2019

When all else fails, listen to the patient: A viewpoint on the use of ecological momentary assessment in clinical trials

Aaron M. Mofsen
Washington University School of Medicine in St. Louis

Thomas L. Rodebaugh
Washington University in St. Louis

Ginger E. Nicol
Washington University School of Medicine in St. Louis

Colin A. Depp
University of California - San Diego

J. P. Miller
Washington University School of Medicine in St. Louis

See next page for additional authors

Viewpoint

When All Else Fails, Listen to the Patient: A Viewpoint on the Use of Ecological Momentary Assessment in Clinical Trials

Aaron M Mofsen[1], DO; Thomas L Rodebaugh[2], PhD; Ginger E Nicol[1], MD; Colin A Depp[3], PhD; J Philip Miller[4], AB; Eric J Lenze[1], MD

[1]Department of Psychiatry, School of Medicine, Washington University in St Louis, St Louis, MO, United States

[2]Department of Psychological and Brain Sciences, Washington University in St Louis, St Louis, MO, United States

[3]Department of Psychiatry, University of California - San Diego, San Diego, CA, United States

[4]Division of Biostatistics, School of Medicine, Washington University in St Louis, St Louis, MO, United States

Corresponding Author:
Colin A Depp, PhD
Department of Psychiatry
University of California - San Diego
9500 Gilman Drive
San Diego, CA, 92093
United States
Phone: 1 8588224251
Email: cdepp@ucsd.edu

Abstract

A major problem in mental health clinical trials, such as depression, is low assay sensitivity in primary outcome measures. This has contributed to clinical trial failures, resulting in the exodus of the pharmaceutical industry from the Central Nervous System space. This reduced assay sensitivity in psychiatry outcome measures stems from inappropriately broad measures, recall bias, and poor interrater reliability. Limitations in the ability of traditional measures to differentiate between the trait versus state-like nature of individual depressive symptoms also contributes to measurement error in clinical trials. In this viewpoint, we argue that ecological momentary assessment (EMA)—frequent, real time, in-the-moment assessments of outcomes, delivered via smartphone—can both overcome these psychometric challenges and reduce clinical trial failures by increasing assay sensitivity and minimizing recall and rater bias. Used in this manner, EMA has the potential to further our understanding of treatment response by allowing for the assessment of dynamic interactions between treatment and distinct symptom response.

(JMIR Ment Health 2019;6(5):e11845) doi:10.2196/11845

KEYWORDS

ecological momentary assessment; mental health; controlled clinical trial; psychiatry; health technology

Introduction

Background

Mental health treatment development and testing has been at an impasse for the past several decades; our clinical trials increasingly fail more often than in other fields [1]. Although the global burden of psychiatric illness continues to be one of the largest contributors to disability worldwide, investment in the discovery of novel pharmacologic agents flows instead toward disease states with identifiable biological targets. These targets remain elusive in psychiatric disorders [2,3]. The central nervous system (CNS) drug development pipeline has become increasingly burdened with late-phase failures [4], contributing to a well-publicized exodus of the pharmaceutical industry from the CNS space. This has resulted in decreased investment in drug discovery [5].

Treatment Failures: Bad Medicine or Bad Measures?

The randomized, placebo-controlled trial is still considered the gold standard test of treatment efficacy. However, over the past 60 years of treatment research in psychiatry, we have observed that treatment effect sizes remain stable, whereas placebo responses rise [6]. Modern clinical trials are difficult to conduct and are fraught with numerous challenges related to cost, regulatory requirements, recruitment difficulties, and other inefficiencies [7,8]. Added to these challenges is the use of imprecise outcome measures, which hinders the ability to detect true separation of active treatment from placebo response [9].

The contribution of poor measures to treatment failures is particularly well-illustrated in antidepressant trials [10-12]. For example, lanicemine, an N-methyl-D-aspartate receptor antagonist differing from ketamine that produces lower psychotomimetic side effects, was thought to show promise in treating depression [13]. Early phase clinical trials showed promising results in rapidly reversing symptoms of treatment resistant depression, but investigators failed to replicate the results in a late phase study [14]. Similarly, basimglurant, a postsynaptic metabotropic glutamate subtype 5 receptor antagonist, showed promise in early phase trials but failed to separate from placebo on the primary outcome measure in a larger phase 2b trial [15]. In both cases, the primary end point was change from baseline to 6 weeks in the Montgomery Asberg Depression Scale (MADRS), which is considered an industry standard in depression treatment research. The authors identified flaws in study design, conduct, and even underlying scientific rationale as possible causes of these late stage failures.

It seems unlikely, given the financial and intellectual resources brought to bear in the early phases of discovery, that investigators could have gotten the scientific rationale so wrong. A more probable explanation for the failed studies might lie in how the primary outcome was determined and measured. Although the MADRS is considered a standard assessment tool in depression research, poor interrater reliability (ie, imprecision of measurement) is one of many limitations to this measure's assay sensitivity.

The Culprit: Faulty Signal Detection

Measurement *assay sensitivity*, as it applies to clinical research, refers to the ability of a symptom assessment measure to detect whether a difference exists between treatment groups [16]. Issues of assay sensitivity are well known in psychiatric treatment research and have been observed with older self-report scales such as the Hamilton Rating Scale for Depression (HAM-D) as well as in newer clinician-administered instruments such as the MADRS. Both measures include several symptom domains but offer only a final summed score. This offers little insight into the specific symptoms underlying the clinical presentation.

Self-report measures may incorporate reporter bias, whereas clinician-administered assessments incorporate bias on the part of the clinician. For example, there may be bias in recruitment or sample ascertainment, such as career patients who serially enroll in research studies for financial reasons and are thus motivated to answer questions in such a way as to increase likelihood of enrollment. Investigators may unconsciously inflate baseline measures of psychiatric symptoms to meet recruitment goals [17-19].

Nonetheless, these arguments fail to explain why academic studies, in which less financial gain accrues to the patient and investigator, also see a high placebo response and failure rate [20]. Regardless, reduced assay sensitivity in clinical trials has the potential to sabotage treatment development at any stage. We submit that these and other depression symptom measures reduce assay sensitivity in 3 primary ways: unnecessary complexity, human error (ie, clinician judgment), and infrequent sampling.

Getting to Precision Assessment

The idea of using technology to increase the accuracy and precision of symptom assessment in clinical trials is gaining momentum. For example, the National Institutes of Health toolbox was designed specifically for this purpose [21]. The Patient-Reported Outcomes Measurement Information System also offers researchers standardized patient-reported outcome (PRO) measurement tools with transparent performance metrics [22]. Self-report measures delivered via mobile technology certainly offer ecological validity and may also prove superior to clinician-administered instruments in large, industry-funded clinical trials. Improved measurement would likely translate into more useful clinical trials. It may even go a long way toward surmounting our present impasse in developing new mental health treatments.

Clearly, we are not the first to contemplate the problem of assay sensitivity in our field. However, public discussion as to why progress in the field of psychometrics has stalled has not extended to industry trials. Open scientific discourse has also been limited on the subject of developing novel, effective, Food and Drug Administration (FDA)–sanctioned instruments, which could be used to track mental health disorder outcomes with greater assay sensitivity. As the success or failure of antidepressant treatment trials often rests solely on the presumed validity and reliability of symptom measures, it should follow that these assessments deserve the same degree of scrutiny regarding assay sensitivity as any laboratory test.

In this viewpoint, we will examine 3 major problem areas we believe the field needs to address in getting to precision assessment: overly complex assessment tools, contributions of human error, and limitations of infrequent sampling. First, we will review the 2 gold standard depression instruments used at present to track psychiatric symptoms in industry-funded drug trials. Next, we will examine the role of clinician assessment and how human involvement in measurement contributes to error. We will then discuss challenges to adequate measurement frequency in obtaining valid self-report data. Finally, we propose a solution to the measurement problem in depression clinical trials. We will explore contributions from the fields of mathematics, human psychology, and computer science to the development of mobile technology–based measures, which we believe may offer significant improvements over traditional symptom assessment.

Problem 1: Needless Complexity Undermines Utility

Key point:

- Overly broad measures that attempt to cover multiple symptoms or symptom domains compromise signal detection. To meaningfully reduce error, consensus on what to measure is needed.

The Problem of Excessive Description

Psychiatric rating scales frequently use diagnostic criteria or descriptive psychopathology to track a patient's progress throughout a clinical trial. The descriptive psychopathology for

a given psychiatric disorder is by nature more expansive than the diagnostic criteria alone, which can be helpful for identifying clinically significant features for treatment targets. This problem is not restricted to mental health research; trials in cardiology have also been compromised by failing to adequately confine outcome measures for meaningful signal detection [23]. In major depression, patients often have irritability, anxiety, and other symptoms in addition to the 9 cardinal symptoms of the disorder. A content analysis by Eiko Fried found 52 symptoms of depression across 7 commonly used depression scales, with a content overlap among all scales of only 32 percent [24].

Take for example the MADRS discussed above [25]. The clinician in using this scale administers a 10-item assessment to a study participant. The change in the total score over time is then used to determine whether the treatment under investigation is effective. The 17-item HAM-D (HAM-D-17) determines efficacy similarly [26]. However, both items assess multiple symptom domains, all considered diagnostic aspects of depression. A recent study by Checkroud et al [27] of over 7000 patients with major depression demonstrates why this approach, as well as any other that relies on indiscriminate use all of the items in a scale to assess primary efficacy outcomes (eg, the HAM-D), may be a problem. In their study, they illustrate how this indiscriminate approach to measurement can jeopardize a potential treatment in late-phase clinical trials. Specifically, they found that consistent antidepressant treatment response was found only for the core emotional symptoms (anergia, dysphoria, anhedonia, feelings of worthlessness, and difficulty concentrating). The detectable signal for treatments shown to be effective is thus obscured by the total score, which is the only score considered when designing trials to determine efficacy. This example highlights how standard rating scales have contributed to treatment failures by introducing unnecessary *complexity*, which reduces measurement specificity.

To further complicate matters, measuring multiple constructs inflates the chance that items tied to each construct will shift unpredictably over time (eg, due to lack of longitudinal factorial invariance) [28]. In this way, depression rating scales are often a mix of sensitive and specific items (dysphoria, anhedonia), nonspecific items (anxiety), and symptoms that may be derived from an unrelated illness (eg, fatigue). Side effects of the treatment itself are also frequently conflated with the items in the primary outcome measure. Moreover, individual items within a scale are often not weighted for relevance. As the success or failure of a treatment rests on a scale's summative score, it follows that some of the score's equally weighed items might be totally irrelevant to the trajectory of the disorder in question [29]. The 24-item HAM-D (HAM-D-24) is more comprehensive than the 17-item version [30]. It was designed to more comprehensively capture relevant symptoms. However, using the HAM-D-24 may conceal treatment effects by introducing items that assess uncommon or diagnostically nonspecific symptoms, such as hypochondriasis or depersonalization. Again, as the total score is used to determine whether or not a treatment is effective, there is a further risk of magnifying irrelevant changes and obscuring important ones.

Less is More

The shortened 6-item HAM-D and MADRS scales, which favor core items such as low mood, anhedonia, and guilt, have both been shown to be more sensitive than HAM-D-17 and the 10-item MADRS, respectively [31]. The shorter 6-item version of the HAM-D [32] was superior to the longer HAM-D-17, 21 and 24 in detecting treatment response to the newer antidepressant vortioxetine versus placebo [33]. Similarly, the buprenorphine/samidorphan combination treatment, which failed to separate from placebo on the primary outcome measure of change from baseline on the MADRS-10 item scale, fared better in separating from placebo using the MADRS-6 item scale [34]. These examples suggest a data reduction approach to symptom assessment focusing on core symptoms is more likely to accurately detect meaningful clinical response. Unfortunately, there is, as of yet, little agreement on which symptoms are most relevant.

Consensus on the most clinically, functionally, or personally relevant features of treatment response or remission is needed to improve signal detection. If we simply wish to use our existing scales more pragmatically, we would take a treatment we know to be effective and choose the individual items from a selected scale that reveal the greatest amount of separation in favor of the proven treatment. We would then use the items from that same scale to determine whether or not an unproven treatment is effective. Alternatively, the field could adopt a universal consensus around measuring the core emotional symptoms of the illness to determine treatment success or failure. This is a difficult and unlikely scenario as we do not have the evidence base at present necessary to establish what exactly these core symptoms might be. In either case, improvement from a functional or pharmacoeconomic perspective may not map well onto any of the items in the measures we currently use. This may force the field to revisit some of its a priori assumptions about clinical relevance. In short, although we can confidently say that our current approach is suboptimal, fixing it will not be so easy.

Problem 2: Human Error Magnifies Measurement Error

Key points:

- Clinician-administered scales compound response bias
- Self-report alone is imperfect but minimizes rater contribution to measurement error

Not All That Glitters is Gold

Psychiatric treatment research has traditionally considered clinician-administered assessments to be the *gold standard* over PRO measures. This stems in part from an inherent belief that the clinician *objectively corrects* for whatever error (eg, errors of omission, exaggeration, expectancy effect, and Hawthorne effect), intentional or otherwise, introduced by the patient. Perhaps somewhat counterintuitively, clinicians may *magnify* the patient's error. A large study evaluating self-report and clinician-administered instruments from the Sequenced Treatment Alternatives to Relieve Depression trial found that self-report measures contributed more to the prediction of

outcomes of clinician-administered instruments than vice versa [35]. The authors of the study also recommended that, in the event that only 1 form of assessment could be used, self-reported outcome measures would be preferable.

Error or bias on the part of the clinician is routine, rather than idiosyncratic. It would be unfair to presume it to be the result of malice or laziness. It may happen unconsciously and even in good faith because clinical judgment is not completely objective. Interviewers are also susceptible to either a positive or negative rater bias depending on whether research participant attributes, often irrelevant to the assessment at hand, are perceived as positive or negative. This can result in sometimes pronounced unconscious alterations of judgment [36] that significantly impact clinical decision making. This has been illustrated in studies finding poor interrater and test-retest reliability in standard clinician-administered assessment measures for depression [3]. The reason for such results may be that clinicians, even when given rules governing the scoring of the assessment at hand, will tend to drift from standard calibrated practice [37]. Whether or not a clinician reliably follows an assessment-related rule depends on the amount of inertia that must be overcome to adopt it, the format in which the rule was originally presented, the number of demands that compete with the rule, and the institutional pressures involved in maintaining compliance with the rule [38].

When all Else Fails, Listen to the Patient

Although the evidence is still far from conclusive, a decent body of literature has elevated the stature of PROs vis-a-vis traditional, clinician-administered rating scales. Self-report assessments represent an improvement over clinician-administered assessments insofar as they eliminate rater bias and reduce the likelihood that participants will feel compelled to give socially desirable responses (a type of response bias) or affirmative answers when interviewed face-to-face [39]. For example, a large meta-analysis of placebo response in 96 antidepressant trials by Mora et al found that clinician-administered instruments were associated with a higher placebo response than PRO measures [40]. Such evidence further supports the idea that clinician-administered scales add error rather than removing or mitigating patient error. In summary, although we place a high value on clinician-administered assessments, clinician objectivity may be more of an appealing myth than reality.

Problem 3: Infrequent Sampling Hurts Sensitivity

Key points:

- Retrospective patient symptom report in the context of a clinical trial may be inaccurate
- Ecologically valid symptom reports collected in real time are needed to interpret treatment effects

(Not So) Total Recall

Self-report also has inherent limitations. This was recognized by Arthur Schopenhauer in the 19th century [41], who observed that one cannot be both the subject and object of accurate perception. Thus, reporting on one's own mood even in the present poses significant challenges and represents an irremediable layer of error. Mehl and Conner have also comprehensively discussed the problem of recall bias in psychological research [42]. In short, asking a participant to provide a retrospective symptom report merely compounds this error by introducing recall bias. In other words, emotional recall bias (unlike the subject-object problem) is a controllable source of error. Neuroscientists have found memory to be frequently unreliable, particularly when the encoding and retrieval of memories occurs during periods of emotional arousal [43]. Memory has many odd biases, not all of which are evident in daily life. For instance, it has been shown that people have a tendency to remember events that ought to be enjoyable, such as a vacation or spending time with one's children, as being more pleasant than they actually were [42]. Thus, asking a respondent to recall something requires filtration through whatever emotional state the subject happens to be in at the time of the assessment, which only compounds this error [44]. Furthermore, respondents are unlikely to accurately create a coherent summary of their emotional states over time.

What is the (Right) Frequency?

Infrequent measurement or sampling in clinical trials tacitly makes the assumption that we know enough about how an illness behaves over time to ask questions with a time frame modifier (eg, "In the last week...") and is associated with measurement error in clinical trials. This has been illustrated in disciplines outside of psychiatry. For example, the Heart Outcomes Prevention Evaluation trial evaluated the effect of the angiotensin-converting enzyme inhibitor ramipril in patients at high risk for adverse cardiovascular events [45]. The study found that ramipril lowered blood pressure assessed via 24-hour ambulatory measurement, whereas office-based blood pressure measurements did not detect the treatment response. Investigators attributed this to a diurnal variation in blood pressure or *white coat hypertension* —phenomena that could not be captured with the limited number of measures obtained during office hours or that were affected by the office visit itself. For this reason, blood pressure assessment in clinical trials has moved to using frequent ambulatory blood pressure sampling to assess treatment efficacy, which has essentially eliminated the placebo response in antihypertensive treatment trials [46,47].

Similar to blood pressure, depressive symptoms also appear to fluctuate throughout the day or in response to specific situations [48]. Mobile technology offers a feasible way to increase sampling frequency, as evidenced by the already rich scientific literature on ambulatory assessment [42]. However, this approach has yet to be fully embraced by industry sponsored studies, where it could be of prime utility. To date, only 1 industry-sponsored study currently underway has attempted to compare daily, ambulatory self-report with a clinician-administered measure [49]. Frequent, in-the-moment self-report also has its limitations. There is no doubt some theoretical limit on high-frequency sampling to the extent that it may, if administered often enough, conflate mood and emotions or succeed in becoming itself a source of negative mood, affect, or emotions [50,51]. However, this issue calls for

careful experimentation with frequency to assess acceptability rather than avoiding frequent sampling altogether.

The State Versus Trait Problem

Symptoms of many psychiatric illnesses are characterized as trait-like in advance of any evidence to support this assumption. However, variation is routinely observed in behaviors studied over time, irrespective of how trait-like they seemed to be (eg, personality traits such as sociability) [52]. For this reason, it is highly probable that important variation is the rule rather than the exception in psychiatric illness. For example, in an individual with major depression, mood might be very depressed at a certain point in the morning and near-normal later that same day [48].

Despite this, we continue to measure mood as a stable trait-like symptom (eg, "in the last 7 days, how has your mood been?"). This is the case for most psychiatric symptom assessments, where dynamic versus stable or trait-like nature of symptoms are poorly described. The only way to ascertain variation or lack thereof is to sample the illness frequently *before finalizing the measure (eg, for use in a treatment study).* In other words, frequent sampling would ideally be used to inform the creation of a scale before using it to track efficacy [52]. Without this approach, scale selection becomes thoughtlessly reflexive [50]. Limited sampling likely further compromises psychiatric research because trait measures require respondents to attempt a summation of states via recall of past experiences, which has been shown to introduce error [53].

Even if the symptoms of psychiatric illness are predominantly trait-like, we would continue to favor frequent sampling, even if this requires us to use a smaller number of items. This is in contrast to classical test theory, from which we take the maxim that adding equally good items to a measure leads to greater reliability and therefore, a better shot at validity [54]. This is based on the ideal circumstance where it is possible to ask a respondent the same question repeatedly, which we cannot do at a single time point without expecting the respondent to become reactive to the question [54,55]. Furthermore, a measure using high-signal items repeatedly over time would better capture any given quality than would a measure with a mix of items with lower signal detection at a single time point [56]. In psychiatric treatment research, we have historically chosen to use a greater number of inferior items at a single time point, even though the maxim we are following was based on equations that are arguably better suited to repeated measurement of a single quality.

Solution: Ecological Momentary Assessment

Overview

Ecological momentary assessment (EMA) is frequent, real time, patient-reported assessment delivered via surveys (eg, "right now, my mood is...") and completed by the patient typically via mobile device to collect information about the patient in a real-world setting [57]. Participants are prompted at prespecified intervals to complete symptom assessments rather than having a prompt dependent upon a passive event (eg, actigraphy and patterns of speech). EMA may overcome the deficiencies inherent in traditional clinician-administered instruments. Evidence from pain studies examining EMA alongside retrospective recall show a consistent discrepancy between the 2 forms of report [58]. A similar discrepancy between real time and retrospective self-report of affect has also been demonstrated [59]. A single item scale measuring mood delivered via EMA outperformed the HAM-D-17 in its ability to predict "current relapse status" in patients with major depressive disorder [60].

Increasing Accuracy in Early Phase Trials

Frequent, real-time EMA sampling has been shown in the same study to both qualify positive findings in clinical trials and detect treatment effects that the HAM-D was unable to detect between groups after 18 weeks of treatment [61]. Frequent real-time sampling has also been shown to unmask differences between treatment responders and nonresponders and to detect treatment effects earlier than clinician-administered assessments [62,63]. Finally, frequent, real-time sampling compared with retrospective assessment has been shown to increase the precision of measurement over time.

An example of how infrequent sampling adversely affects assay sensitivity in clinical trials was recently provided by Moore et al [64]. In this study, the researchers assessed the effects of mindfulness-based stress reduction (MBSR), compared with an attention placebo. For outcome assessments, they measured depressive symptoms, anxiety symptoms, and mindfulness self-ratings in 2 ways: EMA tools delivered to participants electronically via a smartphone 3 times daily for 14 days and traditional paper- and pencil-based measurement tools asking about last week's symptoms (comparable with most outcome measures). The EMA-based outcome assessment resulted in a much lower number needed to treat (NNT) for MBSR than the same outcomes measured using the traditional technique: the NNT for treating depression was 8 using EMA versus 31 using traditional measurement. In other words, EMA captured a treatment effect that was missed by standard self-report assessments. This was also reflected in the smaller SDs for outcomes measured via EMA when averaged over time. In short, frequent ambulatory assessment improves precision.

Increased Understanding of Core Symptom Constructs

EMA may also increase measurement precision by tracking how symptoms of an illness behave and interact over time [65]. This allows investigators to characterize state versus trait-like symptoms and establish the nature of the relationships between symptoms over time. This approach may also be useful because it offers the ability to evaluate interactions between symptoms without first assuming that they are symptoms *of the disorder in question.* This "pragmatic nihilism" [66] or "symptomic" [67] approach differs from how we currently assess psychiatric disorders. Clinician-administered instruments are rated with the built-in assumption that any number of symptoms are all tied to 1 underlying, latent variable (eg, depression). With enough patient-reported EMAs carried out over time, investigators may be able to observe how symptoms interact with one another.

It may also be possible to discern which symptoms are central to the disorder under study and how certain upstream symptoms

may influence a cascade of symptoms downstream. How many EMAs are *enough* depends on the exact questions being asked and the assumptions made in the analysis; however, it is likely that as little as 25 measurements from hundreds of participants or a hundred measurements in even a small number of participants would be a reasonable starting place [68]. Such findings may eventually afford researchers the unique opportunity to stratify clinical trial participants based on *how* they do or do not get better rather than simply whether or not they get better. The approach becomes highly descriptive at the level of the individual, thereby allowing one to answer a host of previously unanswerable questions.

Deconstructing Treatment Response

Another question that might be asked is whether patients responding to an intervention or placebo get better in the same way. In other words, do the *temporal dynamics* of placebo response differ from that observed in drug response? Temporal dynamics here refer to certain discernable patterns in the EMA data that allow a researcher to broadly classify a patient as displaying, for instance, affective inertia (symptoms strongly relate to themselves over time, resulting in less change over time), affective instability (symptoms vary a great deal over time), or inability to differentiate between symptoms (as 1 symptom gets better or worse the rest tend to follow) [69]. This is by no means an exhaustive list of questions that may be asked of the data derived from EMA. It is safe to say EMA has the potential to offer a renaissance of sorts in descriptive psychopathology and may even allow for veritable *personalized medicine* given the types of patterns and points of intervention it is able to reveal.

EMA may also help us detect the phenomenon of regression to the mean. This phenomenon occurs when a baseline assessment of symptoms in a clinical research study is inflated at the initial visit before regressing to where those symptoms normally *live*. This is thought to significantly impact the ability to detect separation whenever it occurs in the placebo group. Using EMA, patients may be monitored in the outpatient setting not simply for clinical research purposes but rather to give the clinician a better idea of whether or not a patient is getting better. This approach appreciates EMA as an instrument that may be used to conduct field research, which is thought to have better "ecological validity" than assessments delivered within the artificial environment of the clinical trial site [42]. Such real-world information could be used to find out where that patient "lives" if a patient is being screened for a clinical research study. Similarly, it is not difficult to envision tailoring inclusion/exclusion criteria to this end. If and when this does take place, CNS research will be indebted to data provided directly by the patient.

Developing Better Interventions

Once individual symptom characteristics are known, targeted interventions can be developed. For instance, if insomnia leads to anergia the following day, which in turn leads to anhedonia, one might examine whether applying an intervention at the onset of insomnia changes the observed course of symptomatology downstream. This sort of intervention is called an ecological momentary intervention (EMI) because it relies on EMA or a just-in-time adaptive intervention. An EMI is an intervention informed by data gathered by EMA. We can already find examples of researchers using EMA data to provide an EMI. For example, EMI has already been shown to be very successful in providing patients with substance use disorders relapse prevention tools precisely when they need it the most [70]. It is conceivable that EMA scales, in addition to providing efficacy outcomes with increased assay sensitivity, may also reveal novel points of intervention in clinical trials.

Multiple methods, including multilevel vector autoregression and multilevel dynamic structural equation modeling, can help researchers examine how individuals may vary from group trends over time [71,72]. This might allow clinicians to tailor a personalized EMI based on a patient's own unique pattern of EMA data. To take this idea further still, EMA may eventually be able to offer the unique ability to evaluate whether a target is being addressed by an intervention via *real-time* lagged mediation rather than post hoc analyses. In other words, we would be able to use real-time lagged mediation to see whether or not we are actually engaging a chosen target precisely when we are attempting to target it.

The use of EMA to gather the data needed to deliver a just-in-time EMI is also consistent with the concept of target engagement raised by the National Institute of Mental Health in an effort to address the declining success of clinical trials in mental health. A target is defined as something "molecular, cellular, circuit, behavioral or interpersonal, commensurate with the intervention," which is expected to be changed in some way by the intervention being studied [73]. The concept of target engagement is closely related to a recent call for a research focus on symptomics or the examination of "symptom-specific effects" [70]. Such a focus, as represented in the example above, may allow us to identify those key symptoms that tend to precede or perhaps even cause other symptoms. Investigating patterns of interaction between symptoms in this way may help us to understand some of the underlying causes of complex psychiatric illnesses.

How Do We Get to Widespread Use of Ecological Momentary Assessment in Clinical Trials?

Understanding and Getting Past Limitations

Although smartphone ownership is not universal, it is increasing, particularly among individuals with psychiatric conditions. John Torous found in a recent survey of 457 individuals with schizophrenia or schizoaffective disorder that greater than half (54%) of such individuals owned a smartphone [74]. Perhaps a greater question then is whether a participant with a smartphone would want to use it to regularly quantify his or her depressive symptoms. User privacy is also becoming an increasingly important issue as faith in *big tech* to safeguard users' privacy has waned in the wake of the numerous scandals. Getting around these limitations may require sponsors to invest in low-cost devices participants can use while enrolled in trials.

Use of EMA in the real world often leads to missing data that have historically made analysis problematic. Users may not be

compliant with the number of surveys they are required to complete in a timely manner, and, as described above, frequency of assessments increase precision only up to a point. Beyond this point, with too frequent assessment, the risk increases of either introducing noise by sampling irrelevant aspects of the human condition or of the assessment itself becoming a negative part of the intervention. Investigators will have to consider an assay sensitivity assessment as part of the startup process to determine how the target population will best respond to EMA.

Although the FDA has made its expectations for PRO measures clear [75], it is not at all clear whether every aspect of FDA guidance will neatly translate to electronic PROs. For example, to what extent, if any, would necessary software updates for an accepted EMA app involve the FDA? FDA guidance for evaluating antidepressant drugs has not been updated since 1977 and explicitly favors selecting scales that have been previously used in drug trials over ones that are novel [76]. This effectively prioritizes tradition over innovation and creates a catch-22 for researchers who might otherwise break with the status quo. Clinician-administered instruments need to be evaluated alongside commensurate EMA-delivered items. This will help us to determine parameters such as the optimal sampling frequency but will likely also be necessary as the FDA typically reports correlation coefficients for established measurement tools [77].

The conceptualization of disorders based on Diagnostic and Statistical Manual of Mental Disorders/International Classification of Diseases criteria has been called into question and may eventually be replaced altogether by Research Domain Criteria [78]. Although EMA is in many ways conducive to a dimensional approach to mental illness, this migration would obviously require a new approach to EMA scale creation and validation. In this case, the role of EMA may be to supplement observable behaviors with self-report.

EMA may not be ideal for detecting rare events, especially if they occur infrequently relative to the sampling frequency (ie, as the sampling frequency decreases so too does the probability of capturing *rare events*). Thus, when and how to apply EMA in clinical trials remains an area requiring additional study and consensus development.

EMA should not be mistaken for a panacea so long as p-hacking, publication bias, and alpha inflation continue to affect the integrity of clinical research. Any scale used to evaluate the efficacy of an intervention in large industry-sponsored clinical trials must be uniform and well-validated. Thus, to create a standard efficacy measure for a given psychiatric disorder, we first must form a consensus about the types of items that should be included in the EMA scales, the frequency and duration of assessments, and the types of analytical approaches that will be used to interpret the data. The FDA would be unlikely to accept an EMA-based primary outcome measure over existing efficacy end point measures without standardization across multiple field trials in different populations. These data should then clearly establish test-retest reliability, external validity, and other parameters necessary to validate an EMA scale.

Conclusions

Moving from clinician-administered rating scales toward real-time patient-reported measures such as EMA offers significant advantages across medical settings. In clinical research studies, EMA may reduce placebo response and increase intervention-placebo separation. EMA also offers an obvious advantage over clinician-administered rating scales in inpatient and community settings given that time, cost, and staff pressures make use of the latter measure impractical. In community and inpatient settings, EMA can be used to identify individual factors leading to relapse, provide a more accurate picture of how a patient has been doing between clinical visits, and link real-world functional outcome measures over time (eg, rates of rehospitalization, days lost because of disability, and likelihood of self-harm) to *scores* on EMA scales. Finally, interventions are rapidly being introduced and delivered via smartphone. EMA may offer the best way to assess intervention acceptability and efficacy, creating the opportunity to personalize treatments with real-time adaptation. For these reasons, EMA is poised not only to replace clinician-administered rating scales in research settings but also to increase accessibility of EMA measures to the patients and health care providers in clinical settings, ultimately allowing real-world clinical settings to contribute meaningful data to research and development of new interventions.

Overall, we believe that the continued use of clinician-administered retrospective self-report assessments in clinical trials contributes significantly to observed treatment failures and squanders innovative potential. As we have described, the instruments currently being used are too broad to adequately assess outcomes, suffer from poor interrater reliability, make inappropriate assumptions about how the illness being studied behaves, and rely on patient recall despite a sizeable body of research, which cautions against this. EMA instruments may play an increasingly important role in addressing the disparity between the need for and investment in novel mental health treatments. Self-report assessment via EMA addresses the limitations of traditional assessment methods but has not yet made its way into large multisite clinical trials sponsored by the industry. Although the FDA's recent efforts to advance mobile technology in clinical trials [79] represents an important first step, iterative testing of standardized EMA-delivered instruments to assess primary outcomes in clinical research is still needed.

Conflicts of Interest

None declared.

References

1. Pankevich DE, Altevogt BM, Dunlop J, Gage FH, Hyman SE. Improving and accelerating drug development for nervous system disorders. Neuron 2014 Nov 5;84(3):546-553 [FREE Full text] [doi: 10.1016/j.neuron.2014.10.007] [Medline: 25442933]
2. Hyman SE. Psychiatric drug development: diagnosing a crisis. Cerebrum 2013 Mar;2013:5 [FREE Full text] [Medline: 23720708]
3. Insel TR, Voon V, Nye JS, Brown VJ, Altevogt BM, Bullmore ET, et al. Innovative solutions to novel drug development in mental health. Neurosci Biobehav Rev 2013 Dec;37(10 Pt 1):2438-2444 [FREE Full text] [doi: 10.1016/j.neubiorev.2013.03.022] [Medline: 23563062]
4. Marder SR, Laughren T, Romano SJ. Why are innovative drugs failing in phase III? Am J Psychiatry 2017 Sep 1;174(9):829-831. [doi: 10.1176/appi.ajp.2017.17040426] [Medline: 28859511]
5. Miller G. Is pharma running out of brainy ideas? Science 2010 Jul 30;329(5991):502-504. [doi: 10.1126/science.329.5991.502] [Medline: 20671165]
6. Leucht S, Leucht C, Huhn M, Chaimani A, Mavridis D, Helfer B, et al. Sixty years of placebo-controlled antipsychotic drug trials in acute schizophrenia: systematic review, Bayesian meta-analysis, and meta-regression of efficacy predictors. Am J Psychiatry 2017 Dec 1;174(10):927-942. [doi: 10.1176/appi.ajp.2017.16121358] [Medline: 28541090]
7. Al-Shahi Salman R, Beller E, Kagan J, Hemminki E, Phillips RS, Savulescu J, et al. Increasing value and reducing waste in biomedical research regulation and management. Lancet 2014 Jan 11;383(9912):176-185 [FREE Full text] [doi: 10.1016/S0140-6736(13)62297-7] [Medline: 24411646]
8. Nutt D, Goodwin G. ECNP Summit on the future of CNS drug research in Europe 2011: report prepared for ECNP by David Nutt and Guy Goodwin. Eur Neuropsychopharmacol 2011 Jul;21(7):495-499. [doi: 10.1016/j.euroneuro.2011.05.004] [Medline: 21684455]
9. Khan A, Brown WA. Antidepressants versus placebo in major depression: an overview. World Psychiatry 2015 Oct;14(3):294-300 [FREE Full text] [doi: 10.1002/wps.20241] [Medline: 26407778]
10. Walsh BT, Seidman SN, Sysko R, Gould M. Placebo response in studies of major depression: variable, substantial, and growing. J Am Med Assoc 2002 Apr 10;287(14):1840-1847. [doi: 10.1001/jama.287.14.1840] [Medline: 11939870]
11. Fava M, Evins AE, Dorer DJ, Schoenfeld DA. The problem of the placebo response in clinical trials for psychiatric disorders: culprits, possible remedies, and a novel study design approach. Psychother Psychosom 2003;72(3):115-127. [doi: 10.1159/000069738] [Medline: 12707478]
12. Iovieno N, Papakostas GI. Correlation between different levels of placebo response rate and clinical trial outcome in major depressive disorder: a meta-analysis. J Clin Psychiatry 2012 Oct;73(10):1300-1306. [doi: 10.4088/JCP.11r07485] [Medline: 23140647]
13. Sanacora G, Johnson MR, Khan A, Atkinson SD, Riesenberg RR, Schronen JP, et al. Adjunctive lanicemine (AZD6765) in patients with major depressive disorder and history of inadequate response to antidepressants: a randomized, placebo-controlled study. Neuropsychopharmacology 2017 Mar;42(4):844-853 [FREE Full text] [doi: 10.1038/npp.2016.224] [Medline: 27681442]
14. Kobak KA, Kane JM, Thase ME, Nierenberg AA. Why do clinical trials fail? The problem of measurement error in clinical trials: time to test new paradigms? J Clin Psychopharmacol 2007 Feb;27(1):1-5. [doi: 10.1097/JCP.0b013e31802eb4b7] [Medline: 17224705]
15. Quiroz JA, Tamburri P, Deptula D, Banken L, Beyer U, Rabbia M, et al. Efficacy and safety of basimglurant as adjunctive therapy for major depression: a randomized clinical trial. JAMA Psychiatry 2016 Jul 1;73(7):675-684. [doi: 10.1001/jamapsychiatry.2016.0838] [Medline: 27304433]
16. Snapinn SM. Noninferiority trials. Curr Control Trials Cardiovasc Med 2000;1(1):19-21 [FREE Full text] [doi: 10.1186/cvm-1-1-019] [Medline: 11714400]
17. Puttagunta PS, Caulfield TA, Griener G. Conflict of interest in clinical research: direct payment to the investigators for finding human subjects and health information. Health Law Rev 2002;10(2):30-32. [Medline: 15739309]
18. McCann DJ, Petry NM, Bresell A, Isacsson E, Wilson E, Alexander RC. Medication Nonadherence, "Professional Subjects," and Apparent Placebo Responders: Overlapping Challenges for Medications Development. J Clin Psychopharmacol 2015 Oct;35(5):566-573 [FREE Full text] [doi: 10.1097/JCP.0000000000000372] [Medline: 26244381]
19. Kobak KA, Leuchter A, DeBrota D, Engelhardt N, Williams JB, Cook IA, et al. Site versus centralized raters in a clinical depression trial: impact on patient selection and placebo response. J Clin Psychopharmacol 2010 Apr;30(2):193-197. [doi: 10.1097/JCP.0b013e3181d20912] [Medline: 20520295]
20. Rutherford BR, Roose SP. A model of placebo response in antidepressant clinical trials. Am J Psychiatry 2013 Jul;170(7):723-733 [FREE Full text] [doi: 10.1176/appi.ajp.2012.12040474] [Medline: 23318413]
21. Gershon RC, Wagster MV, Hendrie HC, Fox NA, Cook KF, Nowinski CJ. NIH toolbox for assessment of neurological and behavioral function. Neurology 2013 Mar 12;80(11 Suppl 3):S2-S6 [FREE Full text] [doi: 10.1212/WNL.0b013e3182872e5f] [Medline: 23479538]
22. Gershon RC, Rothrock N, Hanrahan R, Bass M, Cella D. The use of PROMIS and assessment center to deliver patient-reported outcome measures in clinical research. J Appl Meas 2010;11(3):304-314 [FREE Full text] [Medline: 20847477]

23. Pocock SJ, Stone GW. The primary outcome fails - what next? N Engl J Med 2016 Sep 1;375(9):861-870. [doi: 10.1056/NEJMra1510064] [Medline: 27579636]
24. Fried EI. The 52 symptoms of major depression: lack of content overlap among seven common depression scales. J Affect Disord 2017 Dec 15;208:191-197. [doi: 10.1016/j.jad.2016.10.019] [Medline: 27792962]
25. Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. Br J Psychiatry 1979 Apr;134:382-389. [doi: 10.1192/bjp.134.4.382] [Medline: 444788]
26. Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry 1960 Feb;23:56-62 [FREE Full text] [doi: 10.1136/jnnp.23.1.56] [Medline: 14399272]
27. Chekroud AM, Gueorguieva R, Krumholz HM, Trivedi MH, Krystal JH, McCarthy G. Reevaluating the efficacy and predictability of antidepressant treatments: a symptom clustering approach. JAMA Psychiatry 2017 Apr 1;74(4):370-378 [FREE Full text] [doi: 10.1001/jamapsychiatry.2017.0025] [Medline: 28241180]
28. Fried EI, van Borkulo CD, Epskamp S, Schoevers RA, Tuerlinckx F, Borsboom D. Measuring depression over time...Or not? Lack of unidimensionality and longitudinal measurement invariance in four common rating scales of depression. Psychol Assess 2016 Dec;28(11):1354-1367. [doi: 10.1037/pas0000275] [Medline: 26821198]
29. Hieronymus F, Emilsson JF, Nilsson S, Eriksson E. Consistent superiority of selective serotonin reuptake inhibitors over placebo in reducing depressed mood in patients with major depression. Mol Psychiatry 2016 Apr;21(4):523-530 [FREE Full text] [doi: 10.1038/mp.2015.53] [Medline: 25917369]
30. Guy W. NCDEU Assessment Manual for Psychopharmacology. Washington, DC: US Department of Health, Education, and Welfare; 1976:91-338.
31. Bech P. Rating scales in depression: limitations and pitfalls. Dialogues Clin Neurosci 2006;8(2):207-215 [FREE Full text] [Medline: 16889106]
32. Bech P, Gram LF, Dein E, Jacobsen O, Vitger J, Bolwig TG. Quantitative rating of depressive states. Acta Psychiatr Scand 1975 Mar;51(3):161-170. [Medline: 1136841]
33. Kyle PR, Lemming OM, Timmerby N, Søndergaard S, Andreasson K, Bech P. The validity of the different versions of the Hamilton Depression Scale in separating remission rates of placebo and antidepressants in clinical trials of major depression. J Clin Psychopharmacol 2016 Oct;36(5):453-456. [doi: 10.1097/JCP.0000000000000557] [Medline: 27525966]
34. Carroll J. Endpoint News. 2016. Alkermes plots course to the FDA after its depression drug scores success in last-stand PhIII URL: https://tinyurl.com/y52rk6vm [accessed 2018-08-07] [WebCite Cache ID 71U50UuRV]
35. Uher R, Perlis RH, Placentino A, Dernovšek MZ, Henigsberg N, Mors O, et al. Self-report and clinician-rated measures of depression severity: can one replace the other? Depress Anxiety 2012 Dec;29(12):1043-1049 [FREE Full text] [doi: 10.1002/da.21993] [Medline: 22933451]
36. Nisbett RE, Wilson TD. The halo effect: evidence for unconscious alteration of judgments. J Pers Soc Psychol 1977;35(4):250-256 [FREE Full text] [doi: 10.1037/0022-3514.35.4.250]
37. Dawes RM, Faust D, Meehl PE. Clinical versus actuarial judgment. Science 1989 Mar 31;243(4899):1668-1674. [Medline: 2648573]
38. Grol R, Grimshaw J. From best evidence to best practice: effective implementation of change in patients' care. Lancet 2003 Oct 11;362(9391):1225-1230. [doi: 10.1016/S0140-6736(03)14546-1] [Medline: 14568747]
39. Bowling A. Mode of questionnaire administration can have serious effects on data quality. J Public Health (Oxf) 2005 Sep;27(3):281-291 [FREE Full text] [doi: 10.1093/pubmed/fdi031] [Medline: 15870099]
40. Mora MS, Nestoriuc Y, Rief W. Lessons learned from placebo groups in antidepressant trials. Philos Trans R Soc Lond B Biol Sci 2011 Jun 27;366(1572):1879-1888 [FREE Full text] [doi: 10.1098/rstb.2010.0394] [Medline: 21576145]
41. Schopenhauer A. The World As Will And Representation, Volume 2. New York: Dover Publications; 1966.
42. Mehl MR, Conner TS. Handbook of Research Methods for Studying Daily Life. New York: Guilford Press; 2013.
43. Lacy JW, Stark CE. The neuroscience of memory: implications for the courtroom. Nat Rev Neurosci 2013 Dec;14(9):649-658 [FREE Full text] [doi: 10.1038/nrn3563] [Medline: 23942467]
44. Urban EJ, Charles ST, Levine LJ, Almeida DM. Depression history and memory bias for specific daily emotions. PLoS One 2018;13(9):e0203574 [FREE Full text] [doi: 10.1371/journal.pone.0203574] [Medline: 30192853]
45. Svensson P, de Faire U, Sleight P, Yusuf S, Ostergren J. Comparative effects of ramipril on ambulatory and office blood pressures: a HOPE substudy. Hypertension 2001 Dec 1;38(6):E28-E32. [doi: 10.1161/hy1101.099502] [Medline: 11751742]
46. Pickering TG, Shimbo D, Haas D. Ambulatory blood-pressure monitoring. N Engl J Med 2006 Jun 1;354(22):2368-2374. [doi: 10.1056/NEJMra060433] [Medline: 16738273]
47. O'Brien E, O'Malley K, Cox J, Stanton A. Ambulatory blood pressure monitoring in the evaluation of drug efficacy. Am Heart J 1991 Mar;121(3 Pt 2):999-1006. [doi: 10.1016/0002-8703(91)90611-K] [Medline: 1996533]
48. Peeters F, Berkhof J, Delespaul P, Rottenberg J, Nicolson NA. Diurnal mood variation in major depressive disorder. Emotion 2006 Aug;6(3):383-391. [doi: 10.1037/1528-3542.6.3.383] [Medline: 16938080]
49. Kharasch ED, Neiner A, Kraus K, Blood J, Stevens A, Schweiger J, et al. Bioequivalence and therapeutic equivalence of generic and brand bupropion in adults with major depression: a randomized clinical trial. Clin Pharmacol Ther 2018 Nov 21. [doi: 10.1002/cpt.1309] [Medline: 30460996]

50. Ekkekakis P. The Measurement of Affect, Mood, and Emotion: A Guide for Health-Behavioral Research. First Edition. New York, NY: Cambridge University Press; 2013.
51. Klasnja P, Hekler EB, Shiffman S, Boruvka A, Almirall D, Tewari A, et al. Microrandomized trials: an experimental design for developing just-in-time adaptive interventions. Health Psychol 2015 Dec;34 Suppl:1220-1228. [doi: 10.1037/hea0000305] [Medline: 26651463]
52. Zimmermann J, Woods W, Ritter S, Happel M, Masuhr O, Jaeger U, et al. Integrating structure and dynamics in personality assessment: first steps toward the development and validation of a personality dynamics diary. Psychol Assess 2019 Apr;31(4):516-531 [FREE Full text] [doi: 10.1037/pas0000625] [Medline: 30869961]
53. Russell JA, Carroll JM. On the bipolarity of positive and negative affect. Psychol Bull 1999 Jan;125(1):3-30. [doi: 10.1037/0033-2909.125.1.3] [Medline: 9990843]
54. Nunnally JD, Bernstein IH. Psychometric Theory. New York: Mcgraw-Hill; 1994.
55. Borsboom D. Measuring The Mind: Conceptual Issues In Contemporary Psychometrics. New York, NY: Cambridge University Press; 2019.
56. Embretson SE, Steven PR. Item Response Theory. Mahwah, NJ: CRC Press; 2013.
57. Verhagen SJ, Hasmi L, Drukker M, van Os J, Delespaul PA. Use of the experience sampling method in the context of clinical trials. Evid Based Ment Health 2016 Aug;19(3):86-89 [FREE Full text] [doi: 10.1136/ebmental-2016-102418] [Medline: 27443678]
58. Stone AA, Schwartz JE, Broderick JE, Shiffman SS. Variability of momentary pain predicts recall of weekly pain: a consequence of the peak (or salience) memory heuristic. Pers Soc Psychol Bull 2005 Oct;31(10):1340-1346. [doi: 10.1177/0146167205275615] [Medline: 16143666]
59. Parkinson B, Briner R, Reynolds S, Totterdell P. Time frames for mood: relations between momentary and generalized ratings of affect. Pers Soc Psychol Bull 2016 Jul 2;21(4):331-339. [doi: 10.1177/0146167295214003]
60. van Rijsbergen GD, Burger H, Hollon SD, Elgersma HJ, Kok GD, Dekker J, et al. How do you feel? Detection of recurrent Major Depressive Disorder using a single-item screening tool. Psychiatry Res 2014 Dec 15;220(1-2):287-293. [doi: 10.1016/j.psychres.2014.06.052] [Medline: 25070177]
61. Barge-Schaapveld DQ, Nicolson NA. Effects of antidepressant treatment on the quality of daily life: an experience sampling study. J Clin Psychiatry 2002 Jun;63(6):477-485. [doi: 10.4088/JCP.v63n0603] [Medline: 12088158]
62. Wichers MC, Barge-Schaapveld DQ, Nicolson NA, Peeters F, de Vries M, Mengelers R, et al. Reduced stress-sensitivity or increased reward experience: the psychological mechanism of response to antidepressant medication. Neuropsychopharmacology 2009 Mar;34(4):923-931 [FREE Full text] [doi: 10.1038/npp.2008.66] [Medline: 18496519]
63. Lenderking WR, Hu M, Tennen H, Cappelleri JC, Petrie CD, Rush AJ. Daily process methodology for measuring earlier antidepressant response. Contemp Clin Trials 2008 Nov;29(6):867-877. [doi: 10.1016/j.cct.2008.05.012] [Medline: 18606249]
64. Moore RC, Depp CA, Wetherell JL, Lenze EJ. Ecological momentary assessment versus standard assessment instruments for measuring mindfulness, depressed mood, and anxiety among older adults. J Psychiatr Res 2016 Apr;75:116-123. [doi: 10.1016/j.jpsychires.2016.01.011] [Medline: 26851494]
65. Depp CA, Moore RC, Dev SI, Mausbach BT, Eyler LT, Granholm EL. The temporal course and clinical correlates of subjective impulsivity in bipolar disorder as revealed through ecological momentary assessment. J Affect Disord 2016 Mar 15;193:145-150 [FREE Full text] [doi: 10.1016/j.jad.2015.12.016] [Medline: 26773907]
66. Peters GY, Crutzen R. Pragmatic nihilism: how a Theory of Nothing can help health psychology progress. Health Psychol Rev 2017 Dec;11(2):103-121. [doi: 10.1080/17437199.2017.1284015] [Medline: 28110627]
67. Fried E, Boschloo L, van Borkulo CD, Schoevers R, Romeijn J, Wichers M, et al. Commentary: "Consistent Superiority of Selective Serotonin Reuptake Inhibitors Over Placebo in Reducing Depressed Mood in Patients with Major Depression". Front Psychiatry 2015;6:117 [FREE Full text] [doi: 10.3389/fpsyt.2015.00117] [Medline: 26347663]
68. Schultzberg M, Muthén B. Number of subjects and time points needed for multilevel time-series analysis: a simulation study of dynamic structural equation modeling. Struct Equ Modeling 2018;25:495. [doi: 10.1080/10705511.2017.1392862]
69. Trull TJ, Lane SP, Koval P, Ebner-Priemer UW. Affective dynamics in psychopathology. Emot Rev 2015 Oct;7(4):355-361 [FREE Full text] [doi: 10.1177/1754073915590617] [Medline: 27617032]
70. Trull TJ, Ebner-Priemer U. Ambulatory assessment. Annu Rev Clin Psychol 2013;9:151-176 [FREE Full text] [doi: 10.1146/annurev-clinpsy-050212-185510] [Medline: 23157450]
71. Bringmann LF, Vissers N, Wichers M, Geschwind N, Kuppens P, Peeters F, et al. A network approach to psychopathology: new insights into clinical longitudinal data. PLoS One 2013;8(4):e60188 [FREE Full text] [doi: 10.1371/journal.pone.0060188] [Medline: 23593171]
72. Hamaker EL, Asparouhov T, Brose A, Schmiedek F, Muthén B. At the frontiers of modeling intensive longitudinal data: dynamic structural equation models for the affective measurements from the COGITO study. Multivariate Behav Res 2018 Apr 6:1-22. [doi: 10.1080/00273171.2018.1446819] [Medline: 29624092]
73. Insel T. National Institute of Mental Health. 2013. NIMH's new focus in clinical trials URL: https://tinyurl.com/y65toyyf [accessed 2018-08-07] [WebCite Cache ID 71U6GLinx]

74. Torous J, Chan SR, Yee-Marie TS, Behrens J, Mathew I, Conrad EJ, et al. Patient smartphone ownership and interest in mobile apps to monitor symptoms of mental health conditions: a survey in four geographically distinct psychiatric clinics. JMIR Ment Health 2014;1(1):e5 [FREE Full text] [doi: 10.2196/mental.4004] [Medline: 26543905]
75. Tarver M. US Food and Drug Administration. Development of validated instruments URL: https://tinyurl.com/yy62lghz [WebCite Cache ID 71U6aIPLM]
76. US Food and Drug Administration. 1997. Guidance for industry URL: https://tinyurl.com/y28q9x5v [accessed 2018-08-07] [WebCite Cache ID 71U6cq3O8]
77. Wayback Machine. 2016. Description of the HAMD and the MADRS URL: https://tinyurl.com/yyr65kxm [accessed 2018-08-07] [WebCite Cache ID 71U6Y0K7q]
78. Lupien SJ, Sasseville M, François N, Giguère CE, Boissonneault J, Plusquellec P, Signature Consortium. The DSM5/RDoC debate on the future of mental health research: implication for studies on human stress and presentation of the signature bank. Stress 2017 Dec;20(1):95-111. [doi: 10.1080/10253890.2017.1286324] [Medline: 28124571]
79. Munos B, Baker PC, Bot BM, Crouthamel M, de Vries G, Ferguson I, et al. Mobile health: the power of wearables, sensors, and apps to transform clinical trials. Ann N Y Acad Sci 2016 Jul;1375(1):3-18. [doi: 10.1111/nyas.13117] [Medline: 27384501]

Abbreviations

CNS: central nervous system
EMA: ecological momentary assessment
EMI: ecological momentary intervention
FDA: Food and Drug Administration
HAM-D: Hamilton Rating Scale for Depression
HAM-D-17: 17-item Hamilton Rating Scale for Depression
HAM-D-24: 24-item Hamilton Rating Scale for Depression
MADRS: Montgomery Asberg Depression Scale
MBSR: mindfulness-based stress reduction
NNT: needed to treat
PRO: patient-reported outcome

Edited by J Prescott; submitted 07.08.18; peer-reviewed by J Quiroz, D Kreindler, U Ebner-Priemer, M Brandon; comments to author 19.09.18; revised version received 05.02.19; accepted 03.04.19; published 21.04.19

Please cite as:
Mofsen AM, Rodebaugh TL, Nicol GE, Depp CA, Miller JP, Lenze EJ
When All Else Fails, Listen to the Patient: A Viewpoint on the Use of Ecological Momentary Assessment in Clinical Trials
JMIR Ment Health 2019;6(5):e11845
URL: https://mental.jmir.org/2019/5/e11845/
doi:10.2196/11845
PMID: