# Ecological momentary assessment as a measurement tool in depression trials

Steven D. Targum [a,*], Colin Sauder [b], Miriam Evans [b], John N. Saber [c], Philip D. Harvey [d]

[a] Signant Health, Boston, MA, USA
[b] Adams Clinical LLC, Watertown, MA, USA
[c] EMA Wellness LLC, Norwood, MA, USA
[d] University of Miami School of Medicine, Miami, FL, USA

## ARTICLE INFO

## ABSTRACT

We used ecological momentary assessment (EMA) to track symptoms during a clinical trial. Thirty-six participants with major depressive disorder (MDD) and MADRS scores $\geq 20$ were enrolled in a nonrandomized 6-week open-label trial of commercially available antidepressants. Twice daily, a mobile device prompted participants to self-report the 6 items of the $HamD_6$ sub-scale derived from the Hamilton rating scale for depression ($HamD_{17}$). Morning EMA reports asked "how do you feel now" whereas evening reports gathered a full-day impression. Clinicians who were blinded to the EMA data rated the MADRS, $HamD_{17}$ and $HamD_6$ at screen, baseline and weeks 2,4, and 6. Hierarchical linear modeling (HLM) examined the course of the EMA assessments and convergence between EMA scores and clinician ratings. HLM analyses revealed strong correlations between AM and PM EMA derived $HamD_6$ scores and revealed significant improvements over time. EMA improvements were significantly correlated with the clinician rated $HamD_6$ scores at endpoint and predicted clinician rated $HamD_6$ score changes from baseline to endpoint (p < .001). There was a large correlation between EMA and clinician derived $HamD_6$ scores at each in-person assessment after baseline. Treatment response defined by EMA matched the clinician rated $HamD_6$ treatment responses in 33 of 36 cases (91.7%). EMA derived symptom scores appear to be efficient and valid measures to track daily symptomatic change in clinical trials and may provide more accurate measures of symptom severity than the episodic "snapshots" that are currently used as clinical outcomes. These findings support further investigation of EMA for assessment in clinical trials.

## 1. Introduction

It has been reported that individuals with major depressive disorder (MDD) have greater mood variability and symptom fluctuation (instability) than non-depressed individuals, and that affective dynamics like mood variability and instability may constitute non-specific characteristics of mood disorders (Aan het Rot et al., 2012; Houben et al., 2015; Lamers et al., 2018). It has also been shown that many individuals with MDD experience a diurnal mood variation such that their symptoms fluctuate from day to night (Peeters et al., 2006; Morris et al., 2007; Justice, 2009). In the large STAR*D study of MDD, Morris and colleagues documented the diurnal variation of mood and noted that the symptoms were worse for some individuals in the morning but were worse for others at night (Morris et al., 2007). Given these mood dynamic patterns, it is noteworthy that the baseline and endpoint scores

generally used as the outcome variable to assess symptomatic change during clinical trials of MDD typically rely on just a single timepoint of measurement regardless of any symptom fluctuation. Further, these measurements are rarely collected at a pre-specified time of day in most clinical trials. The baseline and endpoint measures are usually derived from clinician-rated questionnaires that summarize symptom severity based on the patient's retrospective recall of their symptoms and behavior during the past week (or more). Retrospective recall of symptoms or recent behavior may be inaccurate measures because judgments about symptom status may be affected by immediate concurrent events or poorly recalled, particularly in depressed participants (Ben-Zeev et al., 2009; Solhan et al., 2009). Consequently, cross-sectional baseline and endpoint measures may not be reliable reflections of the clinical presentation. The reliability of the baseline measure and the sensitivity to daily symptom changes is particularly

---

important for the evaluation of rapidly acting antidepressants where symptom status 24 h post-randomization may be the endpoint.

Ecological momentary assessments (EMA), also called experience sampling methods (ESM), have been introduced as methods to sample the daily life experience of participants in real-time (Moskowitz and Young, 2006; Granholm et al., 2008; Ebner-Priemer and Trull, 2009; Wenze and Miller, 2010; Aan het Rot et al., 2012; Depp et al., 2012; Bos et al., 2015; Marzano et al., 2015; Armey et al., 2015; Houben et al., 2015; Moore et al., 2016; Lamers et al., 2018; Panaite et al., 2020). EMA/ESM can assess the severity and variability of symptoms, activity, cognitive functioning, and biology in the moment, as frequently during the day as desired, and obviates concerns about retrospective recall (Ebner-Priemer and Trull, 2009). Recent studies have suggested that the most successful methods for adherence include shorter surveys with no impact on sampling density.

EMA/ESM has been explored in studies of MDD participants to study affective dynamics, activity using actigraphy or GPS coordinates, psychophysiology, suicidal ideation, sleep characteristics, emotional reactivity, and the prediction of treatment response or vulnerability to relapse (Barge-Schaapveld et al., 2002; Peeters et al., 2010; Ebner-Priemer and Trull, 2009; Geschwind et al., 2011; Kramer et al., 2014; Armey et al., 2015; Lamers et al., 2018; Panaite et al., 2020). To our knowledge, EMA has not be used to track symptomatic changes during a pharmacological clinical trial to support the clinical assessment of treatment outcome.

We undertook a small study to explore the feasibility of EMA to track the course of clinical symptoms in MDD participants during a clinical trial. A recent review of EMA research in mood disorders noted that these studies have largely focused on group data rather than the course of symptoms in individual participants (Aan het Rot et al., 2012). In this study, we examined both the group and individual participant data through the treatment phases of the clinical trial. We specifically focused on: 1) daily consistency and fluctuation (instability) of EMA ratings; 2) convergence between EMA and clinician derived ratings; and 3) sensitivity of clinical changes based upon the EMA-derived assessments relative to changes detected by comparable clinician ratings.

## 2. Methods

This exploratory EMA study was initiated at a single clinical trial site (Adams Clinical LLC, Watertown MA) in March 2020 as an add-on to an existing site funded study of MDD called TRAIT: Treatment Response After Intervention Trial (Sauder et al., 2019). All potential study participants signed an IRB approved informed consent to participate in the TRAIT study and signed an additional consent if they agreed to participate in the add-on EMA study as well. Participants could withdraw from the EMA component at any time without losing their opportunity to participate in the TRAIT study and still receive antidepressant medication (ADT). Consenting participants were compensated $12.50 per week plus transportation costs for their participation in the TRAIT study and an additional $15 per week if they participated in the EMA study.

### 2.1. Study design

Eligible participants met DSM-5 criteria for MDD based upon the Mini International Psychiatric Interview (MINI) 7.0.2 (updated version for DSM-5) and had a minimum total score of 20 on the Montgomery-Asberg depression rating scale (MADRS) at the screen and baseline visits (Montgomery and Asberg, 1979; APA, 2013; Sheehan et al., 1998). A minimum 6-day screening period was followed by a nonrandomized 6-week open-label study of commercially available antidepressants. Eligible participants returned to the clinic for the baseline visit (day 0) and bi-weekly at weeks 2, 4, and 6 (3 visits after the baseline and initiation of treatment) for medication review, safety assessments, and in-person clinical assessments.

### 2.2. In-person clinical assessments

The clinician-rated Hamilton rating scale for depression ($HamD_{17}$), MADRS, and clinical global impression of severity (CGI-S) rating scale was administered by a trained rater at screen, baseline, and three subsequent study visits (Hamilton, 1960; Guy, 1976; Montgomery and Asberg, 1979). We chose the $HamD_6$ sub-scale derived from the $HamD_{17}$ as our abbreviated rating scale for EMA and extracted the score from the clinician rated $HamD_{17}$ for comparison. (Hamilton, 1960; Bech et al., 1975; Cleary and Guy, 1977). The $HamD_6$ includes six core symptoms of depression: depressed mood, guilt, loss of interest in work and activities, anxiety (psychic), somatic symptoms (general), and psychomotor retardation (Bech et al., 1975). Both the clinician rated and EMA derived $HamD_6$ items are individually scored from 0 - 2 or 0–4 (with increasing symptom severity) for each distinct survey and the total $HamD_6$ score is calculated as the sum of the 6 items. Clinician ratings on the $HamD_6$ have been shown to be highly correlated with the $HamD_{17}$ and the MADRS (O'Sullivan et al., 1997; Hooper and Bakish, 2000).

### 2.3. EMA data

EMA was obtained through the use of a mobile BYOD (bring your own device) strategy that delivered daily queries throughout the screening and treatment phases of the study. Participants were paged twice daily (AM and PM) every day of the study including the in-person clinic visit days. Thus, there were a minimum of 6 days prior to baseline and 42 days after baseline for twice daily EMA prompts to yield a total of 49 EMA rating days. The EMA application was developed by EMA Wellness LLC (Norwood, Massachusetts) and could be used on either an Apple or Android smartphone. Participants received alert prompts to remind them to complete queries between 6 and 10 a.m. and again between 6 and 10 p.m. daily. Participants had 2 ½ hours to complete the survey once it was opened on their mobile device and could return to it anytime within that time frame.

The AM and PM EMA queries included participant self-reports of the six $HamD_6$ symptom items. Each EMA item appeared sequentially and individually on the device touch screen and included a descriptor heading that identified the item followed by anchored response choices reflecting the scoring range of the rating instrument. The wording for the 6 sub-scale items was adapted from the patient self-rating assessment developed by Bech for the $HamD_6$ (2006) but was adjusted for different time-contingent intervals. The morning queries asked for immediate responses (in the moment) similar to the momentary response queries described by other authors for EMA (Ebner-Primer and Trull, 2009; Bos et al., 2015; Armey et al.., 2015; Harvey et al., 2020). The morning descriptor for the sub-scale items read: Choose the statement that describes how you are feeling NOW. The PM response was structured to ask about the severity of the identified symptom during the day and was therefore an experience sample over time (hours of the day) rather than an immediate momentary assessment. The PM response descriptor for the sub-scale items read: Choose the statement that describes how you felt TODAY. Fig. 1 displays two of the EMA morning survey screens as they appeared on the smartphone.

### 2.4. Data analyses

Statistical analysis of the data examined group means from clinician ratings and the average of the available AM and PM EMA derived scores obtained on each day of the in-person assessments (baseline and weeks 2, 4, and 6). We also examined the growth curves for the individual participant daily AM and PM EMA data points to evaluate the course of treatment and variation across individuals and time points in symptoms. We examined the trajectory of treatment responses in the intent to treat (ITT) group composed of all participants who were prescribed antidepressant medication at baseline (day 0) and had at least one post-baseline clinician assessment.
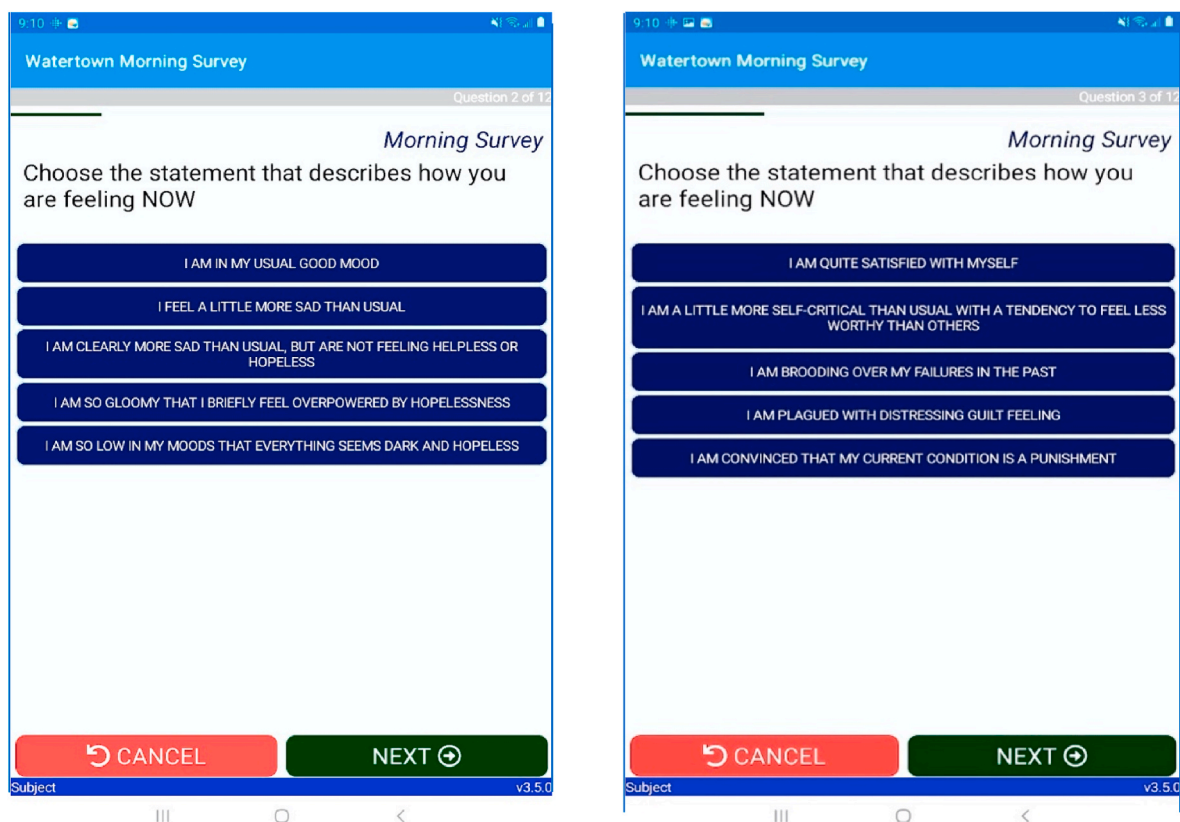
**Fig. 1.** EMA mobile survey: screen shot examples.

Clinician generated ratings (MADRS, HamD$_{17}$, HamD$_6$) administered during in-person clinic visits were compared across the assessment time points with repeated-measures analysis of variance. The EMA derived scores collected on the same day of the in-person assessments was examined in a similar way.

The twice daily EMA data were examined with Hierarchical Linear Modeling (HLM). In the EMA analyses, we used random regression analyses for the intercorrelations of AM and PM EMA scores over the course of the study and Mixed Model Repeated Measures Analysis of Variance (MMRM ANOVA) for the course of the EMA HamD$_6$ growth curves and the association between growth curves for EMA scores and the clinician ratings (GLM Module, SPSS version 26, IBM Corporation, 2019). We entered subject as a random intercept and entered day and time of day and their interaction as repeated measures. We tested all models to ensure that the omnibus test for the fitted model exceeded the fit of the intercept-only model. A statistically significant intercept means that there was significant inter-subject variation. Missing EMA observations up to the point of withdrawal from the study were addressed through use of full-information maximum likelihood procedures. In interpreting the data, a nonsignificant effect of day would mean that there was no change in EMA derived HamD$_6$ scores over the course of the study. Further, a nonsignificant effect of time of day would mean that the AM and PM scores did not differ, and a nonsignificant interaction of day x time of day would mean that AM and PM scores changed equivalently over time.

We performed conventional analyses of the correlations between EMA derived and clinician derived HamD$_6$ scores on the day of the in-person assessments at each of the overlapping visit time points, comparing the correlations between EMA and clinician derived Ham-D$_6$ scores and the correlation between clinician rated HamD$_6$ and HamD$_{17}$ scores. Treatment response was defined as ≥50% improvement of the total MADRS or HamD$_{17}$ score from the baseline to the last observed visit (weeks 2, 4, or 6). Similarly, treatment response for the both the

clinician rated HamD$_6$ and EMA-derived HamD$_6$ scores was defined as ≥50% improvement of the HamD$_6$ score from the baseline to the last reported visit.

## 3. Results

The study was conducted between March and August 2020. Patient recruitment was affected by concerns about possible COVID-19 exposure in ambulatory research patients, but ultimately 49 participants were screened for the study. The total MADRS scores for all screened patients ranged from 22 to 46 with a mean score = 34.4 ± 5.3 (SD) and the total HamD$_{17}$ scores ranged from 12 to 48 (mean = 25.3 ± 7.4). Twelve screened patients did not continue or did not consent to the EMA component of the study, and one patient was discontinued because the MADRS score fell below the entry criterion at the baseline visit. Hence, 36 participants were prescribed antidepressant treatment (ADT) at the baseline (day 0) visit.

Amongst the 36 enrolled participants, there were 19 men and 17 women with ages ranging from 22 to 60 years (mean age = 30.8 ± 8.6 years). Based on the MINI, there were 4 first episode MDD participants and 32 participants (88.9%) with recurrent MDD. Seventeen participants met DSM-5 criteria for MDD, and 19 participants met criteria for MDD with anxious distress. The participants reported current depressive episodes ranging from 2 to 60 months (mean = 12.8 ± 13.1months); 15 participants had been depressed for ≤6 months and 23 had been depressed for ≤ 12 months. Ten participants were taking antidepressant medications at the time of screening whereas 26 participants were not. The mean BMI was 28.2 ± 7.2 (range 19–39.5 kg/m$^2$).

The antidepressant medications prescribed at baseline varied by participant, were required to be different from the participant's previous ADT experience, and included bupropion XL 300 mg, sertraline 50 mg, duloxetine 60 mg, or escitalopram 10 mg.

Participant compliance with the EMA surveys was similar between

the AM and PM assessments over the course of the study. Overall, 2113 of 2836 possible twice-daily EMA prompted surveys were completed over the 49-day study period (74.5%; 74.0% AM and 75.0% PM) by the 36 participants during the study.

### 3.1. Clinician rated and participant rated EMA metrics

Table 1 displays the mean clinician rated MADRS, HamD$_{17}$, and clinician derived HamD$_6$ scores obtained during the 5 scheduled study visits for the study participants. Table 2 displays the paired clinician and participant rated EMA derived HamD$_6$ scores at the screening, baseline, and weeks 2, 4, and 6 visits (n = 150 paired ratings). As shown in Table 2, the mean clinician rated HamD$_6$ scores were significantly higher than the EMA derived HamD$_6$ at several visits, but there was no statistically significant difference between the clinician rated and EMA derived HamD$_6$ score changes from the baseline to the endpoint assessment.

Clinician ratings for the MADRS and HamD$_{17}$ were highly correlated at baseline (r = 0.87, p < .001) and endpoint (r = 0.93, p < .001). In a traditional analysis focusing on convergence between EMA and clinician-rated scores collected at the same time point, we computed Pearson correlations between clinician rated HamD$_6$ scores and averaged AM and PM EMA HamD$_6$ scores at baseline and weeks 2, 4, and 6. The correlations were r = 0.31 at baseline (p = .068), r = 0.54 at week 2 (p < .001), r = 0.66 at week 4 (p < .001), and r = 0.73 at week 6 (p < .001). The correlations between the clinician ratings for the HamD$_{17}$ and HamD$_6$ were r = 0.46 (baseline), r = 0.61 (week 2), r = 0.61 (week 4), and r = 0.61 (week 6) respectively. None of these correlations were significantly different in their magnitude with a Fisher r to z transformation, all z < −0.72, all p > .47. Thus, the Pearson correlations between EMA and clinician ratings on the HamD$_6$ are quite consistent with the intercorrelation of clinician ratings and the correlation increased over the course of the study sharing 50% of the variance at the final assessment.

In the repeated measures analyses, we examined the clinician-rated HamD$_6$ and participant rated EMA derived HamD$_6$ score changes from baseline to the end of the study, using time point (baseline, weeks 2, 4, and 6) as the repeated-measures factor. For all 3 clinician-rated scale analyses, there was a statistically significant omnibus test indicating improvement compared to the random intercept model, all $X^2(3)>9.07$, all p < .03. Also, there was a statistically significant effect of the random intercept in all three models, all $X^2(1)>777.65$, all p < .001, indicating significant between subject's variation in response. All 3 clinician-rated depression rating scales improved over time and all of the time effects were statistically significant ($X^2(3)>9.44$, for all p < .03). Within group contrasts on the clinician rated HamD$_6$ found a significant improvement between baseline and week 2 (p < .01).

For the participant rated EMA derived HamD$_6$ scores on the days of the clinician visits, there were also statistically significant omnibus effects ($X^2(3) >8.60$, p < .03), a significant time effect ($X^2(3) = 7.80$, p =

**Table 1**
Mean clinician rated metrics administered on in-person clinic days.

| | n | MADRS | HamD$_{17}$ | HamD$_6$ |
|---|---|---|---|---|
| SCREEN (all subjects) | 49 | 34.4 ± 5.4 | 25.3 ± 7.4 | 10.9 ± 2.8 |
| SCREEN (enrolled) | 36 | 33.7 ± 5.3 | 20.9 ± 3.9 | 11.8 ± 2.1 |
| Baseline | 36 | 32.1 ± 5.3 | 20.3 ± 4.5 | 11.4 ± 2.2 |
| V1 (week 2) | 36 | 25.7 ± 8.9 | 17.6 ± 5.7 | 9.3 ± 3.4 |
| V2 (week 4) | 25 | 23.5 ± 11.3 | 16.0 ± 7.6 | 8.6 ± 4.3 |
| V3 (week 6) | 17 | 24.9 ± 12.2 | 16.2 ± 8.1 | 8.2 ± 4.5 |
| Change from baseline (last assessment) | 36 | 8.5 ± 12.1 | 4.4 ± 7.4 | 3.2 ± 4.6 |

**Table 2**
Comparison of mean paired clinician and participant rated (EMA derived) HamD$_6$ scores.

| | N | Clinician-rated HamD$_6$ | Participant-rated EMA HamD$_6$ | Pearson Correlation | p Clinician HamD$_6$ vs EMA HamD$_6$ |
|---|---|---|---|---|---|
| SCREEN (enrolled) | 36 | 11.8 ± 2.1 | 8.8 ± 2.9 | 0.235 | <0.001 |
| Baseline | 36 | 11.4 ± 2.2 | 8.3 ± 4.1 | 0.308 | <0.001 |
| V1 (week 2) | 36 | 9.3 ± 3.4 | 5.9 ± 3.3 | 0.543 | <0.001 |
| V2 (week 4) | 25 | 8.6 ± 4.3 | 6.0 ± 4.2 | 0.661 | <0.03 |
| v3 (week 6) | 17 | 8.2 ± 4.5 | 6.0 ± 5.6 | 0.725 | ns |
| Change from baseline (last assessment) | 36 | 3.2 ± 4.6 | 2.5 ± 4.5 | 0.418 | ns |

NOTE: Paired EMA scores are aligned with the day of the corresponding in-person clinician rated HamD$_6$ scores for the specific clinic visit.

.05), and a significant random intercept effect reflecting the significant between subject's variability ($X^2(1) = 265.68$, p < .001).

Fig. 2 displays the AM and PM group mean EMA derived HamD$_6$ scores for all collected data for the days preceding the baseline assessment through to the week 6 assessment. Random regression was used to examine the correlation of AM and PM assessments for the EMA derived HamD$_6$ scores. In the regression model, a random intercept was entered, and day was entered as a repeated measure for the prediction of PM scores by AM scores over the study period. The fitted model for the HamD$_6$ was significantly better than the intercept only-model, $X^2(75) = 434.6$, p < .001. The effect of day was non-significant, $X^2(53) = 62.1$, p = .18. However, the correlation between AM and PM HamD$_6$ scores across the 6-week study period was very significant, $X^2(22) = 920.5$, p $< 4.0 \times 10^{-18}$. These findings indicate that the correlation between EMA derived AM and PM HamD$_6$ scores was consistent across all of the days of the observation period.

### 3.2. Trajectory of EMA scores

In analyzing the time course of EMA derived HamD$_6$ scores as presented in Fig. 2, we examined the effects of visit, time, and visit x time on the scores. The results of these analyses are presented in Table 3. We found a significant omnibus effect and a significant random intercept, meaning that the model improved on the intercept only model but that there was significant between-subjects variability. There was also a statistically significant effect of visit, with scores decreasing over time, but no significant effects of time of day or interaction of visit x time of day, $X^2(42) = 35.5$, p = .89.

We used the same model to predict endpoint clinician rated HamD$_6$ scores with the course of HamD$_6$ scores over the 6-week study period. We found a significant omnibus effect and a significant random intercept. There was also a statistically significant effect of predictive correlation with EMA derived HamD$_6$ scores over the 6-week period and a significant effect of visit, but no significant effect of time of day or interaction of visit x time of day. Thus, decreasing EMA derived HamD$_6$ scores correlated with the clinician rated endpoint HamD$_6$ scores.

In a final predictive analysis, we used the same model to predict change in clinician rated HamD$_6$ scores from baseline to the final assessment. As noted in Table 3, we found a significant omnibus effect and a significant random intercept (p < .001). There was a statistically significant correlation between EMA derived HamD$_6$ scores over the 6-week period and also a statistically significant effect of visit, but no significant effects of time of day or interaction of visit x time of day. Thus, EMA measured decreases in HamD$_6$ scores successfully predicted improvements in the clinician rated HamD$_6$ scores from baseline to endpoint.
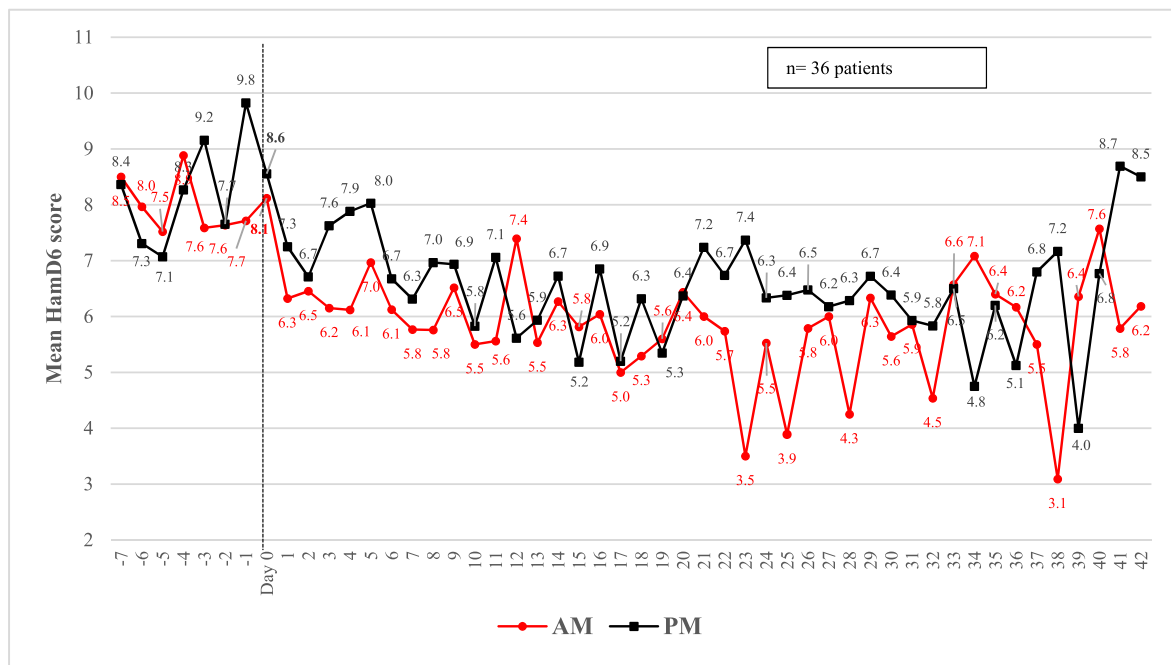
NOTE: Data points indicate mean scores of data that was available at designated timepoint

**Fig. 2.** Daily EMA HamD$_6$ scores in MDD patients:
NOTE: Data points indicate mean scores of data that was available at designated timepoint.

**Table 3**
EMA derived HamD$_6$ scores: Time course and correlations with clinician rated HamD$_6$.

**Time Course of EMA Derived HamD$_6$**

|  | $X^2$ | df | p |
|---|---|---|---|
| Omnibus Test | 135.58 | 107 | 0.03 |
| Intercept | 3153.05 | 1 | <.001 |
| Visit | 97.67 | 42 | <.001 |
| Time of Day | 3.69 | 1 | 0.06 |
| Visit x Time | 35.50 | 42 | 0.89 |

**Daily EMA Prediction of Endpoint Clinician Rated HamD$_6$**

|  | $X^2$ | df | p |
|---|---|---|---|
| Omnibus Test | 357.15 | 66 | <.001 |
| Intercept | 2759.13 | 1 | <.001 |
| EMA HamD$_6$ | 373.38 | 22 | <.001 |
| Visit | 63.9 | 42 | 0.02 |
| Time of Day | 0.1 | 1 | 0.75 |
| Visit x Time | 1.1 | 42 | 0.24 |

**Daily EMA Prediction of Change from Baseline to Endpoint in Clinician Rated HamD$_6$**

|  | $X^2$ | df | p |
|---|---|---|---|
| Omnibus Test | 299.03 | 66 | <.001 |
| Intercept | 213 | 1 | <.001 |
| EMA HamD$_6$ | 285.59 | 22 | <.001 |
| Visit | 74.43 | 42 | 0.002 |
| Time of Day | 0.77 | 1 | 0.38 |
| Visit x Time | 31.21 | 42 | 0.93 |

### 3.3. Score fluctuation and treatment trajectories in individual study participants

The twice daily mean EMA HamD$_6$ score fluctuations displayed in Fig. 2 reflect the real life daily symptomatic changes that occur in depressed individuals. However, group means are not informative about the symptomatic fluctuations of individual study participant's. Fig. 3 displays the individual AM and PM daily EMA HamD$_6$ score trajectories

from 12 study participants. Some, but not all individual participants revealed marked HamD$_6$ score fluctuation from day to night (AM to PM) and/or from day to day.

Two cases displayed in Fig. 4 illustrate the potential of EMA to examine affective dynamics and effectively track the treatment course of depressed individuals beyond the weekly or bi-weekly clinician measurements that are typical of MDD trials.

Participant 8294 was a 28-year-old woman with a diagnosis of MDD with anxious distress who had been depressed for 5 months when she entered the study. Her daily scores revealed marked EMA HamD$_6$ score fluctuation (symptomatic instability) during the screening period and following the initiation of treatment (Fig. 4). The daily EMA score fluctuations reflect a mood instability during the early treatment phase that impedes a meaningful interpretation of the single point in time clinician ratings administered at baseline and week 2. However, the EMA HamD$_6$ score became more stable after 2 weeks and this participant ultimately became a treatment responder (and remitter) based upon both the clinician ratings and EMA scores obtained at week 6. The participant rated EMA HamD$_6$ score changes throughout the study revealed a progressive and sustained improvement that was detected prior to the clinician rating and 3 weeks before the end of the study.

Participant 8301was a 26-year-old woman with a diagnosis of MDD who had been depressed for 6 months prior to entering the study. As shown in Fig. 4, this participant was a treatment non-responder at week 6 whose clinician derived HamD$_6$ scores aligned well with the EMA derived HamD$_6$ scores at each in-person visit. However, the individual AM and PM EMA HamD$_6$ scores fluctuated from day to day displaying the symptomatic instability that may be a characteristic of some depressed individuals. In the five days between days −2 and day 2 of treatment, AM EMA HamD$_6$ scores changed from 13, 10, 12, 16, and 9 yielding an unreliable baseline measurement.

### 3.4. Examination of treatment response

Based on the clinician ratings, there were 11 MADRS, 11 HamD$_{17}$, and 12 HamD$_6$ treatment responders in this group of 36 patients at the last available assessment (endpoint). Ten of the 11 MADRS treatment
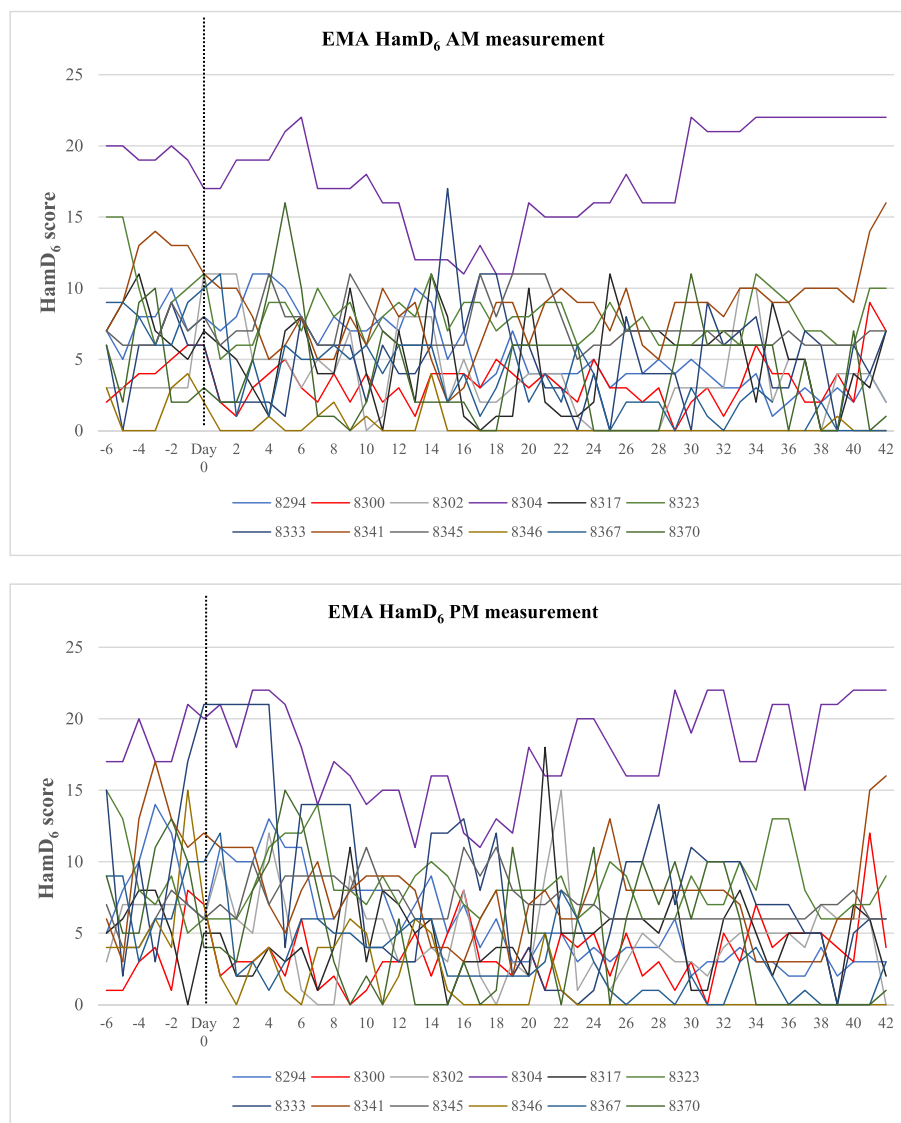
NOTE: Displays individual twice daily EMA HamD$_6$ score trajectories from 12 enrolled participants. Missing data points are interpolated in accordance with MMRM principles.

**Fig. 3.** Individual participant trajectories of EMA HamD$_6$ scores
NOTE: Displays individual twice daily EMA HamD$_6$ score trajectories from 12 enrolled participants. Missing data points are interpolated in accordance with MMRM principles.

responders matched the HamD$_{17}$ responders at that visit (90.9%). Ten of the HamD$_6$ responders matched the HamD$_{17}$ responders (90.9% sensitivity) with a specificity of 92.0% and predictive value of 91.7%.

Eight of the participant-rated EMA-derived HamD$_6$ responses matched the clinician rated HamD$_{17}$ responders (72.7% sensitivity). Ten of the EMA-derived HamD$_6$ responders matched the 12 clinician rated HamD$_6$ responders (83.3% sensitivity) and 22 EMA-derived HamD$_6$ non-responders matched the 24 clinician rated HamD$_6$ non-responders (95.8% specificity) with a resulting predictive value of 91.7%.

## 4. Discussion

We undertook a small study to explore the utility of ecological momentary assessment (EMA) as a method to track treatment outcome during clinical trials of MDD. The study was conducted during the COVID-19 crisis, but we collected enough data to examine the potential use of EMA for clinical trials. The ratio of men to women was higher in this study than in most clinical trials of MDD. We believe that the higher male ratio reported here is related to the COVID-19 crisis because

potential study candidates needed to entertain and consent to participate in a clinical trial that required some in-person clinic visits.

We prompted EMA queries twice per day over 49 days and generated an overall EMA adherence rate of 74.5% that is consistent with other EMA studies (Granholm et al., 2008, 2020; Depp et al., 2012; Moore et al., 2016). There were over 2100 data points collected in the study, and some participants providing over 90% of the possible data. We found that participants were not saturated by daily prompts and that their AM and PM EMA scores remained correlated and consistent with the corresponding clinical scores throughout the study. Granholm et al. (2020) reported that up to 7 EMA samples per day yielded the same adherence rate as this study. It has been suggested that more frequent sampling might generate even better adherence (Depp et al., 2012; Granholm et al., 2020). Other studies have noted that frequent prompts to self-rate did not influence symptomatic changes (a "Hawthorne" effect) that might facilitate a placebo response (Ebner-Priemer and Trull, 2009; Granholm et al., 2020; Santangelo et al., 2013; Harvey et al., in press).

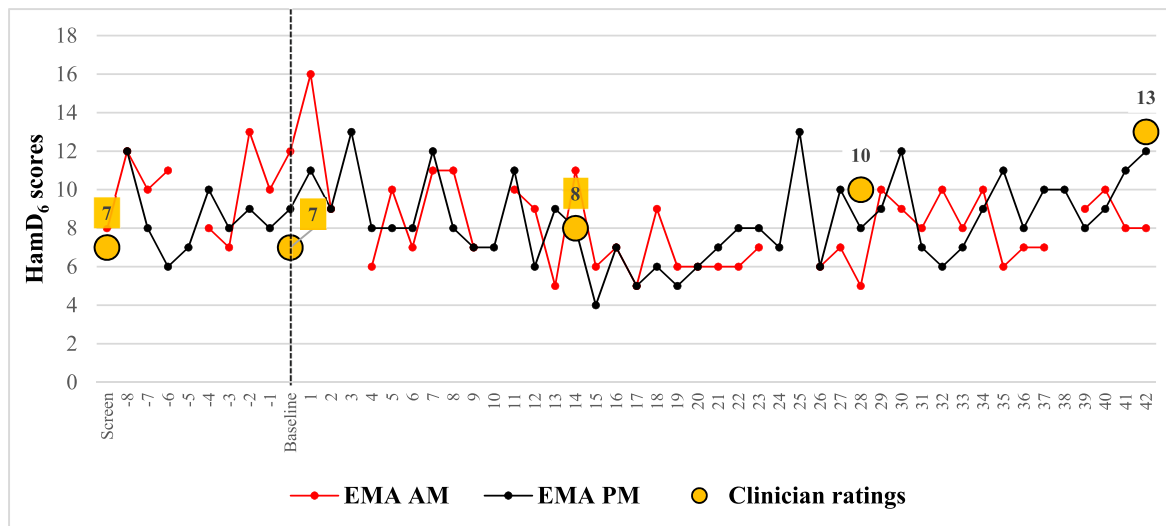Twelve of 36 study participants (33.3%) improved and met the
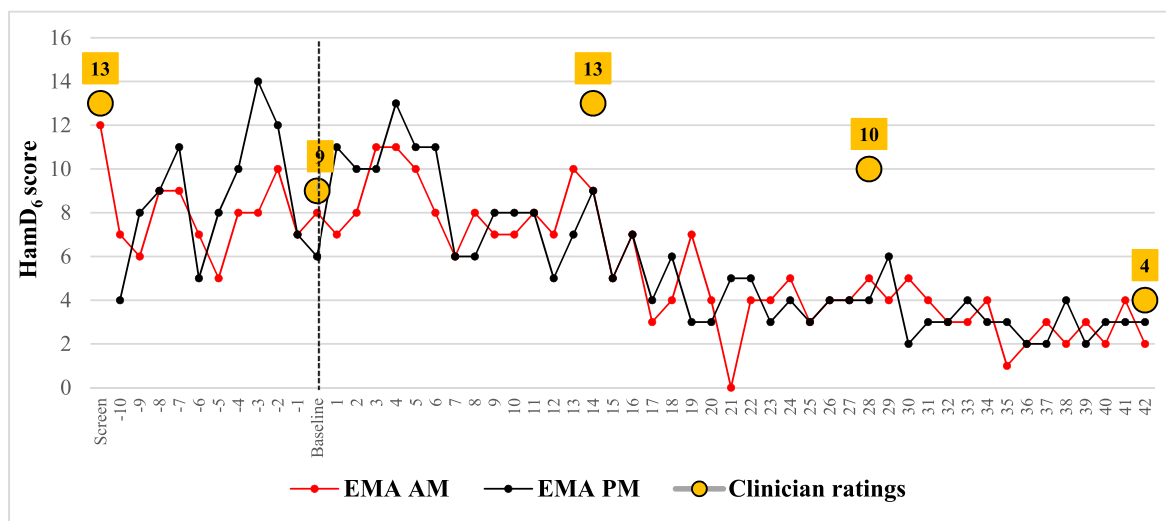
Participant 8301



Fig. 4. Trajectory of twice daily HamD$_6$ scores in individual study participants.

criterion for treatment response on the clinician rated HamD$_6$ sub-scale during the 6-week open label treatment period. Similarly, proportionate improvements in depression scores were found for the EMA derived HamD$_6$ scores. The treatment related course of EMA derived HamD$_6$ scores were highly correlated with the clinician ratings at endpoint, and the treatment related course of EMA derived HamD$_6$ scores was also strongly related to changes in clinician rated depression (p < .001). Further, 33 of the 36 EMA-derived HamD$_6$ treatment response scores matched the clinician rated HamD$_6$ treatment responses at the last assessment (91.7% predictive value).

The AM EMA samplings were true ecological momentary assessments reflecting recency that asked participants how they were feeling "now" whereas the PM samplings asked participants how they had felt during the entire day. Nonetheless, the AM EMA measure of recency was highly correlated with the PM EMA measure that assessed the entire day [$X^2$ (22) = 920.5, p < 4.0 × 10$^{-18}$], and both the AM and PM EMA measures predicted clinical outcome. Thus, the high correlation between the morning and evening EMA scores suggest that the assessments at either time point can yield valid scores. There was a very significant subject level intercept in all analyses, which revealed moderate HamD$_6$

score fluctuation from day to night (AM to PM) and from day to day throughout the study. The presence of fluctuating symptoms (instability) prior to the initiation of treatment had no bearing on the subsequent antidepressant treatment response in this small sample of participants. After baseline, we found that the daily EMA scores effectively documented day to day progressive and sustained symptomatic improvement in the treatment responders, and that a lack of response in the treatment non-responders was highly correlated with the clinician rated scores.

In this study, the AM and PM EMA assessments applied different time-contingencies that were markedly different than the clinician's ratings that used a 7-day recall of the patient's symptom severity at a single point in time. We believe that the EMA methodology may provide more accurate measures of symptom severity than the weekly (or less) "snapshots" that are currently used. It might be argued that the immediacy of the moment could miss the full clinical picture and generate a score that is too limited by symptomatic instability to offer a meaningful reflection of the clinical presentation. A strategy that averages EMA scores across several days might offset this potential limitation, address the likelihood of some missed scores, and more effectively track the

course of treatment. However, an averaging strategy would miss abbreviated symptomatic spikes that reflect the symptomatic instability that may be a characteristic of some depressed individuals. Alternatively, more frequent sampling across the day can broaden the scope of the assessment data collected.

It is clear that the severity and intensity of individual symptoms may fluctuate from assessment to assessment and that diurnal mood changes as well as emotional reactivity to confounding factors or random life events can influence the momentary experience of symptoms. Recent studies of mood dynamics have shown that moment-to-moment mood changes (instability) can occur daily and occur more frequently in depressed than non-depressed individuals (Solhan et al., 2009; Houben et al., 2015; Lamers et al., 2018; Panaite et al., 2020). Thus, fluctuating symptoms can be expected as part of the natural course of depressive illness, and these changes can be assessed (tracked) by more frequent measurements. In our view, the measurement challenge for clinical trials of MDD is not due to this naturally occurring fluctuation of symptoms. Instead, measurement accuracy is challenged by baseline and endpoint measurements that are obtained like "snapshots" at a single fixed point in time and may not be the best indicators of symptom severity. The documentation of daily symptomatic fluctuation illustrates the complexity of tracking symptomatic change in clinical studies and highlights the limitations of a single fixed point in time as a baseline or endpoint measurement. These findings suggest that EMA can be more than an adjunctive measure to clinician assessments and could be a viable substitute in circumstances requiring remote assessments.

This exploratory EMA study has several limitations that must be noted. First, the results require cautious interpretation because only 36 participants were enrolled in a nonrandomized, open-label study. However, both clinician ratings and EMA ratings were sensitive to clinical improvements and these improvements were strongly correlated. The visit-by-visit correlations between EMA and clinician ratings increased in their convergence over the course of the study, probably indicating increased accuracy of clinician ratings. Although we found clear instances of symptom stability and instability in this small sample of participants, we examined too few participants to draw firm conclusions about the meaning or value of these affective dynamics for use in clinical trials. Second, we can affirm that EMA was successfully collected but we cannot affirm the accuracy or reliability of the participant's self-assessment at the time of these assessments. Of course, the same concern about accuracy or reliability applies to self-reports given to clinicians who use these responses for their clinician ratings. EMA scores, like clinician ratings, can be influenced by extraneous factors and it is well known that participant rated scores do not always coincide with clinician ratings (Targum et al., 2013). Third, we cannot assert that the EMA scores are more or less accurate than the periodic clinician ratings that are based on a patient's retrospective recall of the past two weeks. However, it is noteworthy that the EMA-derived HamD$_6$ scores were significantly correlated with the corresponding clinician's scores in this study and that the EMA scores anticipated clinicians' ultimate detection of treatment outcome in most cases.

We believe that the EMA methodology might be a useful tool to inform decisions about subject eligibility for randomized clinical trials, particularly for studies of rapidly acting antidepressants. Our findings suggest that EMA can track the treatment course of depressed individuals more effectively than the weekly or bi-weekly clinician measurements that are typical of MDD trials. It is noteworthy that EMA is a remote tool than can eliminate the need for some in-person clinic visits. Further, the integration of EMA measures that examine mood dynamics, assess negative and positive affect, and use symptomatic rating scales like the HamD$_6$ used in this study might help to differentiate sub-groups of depressed individuals, particularly those who are considering pharmacological study participation (Houben et al., 2015; Panaite et al., 2020). Additional studies that include larger participant samples and examine multiple time points are needed to fully explore the potential utility of the EMA methodology for clinical trials.

## References

Aan het Rot, M., Hogenelst, K., Schoevers, R.A., 2012. Mood disorders in everyday life: a systematic review of experience sampling and ecological momentary assessment studies. Clin. Psychol. Rev. 32 (6), 510–523.

American Psychiatric Association, 2013. Diagnostic and Statistical Manual of Mental Disorders, fifth ed. American Psychiatric Press, Arlington VA.

Armey, M.F., Schatten, H.T., Haradhvala, N., Miller, I.W., 2015. Ecological momentary assessment (EMA) of depression-related phenomena. Current Opin Psychol 1 (4), 21–25.

Barge-Schaapveld, D.Q., Nicholson, N.A., 2002. Effects of antidepressant treatment on the quality of daily life: an experience sampling study. J. Clin. Psychiatr. 63 (6), 477–485.

Bech, P., Gram, L.F., Dein, E., Jacobsen, O., Vitger, J., Bolwig, T.G., 1975. Quantitative rating of depressive states. Acta Psychiatr. Scand. 51, 161–170.

Bech, P., 2006. Rating scales in depression: limitations and pitfalls. Dialogues Clin. Neurosci. 8 (2), 207–215.

Ben-Zeev, D., Young, M.A., Madsen, J.W., 2009. Retrospective recall of affect in clinically depressed individuals and controls. Cognit. Emot. 23, 1021–1040.

Bos, F.M., Schoevers, R.A., Aan het Rot, M., 2015. Experience sampling and ecological momentary assessment studies in psychopharmacology: a systematic review. Eur. Neuropsychopharmacol 25 (11), 1853–1864.

Cleary, P.G.W., Guy, W., 1977. Factor analysis of Hamilton depression scale. Drugs Exp. Clin. Res. 1, 115–120.

Depp, C.A., Kim, D.H., de Dos, L.V., Wang, V., Ceglowski, J., 2012. A pilot study of mood ratings captured by mobile phone versus paper and pencil mood charts in Bipolar disorder. J. Dual Diagn. 8 (4), 326–332.

Ebner-Priemer, U.W., Trull, T.J., 2009. Ecological momentary assessment of mood disorders and mood dysregulation. Psychol. Assess. 21 (4), 463–475.

Geschwind, N., Nicolson, N.A., Peeters, F., van Os, J., Barge-Schaapveld, D., Wichers, M., 2011. Early improvement in positive rather than negative emotion predicts remission from depression after pharmacotherapy. Eur. Neuropsychopharmacol 21 (3), 241–247.

Granholm, E., Loh, C., Swendsen, J., 2008. Feasibility and validity of computerized ecological momentary assessments in schizophrenia. Schizophr. Bull. 34 (3), 507–514.

Granholm, E., Holden, J.L., Mikhael, T., Link, P.C., Swendsen, J., Depp, C., Moore, R.C., Harvey, P.D., 2020. What do people with schizophrenia do all day? Ecological Momentary Assessment of Real-World Functioning in Schizophrenia. Schizophr. Bull. 46 (2), 242–251.

Guy, W., 1976. Early Clinical Drug Evaluation (ECDEU) Assessment Manual for Psychopharmacology. National Institute of Mental Health, Rockville, Md. Publication No. 76-338.

Hamilton, M., 1960. A new depression scale designed to be sensitive to change. Br. J. Psychiatry 134, 382–389.

Harvey PD, Miller ML, Depp C, Moore RC, Pinkham AE. (in press). Capturing clinical symptoms with ecological momentary assessment: convergence of momentary reports of psychotic and mood symptoms with diagnoses and standard clinical assessments. Innov Clin Neurosci.

Harvey, P.D., Pendergrass, C., Nemeroff, C.B., 2020. Challenges in the Assessment of Sleep Impairment in Cognitively Impaired Patients Can Be by Passed through Ecological Momentary Assessment. SAGE Research Methods Cases: Medicine and Health. https://doi.org/10.4135/9781529726428.

Hooper, C.L., Bakish, D., 2000. An examination of the sensitivity of the six-item Hamilton rating scale for depression in a sample of patients suffering from major depressive disorder. J. Psychiatry Neurosci. 25 (2), 178–184.

Houben, M., Van Den Noortgate, W., Kuppens, P., 2015. The relation between short-term emotion dynamics and psychological well-being; a meta-analysis. Psychol. Bull. 141 (4), 901–930.

Ibm Corporation, 2019. Statistical Package for the Social Sciences version 26 (SPSS-26): Armonk, NY.

Kramer, I., Simons, C.J.P., Hartmann, J.A., Menne-Lothmann, C., Viechtbauer, W., Peeters, F., Schruers, K., van Bemmel, A.L., Myin-Germeys, I., Delespaul, P., van Os, J., Wichers, M., 2014. A therapeutic application of the experience sampling method in the treatment of depression: a randomized controlled trial. World Psychiatr. 13, 68–77.

Lamers, F., Swendsen, J., Cui, L., Husky, M., Johns, J., Zipunnikov, V., Merikangas, K.R., 2018. Mood reactivity and affective dynamics in mood and anxiety disorders. J. Abnorm. Psychol. 127 (7), 659–669. https://doi.org/10.1037/abn0000378.

Marzano, L., Bardill, A., Fields, B., Herd, K., Veale, D., Grey, N., Moran, P., 2015. The application of mHealth to mental health: opportunities and challenges. The Lancet Psychiatry 2 (10), 942–948.

Montgomery, S., Asberg, M., 1979. A new depression scale designed to be sensitive to change. Br. J. Psychiatr. 134, 382–389.

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. J. Psychiatr. Res. 75, 116–123.

Morris, D.W., Rush, A.J., Jain, S., Fava, M., Wisniewski, S.R., Balasubramani, G.K., Khan, A.Y., Trivedi, M.H., 2007. Diurnal mood variation in outpatients with major depressive disorder: implications for DSM-V from an analysis of the Sequenced Treatment Alternatives to Relieve Depression Study data. J. Clin. Psychiatr. 68 (9), 1339–1347.

Moskowitz, D.S., Young, S.N., 2006. Ecological momentary assessment: what it is and why it is a method of the future in clinical psychopharmacology. J. Psychiatry Neurosci. 31 (1), 13–20.

O'Sullivan, R.L., Fava, M., Agustin, C., Baer, L., Rosenbaum, J.F., 1997. Sensitivity of the six-item Hamilton depression rating scale. Acta Psychiatr. Scand. 95, 379–384.

Panaite, V., Rottenberg, J., Bylsma, L.M., 2020. Daily affective dynamics predict depression symptom trajectories among adults with major and minor depression. Affective Sci 1, 186–198.

Peeters, F., Berkhof, J., Delespaul, P., Rottenberg, J., Nicolson, N.A., 2006. Diurnal mood variation in major depressive disorder. Emotion 6 (3), 383–391.

Peeters, F., Berkhof, J., Rottenberg, J., Nicolson, N.A., 2010. Ambulatory emotional reactivity to negative daily life events predicts remission from major depressive disorder. Behav. Res. Ther. 48, 754–760.

Santangelo, P.S., Ebner-Priemer, U.W., Trull, T.J., et al., 2013. Experience sampling methods in clinical psychology. In: Comer, J.S., Kendall, P.C. (Eds.), The Oxford Handbook of Research Strategies for Clinical Psychology. Oxford University Press, Oxford, UK, pp. 188–210.

Sauder, C., Aghjayan, A., Engler, J., Evans, M., Vu, A., Domilici, D., Rutrick, D., 2019. The Potential Benefits of Utilizing a Standard of Care Treatment Study to Prepare Participants for Enrollment in a Treatment-Resistant Depression (TRD) Clinical Trial. ISCTM poster presentation, Copenhagen. September 2019.

Sheehan, D.V., Lecrubier, Y., Sheehan, K.H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., Dunbar, G.C., 1998. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J. Clin. Psychiatr. 59 (Suppl. 20), 22–33 quiz 34-57.

Solhan, M.B., Trull, T.J., Jahng, S., Wood, P.K., 2009. Clinical assessment of affective instability: comparing EMA indices, questionnaire reports, and retrospective recall. Psychol. Assess. 21 (3), 425–436.

Targum, S.D., Wedel, P.C., Bleicher, L.S., Busner, J., Daniel, D.S., Robinson, J., Rauh, P., Barlow, C., 2013. A comparative analysis of centralized, site-based, and patient ratings in a clinical trial of Major Depressive Disorder. J. Psychiatr. Res. 47, 944–954.

Wenze, S.J., Miller, I.J., 2010. Use of ecological momentary assessment in mood disorders research. Clin. Psychol. Rev. 30 (6), 794–804.

Wirz-Justice, A., 2009. Diurnal variation of depressive symptoms. 2008. Dialogues Clin. Neurosci. 10 (3), 337–353.