Title: When All Else Fails, Listen to the Patient

Subtitle: E-technology and Patient Self-Report Can Improve Precision in Mental Health Clinical Trials

We are at an impasse in mental health treatment development and testing: our clinical trials are failing far too often. Thus, while the global burden of psychiatric illness continues to increase, investment in novel pharmacologic agents flows instead towards disease states with identifiable biological targets [1]. The Central Nervous System (CNS) pipeline has become increasingly burdened over the years with late-phase failures. This has led to a well-publicized exodus of the pharmaceutical industry from the CNS space. By one estimate, pharmaceutical research and development in psychiatry has declined by 70 percent over the last decade [2].

Many factors have been proposed to underlie the rising difficulties in treatment trials, such as costs, regulatory burdens, and inefficiency; but a key driving factor is poor measurement of outcomes, which has led to an increase in placebo response over time [3]. For example, some investigators have demonstrated that placebo response is substantial and rising in antidepressant drug trials [4]. In a clinical research setting a high placebo response rate results in decreased ability to demonstrate efficacy of an active intervention, be it drug, device, or psychosocial treatment. This high placebo response in clinical trials is not a natural part of the mental illnesses being treated; instead, it reflects imprecision of measurement. Imprecision is the bane of scientific progress, and this reduced assay sensitivity in mental health clinical trials can sabotage treatment development at any stage, including novel development program and expensive phase three projects.

Explanations for this imprecision typically point to problems with recruitment/sample ascertainment (e.g. career patients who enroll in research studies for financial reasons) or cite unscrupulous principal investigators who exaggerate patients' symptoms of psychiatric illness to meet recruitment goals [5]. But these arguments fail to explain why academic studies, in which less financial gain accrues to the patient and investigator, have seen a high placebo response (and failure rate) as well.

There is a more fundamental – and correctable – source of clinical trial failures: the high placebo response rate reflects measurement errors inherent to our outcome measures. These errors can be seen in older instruments (e.g., the Hamilton Depression Scale (HAM-D), created in the 1950s to track Electroconvulsive Therapy outcomes but still widely used as an antidepressant outcome scale) but also in newer ones, such as the Montgomery–Åsberg Depression Rating Scale (MADRS) and the Quick Inventory of Depressive Symptomatology (QIDS). All of these measures have three problems: they have needless complexity, introduce human error from clinician judgment, and sample too infrequently.

In this review, we examine the instruments used at present to track psychiatric illness with an emphasis on depression measurement tools, which are especially instructive for our purposes, and we present research including from basic psychology and computer science, to show that e-
technology based measures can be superior to this status quo. Better measurement will translate into better, more successful clinical trials and surmount our present impasse in developing new mental health treatments.

**Problem # 1: Needless Complexity Undermines Commonly Used Scales**

**Key Points:**
- We're measuring more than we need to or should be measuring
- Over-measuring compromises our ability to detect a signal
- We need to make a decision about what we want to measure

Lengthy, multi-item rating scales are usually the primary outcome to determine whether or not a treatment is effective. For example, in antidepressant studies a clinician will often administer the MADRS to a study participant, and that participant’s MADRS score across a handful of measurements over time will determine whether the intervention in question works. Because the success or failure of a treatment rests on these scales’ validity and reliability, they deserve the same amount of scrutiny regarding assay sensitivity that one would give to a laboratory assay. A recent study by Checkroud and colleagues of over 7,000 patients with major depression illustrates how commonly-used measures can jeopardize a potential treatment in late-phase trials. This study found that the HAM-D measured more than one thing (i.e. contained more than one “factor”), and consistent improvement following treatment with antidepressants we know to be effective was found only for the core emotional symptoms buried within the scale and obscured by the total score [6]. Thus, clinician-administered rating scales have contributed to treatment failures is by introducing unnecessary complexity, thereby reducing measure reducing measure specificity. This problem is not restricted to mental health research; for example, trials in cardiology have also been compromised by failing to adequately restrict their outcome measures [7].

Clinician-administered rating scales mirror the diagnostic criteria or descriptive psychopathology for a given psychiatric disorder and use these criteria to track a patient’s progress throughout a clinical research study. However, the scales do not usually stop at the diagnostic criteria alone. The descriptive psychopathology for a given psychiatric disorder is often more expansive than the diagnostic criteria itself; for example, in major depression, patients often have irritability, anxiety, and other symptoms in addition to the nine cardinal symptoms of the disorder. Clinician-administered rating scales introduce descriptive psychopathology presumably with the aim to cover an illness comprehensively. The problem with designing a scale this way is that there is no accounting for which items clinicians and patients care most about, which items are found in nearly all cases, and which items are rarely found in cases. One study by Eiko Fried found 52 symptoms of depression across seven commonly used depression scales with a content overlap among all scales of 0.32 [8].

One particularly complex scale designed to comprehensively capture an illness is the 24-item Hamilton Depression Scale (HAMD-24). Using the HAMD-24 may conceal treatment effects by introducing items that are not commonly found or are clinically irrelevant such as
hypochondriacal or depersonalization symptoms. As the total score is then used to determine whether or not the treatment is effective, there is a risk of magnifying irrelevant changes and obscuring important ones. This was shown in a recent study suggesting that the six-item HAM-D (HAM-D-6) was superior to the longer HAMD-24, -21, and -17 in its ability to discriminate the effectiveness of an antidepressant medication from placebo [9]. Similarly, the Alkermes compound, ALKS-5461, fared better in separating from placebo with a measurement of core symptoms using the MADRS-6 item scale than the full 10-item scale [10].

Complex rating scales may fare worse than simple rating scales because complex rating scales tend to measure numerous symptom constructs. The items tied to each construct tend to shift unpredictably over time (e.g., due to lack of longitudinal factorial invariance) [11]. This can lead to a conflation of sensitive and specific items (dysphoria, anhedonia), non-specific items (anxiety), symptoms from unrelated illness (e.g., fatigue), and side-effects of the treatment. Moreover, these items are often not weighted for relevance and the success or failure of a treatment rests on a scale's summative score, which equally weights items that are irrelevant to the individual under study.

It makes sense, then, to focus on what we care about tracking, yet there is no agreement on what we care about monitoring. We argue that it is important to develop a consensus on what we care about (i.e., what would be the most clinically, functionally, or personally relevant tracking features of response or remission) before we begin tracking something. There are a number of ways to accomplish this. For instance, if we simply wish to use these scales pragmatically, we would take a treatment we know to be effective and choose the items from a scale that reveals the greatest amount of separation in favor of the proven treatment. We would then use that same scale to determine whether or not an unproven treatment is effective. Alternatively, we could adopt something approaching a universal consensus on what we agree is clinically relevant, which in the case of depression would likely be the core emotional symptoms of the illness, and simply monitor that. On the other hand, improvement from a functional or pharmacoeconomic perspective may not map well on to any items in a measure we currently use, which would call a priori assumptions about clinical relevance into question. In short, it is evident that while we can confidently say we currently measure things the wrong way it is less evident how we should instead go about measuring them.

**Problem #2: Clinical Judgment Magnifies Measurement Error**

**Key Points:**
- Clinician-rated scales simply add a clinician's error to the patient's error
- Self-report scales are preferable to clinician-rated scales

A large study evaluating self-report and clinician-administered instruments from the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial concluded that self-report measures contributed more to the prediction of outcomes of clinician-administered instruments than vice versa and recommended that, in the
event only one form of assessment could be used, self-reported outcome measures would be preferable [12]. This study contradicts a core belief favoring the use of clinician-administered assessments over patient-reported outcome measures: that the clinician is objectively correcting for whatever error (e.g. errors of omission, exaggeration, expectancy effect, Hawthorne effect, etc.), intentional or otherwise, is introduced by the patient. We argue, instead, that the clinician magnifies the patient’s error. Further, this clinician error is routine, not idiosyncratic, and occurs not due to malice or laziness but happens unconsciously and often in good conscience, because clinical judgment is not now and never will be completely objective. This has been demonstrated in studies finding poor inter-rater and test-retest reliability in standard clinician-administered assessment measures for depression [13]. The reason for this may be that clinicians even when given a rule tend to underperform the rule [14]. We may consider the rule in this case to be the proper way to rate a scale.

The evidence suggests that mechanical, or numerical algorithmic, judgment exceeds clinical judgment largely because mechanical judgment, by definition, unfailingly follows a given set of rules. Whether or not a clinician decides to follow a rule depends on the amount of inertia that must be overcome in order 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 among other factors [15]. A decent body of literature has of late elevated the stature of patient reported outcomes vis-a-vis traditional, clinician-administered rating scales. For example, a large meta-analysis of placebo response in 96 antidepressant trials by Rief et al. found not only that clinician-administered instruments were associated with a higher placebo response than patient-reported outcome measures, but also that the oft-cited increase in placebo response over time held true for clinician, but not self-reported ratings [16]. Such evidence is in line with the idea that clinician administered scales add error rather than removing or mitigating patient error. It would seem from all this that we would do well to follow the advice of Sir William Osler and listen to the patient, particularly when all else fails, as it has.

Problem #3: Infrequent Sampling Hurts Sensitivity

Key Points:

- Patients are unlikely to be able to provide accurate retrospective reports
- There is no substitute for frequent in the moment patient reported data
- We need this data to create the measurement before the measurement is used to interpret the data

Typically in clinical trials, outcomes are assessed infrequently; for example, the MADRS or HAMD might be assessed weekly, or even less often, to track treatment response. By sampling infrequently in this way, we are tacitly making the assumption that we know enough about how an illness behaves to feel comfortable monitoring it at that frequency or to ask questions with a time frame modifier (e.g., “In the last week...”). Yet we simply don’t have the data necessary to make this assumption; in fact, this infrequent sampling is due to our reliance on clinician-administered instruments, which are too burdensome to administer frequently. Thus, our
practice of infrequent sampling ignores a long and rich scientific literature on ambulatory assessment.

Again turning to the cardiology field for an example, the importance of not assuming we understand the behavior of a phenomenon in advance of the evidence was shown in the HOPE trial, which evaluated the effect of the ACE inhibitor, Ramipril, in patients at high risk for cardiovascular events. The study found that Ramipril lowered blood pressure assessed via 24-hour ambulatory blood pressure measurements, but not office based blood pressure measurements, owning to a diurnal variation as well as other factors (e.g. “white coat hypertension”) that either could not be captured with a limited number of samplings obtained during office hours or were affected by the office visit itself [17]. For this reason, the hypertension field has moved to using frequent ambulatory blood pressure sampling to assess efficacy, which has essentially eliminated the placebo response in antihypertensive trials [18].

We do not know enough about how psychiatric disorders behave to risk direct extrapolation from the example of blood pressure monitoring, because we rely on limited-sampling instruments, which assume that the symptoms of many psychiatric illnesses are trait-like in advance of any evidence to support this assumption. Even if the symptoms of psychiatric illness are trait-like, we would continue to favor frequent sampling as, for example, measuring the most important or valid symptom 15-20 times will typically offer greater validity and reliability than measuring 20 symptoms of varying degrees of importance or validity using trait-like assessments.

To see why, consider an ideal, completely trait-like illness that shows no variance. We can give no examples of such a thing, but it is useful as a thought experiment. This illness should produce ratings with no variance, and indeed its contribution to (for example) 20 ratings of a single symptom over time will have no variance within a single person. However the ratings themselves will vary, because ratings are influenced by many things that are, for our purposes, considered error. Taking a single measurement risks a source of error impacting all of the items for that person (e.g., how much the person slept that night). It may be helpful to consider the following problem: A distant radio station is transmitting a single tone on an analog frequency. You have 20 radios with various sized antennas, and you have a choice: turn all of the radios on at once, or turn on the radio with the largest antenna on 20 times. Which do you choose? Given that radio signals are subject to transient interference (weather, passing airplanes, etc.), hopefully it is obvious that 20 times with a better antenna is the preferable option. This is akin to the superiority of assessing the one, best item 20 times over assessing 20 items one time. Measuring over 20 time points minimizes the influence of any one source of error. Further, by selecting the radio (for a measure, the item or small set of items) that has higher validity and reliability to begin with, we are maximizing the amount of signal we can expect. The metaphor is not a fanciful one: the information provided by a set of items essentially comes down to the amount of signal versus noise, and measuring using high-signal items repeatedly over time would better capture the level of the trait (because any one source of error would be minimized and possibly averaged away) compared to measuring using a mix of items at a single time point [19].
However, many symptoms in psychiatric illness are not trait-like; they are state-like and vary over time. For example, in an individual with major depression, their mood might be very depressed at a certain point in the morning and near-normal later that same day [20]. Nonetheless, we continue to measure them, contrary to our assumptions and clinical experience, as stable trait-like symptoms (e.g., “in the last 7 days, how has your mood been?”). In short, for most measures of psychiatric illnesses our field has not properly defined the dynamic vs. stable or trait-like nature symptoms, because the only way to do so is to sample the illness frequently before finalizing the measure (e.g., for use in a treatment study). Limited sampling further compromises psychiatric research because trait measures require respondents to attempt a summation of states. This occurs because such measures force a patient to attempt accurate recall of past emotional states. Even if humans were quite good at this operation, it would still be complicated by filtration through the respondent’s emotional state at the time of the assessment. It thus seems quite unlikely that respondents will be particularly good at creating a summary of their states, and the available evidence suggests that there are indeed problems here. For example, evidence from pain studies examining Ecological Momentary Assessment (EMA), or frequent, real time, patient-reported assessment of pain reports alongside retrospective recall shows a rather consistent discrepancy between the two forms of report [21]. And while memory is held sacred in certain settings such as the courtroom, neuroscientists have found memory to be unreliable, particularly when the encoding and retrieval of memories occurs during periods of emotional arousal [22].

An example of how infrequent sampling adversely affects assay sensitivity in clinical trials was recently provided by Moore et al [23]. In this study, the researchers assessed the effects of mindfulness based stress reduction (MBSR), compared to an attention placebo. For outcome assessments, they measured depressive symptoms, anxiety symptoms, and mindfulness self-ratings in two ways: Ecological Momentary Assessment (EMA) tools delivered to participants electronically via a Smartphone three times daily for 14 days, and traditional paper-and-pencil based measurement tools asking about the last week’s symptoms (comparable to most outcome measures). The EMA-based outcome assessment resulted in a number needed to treat for the intervention that was much lower than the same outcomes measured using the traditional technique. In other words, EMA captured a treatment effect that was missed by standard self-report assessments. This was also reflected in the smaller standard deviations for outcomes measured via EMA when averaged over time. This was the first head-to-head proof in an RCT of a long-known scientific fact that has been ignored by our clinical trial field: frequent ambulatory assessment improves precision.

**Solution: Ecological Momentary Assessment**

EMA is frequent, real time, patient-reported assessment (e.g., “right now, my mood is...”) completed by the patient typically via mobile device. EMA may overcome the deficiencies inherent in traditional clinician-administered instruments through its ability to obtain multiple, real-time, patient-reported measures. One example of this superiority is the MBSR study mentioned above. Another study, which nicely synthesizes the various arguments we’ve
advanced here, found that a single item scale, measuring mood on a VAS delivered via EMA, outperformed the HAMD-17 in its ability to predict “current relapse status” in patients with MDD [24].

EMA increases precision, but it may yield a great deal more benefits to clinical research. For instance, data derived from electronically delivered EMA can tell us about the dynamics of the illness being assessed or, in other words, how symptoms of an illness behave and interact with one another over time [25]. This will allow us to answer such questions as which symptoms are state or trait-like or what sorts of causal relationships take place between symptoms over time. This approach is also useful because it agnostically looks at interactions between symptoms without first assuming that they are symptoms of a disorder. This “pragmatic nihilism” [26] or “symptomic” [27] approach differs from how we currently view psychiatric disorders. The current approach, with which the reader is most likely familiar, relies on viewing any number of symptoms as being all tied to one underlying, or latent variable (e.g. depression). With enough patient-reported EMAs carried out over time, it may become possible to observe how symptoms interact and cause one another, whether one symptom is more important or central than another, or how certain upstream symptoms may influence a cascade of symptoms downstream. Such findings will afford researchers the unique opportunity to attempt to stratify a group of patients based on how they do or don’t get better, rather than simply whether or not they get better. The approach becomes highly descriptive at the level of the individual, allowing one to answer a host of previously unanswerable questions. 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 (such as an Ecological Momentary Intervention, or a Just-in-time adaptive intervention) changes the observed course of symptomatology downstream.

This novel use of EMA we have advanced is consistent with the concept of target engagement raised by the NIMH 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 [28]. EMA here offers the unique ability to evaluate whether a target is being addressed by an intervention via real-time lagged mediation rather than post-hoc analyses. 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’ [29]. Such a focus, as represented in the example above, may allow us to identify symptoms driving other symptoms, or, symptomic drivers that may uncover occult underlying causes of complex psychiatric illnesses.

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 discernible 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 one symptom gets better or worse the rest tend to follow) [30]. This is by no means an exhaustive list of questions that may be asked of the data derived from EMA. Suffice it 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. Laura Bringmann and colleagues demonstrated how this might be accomplished using multilevel vector autoregression to reveal individual patterns of symptom interaction that differ from the group [31].

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 information could, however, 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 rather than attempts to move around the patient. In brief, we hope through this discussion to facilitate the uptake of EMA in clinical research and in so doing make the patient the focus of the solution rather than the nidus of the problem.

*How do we get to widespread use of EMA in our clinical trials?*

EMA has several limitations. First, not everyone has a smart phone, although the number of individuals without one seems to become vanishingly small as time goes by. Perhaps a greater question then is whether a participant with a smart phone would want to use it to regularly quantify his or her depressive symptoms. This has become an increasingly important issue as faith in “big tech” to safeguard users’ privacy has waned in the wake of the Cambridge Analytica scandal. Second, we need to evaluate clinician-administered instruments alongside commensurate EMA-delivered items. This is not only desirable, as it 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 [32]. Third, while the FDA has made its expectations for patient reported outcome measures (PROs) clear [33], it is not at all clear whether every aspect of FDA guidance will neatly translate to ePROs. For example, to what extent, if any, would necessary software updates for an accepted EMA app involve the FDA? Of note, the 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. This effectively prioritizes tradition over innovation and creates a catch-22 for researchers who might otherwise break with the status quo [34]. Fourth, use of EMA in the real world often leads to missing data that has historically made analysis problematic, although, for reasons that are outside the scope of the present discussion, there is cause to be hopeful that this may become less of a problem moving forward. Finally, EMA may
not be very good at detecting rare events if they occur infrequently relative to the sampling frequency (i.e. as the sampling frequency decreases so too does the probability of capturing ‘rare events’).

It is important to note that in order to realize these objectives we 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. We cannot advance the field without standardization across multiple field trials in different populations in order to clearly establish test-retest reliability, external validity and other parameters necessary to validate an EMA scale.

We favor moving from clinician-administered rating scales towards patient-reported ecological momentary assessment measures (delivered via smart phone) across all settings: clinical research studies, inpatient hospitalizations, and outpatient community settings. In clinical research studies, EMA will reduce placebo response and increase intervention-placebo separation. EMA also offers an obvious advantage over clinician-administered rating scales in inpatient and community settings where time, cost, and staff pressures make use of the latter measure impracticable. 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 two 15 minute visits at least a month apart, and link real world functional outcome measures over time (e.g. rates of re-hospitalization, days lost due to disability, likelihood of self-harm, etc.) to “scores” on EMA scales. The ability to link a score on a scale to meaningful outcomes cannot be overemphasized as at present no one knows what a 30 on the MADRS means other than “moderate depression.” Finally, interventions are rapidly being introduced and delivered via smart phone. EMA offers the best way to assess the efficacy (or lack thereof) of these interventions unless the outcome is obvious (resumption of heroin use, myocardial infarction, etc.) For these reasons, we believe that EMA will rightly replace clinician-administered rating scales and expand the use of assessment measures to the benefit of patients, all those who care for them, and research and development of new interventions.

Conclusion

In sum, clinical trials are wasting money and innovative potential by relying on poor measurement tools. As we have described, the instruments currently being used are too broad to adequately assess outcomes, suffer from poor inter-rater reliability, make inappropriate assumptions about how the illness being studied behaves, and rely on patient recall despite a sizeable body of research in psychology cautioning against this. Each one of these problems may be solved using EMA delivered via participants' own smartphones. We encourage the FDA to take a permissive stance towards allowing the development and iterative testing of standardized EMA-delivered instruments to assess outcomes in clinical research, as these instruments will be necessary to bridge the disparity between the need for and investment in novel mental health treatments.
1. Innovative solutions to novel drug development in mental health
   PMID: 23563062


3. Placebo response in studies of major depression: variable, substantial, and growing.
   B. Timothy Walsh, Stuart N. Seidman, Robyn Sysko, Madelyn Gould
   JAMA. 2002 Apr 10; 287(14): 1840–1847
   PMID: 11939870

4. The problem of the placebo response in clinical trials for psychiatric disorders: culprits, possible remedies, and a novel study design approach.
   Maurizio Fava, A. Eden Evins, David J. Dorer, David A. Schoenfeld
   PMID: 12707478

5. Site versus centralized raters in a clinical depression trial: impact on patient selection and placebo response.
   Kenneth A. Kobak, Andrew Leuchter, David DeBrota, Nina Engelhardt, Janet B. W. Williams, Ian A. Cook, Andrew C. Leon, Jonathan Alpert
   PMID: 20520295

6. Reevaluating the Efficacy and Predictability of Antidepressant Treatments: A Symptom Clustering Approach
   Adam M. Chekroud, Ralitza Gueorguieva, Harlan M. Krumholz, Madhukar H. Trivedi, John H. Krystal, Gregory McCarthy
   PMID: 28241180

7. The Primary Outcome Fails - What Next?
   Stuart J. Pocock, Gregg W. Stone
   PMID: 27579636

8. The 52 symptoms of major depression: Lack of content overlap among seven common depression scales.
   Eiko I. Fried
PMID: 27792962

9. 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. Phillip Raphael Kyle, Ole Michael Lemming, Nina Timmerby, Susan Søndergaard, Kate Andreasson, Per Bech
PMID: 27525966

http://www.webcitation.org/71U50UuRV.

Psychol Assess. 2016 Jan 28 Published online 2016 Jan 28. doi: 10.1037/pas0000275
PMID: 26821198

12. SELF-REPORT AND CLINICIAN-RATED MEASURES OF DEPRESSION SEVERITY: CAN ONE REPLACE THE OTHER?
Rudolf Uher, Roy H. Perlis, Anna Placentino, Mojca Zvezdana Dernovšek, Neven Henigsberg, Ole Mors, Wolfgang Maier, Peter McGuffin, Anne Farmer
Depress Anxiety. Author manuscript; available in PMC 2013 Aug 23.
Published in final edited form as: Depress Anxiety. 2012 Dec; 29(12): 1043–1049.
Published online 2012 Aug 29. doi: 10.1002/da.21993
PMID: 22933451

PMID: 12707478

R. M. Dawes, D. Faust, P. E. Meehl
PMID: 2648573

15. Lessons learned from placebo groups in antidepressant trials Meiike Shedden Mora, Yvonne Nestoriuc, Winfried Rief
PMID: 21576145

P. Svensson, U. de Faire, P. Sleight, S. Yusuf, J. Oestergren
PMID: 11751742
17. Ambulatory blood-pressure monitoring.
Thomas G. Pickering, Daichi Shimbo, Donald Haas
PMID: 16738273

18. Ambulatory blood pressure monitoring in the evaluation of drug efficacy.
E. O’Brien, K. O’Malley, J. Cox, A. Stanton
PMID: 1996533


20. Diurnal mood variation in major depressive disorder.
Peeters F, Berkhof J, Delespaul P, Rottenberg J, Nicolson NA.
PMID: 16938080

21. Variability of momentary pain predicts recall of weekly pain: a consequence of the peak (or salience) memory heuristic.
Arthur A. Stone, Joseph E. Schwartz, Joan E. Broderick, Saul S. Shiffman
PMID: 16143666

22. The Neuroscience of Memory: Implications for the Courtroom
Joyce W. Lacy, Craig E. L. Stark
Nat Rev Neurosci. Author manuscript; available in PMC 2014 Oct 2.
Published in final edited form as: Nat Rev Neurosci. 2013 Sep; 14(9): 649–658. Published online 2013 Aug 14. doi: 10.1038/nrn3563
PMID: 23942467

Gerard D. van Rijsbergen, Huibert Burger, Steven D. Hollon, Hermien J. Elgersma, Gemma D. Kok, Jack Dekker, Peter J. de Jong, Claud L. H. Bockting
PMID: 25070177

24. Ecological momentary assessment versus standard assessment instruments for measuring mindfulness, depressed mood, and anxiety among older adults
Raeanne C. Moore, Colin A. Depp, Julie Loebach Wetherell, Eric Lenze
J Psychiatr Res. Author manuscript; available in PMC 2017 Apr 1.
Published in final edited form as: J Psychiatr Res. 2016 Apr; 75: 116–123.
PMID: 26851494

25. The temporal course and clinical correlates of subjective impulsivity in bipolar disorder as revealed through ecological momentary assessment
Colin A. Depp, Raeanne C. Moore, Sheena I. Dev, Brent T. Mausbach, Lisa T. Eyler, Eric L. Granholm
J Affect Disord. Author manuscript; available in PMC 2016 Jun 21.
Published in final edited form as: J Affect Disord. 2016 Mar 15; 193: 145–150.
PMID: 26773907
Gjalt-Jorn Ygram Peters, Rik Crutzen
PMID: 28110627

27. Symptomics as a new research paradigm in psychiatry.
Fried, Eiko & Boschloo, Lynn & van Borkulo, Claudia & Schoevers, Robert & Romeijn, Jan-Willem & Wichers, Marieke & Jonge, Peter & Nesse, Randolph & Tuerlinckx, Francis & Borsboom, Denny.
Frontiers in psychiatry. 2015. 6. 117. 10.3389/fpsyt.2015.00117

28. Insel, T. NIMH’s new focus in clinical trials. 2013.
http://www.webcitation.org/71U6GLinx

29. Symptomics as a new research paradigm in psychiatry.
Fried, Eiko & Boschloo, Lynn & van Borkulo, Claudia & Schoevers, Robert & Romeijn, Jan-Willem & Wichers, Marieke & Jonge, Peter & Nesse, Randolph & Tuerlinckx, Francis & Borsboom, Denny.
Frontiers in psychiatry. 2015. 6. 117. 10.3389/fpsyt.2015.00117

30. Affective Dynamics in Psychopathology
Timothy J. Trull, Sean P. Lane, Peter Koval, Ulrich W. Ebner-Priemer
PMID: 27617032

31. A Network Approach to Psychopathology: New Insights into Clinical Longitudinal Data
Laura F. Bringmann, Nathalie Vissers, Marieke Wichers, Nicole Geschwind, Peter Kuppens, Frenk Peeters, Denny Borsboom, Francis Tuerlinckx
PMID: 23593171

http://www.webcitation.org/71U6Y0K7q

33. Tarver, M. Development of validated instruments.