Effects of intense assessment on statistical power in randomized controlled trials: Simulation study on depression

Raphael Schuster, PhD 1*; Manuela Larissa Schreyer, PhD 2; Tim Kaiser, PhD. 1; Thomas Berger, Prof. 3; Jan Philipp Klein, PhD 4; Steffen Moritz, Prof; 5, Anton-Rupert Laireiter, Prof. 1 6; Wolfgang Trutschnig, Ass. Prof 2

1 Department of Psychology, University of Salzburg, Austria

2 Department of Mathematics, University of Salzburg, Austria

3 Department of Clinical Psychology and Psychotherapy, University of Berne, Switzerland

4 Department of Psychiatry and Psychotherapy, Lübeck University, Lübeck, Germany

5 Department of Psychiatry and Psychotherapy, University Medical Center Hamburg-Eppendorf, Germany

6 Faculty of Psychology, University of Vienna, Austria

*Corresponding author:

Dr. Raphael Schuster

Department of Psychology, University of Salzburg

Hellbrunnerstraße 34, 5020 Salzburg, Austria

raphael.schuster@sbg.ac.at

Abstract

Smartphone-based devices are increasingly recognized to assess disease symptoms in daily life (e.g. ecological momentary assessment, EMA). Despite this development in digital psychiatry, clinical trials are mainly based on point assessments of psychopathology. This study investigated expectable increases in statistical power by intense assessment in randomized controlled trials (RCTs).

A simulation study, based on three scenarios and several empirical data sets, estimated power gains of two- or fivefold pre-post-assessment. For each condition, data sets of various effect sizes were generated, and AN(C)OVAs were applied to the sample of interest (N=50-N=200).

Power increases ranged from 6% to 92%, with higher gains in more underpowered scenarios and with higher number of repeated assessments. ANCOVA profited from a more precise estimation of the baseline covariate, resulting in additional gains in statistical power. Fivefold pre-post EMA resulted in highest absolute statistical power and clearly outperformed traditional questionnaire assessments. For example, ANCOVA of automatized PHQ-9 questionnaire data resulted in absolute power of 55 (for N=200 and d=0.3). Fivefold EMA, however, resulted in power of 88.9. Non-parametric and multi-level analyses resulted in comparable outcomes.

Besides providing psychological treatment, digital mental health can help optimizing sensitivity in RCT-based research. Intense assessment appears advisable whenever small sample sizes or small treatment effects are expected, or when applying optimization problems (e.g. machine learning). Simulations for various effects and a short guide for popular power software are provided for study planning. While feasibility of weekly assessment is established, the strategy of pre-post-EMA to boost statistical power needs to be further tested.

Statistical power in clinical research

Statistical power is the probability to actually detect the phenomenon one is looking for (given the phenomenon exists). Statistical power is therefore a key component of every statistical study. As known from the continuing debate in neuroscience and psychological research (and the related replication crisis), many studies remain underpowered (Button et al., 2013; Halpern, Karlawish, & Berlin, 2002) - implying a high chance of overlooking effects. An assessment of effect sizes in the field of psychology and neuroscience revealed a median power of 0.12 to detect small effects and a power of 0.44 for medium effects (Szucs & Ioannidis, 2017). Considering the convention for statistical power (from 0.8 to 0.9), the reported lack can be classified as substantial. Finally, low statistical power also reduces the likelihood of statistically significant findings to actually reflect true effects.

The situation in clinical research is comparable, where interested patients potentially volunteer in trials with restricted clinical value, or where studies fail in later stages of the admission process (Halpern, Karlawish, & Berlin, 2002; Khan, Fahl, & Brown, 2018). Even though this phenomenon does not present uniformly (Maddock & Rossi, 2001; Marszalek, Barber, Kohlhart, & Cooper, 2011), the practice of underpowered studies in the clinical field can be described as widespread and hard to change –ultimately increasing the risk to aggregate false findings in meta analytic evaluations (Califf et al., 2012; Maxwell, 2004; Roozenbeek, Lingsma, Steyerberg, & Maas, 2010; Wampold et al., 2017), or increasing the probability of failed attempts when replicating previous studies (replication crisis).

As resources in clinical research are restricted, diverse strategies to optimize statistical power have been developed. Those techniques can be classified into such strategies to maintain and such strategies to increase statistical power. Besides more elaborated procedures, a quantity of generic strategies are suggested in literature (Hansen, & Collins, 1994; Harrison, 2009; Roozenbeek et al., 2009). Prominent examples are: i) Maintaining sample size (e.g. preventing attrition and missing data), ii) maximizing effect size (e.g. maintaining program integrity, prognostic targeting), or iii) reducing variance (e.g. investigating homogenous populations). Even though a conscious consideration of these strategies may help to boost power, some of the described techniques entail important disadvantages. For

example, a highly homogenous population can restrict validity of findings, and prognostic targeting may result in considerable extra work.

With the intention of finding efficient solutions, scientists also developed advanced statistical methods to increase power in clinical research. The most prominent strategies are: i) Imputation of missing data, ii) repeated measurements, iii) covariate adjustment, or iv) linear mixed models (LMM). Most of these techniques helped to improve clinical research. However, while some of these techniques increase the accuracy of a determined model (e.g. repeated measurements), others require additional assumptions, which potentially are prone to introducing further bias. For example, covariance adjustment leads to biased results, if the imputed covariate is not equally distributed over trial conditions (e.g. no true randomization) (Harrison, 2009; van Breukelen, 2006; Zhang et al., 2014). Additionally, the exact influence of a given covariate is not always clear beforehand, and, thus, covariate adjustment sometimes is of limited value for a priori power or sample size calculations (Pocock, Assmann, Enos, & Kasten, 2002; Raab, Day, & Sales, 2000).

RCTs do not adequately capture intraindividual variation

Randomized clinical trials (RCTs) are mainly conducted using point assessments of psychopathology (e.g. questionnaires applied on a random day at study onset). Recent studies suggest, however, that many psychological constructs (e.g. depressed mood) show substantial intraindividual variation when measured over time (e.g. different days). Even after improving the test length, such fluctuations remain undetected if point assessments are used. Fisher, Medaglia and Jeronimus (2018) showed that the intraindividual variance of depressive symptoms and anxiety is three times higher than interindividual variation. Together with other studies (Pfeiffer et al., 2015), the authors conclude that future research should attempt to capture intraindividual variance more extensively. Implementing EMA or other intense assessment strategies into RCTs can be seen as a pragmatic approach to this, as intraindividual fluctuations are captured over several days, while the research design stays within well-established practices (cf. *Figure 1*).

********** Figure 1 about here **********

Intense assessments are an increasing practice in clinical research. Historically, the (practical) costs of multiple assessments were high, and the investigated impact of single added measurements was relatively small compared to other factors such as sample size or study duration (Moerbeek, 2008; Venter, Maxwell, & Bolig, 2002). Technological advances of the past years made intense assessments less effortful, leading to a clear trend towards EMA, or other forms of time series-based analyses (Bhugra et al., 2017; Holmes et al., 2016). Additionally, other forms of automatized intense assessment have proven feasible in Internet interventions (e.g. weekly or bi-weekly assessments). This is indicated by the umbrella-term *digital mental health,* covering both the provision and the evaluation of digital treatment.

Considering the afore-mentioned aspects, this article investigates the impact of implementing intense assessment by short questionnaires or short EMA (for the purpose of this article abbreviated by sEMA) into RCT-based research designs. While conducting an empiric study would provide first-hand data, simulations yield the advantage to effortlessly test the underlying assumptions independently from the specific study context. In order to optimize study validity, we implemented first-hand data from our lab together with several external sources. The aim of this simulation study is to infer the extent to which intense assessment can contribute to increased power in RCTs. In this regard, the following scenarios (*Table 1*) will be simulated in order to test questionnaire-based point assessments of psychopathology in comparison with intense assessment: In Scenario 1 (standard scenario) we assumed that an average psychological short questionnaire (high correlation) is being applied once, twice, or five times at pre- and at post-measurement. In Scenario 2 those parameters were confirmed by empiric data (Klein et al., 2016; Nuij et al., 2018) on one frequently used short questionnaire - the automatized version of the PHQ-9 - which has been validated for digitalized application as well (Erbe, Eichert, Rietz, & Ebert, 2016). In Scenario 3 we assumed that sEMA (low correlation) is being used to assess state-like depressiveness or depressed mood (Torous & Powell, 2015) instead of applying the

investigated automatized questionnaire. Therefore, the correlation of the single pre- or post-assessments in this scenario is considerably lower.

********** Table 1 about here **********

METHOD

Parameter estimation

The parameters in Scenario 1 corresponded to average reliabilities of frequent (automatized) depression questionnaires in the field of clinical psychology (Drake, Csipke, & Wykes, 2013; Löwe et al., 2004; Vittengl, Clark, Kraft, & Jarrett, 2005). Thus, Scenario 1 will be referred to as standard scenario. In Scenario 2, modeling was based on two data sets. The first data set was the EVIDENT trial (N = 1013) (Klein et al., 2016), a multicenter trial on the effects of online depression treatment. In this trial, PHQ-9 was automatically applied biweekly during treatment course. The time lag between two of the repeated assessments varied, providing a fine-grained gradient of real world correlations. The data also allowed us to model a learning effect (increased correlations over time). Thus, Scenario 2 constitutes an empirically informed analogue to Scenario 1. The estimated PHQ-9 parameters were confirmed by a further data set from a pilot study provided by Nuij and colleagues (2018). This second study investigated smartphone-based self-monitoring by applying automatized PHQ-9 items in a university sample, and correlations of both studies only deviated marginally ($r_{diff} < 0.1$).

For EMA data in Scenario 3 we set correlation to r = 0.4, as provided by data from our lab's research on high frequency time series (Kaiser, & Laireiter, 2019), as well as data from Fisher and Colleagues (2017). Fisher and Colleagues assessed 40 individuals with generalized anxiety disorder (GAD), major depression (MDD), or comorbid GAD and MDD over a period of 30 days. Daily correlations of GAD and MDD scales (based on DSM-V criteria) ranged from r = 0.36 (SD for r = 0.19) for MDD to r = 0.44 (SD for r = 0.20) for GAD. Additionally, average scale values fluctuated, but neither increased nor decreased over the course of time (adjusted R^2 = 0.056 - 0.007), suggesting no reactive

measurement due to multiple assessments. Although consistent with ongoing research from other EMA studies, the correlation of $r = 0.4$ represents an approximation, which in practice depends on potential disease subtypes (e.g. melancholic vs. bipolar) and the severity of a given syndrome; as well as the exact EMA instruction (e.g. "how do you feel at the moment" vs. "how do you feel today") and the item wording. In order to account for this complexity, we present results for higher correlations in *Appendix 1*.

Data simulation

Simulations and graphs were produced using the R packages copula, reshape, and ggplot2. In a first step, the respective covariance structure was extracted from the given real-world data set. As example, Section 1 of *Appendix 1* provides the process of data extraction for PHQ-9 questionnaire data (Scenario 2) based on N =1013 real world patients. The same procedure was applied for EMA data in Scenario 3.

In a next step, we used Clayton und Frank copula to implement the respective covariance structure into the simulation model. Copulas are common mathematical functions that connect joint distributions and their one-dimensional marginal distributions, representing the desired covariance structure and providing the mathematical base for generation of data sets. To assure correctness of generated data sets, Bernstein estimator indicated goodness-of-fit for each Scenario 1-3 (cf. *Figure 2* of *Appendix 1*), and the three assessment types: single-, two-, or fivefold. Bernstein estimators are polynomials for estimating fit of smooth distributions (Leblanc, 2012) on a closed interval (e.g. statistical power between 0 % and 100 %). Again, PHQ-9 data from Scenario 2 serves as example for this simulation step (*Appendix 1*, Section 2), with the corresponding dependence structure for pre-to-post-assessments (Section 2.1), as well as repeated pre- and repeated post-measures (Section 2.2). *Figures 2* and *3* of *Appendix 1* demonstrate the fit between empiric data (blue lines) and simulation model by Bernstein estimator (red curve). After fitting the model, final patient data sets can be generated. For our example, this resulted in smooth slightly skewed distributions (*Figure 4* of *Appendix 1*). We produced 1000 virtual RCTs for 62 different effect sizes and the sample sizes of N = 50, 100, 150, 200.

In a last step, the statistical model of interest (ANOVA or ANCOVA; and LMM or non-linear Bootstrap permutation test for additional analyses) was performed on the 62*1000 virtual RCTs of Scenarios 1 - 3. Whenever applicable, the generated pre- and post-values (two-, or fivefold) were averaged for each simulated patient. For example, if a simulated patient would score 9, 13, 14, 8, 10 on the PHQ-9 at pre-measurement, the resulting value would be 10.8 scale points. This process led to the intended reduction of within-subject error variance in the applied statistical model.

Finally, single results were logged, and statistical power was calculated as the proportion of significant results over all conducted tests (e.g. 800 significant results over 1000 applied AN(C)OVAs: power = 80%). Corresponding results were printed by power curves mapping effect size (x-axis) and achieved power (y-axis). Additional power curves are provided in Sections 3 - 5 of *Appendix 1*.

RESULTS

Standard scenario

In Scenario 1 we tested the influence of multiple assessments on achievable power in questionnaire-based RCTs. *Figure 2* depicts power curves as a function of sample size and number of assessments. Accordingly, ANOVA without multiple assessments resulted in lowest power (e.g. 53 % for $N = 50$, $d = 0.63$), while ANOVA with fivefold questionnaire assessments (e.g. 65 % for $N = 50$, $d = 0.63$) and ANCOVA without multiple assessments (e.g. 63 % for $N = 50$, $d = 0.63$) resulted in comparable power. With a clearly discernible difference, power was highest for ANCOVA with 5 pre-post-assessments (e.g. 81 % for $N = 50$, $d = 0.63$). Furthermore, increases in sample size resulted in higher power (steeper curves for bigger samples), but the proportion of gained power remained constant (green line). This indicates, that multiple questionnaire-based assessments yield advantages independently of the respective sample size. Twofold pre-post-assessments, however, resulted in only marginal power increases.

********** Figure 2 about here **********

Empiric short questionnaire scenario

In Scenario 2, we tested the influence of multiple assessments on gained power based on a model implementing empiric parameters (automatized PHQ-9 assessments) from two external sources. Results (*Figure 3*) coincided with Scenario 1, and, thus, support the validity of standard scenario (e.g. simple ANOVA 49 %; fivefold pre-post-assessments ANCOVA 82 %).

********** Figure 3 about here **********

Empiric EMA scenario

In Scenario 3, we tested the influence of short EMA assessments on power in RCTs. Due to their weaker auto-correlation, potential power gains are (per sé) expected to be higher in this scenario. The results are depicted in *Figure 4*, where fivefold sEMA of ANOVA (e.g. 78 % for $N = 50$, $d = 0.63$) already outperformed standard ANCOVA. Power was highest in fivefold sEMA combined with baseline ANCOVA (e.g. 94 % for $N = 50$, $d = 0.63$), and lowest if only twofold sEMA was applied.

********** Figure 4 about here **********

Comparison of absolute power

Additionally, the absolute power of both strategies can be compared. *Table 2* presents proportions of relative and absolute power gains for Scenario 2 (empiric PHQ-9 data) and Scenario 3 (empiric sEMA). Relative increases in power ranged from 6% to 92%, with highest increase rates for more severely underpowered studies. Importantly, sEMA outperformed point assessments of psychopathology in terms of absolute statistical power. For example, ANCOVA of simple PHQ-9

questionnaire data resulted in an absolute statistical power of 55 (for $N = 200$ and $d = 0.3$). Fivefold sEMA with baseline as covariate, however, resulted in power of 88.9 to detect a comparable effect in a comparable sample.

********** Table 2 about here **********

Additional findings

In order to test the robustness of findings, we conducted additional simulations based on non-parametric tests and simple linear mixed models (LMM). As a proof of concept, and to avoid redundancy, the corresponding findings are presented in *Appendix 1*. Non-parametric and parametric tests yielded comparable results, indicating good robustness independent of scaling (ordinal vs. interval data). LMM led to comparable effects as obtained in ANCOVA, if baseline was used as covariate. A plot of apriori (predefined) versus observed effect sizes is provided in *Appendix 1*, Section 6. This plot indicates that averaging across twofold or fivefold pre- or post-assessments (to achieve the intended variance reduction) did not bias the results (e.g. overestimation of true effect).

DISCUSSION

This study examined the effects of intense pre-post-assessment on achievable statistical power in RCTs. It is based on the assumption that repeated assessments will allow more precise estimation of psychopathology, reducing unrespectable variance within subjects (in terms of time-related fluctuations). Reduced error variance increases the proportion of explainable to unexplainable variance, resulting in increased statistical power, and, thus, higher sensitivity to changes. To test the magnitude of expectable power increases, three scenarios were simulated.

Principal findings indicate that RCTs with intense assessment lead to power gains beyond standard methods of point assessment of psychopathology. A simulation based on empiric parameters from two external sources (Scenario 2) coincided with the corresponding standard scenario (Scenario 1; assessment by short questionnaire), indicating high generalizability of presented findings. Furthermore, short pre-post EMA (sEMA) resulted in highest absolute statistical power when compared to automatized point assessments. Thus, findings suggest that sEMA or comparable forms of intensive repeated assessment may be well suited for implementation into RCT-based research, as they can outperform standard methods. In the wider perspective, multiple assessments could provide a strategy to tackle the problem of underpowered studies in clinical research (Khan, Fahl, & Brown, 2018; Roozenbeek, Lingsma, Steyerberg, & Maas, 2010; Szucs & Ioannidis, 2017), as small sample size situations exhibited highest improvements.

Questionnaire-like data (high correlation)

The principal study results indicated a clear superiority of fivefold over twofold pre-post-assessments, with the latter leading to marginal power increases (cf. *Figure 2*). This finding is in line with studies indicating only small power gains through occasional repeated assessments (Moerbeek, 2008; Venter, Maxwell, & Bolig, 2002).

For fivefold pre-post-assessments, power gains of ANOVA were comparable to applying point assessments and ANCOVA with baseline as covariate (cf. *Figure 2*). So far, baseline ANCOVA without intense assessments would be indicated as it constitutes the most efficient way to optimize power (van Breukelen, 2006; Zhang et al., 2014). However, as multiple pre-assessment provides a more precise estimation of the investigated construct (e.g. depressed mood), the precision of the baseline covariate also improved. According to our simulation, the combination of intense assessment and ANCOVA led to substantial power gains. Contrary to the sometimes unknown influence of additional third variables in ANCOVA (Harrison, 2009; Pocock, Assmann, Enos, & Kasten, 2002), potential sample size reductions can be approximated by standard parameters (e.g. retest reliability). This means that multiple assessment is applicable for a priori sample size calculation, and could

thereby help to reduce the costs of conducting clinical research. *Appendix 2* provides a pragmatic guide on how expectable sample reductions can be approximated by the popular G*Power software.

EMA-like data (low correlation)

Even though EMA and other time series-based procedures are increasingly used in clinical research (Bhugra et al., 2017; Holmes et al., 2016), the practice of implementing them into RCTs to improve statistical power is not widespread. However, first empirical evidence exists. For example, a recent study on the comparison of EMA-based and paper-pencil measures of depression and anxiety reported a 25-50% improvement of change sensitivity (number needed to treat, NNT) with 10 pre- and 10 post-assessments (Moore, Depp, Wetherell, & Lenze, 2016). In this context, previous simulations revealed that more weakly related constructs can result in higher statistical power in multiple assessment situations (Basagana, & Spielman, 2011, p. 61). Further supportive evidence comes from a medical study on irritable bowel syndrome (IBS), in which sensitivity of retrospective symptom rating was improved by EMA (Vork et al., 2019). In this study, 10 assessments were taken during seven consecutive days per measurement period. Taken together, recent empiric findings support the assumption that change in psychological constructs could rather be evaluated by time series then by point assessments of psychopathology (Fisher, Medaglia, & Jeronimus, 2018; Moore, Depp, Wetherell, & Lenze, 2016, Vork et al., 2019). Such intense assessments could simultaneously serve to investigate temporal dynamics of disease symptoms (or syndromes) (Bos, Schoevers, & Rot, 2015), to improve classification (Pfeiffer et al., 2015), *and* to improve statistical power in clinical trials.

Regarding the optimal number of assessments per measurement period (e.g. baseline, post-assessment, and follow-up) a range of 5 to 10 assessments appears advisable, to balance expectable gains in statistical power against burden of assessment. While the above mentioned empiric studies (Moore et al., 2016; Vork et al., 2019) implemented 10 assessments, our computer simulation suggests reasonable improvements with 5 valid assessments. Therefore, 6 to 7 assessments per measurement period appear advisable, if 20 – 30 % missing data are being taken into account. In this context, common statistical power calculators can be helpful for providing a rough estimation of expectable power increases. We therefore provide a pragmatic guide on how sample reductions can be

approximated by the popular G*Power software in *Appendix 2*. At this, restrictions exists as many open software calculators only allow to specify one single value in the covariance matrix, assuming equal correlations between groups or within repeated measurements. Furthermore, models usually entail equidistance between assessments (e.g. continuous weekly or bi-weekly point assessments over the treatment period), leading to differences in statistical model and empiric data (including additional variance caused by treatment). In contrast, correlations of repeated assessments between and within each measurement period of sEMA will deviate considerably. *Appendix 2* therefore features a table to estimate deviation between G*Power and our model. Independently from the specific assessment strategy (questionnaire- or EMA-based) the overall convergence of both models was sufficient to allow sample size planning.

Pros & cons of sEMA

Summing up, sEMA (and multiple applications of short questionnaires) might constitute a promising approach to tackle some current problems in clinical research by blending EMA and RCT-based paradigms. While daily EMA assessments over the entire study period may quickly overload patients, a limited number of assessments in the pre- and post-phase of a clinical study seems much more feasible (Verhagen et al., 2016). Nevertheless, this additional effort results in significant increase in data quality. Especially small trials in early research stages and lager trials with active comparators (e.g. testing against the gold-standard treatment) could benefit from selecting a set of items to be assessed intensely. In this regard, feasibility of sEMA might be higher for the suggested research context (e.g. multiple baseline designs), and lower for standard application in routine care. On the other hand, many routine Internet-based treatments feature intense assessment in terms of weekly monitoring, and comparable approaches are being implemented into routine blended treatment (Lutz, Rubel, Schwartz, Schilling, & Deisenhofer, 2019). Thus, weekly assessments can be a useful alternative to optimize statistical power. Finally, the decision between those two forms of intense assessment depends on study purpose (e.g. focus on treatment process versus outcome, or type of mediation analysis, or assessment intensity of process variables).

As a related topic, EMA has been suspected to not only measure, but also influence symptoms of mental health. Until now, the exact circumstances of "reactive measurement" in psychiatric research are unknown (Mehl & Conner, 2011; Schrimsher, & Filtz, 2011), and corresponding studies are still ongoing (van Ballegooijen et al., 2016). Experts previously suggested that EMA may be contraindicated for patients with severe psychiatric conditions (Rot, Hogenelst, & Schoevers, 2012), or with high social desirability (e.g. alcohol intake) (Johnson et al., 2009). Amongst other strategies to counteract potential reactivity (and to increase engagement) (Sandstrom, Lathia, Mascolo, & Rentfrow, 2016), Torous and colleagues (2015) investigated the benefits of item-shuffling. Regarding the empiric data of the current study, we did not find any reactive measurement in terms of increased or decreased scale values over the course in time. However, in order to test the potential impact of another form of reactivity (increased auto-correlations), such a learning effect was implemented into Scenario 2. This effect did not excerpt any relevant influence on the presented results. To consider pros- and cons of intense assessment an overview of relevant aspects is provided in *Table 3*.

********** Table 3 about here **********

Recommendations for applied researchers:

- Use intense assessment regimes in RCTs in order to optimize statistical power.
- Apply the same regime in the experimental and control group to assure study validity.
- Choose at best 10 valid repeated assessments for the active treatment period (pre-to-post).
- Account for missing data.
- Consider EMA as a useful alternative to classic point assessments of psychopathology.
- Consider sEMA whenever psychopathology needs to be assessed with maximum precision before and after treatment (e.g. to estimate treatment effects).
- Consider sEMA for process studies, for example relating psychological constructs to physiological point assessments (e.g. EEG or fMRI), or treatment moderators and mediators, or optimization problems (e.g. machine learing).
- Use information featured in figures and *Appendix 1* & *2* for sample size planning.

- Be aware of restrictions (e.g. limited number of constructs to be assessed) and possible risks of intense assessment (e.g. higher burden for patients).

Strengths and limitations

This study has several noteworthy strengths and limitations. Amongst its most important strengths, reported findings are based on numerous data sets and simulations and therefore reproducible and well interpretable, providing insights which are merely independent from fluctuations in single trials. Additionally, a replication modeled after standard parameters (Scenario 1) and complementing analyses based on non-parametric tests and basic linear mixed models supported principal findings. Furthermore, the simulation process was carried out by four authors (RS, MS, TK, WT), resulting in a high degree of mutual control in a multidisciplinary team. During the process, two models were developed independently, and integrated stepwise.

Regarding the study limitations, further evidence from empiric research is warranted for some forms of intense assessment. Due to the novelty of sEMA, only scarce empiric evidence for its positive impact on statistical power exists (Moore et al., 2016; Vork et al., 2019). Additionally, findings may not account for psychiatric conditions with more complex symptom dynamics (e.g. PTSD or eating disorders), or whenever strong patient reactivity is being expected. As a further limitation, the presented simulations did not include missing data or dropout. Therefore, the calculated effects will be lower at an increasing dropout rate. On the other hand, intense assessments damper the impact of missing out single assessments – which is equivalent to dropout in RCTs. At this, the investigated maximum of five assessments per measurement occasion was set somewhat arbitrary, with eight or ten assessments constituting a feasible alternative. The specific impact of missing data will depend on study context. While data missing at random (e.g. 20 % missing data) has relatively small impact on reported findings and easily can be compensated by one extra assessment (e.g. 6 instead of 5), scenarios with very low participant engagement at post treatment (e.g. with high proportions of dropout) will lead to reduction in beneficial effects on statistical power. As a last limitation, more complex methods of time series analysis would have been applicable as well. One such strategy

includes complexity or entropy measures from non-linear time series, as investigated by our workgroup (Kaiser & Laireiter, 2018). Alternatively, hierarchical methods (LMM) are one suggested standard method in classic and recent literature (Schwartz, & Stone, 1998; Bolger & Laurenceau, 2013). To speed up the simulation process, our principal analysis was based on AN(C)OVA, with complementary analyses based on LMM and permutation tests. At this, the conducted bias analysis indicated robustness of findings (cf. *Appendix 1*). One possible explanation is that the method of averaging over multiple assessments constitutes a data aggregation and not a disaggregation procedure (Nezlek, 2001), which optimize other statistical requirements, such as the underlying normal distribution or homogeneity of variances, of the simulation process.

Conclusion

To sum up, intense assessment strategies indicate clear superiority of multiple assessments over frequently used point assessments of psychopathology. At this, time series-based procedures (such as EMA) can outperform classic point assessments by a comparably low number of repeated assessments. This is because psychological constructs underlie natural fluctuations which cannot be addressed by means of test extension. As automatization has made multiple assessments less effortful, intense assessment strategies are seen more frequently in clinical and in research context (e.g. weekly assessment, or multiple baseline assessment). Further ongoing evidence on sEMA's feasibility is promising, but more research in diverse populations is needed. As clinical research suffers from underpowered studies, intense assessment strategies should find more recognition in RCT-based designs.

Acknowledgements

We want to thank Wouter van Ballegooijen and his research team for sharing of EMA data. Furthermore, we want to thank Aaron Fisher for providing EMA data via online repositories.

References

Basagana S, Spielman D. 2011. The Design of Observational Longitudinal Studies. https://cdn1.sph.harvard.edu/wp-content/uploads/sites/271/2012/08/optitxs-The-Design-of-Observational-Longitudinal-Studies.pdf. Archived at: http://www.webcitation.org/72avnnPNJ

Bhugra, D., Tasman, A., Pathare, S., Priebe, S., Smith, S., Torous, J., ... & First, M. B. (2017). The WPA-lancet psychiatry commission on the future of psychiatry. The Lancet Psychiatry, 4(10), 775-818.

Bolger, N., & Laurenceau, J. P. Intensive longitudinal methods: an introduction to diary and experience sampling research. 2013. New York: Guilford Press.

Bos, F. M., Schoevers, R. A., & aan het Rot, M. (2015). Experience sampling and ecological momentary assessment studies in psychopharmacology: A systematic review. European Neuropsychopharmacology, 25(11), 1853-1864.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience, 14(5), 365.

Califf, R. M., Zarin, D. A., Kramer, J. M., Sherman, R. E., Aberle, L. H., & Tasneem, A. (2012). Characteristics of clinical trials registered in ClinicalTrials. gov, 2007-2010. Jama, 307(17), 1838-1847.

Drake, G., Csipke, E., & Wykes, T. (2013). Assessing your mood online: acceptability and use of Moodscope. Psychological medicine, 43(7), 1455-1464.

Erbe, D., Eichert, H. C., Rietz, C., & Ebert, D. (2016). Interformat reliability of the patient health questionnaire: Validation of the computerized version of the PHQ-9. Internet Interventions, 5, 1-4.

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. Proceedings of the National Academy of Sciences, 201711978.

Fisher, A. J., Reeves, J. W., Lawyer, G., Medaglia, J. D., & Rubel, J. A. (2017). Exploring the idiographic dynamics of mood and anxiety via network analysis. Journal of Abnormal Psychology, 126(8), 1044-1056.

Halpern, S. D., Karlawish, J. H., & Berlin, J. A. (2002). The continuing unethical conduct of underpowered clinical trials. Jama, 288(3), 358-362.

Hansen, W. B., & Collins, L. M. (1994). Seven ways to increase power without increasing N. NIDA research monograph, 142, 184-184.

Harrison, D. A. (2009). Increasing power in randomized controlled trials. Critical care medicine, 37(10), 2840-2841.

Holmes, E. A., Bonsall, M. B., Hales, S. A., Mitchell, H., Renner, F., Blackwell, S. E., ... & Di Simplicio, M. (2016). Applications of time-series analysis to mood fluctuations in bipolar disorder to promote treatment innovation: a case series. Translational psychiatry, 6(1), e720.

Johnson, E. I., Grondin, O., Barrault, M., Faytout, M., Helbig, S., Husky, M., ... & Swendsen, J. (2009). Computerized ambulatory monitoring in psychiatry: a multi-site collaborative study of acceptability, compliance, and reactivity. International journal of methods in psychiatric research, 18(1), 48-57.

Kaiser, T., & Laireiter, A. R. (2018). Daily dynamic assessment and modelling of intersession processes in ambulatory psychotherapy: A proof of concept study. Psychotherapy Research, 1-12.

Kaiser, T., & Laireiter, A-R. (2019). Process-symptom-bridges in psychotherapy: an idiographic network approach. Psychotherapy Research.

Khan, A., Fahl, M. K., & Brown, W. A. (2018). The Impact of Underpowered Studies on Clinical Trial Results. The American journal of psychiatry, 175(2), 188.

Klein, J. P., Berger, T., Schröder, J., Späth, C., Meyer, B., Caspar, F., ... & Hautzinger, M. (2016). Effects of a psychological internet intervention in the treatment of mild to moderate depressive symptoms: results of the EVIDENT study, a randomized controlled trial. Psychotherapy and psychosomatics, 85(4), 218-228.

Leblanc, A. (2012). On estimating distribution functions using Bernstein polynomials. *Annals of the Institute of Statistical Mathemathics*, *64*(5), 919-943.

Löwe, B., Unützer, J., Callahan, C. M., Perkins, A. J., & Kroenke, K. (2004). Monitoring depression treatment outcomes with the patient health questionnaire-9. Medical care, 1194-1201.

Lutz, W., Rubel, J. A., Schwartz, B., Schilling, V., & Deisenhofer, A. K. (2019). Towards integrating personalized feedback research into clinical practice: Development of the Trier Treatment Navigator (TTN). *Behaviour research and therapy*, 120, 103438.

Maddock, J. E., & Rossi, J. S. (2001). Statistical power of articles published in three health-psychology related journals. Health psychology, 20(1), 76-78.

Marszalek, J. M., Barber, C., Kohlhart, J., & Cooper, B. H. (2011). Sample size in psychological research over the past 30 years. Perceptual and motor skills, 112(2), 331-348.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: causes, consequences, and remedies. Psychological methods, 9(2), 147.

Mehl, M. R., & Conner, T. S. (2011). Handbook of research methods for studying daily life. New York: Guilford Press.

Moerbeek, M. (2008). Powerful and cost-efficient designs for longitudinal intervention studies with two treatment groups. Journal of Educational and Behavioral Statistics, 33(1), 41-61.

Moore, R. C., Depp, C. A., Wetherell, J. L., & Lenze, E. J. (2016). Ecological momentary assessment versus standard assessment instruments for measuring mindfulness, depressed mood, and anxiety among older adults. Journal of psychiatric research, 75, 116-123.

Nezlek, J. B. (2001). Multilevel random coefficient analyses of event-and interval-contingent data in social and personality psychology research. Personality and Social Psychology Bulletin, 27(7), 771-785.

Nuij, C., van Ballegooijen, W., Ruwaard, J., de Beurs, D., Mokkenstorm, J., van Duijn, E., ... & Kerkhof, A. (2018). Smartphone-based safety planning and self-monitoring for suicidal patients: Rationale and study protocol of the CASPAR (Continuous Assessment for Suicide Prevention And Research) study. Internet Interventions, 13, 16-23.

Pfeiffer, P. N., Bohnert, K. M., Zivin, K., Yosef, M., Valenstein, M., Aikens, J. E., & Piette, J. D. (2015). Mobile health monitoring to characterize depression symptom trajectories in primary care. Journal of affective disorders, 174, 281-286.

Pocock, S. J., Assmann, S. E., Enos, L. E., & Kasten, L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. Statistics in medicine, 21(19), 2917-2930.

Raab, G. M., Day, S., & Sales, J. (2000). How to select covariates to include in the analysis of a clinical trial. Controlled clinical trials, 21(4), 330-342.

Roozenbeek, B., Lingsma, H. F., Steyerberg, E. W., & Maas, A. I. (2010). Underpowered trials in critical care medicine: how to deal with them? Critical Care, 14(3), 423.

Roozenbeek, B., Maas, A. I., Lingsma, H. F., Butcher, I., Lu, J., Marmarou, A., ... & Steyerberg, E. W. (2009). Baseline characteristics and statistical power in randomized controlled trials: selection, prognostic targeting, or covariate adjustment? Critical care medicine, 37(10), 2683-2690.

Rot, M., Hogenelst, K., & Schoevers, R. A. (2012). Mood disorders in everyday life: a systematic review of experience sampling and ecological momentary assessment studies. Clinical psychology review, 32(6), 510-523.

Saeb, S., Zhang, M., Kwasny, M., Karr, C. J., Kording, K., & Mohr, D. C. (2015, May). The relationship between clinical, momentary, and sensor-based assessment of depression. In Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2015 9th International Conference on (pp. 229-232). IEEE.

Sandstrom, G. M., Lathia, N., Mascolo, C., & Rentfrow, P. J. (2016). Opportunities for smartphones in clinical care: the future of mobile mood monitoring. The Journal of clinical psychiatry, 77(2), e135.

Schrimsher, G. W., & Filtz, K. (2011). Assessment reactivity: Can assessment of alcohol use during research be an active treatment?. Alcoholism treatment quarterly, 29(2), 108-115.

Schwartz, J. E., & Stone, A. A. (1998). Strategies for analyzing ecological momentary assessment data. Health Psychology, 17(1), 6-16.

Szucs, D., & Ioannidis, J. P. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. PLoS biology, 15(3), e2000797.

Torous, J., & Powell, A. C. (2015). Current research and trends in the use of smartphone applications for mood disorders. Internet Interventions, 2(2), 169-173.

Torous, J., Staples, P., Shanahan, M., Lin, C., Peck, P., Keshavan, M., & Onnela, J. P. (2015). Utilizing a personal smartphone custom app to assess the patient health questionnaire-9 (PHQ-9) depressive symptoms in patients with major depressive disorder. JMIR mental health, 2(1).

van Ballegooijen, W., Ruwaard, J., Karyotaki, E., Ebert, D. D., Smit, J. H., & Riper, H. (2016). Reactivity to smartphone-based ecological momentary assessment of depressive symptoms (MoodMonitor): protocol of a randomised controlled trial. BMC psychiatry, 16(1), 359.

Van Breukelen, G. J. (2006). ANCOVA versus change from baseline had more power in randomized studies and more bias in nonrandomized studies. Journal of clinical epidemiology, 59(9), 920-925.

Venter, A., Maxwell, S. E., & Bolig, E. (2002). Power in randomized group comparisons: The value of adding a single intermediate time point to a traditional pretest-posttest design. Psychological Methods, 7(2), 194.

Verhagen, S. J., Hasmi, L., Drukker, M., van Os, J., & Delespaul, P. A. (2016). Use of the experience sampling method in the context of clinical trials. Evidence-based mental health, 19(3), 86-89.

Vittengl, J. R., Clark, L. A., Kraft, D., & Jarrett, R. B. (2005). Multiple measures, methods, and moments: a factor-analytic investigation of change in depressive symptoms during acute-phase cognitive therapy for depression. Psychological Medicine, 35(5), 693-704.

Vork, L., Mujagic, Z., Drukker, M., Keszthelyi, D., Conchillo, J. M., Hesselink, M. A., ... & Kruimel, J. W. (2019). The Experience Sampling Method—Evaluation of treatment effect of escitalopram in IBS with comorbid panic disorder. Neurogastroenterology & Motility, 31(1), e13515.

Wampold, B. E., Flückiger, C., Del Re, A. C., Yulish, N. E., Frost, N. D., Pace, B. T., ... & Hilsenroth, M. J. (2017). In pursuit of truth: A critical examination of meta-analyses of cognitive behavior therapy. Psychotherapy Research, 27(1), 14-32.

Zhang, S., Paul, J., Nantha-Aree, M., Buckley, N., Shahzad, U., Cheng, J., ... & Avram, V. (2014). Empirical comparison of four baseline covariate adjustment methods in analysis of continuous outcomes in randomized controlled trials. Clinical epidemiology, 6, 227.

Figure 1. Different slopes of improvement as a function of measurement day during pre- and post-assessment.

Figure 2. Power for standard scenario.

Figure 3. Power for EMA data.

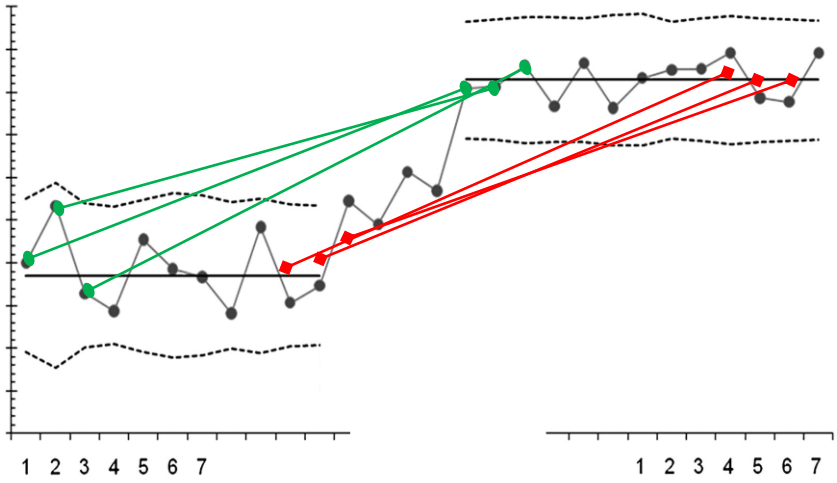
Figure 4. Power modeled according to empiric data.

Table 1. Scenarios to test the impact of sEMA

Table 2. Achieved power through intense pre-post-assessment (sEMA)

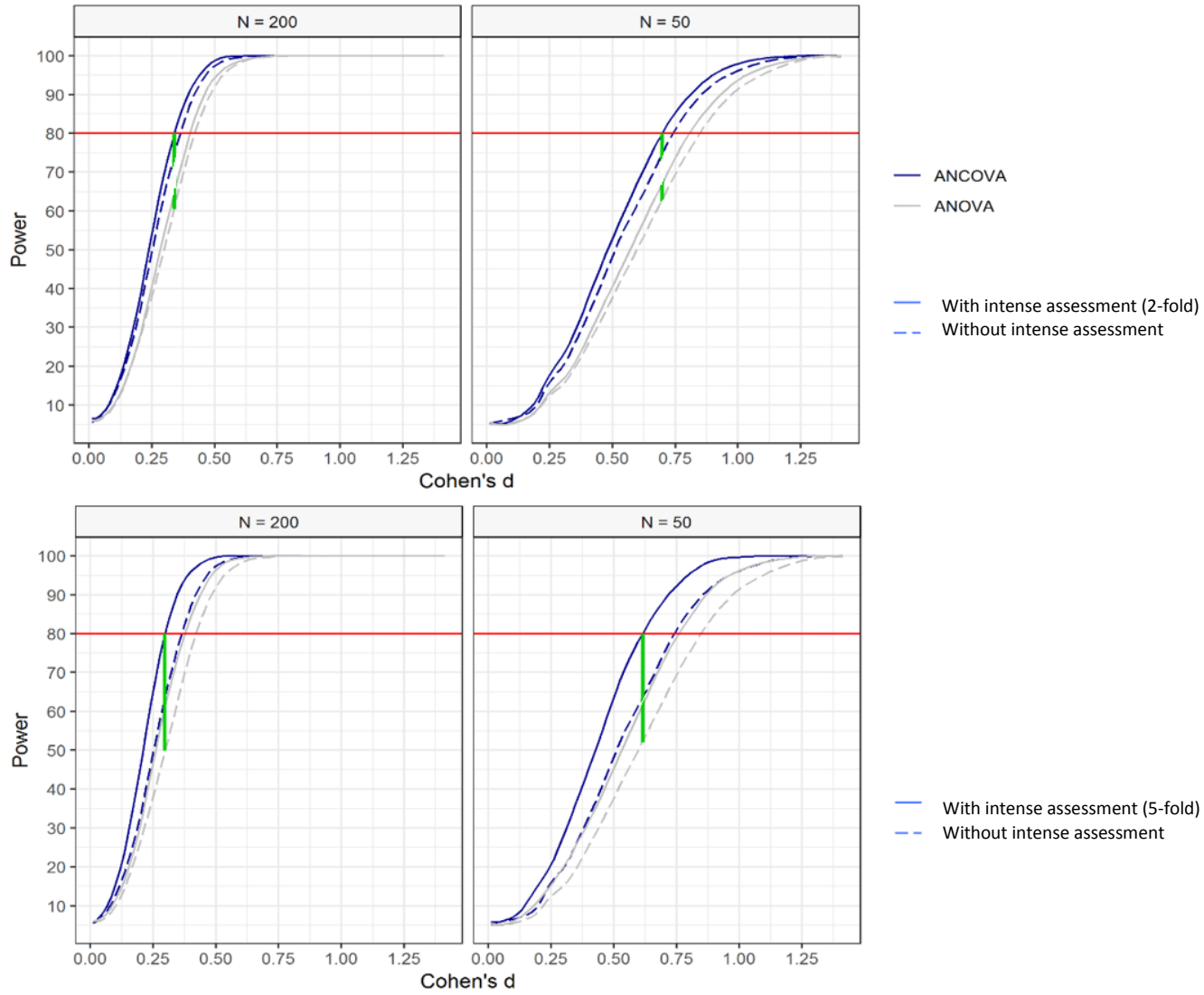
Table 3. Advantages and disadvantages of sEMA

Figure 1. Different slopes of improvement as a function of measurement day introduce measurement error.



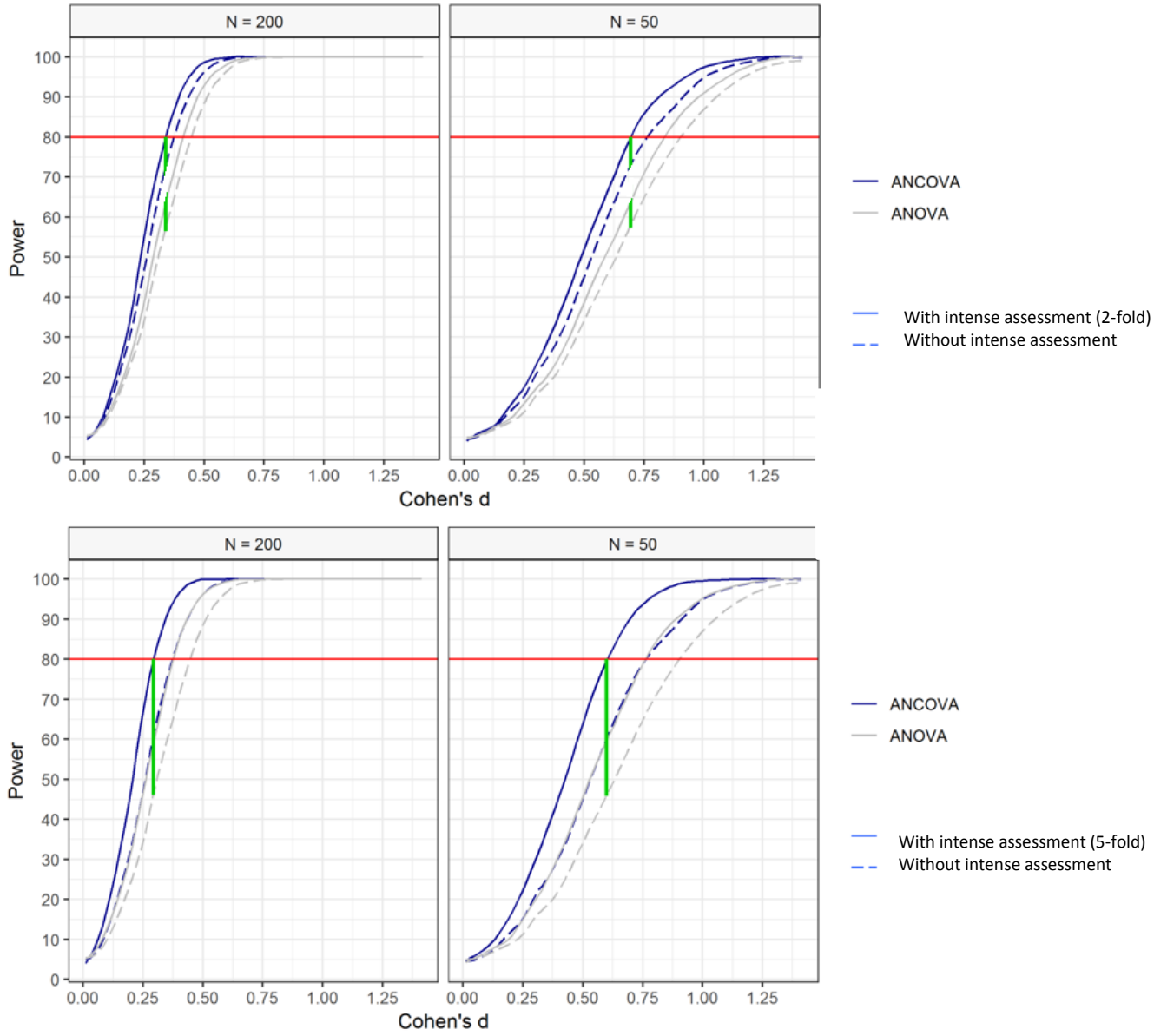
Note. Point assessments of psychopathology (by standard questionnaires) introduce measurement error as symptoms fluctuate over time. For example, for a questionnaire with 16 items and a standard deviation of *SD* = 5, a fluctuation of 1 point on 2 items of a given Likert scale would result in 40% fluctuation of *SD*. This imprecision increases if both, pre- and post-assessment, are affected equally. Green lines represent three slopes of single point assessments. Red lines represent averaged slopes over a moving window of three measurement occasions.

Figure 2. Power curves mapping effect size (x-axis) and achieved power (y-axis) for Scenario 1 (standard scenario).



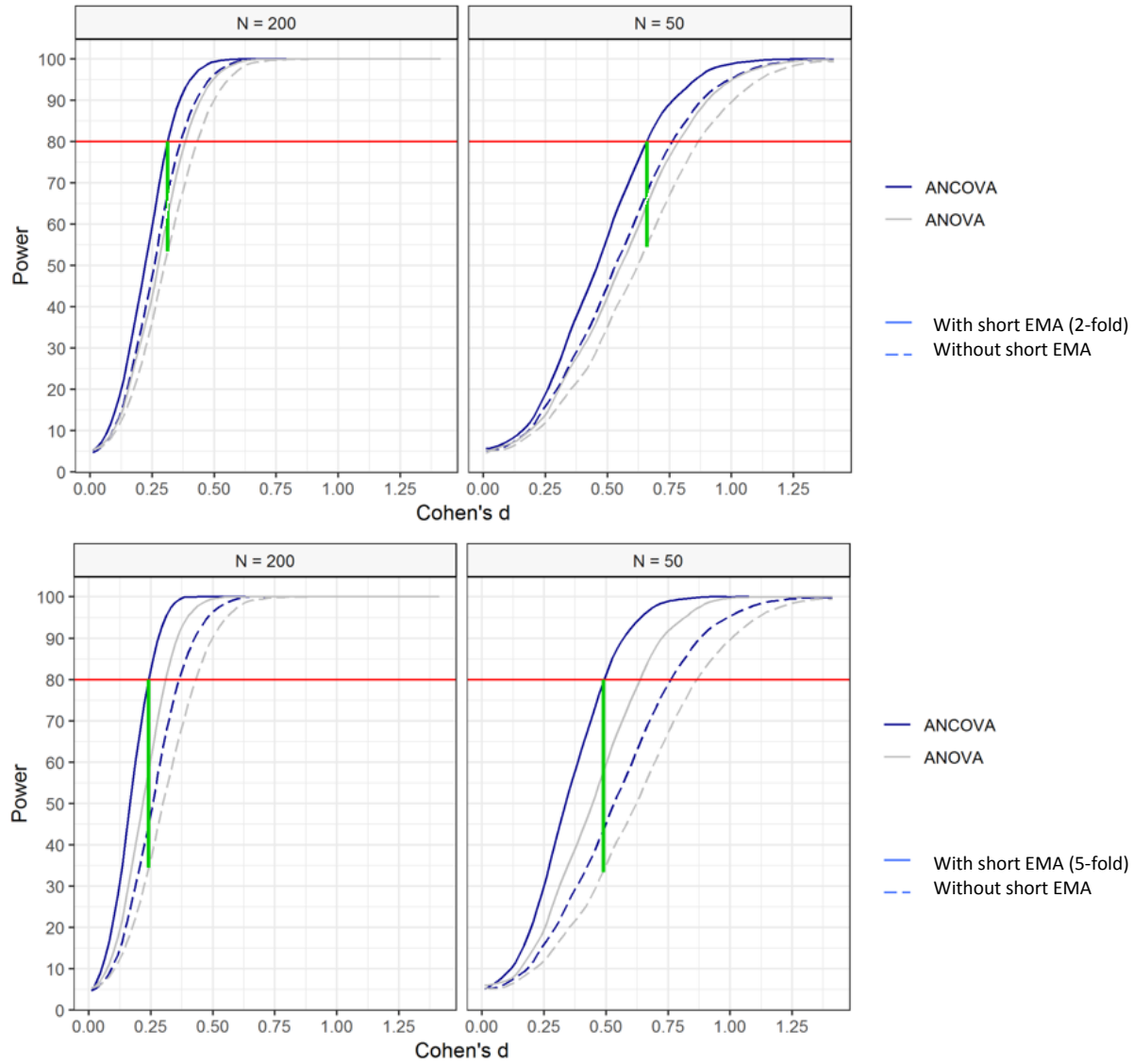
Note. Green line = power gain; red line = 80% power level; dashed line: standard AN(C)OVA; solid line: intense assessment.

Figure 3. Power curves mapping effect size (x-axis) and achieved power (y-axis) for Scenario 2 (empiric data based on automatized PHQ-9 assessments).



Note. Green line = power gain; red line = 80% power level; dashed line: standard AN(C)OVA; solid line: intense assessment.

Figure 4. Power curves mapping effect size (x-axis) and achieved power (y-axis) for Scenario 3 (empiric EMA data).



Note. Green line = power gain; red line = 80% power level; dashed line: standard AN(C)OVA; solid line: intense assessment.

Table 1. Scenarios to test the impact of intense assessment

	Scenario 1 (standard scenario)	Scenario 2 (emp. trial data)	Scenario 3 (emp. EMA data)
Assessment method	Average questionnaire	Automatized PHQ-9	Automatized EMA
Reliability of repeated pre-assessments (*r*)	0.7	≈ 0.4 - 0.65	0.4
Reliability of repeated post-assessments (*r*)	0.7	≈ 0.4 - 0.65	0.4
Quantity of pre-assessments	2 or 5	2 or 5	2 or 5
Quantity of post-assessments	2 or 5	2 or 5	2 or 5

Abbreviations: EMA = ecological momentary assessment; PHQ-9 = Patient Health Questionnaire (depression); *r* = auto-correlation.

Table 2. Achieved power through intense pre-post-assessment by automatized short questionnaires or sEMA

	ANOVA			ANCOVA		
	Standard pre-post	Twofold pre-post	Fivefold pre-post	Standard pre-post	Twofold pre-post	Fivefold pre-post
Simulation	Power (%)	Power (%*)	Power (%*)	Power (%)	Power (%*)	Power (%*)
Scenario 2 (automatized PHQ-9)						
$N = 50$; $d = 0.8$[a]	**68.0 (100)**	72.1 (106)	76.7 (113)	**79.8 (100)**	83.8 (105)	90.4 (113)
$N = 100$; $d = 0.5$[a]	**58.3 (100)**	61.7 (106)	66.7 (114)	**70.1 (100)**	76,7 (109)	83.8 (120)
$N = 200$; $d = 0.3$[a]	**43.8 (100)**	47.9 (109)	52.9 (121)	**55.0 (100)**	60.4 (110)	71,3 (130)
Scenario 3 (automatized EMA)						
$N = 50$; $d = 0.8$[a]	58.2 (100)	71.9 (123)	**92.8 (159)**	75.2 (100)	89.5 (119)	**99.4 (132)**
$N = 100$; $d = 0.5$[a]	51.3 (100)	64.4 (125)	**87.2 (169)**	66.9 (100)	82.3 (123)	**97.5 (146)**
$N = 200$; $d = 0.3$[a]	38.4 (100)	50.0 (130)	**73.9 (192)**	51.4 (100)	67.5 (131)	**88.9 (173)**

Note. Fivefold sEMA (columns 3 and 6 of Scenario 3) clearly outperforms questionnaire-based point assessments of psychopathology (columns 1 and 4 of Scenario 2) in terms of absolute statistical power.
Abbreviations: sEMA = intense pre-post-Ecological Momentary Assessment; %* = increase in percent relative to reference; *N* = number of participants. [a] Cohen's *d*.

Table 3. Advantages and disadvantages of intense assessment

Advantage	Disadvantage
Fits recent trends in clinical research	More feasibility research needed
Increases measurement precision	May act as intervention
Reduces impact of missing assessments	Increases burden for participants
Provides additional information on disease dynamics	Applicability decreases with number of items
Improves triangulation of data sources (e.g. neuroscience)	
Increases statistical power / reduces required sample size	