Available online at www.sciencedirect.com

ScienceDirect

Behavior Therapy 57 (2026) 234–249

Behavior Therapy

www.elsevier.com/locate/bt

# Early Improvement Predicts Treatment Response in Depression: An Ecological Momentary Assessment Study in an Inpatient and Day Clinic Setting

**Jeanette Tamm**
Max-Planck-Institute of Psychiatry, Munich
Ludwig-Maximilians-University of Munich

**Keisuke Takano**
Ludwig-Maximilians-University of Munich
Human Informatics and Interaction Research Institute,
National Institute of Advanced Industrial Science and Technology, Ibaraki, Japan

**Leah Just**
Max-Planck-Institute of Psychiatry, Munich

**Thomas Ehring**
Ludwig-Maximilians-University of Munich
DZPG (German Center for Mental Health), partner site Munich

**Tabea Rosenkranz**
Ludwig-Maximilians-University of Munich

**OPTIMA Study Group**

**Johannes Kopf-Beck**
Max-Planck-Institute of Psychiatry, Munich
Ludwig-Maximilians-University of Munich
DZPG (German Center for Mental Health), partner site Munich

We express our gratitude to all patients of the hospital of the Max Planck Institute of Psychiatry in Munich who participated in our study, as well as to the OPTIMA study group.
This study was sponsored by the Max Planck Institute of Psychiatry in Munich, Germany, where the authors JT and JKB were employed. The research publication was supported by LMUexcellent, funded by the Federal Ministry of Education and Research (BMBF) and the Free State of Bavaria under the Excellence Strategy of the Federal Government and the Länder. The funders did not influence the collection, management, analysis and interpretation of the data, or preparation, review, or approval of the manuscript.
The authors declare no conflicts of interest.
Ethical approval for the study protocol was obtained from the Institutional Ethical Committee of the Faculty of Medicine at LMU Munich (Project number 17–395). Privacy rights of human subjects have been observed and participants provided written informed consent prior to study assessments and randomization. Further details of the study procedures are described in the study protocol (Kopf-Beck et al., 2020). The dataset utilized in this study, comprising clinical data, is not publicly accessible to protect participants' rights to privacy and data protection, but will be provided by the corresponding author upon reasonable request. This also applies to the materials and the analysis code.
Address correspondence to Jeanette Tamm, Ph.D., Max Planck Institute of Psychiatry, Kraepelinstr. 2-10, 80804 Munich, Germany.
e-mail: jeanette_tamm@psych.mpg.de.

**Predicting treatment response through early improvement can reduce patients' time in ineffective treatments before considering alternatives. However, for psychological interventions, there is no consensus on what time window and improvement rate early in the treatment is the most informative for distinguishing treatment responders from nonresponders. This study investigated these aspects in an inpatient and day clinic setting among severe depressed patients who perceived intensive psychological treatment and compared Weekly Questionnaire Assessments (WQA) and Ecological Momentary Assessment (EMA) regarding their power to predict treatment response through early improvement. Fifty-two depressed patients were randomly assigned to one of three intensive 7-week psychological interventions (two individual and two group sessions per week) applied in an inpatient or day clinic setting. Early improvement was assessed three times daily using EMA and weekly using questionnaires (BDI-II). Linear Regression Models and Receiver Operating Characteristic Analyses were conducted to predict treatment response (BDI-II improvement from pre- to postintervention $\geq$50%) in patients who received a full course of treatment. Moreover, ratios of true negative/false negative predictions were calculated to explore the predictive value of different early improvement definitions: 10%, 20%, 30%, or 40% improvement after 1, 2, 3, or 4 treatment weeks. Both EMA and WQA significantly predicted treatment response after 3 weeks with AUC values of 73% (EMA) and 77% (WQA). A WQA-assessed 10% improvement after 4 weeks yielded the highest ratio of true negative/false negative predictions, with a true negative rate of 22% and a false negative rate of 0%. 10% improvement in depressive symptoms assessed with WQA after 3 to 4 weeks of treatment was the best predictor in our study. Further research is needed to validate the results. This trial design is registered with osf.io/9fuhn.**

*Keywords:* depression; early improvement; treatment prediction; ecological momentary assessment; cognitive behavioral therapy

IT IS WELL established that there are several effective psychological interventions for the treatment of depression (Cuijpers, 2015). Besides cognitive behavioral therapy (CBT), interventions such as schema therapy (ST) have proven to effectively reduce depression (Kopf-Beck et al., 2024). However, despite these advancements, meta-analytic findings show that almost 60% of depressed patients do not adequately respond to these treatments (Cuijpers, Karyotaki, et al., 2021). This is commonly defined as a reduction of depressive symptoms from baseline to the end of treatment by at least 50% (Rush et al., 2006). Combined treatments of psychotherapy and pharmacotherapy show better outcomes (Cuijpers, Quero, et al., 2021), but there is still room for improvement. This means, although different interventions yield similar average effects, treatment responses are highly variable on the individual level. Assuming that nonresponders of one treatment might respond to other interventions (Gloster et al., 2020; McKay et al., 2010), it is essential to develop decision rules that can guide clinicians in tailoring treatments to individual patients.

Besides efforts to develop personalized interventions by allocating patients to optimal treatments from the outset of therapy (DeRubeis et al., 2014; Huibers et al., 2015), it is essential to investigate whether and how nonresponse to an ongoing treatment can be predicted early on. Treatment prediction based on early improvement monitoring is a promising approach, as it may reduce the amount of time patients spend in ineffective treatment before considering alternatives. In this way, early treatment prediction may lead to better clinical outcomes (Schaffer et al., 2013) and help save scarce resources of the mental health care system (Richards, 2012). Additionally, it could be used in innovative trial designs such as the "leapfrog" method (Blackwell et al., 2019), which involves rapidly testing and modifying treatments based on early indicators of effectiveness.

Indeed, there is robust and replicated evidence that early improvement in therapy is a reliable prognostic indicator for treatment outcome in depression (Rubel et al., 2015). For pharmacological treatment, the absence of early improvement within the initial 2 weeks of therapy has been identified as a strong indicator of treatment nonresponse (Szegedi et al., 2009). Additionally, the absence of early improvement within the first 4 weeks has been established as a guideline for clinical decisions regarding medication change (Gautam et al., 2017). For psychological treatments, however, such guidelines are lacking. While compelling evidence suggests that early improvements can predict treatment outcomes in psychological interventions as well, studies exhibit significant heterogeneity and lack a standardized definition for predictive early improvement in depression (Beard & Delgadillo, 2019; Li et al., 2023). Importantly, earlier studies have used very different time windows, classified as "early," and different definitions for the rate of symptomatic change, classified as "improvement." For example, "early" time windows in earlier research could

encompass 2, 4, 6, or 8 weeks of treatment, and early improvement has been defined either using cut-offs for symptom reduction (e.g., >25%, Gois et al., 2014), the achievement of reliable or clinically significant improvement (Rubel et al., 2015), or the occurrence of sudden gains (Hunnicutt-Ferguson et al., 2012).

Moreover, to our knowledge, except for one online intervention study (Schibbye et al., 2014), none of the existing studies has used Ecological Momentary Assessment (EMA) as an alternative technique to monitor early improvement. However, EMA might be better suited for this purpose. It is well known that depressive symptoms naturally fluctuate during the day and from day to day (Chen et al., 2022; Takano & Tanno, 2011). Furthermore, the human recall of past affective experiences is biased, especially in depression (Colombo et al., 2020). Consequently, it is questionable whether a single retrospective point assessment can accurately and reliably capture the true symptom severity that patients experience over the past week or weeks. EMA may therefore assess change in depression more reliably, especially when observing short time windows (Moore et al., 2016; Tamm et al., 2024).

Thus, we assessed early improvement in depression comparing EMA and Weekly Questionnaire Assessment (WQA) and investigated their predictive value for identifying treatment responders and nonresponders across three different 7-week psychological interventions. In addition, we compared four different time windows (1, 2, 3, or 4 weeks after treatment initiation) and four different definitions of early improvements operationalized as symptomatic change rates (minimum improvement of 10%, 20%, 30% or 40%). In this way, we addressed the following research questions: (1a) At which time point after treatment initiation does early improvement in depressive symptoms significantly predict treatment response? (1b) Do both WQA and EMA of early improvement significantly predict treatment response at these time points? (2) How predictive are different definitions of early improvement in terms of the defined time window and symptom cutoff, as assessed by EMA versus WQA? While other authors aimed to maximize the sum of sensitivity (rate of correctly identified responders, i.e., true positive rate) and specificity (rate of correctly identified nonresponders, i.e., true negative rate) using the Youden's-index (Crits-Christoph et al., 2001), we took a different approach. We decided to evaluate our predictors based on a reverse version of the negative likelihood ratio (Bolin & Lam, 2013). This approach takes the following critical implications for clinical decision-making into account: we defined the primary goal of treatment prediction through early improvement as identifying a high number of nonresponders, in order to consider alternative interventions early in treatment. The primary goal is therefore to achieve a high true negative rate. At the same time, it is particularly important to avoid the disruption of effective treatments, which means targeting a low false negative rate. Although erroneously recommending the continuation of a treatment to a nonresponder would also be unbeneficial, we deemed these scenarios less critical since they reflect the current clinical reality without decision rules anyway. We therefore decided that our best definition of early improvement would be the highest ratio between the true negative rate (i.e., specificity) and the false negative rate (i.e., 1-sensitivity rate), which is a reverse of the negative likelihood ratio described in the literature (Bolin & Lam, 2013).

We hypothesized that (1a) early improvement in depressive symptoms predicts treatment response within the first 4 treatment weeks, (1b) respectively of whether the early improvement is assessed with EMA or WQA. Research question (2) investigating the predictive value of different definitions of early improvement was kept exploratory. For our analyses we utilized data from the OPTIMA study (Kopf-Beck et al., 2024), a clinical trial investigating the effectiveness of three psychological interventions, cognitive behavioral therapy (CBT), schema therapy (ST), and individual supportive therapy (IST). The study included moderately to severely depressed patients treated in an inpatient or day clinic setting.

## Materials and Methods

### PARTICIPANTS

The sample size of the study was designed for its primary analysis on treatment effects between intervention conditions assessed by EMA, which is described elsewhere (Tamm et al., 2024). Here, we present a secondary analysis of the data. For the primary analysis, a required sample size around 20 per condition, i.e., $n = 60$ patients in total, was targeted.

The inclusion criteria were an age between 18 and 65 years, a diagnosis of a major depressive disorder, single episode or recurrent, moderate or severe without psychotic symptoms diagnosed by clinical assessment. Additionally, patients had to own a smartphone and agree to the EMA. Participants were compensated according to their EMA response rate.

Exclusion criteria included the presence of psychotic symptoms in depression (F32.3, F33.3), acute suicidality, a lifetime history of any psychotic or bipolar disorder, an IQ below 80 and/or severe learning disabilities, severe neurological or internal diseases, current withdrawal syndromes from alcohol or illicit drugs, concomitant organic mental disorders, severe mutism or stupor, mental disorders secondary to medical conditions or substance use disorder, pregnancy or lactation, and language barriers to psychotherapy. Other co-occurring mental health conditions were not exclusionary and are reported in Supplementary Table 3.

Dropouts were defined as enrolled patients of whom incorrect in- or exclusion criteria emerged during the study, who left the clinic before completing the intervention, or missed more than six (22%) intervention sessions. Dropouts were excluded from analyses due to incomplete data for the calculation of their responder status.

### DESIGN AND PROCEDURES

The data was gathered at the Max Planck Institute of Psychiatry in Munich, Germany, within the OPTIMA study (Identifier on clinicaltrials.gov: NCT03287362). The OPTIMA study is a monocentric, prospective, rater-blinded, parallel-group, block-randomized clinical trial incorporating repeated measures and three distinct intervention conditions (CBT, ST, and IST). Ethical approval for the study protocol was obtained from the Institutional Ethical Committee of the Faculty of Medicine at LMU Munich (Project number 17–395). Participants provided written informed consent prior to study assessments and randomization. Further details of the study procedures are described in the study protocol (Kopf-Beck et al., 2020).

### INTERVENTIONS

Participants of the OPTIMA study were randomly assigned to CBT, schema therapy (ST), or individual supportive therapy (IST). Each intervention condition lasted 7 weeks and comprised two single (50 minutes each) and two group sessions (100 minutes each) per week. CBT was based on Beck's cognitive model of depression and followed evidence-based treatment guidelines (Hautzinger, 2013). The ST manual (Egli et al., 2019) integrated cognitive, psychodynamic, gestalt, and experiential techniques to address underlying maladaptive schemas. IST served as an active control condition, focusing on common therapeutic factors within a biopsychosocial framework without applying disorder-specific techniques (Markowitz, 2022). Therapists were clinical psychologists or psychiatrists who received standardized treatment-specific training. Fidelity was ensured through monthly treatment-specific expert supervision based on video recordings of therapy sessions. Adherence ratings of a random subsample of sessions showed high to very high treatment adherence and significant differences between treatment arms regarding the execution of specific interventions. Detailed descriptions of the interventions, therapist training, and fidelity procedures are provided in the study protocol and the primary analysis of the OPTIMA trial (Kopf-Beck et al., 2020; Kopf-Beck et al., 2024).

It is important to note that concurrent pharmacotherapy or additional treatments that are common in inpatient or day clinic settings, such as occupational therapy or case management, were not regulated. Decisions on this were at the discretion of the attending psychiatrist but were meticulously documented as control variables. Concomitant care conditions showed no significant effects on treatment differences between intervention conditions in the primary analysis of the OPTIMA study (Kopf-Beck et al., 2024).

### MEASURES

Details about all measures performed in the OPTIMA study, including a range of questionnaires, clinical interviews, imaging, and tests, are provided in the study protocol (Kopf-Beck et al., 2020). Here, we report only the measures included in our analysis, which encompass two primary outcome variables: a questionnaire and an EMA score assessing depressive symptoms. An assessment plan is provided in Supplementary Table 1.

#### EMA

EMA was administered continuously throughout the entire intervention period, starting immediately after patients' enrollment in the study. It involved three prompts per day and was signal-contingent, meaning patients were automatically prompted by their device. During the app onboarding process, patients reported their individual approximate wake-up times. Based on these times, each day was divided into three phases: morning = 2 hours before to 5 hours after wake-up time, noon = 5 to 10 hours after wake-up time, and evening = 10 hours after to 2 hours before wake-up time. Within each phase, one EMA prompt was emitted with a semirandomized reminder (signal). Prompts could only be completed within their assigned phase. Average times between patients' responses were calculated: M = 304.8 min ($SD$ = 94.47 min) from morning

to noon, M = 339.5 min (*SD* = 106.51 min) from noon to evening and M = 809.3 min (*SD* = 109.63 min) from evening to morning, showing no temporal clustering of responses. To counteract sequence effects, the order of EMA items was randomized across prompts. This protocol, including baseline, resulted in a total of 168 EMA prompts (56 days * 3 prompts per day).

The EMA score of depressive symptoms is a sum score of four EMA items, formulated by JT and JKB in collaboration with clinicians and aligning with diagnostic criteria for major depression (ICD-10). The items encompass three core symptoms of depression (loss of interest, withdrawal, and psychomotor agitation/inhibition) and current mood (the item wordings are provided in Supplementary Table 2). The response scale of the items "loss of interest," "withdrawal," and "psychomotor agitation/inhibition" was two-tiered: Initially, participants responded to a binary *Yes-No* scale. If *Yes* was chosen, participants provided further feedback using a 5-point Likert scale, gauging their level of agreement (*not at all, a bit, moderately, considerably, very much*). Conversely, if *No* was selected, the Likert scale was omitted, and the subsequent item followed. The 'mood' item was rated by selecting one of five emojis (*very good, good, moderate, bad, very bad*) that best represented their current mood.

The internal reliability of the EMA total score was assessed within the primary analysis of the data (Tamm et al., 2024) using the "multilevel.reliability()" function from the "psych" package in R, showing good within-participant reliability (Rc = 0.79) and excellent between-participant reliability (RkF = 0.91). Rc reflects the consistency of repeated measurements within individuals over time, indicating reliable detection of symptom changes, while RkF quantifies the reliability of average scores between participants, reflecting differentiation between individuals. These indices were chosen because they capture distinct and complementary aspects of reliability critical for interpreting our EMA data and associated statistical power.

### *Weekly Questionnaire Assessments*

The corresponding questionnaire scores of depressive symptoms were assessed weekly with the German version of the Beck Depression Inventory II (BDI-II; Hautzinger et al., 2009), resulting in a total of 8 assessment points, including baseline.

## DEFINITION OF THE INDEPENDENT VARIABLE EARLY IMPROVEMENT

We explored the predictive potential of four distinct time windows and four distinct change rates of early improvement in depressive symptoms: a minimum improvement of 10%, 20%, 30% or 40% by Week 1, 2, 3, or 4 after treatment initiation assessed with EMA or WQA. As stated before, the EMA score was derived by aggregating four EMA items, while the WQA score was the BDI-II sum score. The four time windows were selected based on previous studies of early improvement, reviewed by Beard and Delgadillo (2019). Prior research focused on the 2-week and 4-week time windows. We decided against time windows longer than 4 weeks, as Rubel et al. (2015) have shown that most change in patients' progress occurs by the third treatment session, which corresponds to a treatment duration of less than 1 week in our study setup. Similar observations derive from pharmacological studies (Schaffer et al., 2013). In recognizing the ecological benefits of "early" treatment prediction in guiding clinical practice, we expanded our investigation to include Weeks 1 and 3, alongside Weeks 2 and 4, aiming to shed light on the earliest point at which early improvements can predict psychological treatment outcomes. The symptom cutoffs were chosen based on considerations that they should be smaller than the treatment response rate (50% improvement), reflect change rates similar to those investigated by other studies (e.g., Gois et al., 2014), and have equal intervals between each cutoff.

## DEFINITION OF THE DEPENDENT VARIABLE TREATMENT RESPONSE

In line with other studies (Keller, 2003; Rush et al., 2006), treatment response was defined as a reduction of greater than or equal to 50% in the BDI-II total score from baseline to the end of the intervention. For the definition of treatment response, we also considered other types of outcomes, e.g., quality of life. However, we ultimately decided against including them as global mental health indices such as quality of life do not change as markedly within a short duration of psychotherapy (Trivedi et al., 2006).

## STATISTICAL ANALYSES

Data was analyzed using R, version 4.2.2 (R Core Team, 2020). Patients with insufficient EMA data to calculate early improvement were excluded from the analyses. To analyze a change rate in the first treatment week, at least two data points are needed. For analyzing change rates over 2, 3 or 4 weeks, at least one additional data point per subsequent week is needed. Therefore, patients with fewer than two EMA prompts in the first week or fewer than one EMA prompt in the

second, third or fourth week were excluded. For plausibility, we examined the person-level standard deviations for each EMA item throughout the trial period and excluded the two patients exhibiting a standard deviation of zero in at least one EMA item following previous research (Rosenkranz et al., 2020). For all analyses, we defined significance as $p = 0.05$.

*Research Question 1a and 1b*

We hypothesized that (1a) early improvement in depressive symptoms would predict treatment response within the first 4 treatment weeks, (1b) regardless of whether early improvement is assessed with EMA or WQA. To test these hypotheses, we conducted a multistep analysis using linear regression models (LRMs) and logistic regression models. For each participant, we fitted an LRM with time predicting EMA-assessed or WQA-assessed depressive symptoms. The models were estimated for each of the four time-window conditions (i.e., 1, 2, 3, and 4 weeks)—therefore, we obtained four pairs of individual participants' estimates of the intercept and slope (of time), each for EMA and WQA.

In the second step, an Improvement Rate Score was calculated for each patient and each time-window condition. The Improvement Rate was operationalized as the ratio of symptom change over a given time window (slope * time window) being divided by the baseline symptom level (intercept).

Lastly, we conducted binary logistic regressions to predict responder status (responder vs. nonresponder) after 7 weeks of treatment through individual baseline BDI-II and the Improvement Rate Score as independent variables.

*Research Question 2*

To determine the "best" definition of early improvement, considering both the designated time window and the symptom cutoff, we first performed Receiver Operating Characteristic (ROC) analysis for each variable (EMA or WQA) and time window. The ROC analysis is a method commonly used to evaluate the diagnostic performance of a test or model by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity) across different threshold values. The Area Under the Curve (AUC) derived from the ROC curve serves as a summary measure of discriminative ability, where higher AUC values indicate better discrimination between responder and nonresponder groups (Hanley & McNeil, 1982). To evaluate significant differences between the AUCs of the ROC curves, model comparison analyses were employed using the DeLong's test (DeLong et al., 1988).

Next, we analyzed the true negative rate (specificity), false negative rate (1-sensitivity), and the ratio between the two for each predefined combination of variable (EMA or WQA), time window (1, 2, 3, or 4 weeks) and symptom cutoff (minimum improvement of 10%, 20%, 30%, or 40%). In the literature, the inverted form of this ratio ((1-sensitivity)/specificity) is described as the "negative likelihood ratio (LR-)," displaying the likelihood of a patient testing negative while having a disease (false negative rate) divided by the likelihood of a patient testing negative while not having a disease (true negative rate; Bolin & Lam, 2013). For our research, however, a reverse LR- (i.e. specificity/(1 – sensitivity)), which we called the "TNFN" (true negative to false negative) ratio for clarity is better suited and easier to interpret. In our context specificity (the true negative rate) denotes the rate of nonresponders correctly identified, while 1-Sensitivity (the false negative rate) denotes the rate of responders falsely classified as nonresponders. The TNFN ratio therefore gives us the likelihood of a patient having no early improvement and who does not respond to treatment (true negative rate) divided by the likelihood of a patient having no early improvement but who does respond to treatment (false negative rate). Defined as such, the highest TNFN ratio displays our "best" predictor—that is, our "best" definition of early improvement. To determine whether the TNFN ratio results in more true negative than false negative predictions given the responder rate in our sample, we also calculated a weighted TNFN ratio (wTNFN ratio) by multiplying the TNFN ratio with the ratio of the nonresponder to responder rate: wTNFN ratio (RR) = TNFN ratio * ((1 – RR)/RR). Note that the responder rate (RR) must lie between zero and one. A higher number of true negative to false negative predictions is indicated by wTNFN > 1.

Finally, to improve the comparability of our results to other studies, we also calculated Youden's indices (Youden, 1950), which serve as composite measures taking the sum of specificity (true positive rate) and sensitivity (true negative rate) into account (Youden's index = sensitivity + specificity – 1).

## Results

### SAMPLE DESCRIPTION

Patients were recruited between August 2019 and December 2020. Initially, 137 patients were evaluated for eligibility, with 106 meeting the inclusion criteria, expressing willingness to engage in EMA,

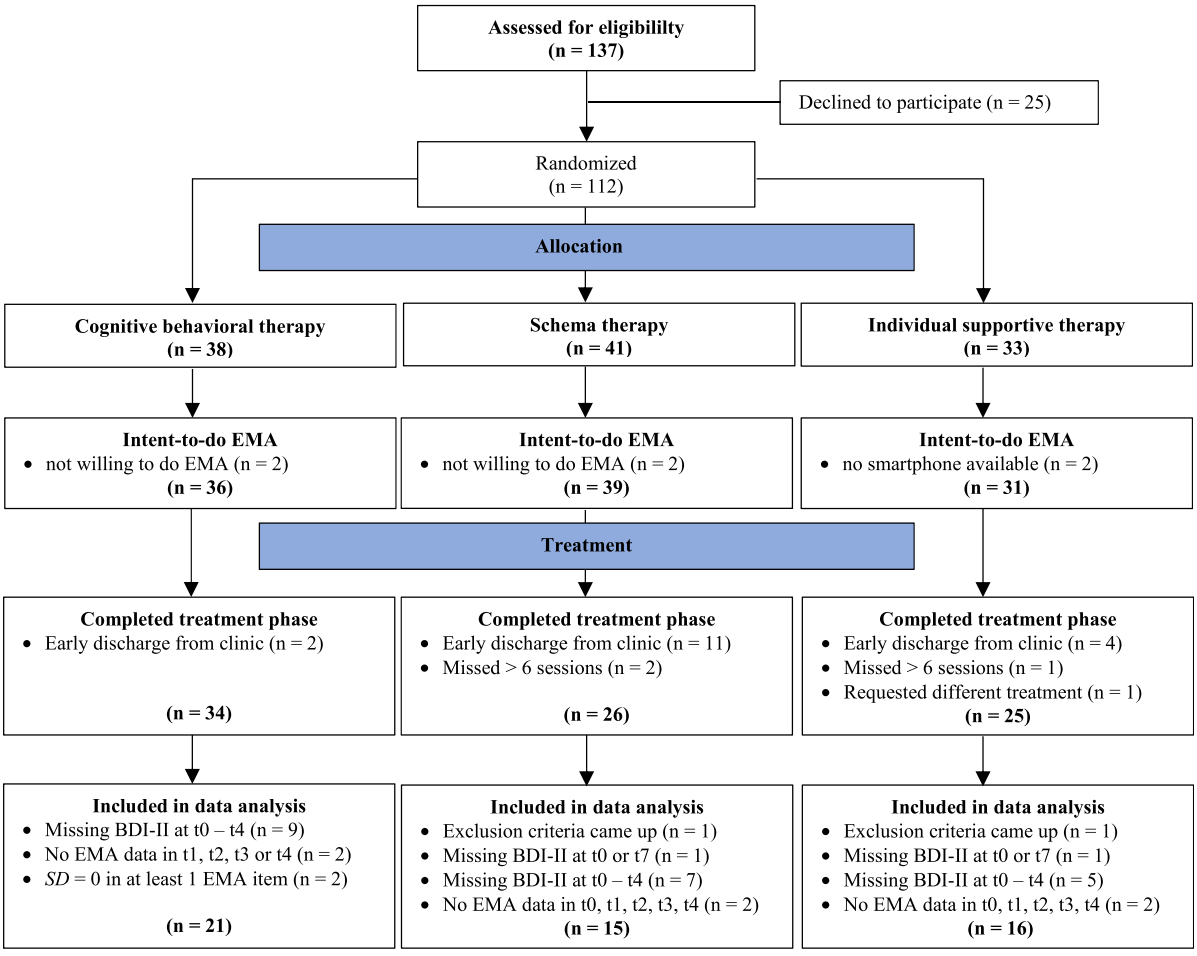


FIGURE 1 Flow Diagram. *Note*. EMA: Ecological Momentary Assessment; BDI-II; Beck's Depression Inventory II.

and being randomly assigned to the intervention conditions (ST: 39, CBT: 36, IST: 31). Details are provided in the CONSORT flow chart in Figure 1. Reasons for dropout included early discharge from the clinic prior to completing the intervention trial (17), missing more than six intervention sessions (3), and requesting a different intervention (1). Before conducting analyses, we further excluded data from patients due to ineligible diagnoses identified during the study (2), missing data in the BDI-II baseline or postintervention assessment (2), missing BDI-II data in the first 4 intervention weeks (11), missing EMA data in the first 4 intervention weeks (6), or lack of data integrity in the EMA (2). Finally, the statistical analyses included the data of $N = 52$ patients.

As shown in Table 1, patient's age in the analyzed cohort was M = 40.5 years ($SD = 11.71$), 57.69% were female ($n = 30$), 84.62% were German ($n = 44$), and 51.92% had qualifications for university entrance ($n = 27$). Patients exhibited severe levels of depression on average at baseline (BDI-II: $M = 32.58$, $SD = 8.5$). Patients' EMA response rate in the analyzed sample was M = 64.67% ($SD = 23.35\%$). Further descriptive statistics such as comorbidities are provided in Supplementary Table 3. Distributions of the EMA response rates across intervention weeks are provided in Supplementary Figure 1.

### RESEARCH QUESTIONS 1A AND 1B

To test whether early improvement predicts treatment response within the first 4 treatment weeks regardless of whether early improvement is assessed with EMA or WQA, we fitted logistic regression models, predicting treatment response by baseline depression and early improvement rates in depressive symptoms assessed via EMA versus WQA. Intervention condition was not a significant covariate in our models and therefore not included in further analyses. For each variable (EMA and WQA), we ran four different models denoting the four time windows (Table 2). For both EMA and WQA, early improvement rates of depressive symptoms significantly predicted treatment responder status not after 1 or 2 weeks

Table 1
Descriptive Statistics of the Sample

| Characteristic | Total (*N* = 52) | | Responder Status: Responder (*N* = 25) | | Responder Status: Nonresponder (*N* = 27) | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | N | % | N | % | N | % | $Chi^2$ | df | *p* |
| **Intervention condition** | | | | | | | 1.57 | 2 | 0.455 |
| CBT | 21 | 40.38 | 10 | 40.00 | 11 | 40.74 | | | |
| ST | 15 | 28.85 | 9 | 36.00 | 6 | 22.22 | | | |
| IST | 16 | 30.77 | 6 | 24.00 | 10 | 37.04 | | | |
| **Gender** (female) | 30 | 57.69 | 15 | 60.00 | 15 | 55.56 | <0.00 | 1 | 0.966 |
| **Nationality** (German) | 44 | 84.62 | 22 | 88.00 | 22 | 81.48 | 1.07 | 2 | 0.586 |
| **Qualification for University entrance** | 27 | 51.92 | 14 | 56.00 | 13 | 48.15 | 2.96 | 2 | 0.227 |
| **Income** | | | | | | | 6.21 | 3 | 0.102 |
| Low income | 20 | 38.46 | 8 | 32.00 | 12 | 44.44 | | | |
| Middle income | 19 | 36.54 | 11 | 44.00 | 8 | 29.63 | | | |
| High income | 9 | 17.31 | 6 | 24.00 | 3 | 11.11 | | | |
| not specified | 4 | 7.69 | 0 | 0.00 | 0 | 0.00 | | | |
| | **Mean** | **SD** | **Mean** | **SD** | **Mean** | **SD** | **t or W** | **df** | ***p*** |
| **Age (years)** | 40.5 | 11.71 | 43.12 | 12.04 | 38.07 | 11.06 | −1.57 | 48.70 | 0.123 |
| **EMA response rate** | 64.67 | 23.35 | 67.84 | 20.44 | 61.73 | 25.80 | −0.95 | 48.87 | 0.347 |
| **Baseline Symptoms (BDI-II)** | 32.58 | 8.50 | 32.84 | 8.42 | 32.33 | 8.73 | −0.21 | 49.91 | 0.832 |

*Note.* ST = Schema Therapy; CBT = Cognitive Behavioral Therapy; IST = Individual Supportive Therapy; BDI-II = Beck's Depression Inventory II; EMA = Ecological Momentary Assessment.

of treatment (1 week: EMA: z = −0.38, 95% CI = [−1.79 to 1.14], $p = 0.704$; WQA: z = −1.67, 95% CI = [−7.14 to 0.15], $p = 0.095$; 2 weeks: EMA: z = −0.84, 95% CI = [−1.59 to 0.52], $p = 0.399$; WQA: z = −1.87, 95% CI = [−5.83 to −0.15], $p = 0.062$), but after 3 and 4 weeks of treatment (3 weeks: EMA: z = −2.33, 95% CI = [−4.49 to −0.51], $p = 0.02$; WQA: z = −2.32, 95% CI = [−7.65 to −0.94], $p = 0.02$; 4 weeks: EMA: z = −2.43, 95% CI = [−4.09 to −0.62], $p = 0.015$; WQA: z = −3.11, 95% CI = [−10.53 to −2.7], $p = 0.002$). These results indicate that early improvement in depressive symptoms, as assessed by both EMA and WQA, is a significant predictor of treatment response after three to four weeks of treatment. Supplementary Figure 2 shows the improvement rate distributions of the four time windows separately for responders and nonresponders and the two variables EMA and WQA.

### RESEARCH QUESTION 2

Next, we computed ROC-analyses and calculated the AUC values of each combination of variable (EMA versus WQA) and time window (1, 2, 3, or 4 weeks). All four time windows yielded AUC values exceeding 50%, indicating predictions better than chance on average (Figure 2). The highest AUC value was reached by WQA after 4 weeks of treatment (AUC = 0.77), which was slightly but not significantly larger than its AUC value after 3 weeks of treatment (AUC = 0.7; p = 0.143, 95% CI = [−0.17 to 0.02]) or the AUC value of EMA after 4 weeks of treatment (EMA: AUC = 0.73; $p = 0.59$, 95% CI = [−0.19 to 0.11]). In summary, these results indicate that both WQA and EMA provide comparable predictions, which are better than chance and become particularly significant after 3 to 4 weeks of treatment. For the sake of completeness and comparability, comparison analyses between all AUC values and the maximum Youden's-index of each time-window are reported in Supplementary Table 4.

Finally, we explored which combination of time window and symptom cutoff yields the optimal definition of early improvement. Figure 3 shows the distribution of true positive, true negative, false negative and false positive predictions for each definition. Additionally, for each definition we calculated (Table 3): (a) the sensitivity, (b) the specificity, (c) the Youden-index (y), i.e., the percentage of correct predictions, (d) our predefinied TNFN ratio, i.e., the ratio between the true negative to false negative rate (specificity/(1-sensitivity)), and (e) the weighted TNFN ratios (wTNFN), taking the responder rate found in our sample of RR = 48% into account (wTNFN ratio (RR) = TNFN ratio*((1 − RR)/RR)). All but two definitions of early improvement yielded pre-

Table 2
Logistic Regression Models Predicting Treatment Response Based on Early Improvement Rates Measured Within Four Different Time Windows With EMA or WQA

| DV/predictors | Estimates | SE | z (*p*) | 95% CI | AIC | BIC |
|---|---|---|---|---|---|---|
| **Early Improvement in EMA – Depressive symptoms – 1 week** | | | | | | |
| Intercept | −0.26 | 1.12 | −0.23 (0.817) | [−2.52 to 1.95] | 77.82 | 83.67 |
| Baseline BDI | 0.01 | 0.03 | 0.17 (0.866) | [−0.06 to 0.07] | | |
| Early Improvement | −0.27 | 0.72 | −0.38 (0.704) | [−1.79 to 1.14] | | |
| **Early Improvement in EMA – Depressive symptoms – 2 week** | | | | | | |
| Intercept | −0.16 | 1.13 | −0.15 (0.884) | [−2.43 to 2.07] | 77.19 | 83.05 |
| Baseline BDI | 0 | 0.03 | 0.09 (0.93) | [−0.06 to 0.07] | | |
| Early Improvement | −0.43 | 0.51 | −0.84 (0.399) | [−1.59 to 0.52] | | |
| **Early Improvement in EMA – Depressive symptoms – 3 week** | | | | | | |
| Intercept | −0.14 | 1.17 | −0.12 (0.906) | [−2.5 to 2.19] | 71.52 | 77.37 |
| Baseline BDI | −0.01 | 0.04 | −0.15 (0.883) | [−0.08 to 0.06] | | |
| Early Improvement | −2.34 | 1 | −2.33 (0.02) | [−4.49 to −0.51] | | |
| **Early Improvement in EMA – Depressive symptoms – 4 week** | | | | | | |
| Intercept | 0.51 | 1.22 | 0.42 (0.678) | [−1.91 to 2.97] | 69.49 | 75.34 |
| Baseline BDI | −0.02 | 0.04 | −0.62 (0.537) | [−0.1 to 0.05] | | |
| Early Improvement | −2.12 | 0.87 | −2.43 (0.015) | [−4.09 to −0.62] | | |
| **Early Improvement in WQA – BDI – 1 week** | | | | | | |
| Intercept | 0.3 | 1.24 | 0.24 (0.807) | [−2.15 to 2.77] | 74.49 | 80.35 |
| Baseline BDI | −0.02 | 0.04 | −0.52 (0.602) | [−0.1 to 0.05] | | |
| Early Improvement | −3.05 | 1.83 | −1.67 (0.095) | [−7.14 to 0.15] | | |
| **Early Improvement in WQA – BDI – 2 week** | | | | | | |
| Intercept | 0.04 | 1.24 | 0.03 (0.977) | [−2.43 to 2.5] | 73.6 | 79.46 |
| Baseline BDI | −0.02 | 0.04 | −0.46 (0.643) | [−0.09 to 0.06] | | |
| Early Improvement | −2.67 | 1.43 | −1.87 (0.062) | [−5.83 to −0.15] | | |
| **Early Improvement in WQA – BDI – 3 week** | | | | | | |
| Intercept | −0.61 | 1.22 | −0.5 (0.615) | [−3.12 to 1.76] | 70.89 | 76.75 |
| Baseline BDI | −0.01 | 0.04 | −0.37 (0.709) | [−0.09 to 0.06] | | |
| Early Improvement | −3.94 | 1.7 | −2.32 (0.02) | [−7.65 to −0.94] | | |
| **Early Improvement in WQA – BDI – 4 week** | | | | | | |
| Intercept | −1.92 | 1.37 | −1.4 (0.161) | [−4.8 to 0.71] | 62.93 | 68.78 |
| Baseline BDI | 0 | 0.04 | −0.11 (0.911) | [−0.08 to 0.07] | | |
| Early Improvement | −6.11 | 1.97 | −3.11 (0.002) | [−10.53 to −2.7] | | |

*Note.* EMA = Ecological Momentary Assessment; WQA= Weekly questionnaire assessment; Sample size in all models is 52. Early Improvement is defined as Improvement Rate.

dictions better than chance ($y > 0$) and wTNFN ratios >1, indicating higher numbers of true negative to false negative predictions. The optimal definition of early improvement was a WQA-assessed 10% improvement by Week 4, reaching a true negative rate (specificity) of 22% in combination to a false negative rate of 0%. Accordingly, the TNFN ratio and wTNFN ratio resulted in infinite values. Specifically, this definition of early improvement resulted in 48% true positive ($n = 25$), 12% true negative ($n = 6$), 0% false negative ($n = 0$) and 40% false positive ($n = 21$) predictions in our sample (Table 3).

## Discussion

The aim of this study was to investigate whether early improvements in depressive symptoms measured by both EMA and WQA could significantly predict treatment response versus nonresponse and which specific definitions of early improvement may be most useful to support clinical decisions.

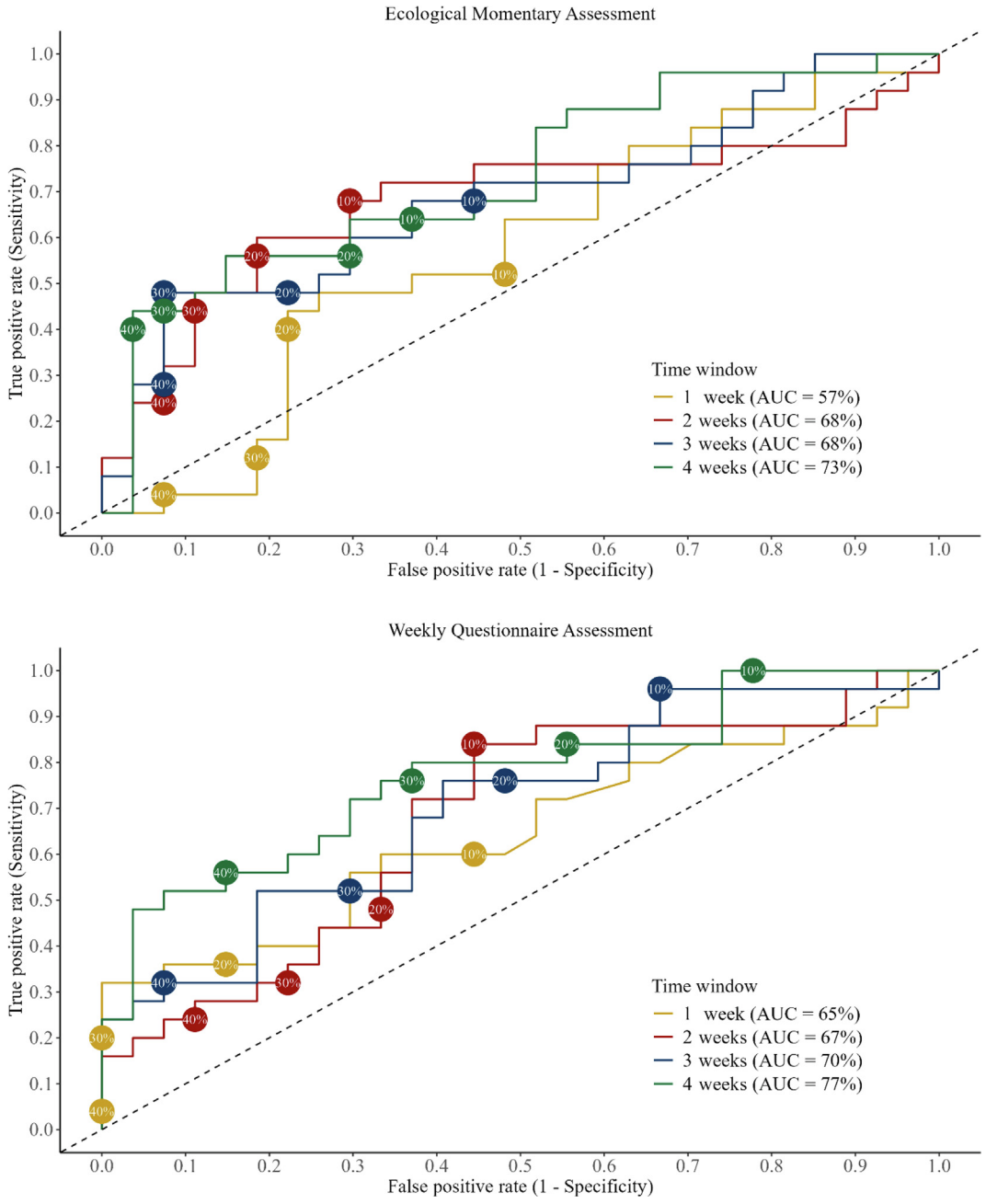


**FIGURE 2** Receiver Operating Characteristic (ROC) plots for the prediction of treatment response after 7 weeks through early improvement in depressive symptoms measured by ecological momentary assessment or weekly questionnaire assessment within four different time windows. *Note*. $n = 52$. The ROC curves illustrate the prediction accuracy of early improvement measured with EMA and WQA within four distinct time windows: first treatment week = yellow, first two weeks = red, first three weeks = blue, and first four weeks = green. Dots mark specifically investigated rates of early improvement (≥10%, 20%, 30%, and 40%).

### TREATMENT PREDICTION THROUGH EMA AND WQA

Confirming our first hypotheses, we found that early improvements assessed by both EMA and WQA serve as robust predictors of treatment response. Three weeks after treatment initiation, both assessment techniques reached significant predictions of responder status. This corresponds to a treatment dose of 12 sessions (2 individual and 2 group sessions per week) and about halftime in our study setup. We observed that the prediction accuracies of EMA and WQA did not significantly differ from each other. This underlines the predictive power of EMA. Especially notable is


FIGURE 3 Rates of false and correct predictions of treatment response based on different definitions of early improvement. *Note*. $n = 52$. Following definitions of early improvement were examined: an improvement in depressive symptoms of ≥10%, 20%, 30% or 40% after 1, 2, 3 or 4 weeks of treatment as measured by EMA or WQA. All but one definition predict treatment response better than chance (dashed line). For clinical decision-making it is particularly important that the number of true negatives (light green) is high and, above all, higher than the number of false negatives (light red).

that the early improvement assessed via WQA and the outcome assessment were homogeneous in measurement (both used the BDI-II), whereas the EMA assessment differed from the outcome assessment in both general methodology (EMA vs. retrospective assessment) as well as specific items used.

Our results align with previous studies showing that early improvement after 4 weeks of treatment is predictive (Beard & Delgadillo, 2019) of treatment response. However, given the intensity and type of the comprehensive psychiatric care program, including two individual and two group sessions of psychotherapy per week, it would have also been reasonable to expect significant predictions as early as 1 or 2 weeks into treatment. Studies on early change patterns show that most change in patients' progress occurs early in treatment, e.g., within the first four sessions (Rubel et al., 2015), which corresponds to a treatment duration of 1 week in our study setup. Besides, studies on dose-effect relationships show that frequent treatment schedules, like in our study (two group and two single sessions) are more effective for the treatment of depression than schedules with only one session per week (Cuijpers et al., 2013). However, although early improvement within the initial weeks of treatment is common, this does not necessarily mean that it is also predictive for the final treatment outcome. Such early changes could be influenced by nonspecific factors related to study participation, such as expectation biases (Targum, 2020). Besides, Lutz et al. (2017) showed that the pattern of patients' early improvement adds beneficial information to the simple measure of "how much" they improve. Specifically, they found three different patterns of early changes within the initial 6 weeks of a web-based intervention study for the treatment of depression. The trial started with a screening at baseline and a registration after 2 weeks, followed by 4 weeks of intervention. The three patterns found were: continuous improvement starting directly after screening, improvement starting after registration (i.e., with a 2-week delay) and linear

Table 3
Prediction Metrics of Different Definitions of Early Improvement in Depressive Symptoms Measured by EMA or WQA on Treatment Response to a 7-Week Psychological Treatment

| | (Specificity, Sensitivity, Youden-Index), TNFN, wTNFN, | | | |
|---|---|---|---|---|
| | **10% improvement** | **20% improvement** | **30% improvement** | **40% improvement** |
| **Early Improvement assessed with Ecological Momentary Assessment** | | | | |
| **1 week** | (52%, 52%, 0.04)<br>TNFN = **1.08**<br>wTNFN = 1.17 | (40%, 78%, 0.18)<br>TNFN = **1.3**<br>wTNFN = 1.4 | (12%, 81%, -0.07)<br>TNFN = **0.93**<br>wTNFN = 1 | (4%, 93%, -0.03)<br>TNFN = **0.96**<br>wTNFN = 1.04 |
| **2 weeks** | (68%, 70%, 0.38)<br>TNFN = **2.2**<br>wTNFN = 2.38 | (56%, 81%, 0.37)<br>TNFN = **1.85**<br>wTNFN = 2 | (44%, 89%, 0.33)<br>TNFN = **1.59**<br>wTNFN = 1.71 | (24%, 93%, 0.17)<br>TNFN = **1.22**<br>wTNFN = 1.32 |
| **3 weeks*** | (68%, 56%, 0.24)<br>TNFN = **1.74**<br>wTNFN = 1.88 | (48%, 78%, 0.26)<br>TNFN = **1.5**<br>wTNFN = 1.62 | (48%, 93%, 0.41)<br>TNFN = **1.78**<br>wTNFN = 1.92 | (28%, 93%, 0.21)<br>TNFN = **1.29**<br>wTNFN = 1.39 |
| **4 weeks*** | (64%, 63%, 0.27)<br>TNFN = **1.75**<br>wTNFN = 1.89 | (56%, 70%, 0.26)<br>TNFN = **1.6**<br>wTNFN = 1.73 | (44%, 93%, 0.37)<br>TNFN = **1.65**<br>wTNFN = 1.79 | (40%, 96%, 0.36)<br>TNFN = **1.6**<br>wTNFN = 1.73 |
| **Early Improvement assessed with Weekly Questionnaire Assessment** | | | | |
| **1 week** | (60%, 56%, 0.16)<br>TNFN = **1.39**<br>wTNFN = 1.5 | (36%, 85%, 0.21)<br>TNFN = **1.33**<br>wTNFN = 1.44 | (20%, 100%, 0.2)<br>TNFN = **1.25**<br>wTNFN = 1.35 | (4%, 100%, 0.04)<br>TNFN = **1.04**<br>wTNFN = 1.13 |
| **2 weeks** | (84%, 56%, 0.4)<br>TNFN = **3.47**<br>wTNFN = 3.75 | (48%, 67%, 0.15)<br>TNFN = **1.28**<br>wTNFN = 1.38 | (32%, 78%, 0.1)<br>TNFN = **1.14**<br>wTNFN = 1.24 | (24%, 89%, 0.13)<br>TNFN = **1.17**<br>wTNFN = 1.26 |
| **3 weeks*** | (96%, 33%, 0.29)<br>TNFN = **8.33**<br>wTNFN = 9 | (76%, 52%, 0.28)<br>TNFN = **2.16**<br>wTNFN = 2.33 | (52%, 70%, 0.22)<br>TNFN = **1.47**<br>wTNFN = 1.58 | (32%, 93%, 0.25)<br>TNFN = **1.36**<br>wTNFN = 1.47 |
| **4 weeks*** | (100%, 22%, 0.22)<br>TNFN = **∞**<br>wTNFN = ∞ | (84%, 44%, 0.28)<br>TNFN = **2.78**<br>wTNFN = 3 | (76%, 63%, 0.39)<br>TNFN = **2.62**<br>wTNFN = 2.83 | (56%, 85%, 0.41)<br>TNFN = **1.94**<br>wTNFN = 2.09 |

*Note.* $n = 52$; The table compares the following definitions of early improvement in predicting treatment response: ≥10%, 20%, 30%, or 40% improvement after 1, 2, 3, or 4 weeks of treatment, measured with Ecological Momentary Assessment or Weekly Questionnaire Assessment. In brackets the metrics sensitivity, specificity, and Youden's-index are reported. TNFN denotes the ratio of the true negative to false negative rate (specificity/1 − sensitivity). wTNFN denotes the weighted TNFN ratio taking the responder rate (RR) of the investigated sample, which was 48% into account (TNFN * ((1 − RR)/RR)). Asterisks (*) highlight time windows that significantly predicted treatment response (for details see Table 2). The cell colors indicate the size of the TNFN ratios (bold) in relation to each other reaching from light green (lower values) to dark green (higher values).

deterioration starting directly after screening. Interestingly, the delayed improvement was the most accurate predictor of treatment outcome, followed by the direct improvement. Therefore, the methodological approach of our study, in which we estimated change rates using linear regressions involving multiple assessment points rather than simply calculating pre-post differences between two time points, may have positively impacted the accuracy of our predictions of both EMA and WQA. Another notable finding is that, unlike some other studies (e.g., Li et al., 2023), we found that baseline depression severity (BDI-II scores) did not significantly predict treatment response in our models, suggesting that early improvement provides predictive information beyond initial symptom levels.

### EARLY IMPROVEMENT DEFINITION

The second aim of our study was to explore different definitions of early improvement, combining the four different time windows with four different symptom cutoffs. Comparing the AUC values of our ROC-analyses revealed that the significant prediction accuracy of the 3-week time window did not significantly improve when extended to 4 weeks. Additionally, the ROC-analyses revealed that almost all investigated definitions of early improvement yielded prediction accuracies above chance, as indicated by positive Youden's indices (Table 3). For the symptom cutoffs, we observed that WQA reached the highest ratios between the true negative and false negative predictions with a symptom improvement of $\geq$10%. However, both EMA and WQA achieved ratios >1, indicating more true negative than false negative predictions.

This suggests that small cutoffs such as a $\geq$10% improvement assessed with WQA 3 to 4 weeks after treatment initiation might serve as a suitable predictor of treatment response that could guide clinicians in their decision about whether to continue or change a psychological treatment in the sense of stepped care or modularized therapy. The definition identified responders with a sensitivity of 100% and nonresponders with a specificity of 22% without taking any false negative prediction into account. This means after 4 weeks in our study, 0 patients (0%) would have been falsely recommended to change their treatment, while 6 patients (12%) would have been correctly advised to do so and therefore spent 3 weeks (42%) less in an ineffective treatment before considering alternatives. In total, this would have saved 18 weeks ($\sim$4.5 months) that our clinicians spend treating patients without the prospect of a response. From a clinical and economic perspective, examining the ratio between the true negative and false negative predictions is a promising approach to advancing the development of decision rules for treatment allocation in the context of stepped care and personalized therapy.

### LIMITATIONS

Despite the robustness of our findings, several limitations warrant consideration. First, the lack of validation for the identified definitions of early improvement underscores the need for further research to confirm our results. The time windows and improvement rates used to define early improvement were chosen pragmatically. Additionally, the modest sample size and unique therapy setting of our study, involving intensive inpatient or day clinic treatment with four therapy sessions per week and optional concomitant care, limits the generalizability of our findings to other treatment contexts. Given the intensive treatment dose over the 7-week course in a combined group and individual session format, including concomitant inpatient and day clinic care, the question remains as to whether the results (3 weeks or 12 sessions) are transferable to, for example, an outpatient setting. Therefore, future studies should aim to replicate our findings across diverse treatment durations and intensities, less severe levels of depression, and different settings. Moreover, the results need to be validated in a sample without concomitant care. Although concomitant pharmacotherapy and additional treatments were documented, our sample size did not allow for a sufficiently powered analysis of their potential effects on treatment response. Due to the research design, detected change in symptoms cannot be disentangled from psychotherapy effects. Therefore, future studies with larger samples should further examine their impact. Similarly, the potential differentiating effect of various psychological treatments as a between-subjects variable could not be robustly examined due to limited sample size. Since our study was underpowered to detect potential moderation effects, treatment condition was therefore not included as a covariate in the reported models. However, exploratory analyses testing interactions with early improvement yielded no significant results. Future research with larger samples is needed to investigate whether the predictive value of early change differs across therapeutic approaches.

Exploring alternative methods for assessing responder status could provide deeper insights into the comparative effectiveness of EMA and WQA assessments. In our study, we measured treatment

response with the same measure (namely the BDI-II) as the WQA-assessed early improvement rates. This confounds the comparisons between our EMA and WQA predictions, as measurement instruments may predict their own outcomes more accurately than others. Moreover, this opens the question which measurement instrument might generate the most valid and reliable assessment of change in depressive symptoms. When comparing EMA and WQA, it is also important to consider that multiple EMA prompts per day might cause higher patient burden than WQA (van Genugten et al., 2020). Therefore, it could be promising to develop EMA approaches that track depression-related behaviors passively by using wearables (e.g., sleep, activity, stress) and formulate EMA items neutrally to avoid systematically drawing patients' attention to negative aspects. Moreover, the substantial proportion of missingness in our sample is a limitation of our study. More than 50% of the originally recruited sample had to be excluded due to dropout and/or missing data, which limits our conclusions to patients who completed our treatment and assessments. However, exploratory analyses showed that BDI-II scores at the final available assessment did not significantly differ between patients who completed treatment (M = 16.81) and were included in the analysis sample and those who were excluded, for example due to dropout during treatment (M = 19.72, t = 1.45, $p$ = .15).

Finally, it is important to note that our wTNFN ratio depends on the responder rate of the population. As our prediction analyses were retrospective, we knew the exact responder rate of our sample. In a prospective study, the responder rate would have to be estimated. Especially when high responder rates are expected, it is important to prove which TNFN ratios reach their target. For example, when our TNFN ratios are transferred to a sample with 90% responders, the TNFN ratio of a WQA-assessed 10% improvement after Week 3 of TNFN = 8.33 results in a wTNFN <1 (wTNFN(RR = 0.9): 8.33 * ((1 − 0.9)/0.9) = 0.93). This means that in this sample the TNFN ratio would not have resulted in more true negative than false negative predictions as the rate of false negative predictions increases proportionally to the rate of responders.

## Conclusion

In conclusion, our study highlights the potential of early improvements in depressive symptoms, measured by EMA or WQA, serving as robust predictors of treatment response in intensive psychological treatments. The insight that after 3 weeks of treatment WQA was able to detect nonresponders with a specificity of 33% while taking a false negative rate of only 4% into account, coupled with the importance of small cutoffs to yield such high ratios between the specificity and false negative rate, offers valuable implications for further research and the development of clinical decision-rules. Moreover, they can be used in trial designs that aim to speed up the development of psychological treatments such as the "leapfrog" method (Blackwell et al., 2019). By identifying early predictors of treatment response, our findings could help streamline this process, making it more efficient. The validation of EMA as an alternative tool to WQA inspires further research investigating its potential when different definitions of treatment response are used. However, we recognized that there is little research on how effective nonresponders of one intervention can be treated with another, which is an important assumption of treatment prediction research in any way. Finally, further validation studies are warranted to confirm our findings and support their integration into clinical practice.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.beth.2025.08.002.

## References


Beard, J. I. L., & Delgadillo, J. (2019). Early response to psychological therapy as a predictor of depression and anxiety treatment outcomes: A systematic review and meta-analysis. *Depression and Anxiety, 36*(9), 866–878. https://doi.org/10.1002/da.22931.

Blackwell, S. E., Woud, M. L., Margraf, J., & Schönbrodt, F. D. (2019). Introducing the leapfrog design: A simple Bayesian adaptive rolling trial design for accelerated treatment development and optimization. *Clinical Psychological Science, 7*(6), 1222–1243. https://doi.org/10.1177/2167702619858071.

Bolin, E., & Lam, W. (2013). A review of sensitivity, specificity, and likelihood ratios: Evaluating the utility of the electrocardiogram as a screening tool in hypertrophic cardiomyopathy. *Congenital Heart Disease, 8*(5), 406–410. https://doi.org/10.1111/chd.12083.

Chen, Z., Zhao, S., Tian, S., Yan, R., Wang, H., Wang, X., Zhu, R., Xia, Y., Yao, Z., & Lu, Q. (2022). Diurnal mood variation symptoms in major depressive disorder associated with evening chronotype: Evidence from a neuroimaging study. *Journal of Affective Disorders, 298*, 151–159. https://doi.org/10.1016/j.jad.2021.10.087.

Colombo, D., Suso-Ribera, C., Fernández-Álvarez, J., Cipresso, P., Garcia-Palacios, A., Riva, G., & Botella, C. (2020). Affect recall bias: Being resilient by distorting reality. *Cognitive Therapy and Research, 44*, 906–918. https://doi.org/10.1007/s10608-020-10122-3.

Crits-Christoph, P., Connolly, M. B., Gallop, R., Barber, J. P., Tu, X., Gladis, M., & Siqueland, L. (2001). Early improvement during manual-guided cognitive and dynamic

psychotherapies predicts 16-week remission status. *The Journal of Psychotherapy Practice and Research, 10*(3), 145–154.

Cuijpers, P. (2015). Psychotherapies for adult depression: Recent developments. *Current Opinion in Psychiatry, 28*(1), 24–29. https://doi.org/10.1097/YCO.0000000000000121.

Cuijpers, P., Huibers, M., Ebert, D. D., Koole, S. L., & Andersson, G. (2013). How much psychotherapy is needed to treat depression? A metaregression analysis. *Journal of Affective Disorders, 149*(1–3), 1–13. https://doi.org/10.1016/j.jad.2013.02.030.

Cuijpers, P., Karyotaki, E., Ciharova, M., Miguel, C., Noma, H., & Furukawa, T. A. (2021). The effects of psychotherapies for depression on response, remission, reliable change, and deterioration: A meta-analysis. *Acta Psychiatrica Scandinavica, 144*(3), 288–299. https://doi.org/10.1111/acps.13335.

Cuijpers, P., Quero, S., Noma, H., Ciharova, M., Miguel, C., Karyotaki, E., Cipriani, A., Cristea, I. A., & Furukawa, T. A. (2021). Psychotherapies for depression: A network meta-analysis covering efficacy, acceptability and long-term outcomes of all main treatment types. *World Psychiatry, 20*(2), 283–293. https://doi.org/10.1002/wps.20860.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics, 44*(3), 837–845. https://doi.org/10.2307/2531595.

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

Egli, S., Frieß, E., Graf, P., Hoehn, D., Kopf-Beck, J., Rein, M. L., Ruderer, A., Szesny-Mahlau, N., Zimmermann, P., & Keck, M. E. (2019). *Schematherapie der Depression: Ein Behandlungskonzept für das (teil-)stationäre Setting*. Hogrefe.

Gautam, S., Jain, A., Gautam, M., Vahia, V. N., & Grover, S. (2017). Clinical practice guidelines for the management of depression. *Indian Journal of Psychiatry, 59*(Suppl 1), S34–S50. https://doi.org/10.4103/0019-5545.196973.

Gloster, A. T., Rinner, M. T. B., Ioannou, M., Villanueva, J., Block, V. J., Ferrari, G., Benoy, C., Bader, K., & Karekla, M. (2020). Treating treatment non-responders: A meta-analysis of randomized controlled psychotherapy trials. *Clinical Psychology Review, 75*, 101810. https://doi.org/10.1016/j.cpr.2019.101810.

Gois, C., Dias, V. V., Carmo, I., Duarte, R., Ferro, A., Santos, A. L., Sousa, F., & Barbosa, A. (2014). Treatment response in type 2 diabetes patients with major depression. *Clinical Psychology & Psychotherapy, 21*(1), 39–48. https://doi.org/10.1002/cpp.1817.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*(1), 29–36. https://doi.org/10.1148/radiology.143.1.7063747.

Hautzinger, M. (2013). *Kognitive Verhaltenstherapie bei Depressionen* (7., überarbeitete Auflage). Beltz.

Hautzinger, M., Keller, F., & Kühner, C. (2009). *BDI-II: Beck-Depressions-Inventar. Revision* (2. Auflage). Pearson Assessment.

Huibers, M. J. H., Cohen, Z. D., Lemmens, L. H. J. M., Arntz, A., Peeters, F. P. M. L., 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. https://doi.org/10.1371/journal.pone.0140771.

Hunnicutt-Ferguson, K., Hoxha, D., & Gollan, J. (2012). Exploring sudden gains in behavioral activation therapy for major depressive disorder. *Behaviour Research and Therapy, 50*(3), 223–230. https://doi.org/10.1016/j.brat.2012.01.005.

Keller, M. B. (2003). Past, present, and future directions for defining optimal treatment outcome in depression: Remission and beyond. *JAMA, 289*(23), 3152–3160. https://doi.org/10.1001/jama.289.23.3152.

Kopf-Beck, J., Müller, C. L., Tamm, J., Fietz, J., Rek, N., Just, L., Spock, Z. I., Weweck, K., Takano, K., Rein, M., Keck, M. E., & Egli, S. (2024). Effectiveness of schema therapy versus cognitive behavioral therapy versus supportive therapy for depression in inpatient and day clinic settings: A randomized clinical trial. *Psychotherapy and Psychosomatics, 93*(1), 24–35. https://doi.org/10.1159/000535492.

Kopf-Beck, J., Zimmermann, P., Egli, S., Rein, M., Kappelmann, N., Fietz, J., Tamm, J., Rek, K., Lucae, S., & Brem, A.-K. (2020). Schema therapy versus cognitive behavioral therapy versus individual supportive therapy for depression in an inpatient and day clinic setting: Study protocol of the OPTIMA-RCT. *BMC Psychiatry, 20*, 1–19. https://doi.org/10.1186/s12888-020-02880-x.

Li, F., Jörg, F., Merkx, M. J. M., & Feenstra, T. (2023). Early symptom change contributes to the outcome prediction of cognitive behavioral therapy for depression patients: A machine learning approach. *Journal of Affective Disorders, 334*, 352–357. https://doi.org/10.1016/j.jad.2023.04.111.

Lutz, W., Arndt, A., Rubel, J., Berger, T., Schröder, J., Späth, C., Meyer, B., Greiner, W., Gräfe, V., Hautzinger, M., Fuhr, K., Rose, M., Nolte, S., Löwe, B., Hohagen, F., Klein, J. P., & Moritz, S. (2017). Defining and predicting patterns of early response in a web-based intervention for depression. *Journal of Medical Internet Research, 19*(6), e2064. https://doi.org/10.2196/jmir.7367.

Markowitz, J. C. (2022). Supportive evidence: Brief supportive psychotherapy as active control and clinical intervention. *American Journal of Psychotherapy, 75*(3), 122–128. https://doi.org/10.1176/appi.psychotherapy.2021.20210041.

McKay, D. E., Abramowitz, J. S., & Taylor, S. E. (2010). *Cognitive-behavioral therapy for refractory cases: Turning failure into success*. American Psychological Association.

Moore, R. C., Depp, C. A., Wetherell, J. L., & Lenze, E. J. (2016). Ecological momentary assessment versus standard assessment instruments for measuring mindfulness, depressed mood, and anxiety among older adults. *Journal of Psychiatric Research, 75*, 116–123. https://doi.org/10.1016/j.jpsychires.2016.01.011.

R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing https://www.R-project.org/.

Richards, D. A. (2012). Stepped care: A method to deliver increased access to psychological therapies. *Canadian Journal of Psychiatry, 57*(4), 210–215. https://doi.org/10.1177/070674371205700403.

Rosenkranz, T., Takano, K., Watkins, E. R., & Ehring, T. (2020). Assessing repetitive negative thinking in daily life: Development of an ecological momentary assessment paradigm. *PLOS ONE, 15*(4), e0231783. https://doi.org/10.1371/journal.pone.0231783.

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.

Rush, A. J., Kraemer, H. C., Sackeim, H. A., Fava, M., Trivedi, M. H., Frank, E., Ninan, P. T., Thase, M. E., Gelenberg, A. J., & Kupfer, D. J. (2006). Report by the ACNP Task Force on response and remission in major depressive disorder. *Neuropsychopharmacology, 31*(9), 1841–1853. https://doi.org/10.1038/sj.npp.1301131.

Schaffer, A., Kreindler, D., Reis, C., & Levitt, A. J. (2013). Use of mental health telemetry to enhance identification and predictive value of early changes during augmentation treatment of major depression. *Journal of Clinical Psychopharmacology, 33*(6), 775–781. https://doi.org/10.1097/JCP.0b013e31829e8359.

Schibbye, P., Ghaderi, A., Ljótsson, B., Hedman, E., Lindefors, N., Rück, C., & Kaldo, V. (2014). Using early change to predict outcome in cognitive behaviour therapy: Exploring timeframe, calculation method, and differences of disorder-specific versus general measures. *PLOS ONE, 9*(6), e100614. https://doi.org/10.1371/journal.pone.0100614.

Szegedi, A., Jansen, W. T., van Willigenburg, A. P. P., van der Meulen, E., Stassen, H. H., & Thase, M. E. (2009). Early improvement in the first 2 weeks as a predictor of treatment outcome in patients with major depressive disorder: A meta-analysis including 6562 patients. *Journal of Clinical Psychiatry, 70*(3), 344–353. https://doi.org/10.4088/jcp.07m03780.

Takano, K., & Tanno, Y. (2011). Diurnal variation in rumination. *Emotion, 11*(5), 1046–1058. https://doi.org/10.1037/a0022757.

Tamm, J., Takano, K., Just, L., Ehring, T., Rosenkranz, T., & Kopf-Beck, J. (2024). Ecological momentary assessment versus weekly questionnaire assessment of change in depression. *Depression and Anxiety, 2024*(1), 9191823. https://doi.org/10.1155/2024/9191823.

Targum, S. D. (2020). Baseline reliability and early response in clinical trials of major depressive disorder. *Medical Research Archives, 8*(10). https://doi.org/10.18103/mra.v8i10.2241.

Trivedi, M. H., Rush, A. J., Wisniewski, S. R., Warden, D., McKinney, W., Downing, M., Berman, S. R., Farabaugh, A., Luther, J. F., & Nierenberg, A. A. (2006). Factors associated with health-related quality of life among outpatients with major depressive disorder: A STAR*D report. *The Journal of Clinical Psychiatry, 67*(2), 185–195.

van Genugten, C. R., Schuurmans, J., Lamers, F., Riese, H., Penninx, B. W., Schoevers, R. A., Riper, H. M., & Smit, J. H. (2020). Experienced burden of and adherence to smartphone-based ecological momentary assessment in persons with affective disorders. *Journal of Clinical Medicine, 9*(2), 322. https://doi.org/10.3390/jcm9020322.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer, 3*(1), 32–35.

Received: December 7, 2024
Accepted: August 7, 2025
Available online: 11 August 2025