.2019.1566676

Rubel, J., Lutz, W., Kopta, S. M., Köck, K., Minami, T., Zimmermann, D., & Saunders, S. M. (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. https://doi.org/10.1037/pas0000060

Sapyta, J., Riemer, M., & Bickman, L. (2005). Feedback to clinicians: Theory, research, and practice. *Journal of Clinical Psychology*, *61*(2), 145–153. https://doi.org/10.1002/jclp.20107

Schlagert, H., & Hiller, W. (2017a). Merkmale und prädiktiver Wert von früher Verschlechterung in der ambulanten Psychotherapie. *Zeitschrift für Klinische Psychologie und Psychotherapie*, *46*(1), 11–22. https://doi.org/10.1026/1616-3443/a000400

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

Sembill, A., Vocks, S., Kosfelder, J., & Schöttke, H. (2019). The phase model of psychotherapy outcome: Domain-specific trajectories of change in outpatient treatment. *Psychotherapy Research*, *29*(4), 541–552. https://doi.org/10.1080/10503307.2017.1405170

Spielmans, G. I., Masters, K. S., & Lambert, M. J. (2006). A comparison of rational versus empirical methods in the prediction of psychotherapy outcome. *Clinical Psychology & Psychotherapy*, *13*(3), 202–214. https://doi.org/10.1002/cpp.491

Stevenson, M., & Sergeant, E. (2022). *epiR: Tools for the analysis of epidemiological data*. https://cran.r-project.org/web/packages/epiR/index.html

Tharwat, A. (2021). Classification assessment methods. *Applied Computing and Informatics*, *17*(1), 168–192. https://doi.org/10.1016/j.aci.2018.08.003

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, *58*(1), 267–288.

Wise, E. A., Streiner, D. L., & Gallop, R. J. (2016). Predicting individual change during the course of treatment. *Psychotherapy Research*, *26*(5), 623–631. https://doi.org/10.1080/10503307.2015.1104421

Wittchen, H.-U., Wunderlich, U., Gruschwitz, S., & Zaudig, M. (1997). *SKID I. Strukturiertes Klinisches Interview für DSM-IV. Achse I: Psychische Störungen. Interviewheft und Beurteilungsheft.* Hogrefe.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32–35. https://doi.org/10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3