# Reliable Change in Times of Ecological Momentary Assessment: Adaptation and Expectable Increases in Classification Accuracy

## Masterarbeit

eingereicht von

Stephan Bartholdy, B.Sc.

Matrikelnummer: 01522628

Zur Erlangung des

Mastergrades MSc

an der Naturwissenschaftlichen Fakultät

der Paris–Lodron–Universität Salzburg

Gutachter:

Ao. Univ.–Prof. Dr. Anton–Rupert Laireiter

Dr. Raphael Schuster

Fachbereich Psychologie

Salzburg, September 2021

# Acknowledgements

I want to thank my supervisor, Ao. Univ.–Prof. Dr. Anton–Rupert Laireiter, for always offering his advice and expertise and allowing me to work as flexibly and as freely as I needed to.

I also want to thank my second supervisor, Dr. Raphael Schuster, for his unlimited help and kind encouragement throughout the whole process of writing this thesis. His level of interest in the topic and his constructive feedback were inspiring to me and helped me very much.

Apart from the excellent supervision by both of them, I´m thankful to the whole team of authors consisting of Raphael Schuster, Manuela Larissa Schreyer, Tim Kaiser, Thomas Berger, Jan Philipp Klein, Steffen Moritz, Anton-Rupert Laireiter, and Wolfgang Trutschnig for providing me with data which they generated for their own recent study on the statistical power of intense pre-post assessment approaches.

On a side note, I want to acknowledge that this thesis was written using the *Salzburgthesisdown* template[1] by Veronika Priesner. Based on the *Thesisdown* package[2] (Ismay & Solomon, 2020), this format allows for the preparation and formatting of theses using a combination of R code, Markdown and LaTeX syntax.

Finally, I want to thank my parents Laura and Frank, as well as my grandparents, for supporting me unconditionally throughout every step of my life and education. I dedicate this thesis, being the most difficult task of my life yet, to you, because without you I would not have been able to fulfill any of the dreams nor achieve any of the goals that I had in my life until today.

[1] `https://github.com/irmingard/salzburgthesisdown`
[2] `https://github.com/ismayc/thesisdown`

# Table of Contents

# List of Tables

# List of Figures

# Abstract

Digital mental health and precision medicine practices introduce new ways of data acquisition and analysis into the field of psychiatric research, which have the potential to advance the individualization of psychotherapy. Within this line of research, the present study addresses two important problems which affect the validity of empirical findings in psychiatric research: (1) the limited precision of single-point assessment strategies for measuring psychological symptoms, and (2) the lack of truly individualized measures for assessing clinically significant change on the level of single patients. These issues could be minimized by applying individualized change criteria to ecological momentary assessment data. The two aims of this study were (1) to investigate the theoretically expected increases in sensitivity and specificity of clinical change estimates in clinical trials resulting from the use of multiple baseline and follow-up assessments, as practically implemented in ecological momentary assessment, as well as (2) to evaluate the psychometric quality of methods for determining meaningful change in longitudinal clinical trials. An exploratory simulation study was conducted, in which three clinical change methods (i.e. Percentage Change, Reliable Change Index, and the Individualized Reliable Change Index as a proposed adaptation of the RCI) were compared for both a classical questionnaire format and an EMA format of the PHQ-9 scale for depressive symptoms. The results suggest an accuracy advantage of repeated-assessment over single-point assessment approaches, with accuracy increases ranging from 17–22% in questionnaire scenarios and generally higher accuracy levels in EMA scenarios. Furthermore, individualized clinical change criteria generally resulted in higher sensitivity and specificity levels than criteria calculated from group-level information (i.e. assuming linear change), especially in repeated-assessment scenarios. Implications, strengths and limitations are discussed.

*Keywords:* digital mental health, EMA, precision, reliable change, percentage change

# Zusammenfassung

Digital Mental Health und Präzisionsmedizin führen neue Wege der Datenerhebung und -analyse in die psychiatrische Forschung ein, die das Potenzial haben, die Individualisierung von Psychotherapie voranzutreiben. Im Rahmen dieses Forschungszweiges befasst sich die vorliegende Arbeit mit zwei wichtigen Problemen, die die Validität empirischer Befunde in der psychiatrischen Forschung beeinträchtigen: (1) die begrenzte Messgenauigkeit von Strategien zur Messung psychiatrischer Symptome mit nur 2 Messzeitpunkten und (2) das Fehlen tatsächlich individualisierter Maße zur Beurteilung klinisch bedeutsamer Veränderungen in der Therapie-Wirksamkeitsforschung. Diese Probleme könnten durch die Anwendung individualisierter Veränderungskriterien in Ecological Momentary Assessment-Daten minimiert werden. Die beiden Ziele dieser Studie waren (1) die Untersuchung des theoretisch erwarteten Zugewinns an Sensitivität und Spezifität von Klassifikationsmethoden für klinisch signifikante Veränderungen durch die Verwendung von multiplen Baseline- und Follow-Up-Assessments, wie sie mittels EMA praktisch umgesetzt werden, sowie (2) die Evaluation der psychometrischen Qualität von Methoden zur Bestimmung klinisch bedeutsamer Veränderungen in längsschnittlichen klinischen Studien. Es wurde eine explorative Simulationsstudie durchgeführt, in der drei Methoden der klinischen Signifikanz (Percentage Change, Reliable Change Index und der Individualized Reliable Change Index als vorgeschlagene Anpassung des RCI) sowohl für ein klassisches Fragebogenformat als auch für ein EMA-Format der PHQ-9-Skala für depressive Symptome verglichen wurden. Die Ergebnisse deuten auf einen Genauigkeitsvorteil für wiederholte Messungen gegenüber Single-Point-Assessments hin, bei einem Genauigkeitszuwachs von 17–22% in Fragebogen-Szenarien und allgemein höheren Genauigkeits-Levels in EMA-Szenarien. Darüber hinaus führten individualisierte Maße für klinische Signifikanz im Allgemeinen zu höherer Sensitivität und Spezifität als Methoden, die auf Gruppenebene berechnet wurden (unter der Annahme linearer Veränderung), insbesondere in Szenarien mit wiederholten Messungen. Implikationen, Stärken und Grenzen der Studie werden diskutiert.
*Stichwörter:* digital mental health, EMA, Präzision, reliable change, percentage change

# Chapter 1

# Introduction

The most common paradigm in clinical research is the pre-post design, in which a construct is repeatedly measured over time to evaluate changes in this construct and reach conclusions about the influences that led to the observed changes. The research process then includes statistical tests for the hypothesis, which, in most studies, would regard the efficacy or effectiveness of a therapy or a type of medication. The validity of the resulting inferences strongly depends on the operationalization of the construct, as well as on the methods used to evaluate changes in test scores (Estrada et al., 2018). These approaches for evaluating clinically significant changes over time can vary strongly in their definitions and outcomes, and three of them will be explored in detail within this study (i.e. the Reliable Change Index, the Percentage Change method, and an Individualized Reliable Change Index).

The present thesis aims to address two important problems which affect the validity of certain types of empirical findings in psychiatric research: (1) the limited precision of single-point assessment strategies for measuring psychological symptoms, and (2) the lack of assessment methods for clinically significant changes on an individual level.

Starting with a theoretical introduction into both problems, the methodological analyses in this thesis will explore and test possible solutions for increasing the validity of research designs.

# Chapter 2

# Theoretical Background

Before diving into the empirical methodology of this thesis, some important theoretical concepts ought to be clarified.

## 2.1 Assessment of Treatment Effectiveness on a Group Level vs. Individual Level

Digital mental health and precision medicine practices introduce new ways of data acquisition and analysis into the field of psychiatric research, which have the potential to advance the individualization of psychotherapy. With the rise of digitally assisted interventions in clinical research and practice and the popularization of machine learning algorithms in precision medicine (e.g., treatment outcome prediction models), an emphasis on validity and data quality is necessary to achieve the goals of individualization in person-centered treatments.

In order to evaluate new treatments in psychiatric research, patients are commonly monitored over the course of a therapy. This enables researchers to observe changes in their symptoms over time, aiming to find effects which can be attributed to the applied treatment. Then, the crucial part of the analysis is determining from which perspective the observed changes should be analyzed and interpreted to evaluate potential treatment effects. While most treatment outcome studies report standardized, group-level measures of treatment efficacy, it is often overlooked that individualized measures could give much more valuable insights on within-subjects treatment effects (Lambert, 2013, p. 149).

Regarding the strength of evidence in study designs in general, the "*gold standard*" of empirical clinical research are randomized-controlled trials (RCTs). They are

characterized by the following components: experimental conditions and control conditions, random assignment of participants to conditions, and commonly double blinding (i.e., neither participants nor experimenters know which group each participant is assigned to).

These characteristics provide a necessary basis for researchers to be able to reach conclusions about the effects of an investigated treatment , but there are other methodological problems on a deeper level that need to be considered to ensure sufficiently valid conditions for the type of causal inferences that researchers aim for. Well-studied, but often overlooked problems of commonly applied research designs in RCTs include, i.a., (1) limited precision of the predominantly used single-point assessment strategies for measuring psychological symptoms, and (2) a lack of methods for evaluating clinically significant symptom changes on a truly individualized basis.

1. The first problem is caused by the common practice in clinical studies to assess psychological symptoms, which are often fluctuating in nature, only on single-point occasions, e.g., pre- and post-treatment and on a later follow-up assessment. This practice introduces a certain amount of measurement error and, thus, imprecision (Fisher et al., 2018; Pfeiffer et al., 2015; Schuster et al., 2020).
2. The second problem is concerned with the specific methods used to determine the clinical effectiveness of a treatment. Estrada et al. (2018) introduced a useful terminology for the distinction between two types of approaches to clinically significant change: *average-based change approaches (ABC)* and *individual-based change approaches (IBC). ABCs* examine clinically relevant changes only on a group-level (i.e. between subjects), whereas *IBCs* directly identify individuals who show meaningful changes (i.e. within subjects). Changes on a group-level are typically measured as average differences between the group´s pre-treatment and post-treatment distributions of test scores, implemented in statistical hypothesis tests, which result in effect sizes (e.g., Cohen´s $d$, Hedge´s $g$) and corresponding significance levels. Individual changes, on the other hand, are examined using a variety of different clinical significance methods, which may include standardized pre-post differences and standard errors or variations of linear regressions (Anderson et al., 2014; Estrada et al., 2018; Ferrer & Pardo, 2014). An example for an IBC approach is the popular Reliable Change Index (Jacobson et al., 1984; Jacobson & Truax, 1991), which determines individual treatment responses by evaluating within-subjects pre-post differences, while also relying on group-level variance estimates.

The issues mentioned above could be minimized by applying individualized change criteria to ecological momentary assessment data, as part of the increasingly widespread paradigm of digital mental health studies.

## 2.2 Ecological Momentary Assessment

Ecological Momentary Assessment (EMA), also known as Ambulatory Assessment (e.g., see Fahrenberg et al., 2007), Experience Sampling Method (ESM) (e.g., see Csikszentmihalyi & Larson, 2014; Vork et al., 2019), or Real-Time Data Capture (RTDC) (e.g., see Stone & Broderick, 2007), is the repeated assessment of self-reported symptoms, behavioral or physiological variables via short scales or questionnaires, commonly presented on mobile devices, in order to measure the construct directly in the subject´s natural environment (Ebner-Priemer & Trull, 2009). The method gained popularity as part of the broad concept of Digital Mental Health, which includes methods for the assessment of psychopathology as well as for treatment assistance with digital tools.

EMA is especially suitable for accurately capturing psychological constructs with high intra-individual variability over time (e.g., depression, anxiety, or craving). Despite oftentimes high sampling frequencies, it can be applied efficiently, as it enables highly informative insights from data that is gathered at a minimal cost and effort. Longitudinal EMA designs typically consist of a small number of items that require minimal effort and time for the respondents to answer (Rot et al., 2012; Shiffman et al., 2008). As these short self-reports can be presented in smartphone apps or computer programs, this approach forms a powerful, yet feasible opportunity to study the progression of affective states and behaviors on an individual level.

This diagnostic format could be seen as a bridge between empirical research and clinical practice: Current research in psychiatry has a tendency towards producing group-level based evidence, while the actual treatment of patients is much more concerned with their individual psychopathology. Any practical work with them is in itself a study of processes within each person. EMA formats can benefit both fields by facilitating a deeper understanding of complex psychological processes. For instance, EMA is applied in psychiatric research and therapy, e.g., to capture mood instability in bipolar disorders (e.g., Holmes et al., 2016) or fluctuating symptoms of depression (e.g., Armey et al., 2015; Silk et al., 2011). Through this form of repeated measurement, it is possible to assess relevant information at random or non-random times of the day or week (e.g., directly after panic attacks in patients with panic disorders, or every

morning in depressive patients), while always embedded in the participant´s normal environment and everyday life, instead of in a laboratory, a clinic, or a counseling center. It therefore has the inherent advantage of eliminating lab-specific response tendencies, which is certainly also coupled with the disadvantages of introducing other, environment-specific sources of bias, and possibly the risk of a lower response rate than usually obtained in settings with personally given instructions.

Assessing symptoms over time through EMA may capture more accurately each individual´s treatment responses, but to be able to interpret the course of multiple individual observations, it is important to consider that symptom trajectories show higher variability when captured with EMA than with classical questionnaire assessments. The presence of individual fluctuation between single assessments raises the essential question how to interpret these score changes, and furthermore, what to conclude about the effectiveness of the given treatment for a given individual on the basis of this data. Several questions need to be answered in the process of developing an analytic strategy.

## 2.3 Methods for the Classification of Meaningful Change in Clinical Research

Psychiatric research and practice in any context that involves repeated testing of subjects regarding some measurable construct can benefit from being able to determine if a specific change in test scores over time could be attributed to measurement error alone or if it exceeds this error interval significantly and could therefore be attributed to another influence, e.g., an intervention. The application of this idea and other highly relevant clinical change methods will be explained in detail in Chapter 3.

The broad term clinical significance comprises a variety of different approaches to assess how practically meaningful a change in symptom severity is. Quoting Bauer et al. (2004), “clinically significant change refers to meaningful change in individual patient functioning during psychosocial or medical interventions”. There are many variations of the same broad concept of meaningful differences (e.g., the Clinically Significant Improvement *CSI*, Clinically Significant Difference *CSD*, the Minimal Detectable Change *MDC*, the Minimal Clinically Important Difference *MCID*, and the Minimal Important Difference *MID*) and there are many more approaches to calculating estimates which represent the extent of individual change relative to some error variance, as well as to other subjects. The variations of these individual

criteria and formulas can be roughly divided into two approaches, which in practice are not always mutually exclusive: (1) *distribution-based* methods (i.e. interpreting score changes in relation to the underlying distribution of test scores in a relevant sample) and (2) *anchor-based* methods (i.e. involving external criteria as references for clinically meaningful change, e.g., a cutoff criterion of 2 between-subjects standard deviations for significance) (Estrada et al., 2018; Haley & Fragala-Pinkham, 2006). The clinical significance approaches included in this study´s analyses represent both of these categories separately and in combination.

## 2.4 Estimating the Precision of Clinical Significance Methods

The clinical significance methods investigated in this study are both statistical tests and diagnostic criteria, which makes their performance dependent not only on their ability to detect existing meaningful changes, but also on their ability to detect cases without meaningful changes. To evaluate the performance of classification methods (e.g., bio-medical tests for medical conditions), different commonly used metrics can be calculated from their classifications, if reference classifications from another method are available for comparison. The classes resulting from the investigated method can then, for each class respectively, be categorized into true positive (TP), true negative (TN), false positive (FP), and false negative (FN) classifications (i.e., in a contingency table or confusion matrix). Naturally, in the optimal case, the *true* classes of the observed object are known and can be used as a reference, resulting in the most precise and meaningful estimates for the performance of the classification method. The labels *positive* and *negative* are not necessary, but describe the most common binary classification problems. Regardless of the number of classes, there exist many performance measures which can be computed from the frequencies of TP, TN, FP, and FN cases. For the present study, the following quality measures will be relevant (e.g., see Berthold et al., 2020, p. 107).

*Sensitivity*, also known as *recall* or *true-positive rate*, is the probability of a given method to correctly identify positive cases.

$$Sensitivity = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{2.1}$$

*Specificity*, also known as *selectivity* or *true-negative rate*, is the probability of a given

method to correctly identify negative cases.

$$Specificity = \frac{TN}{N} = \frac{TN}{TN + FP} \tag{2.2}$$

## 2.5 Purpose of the Study

The present thesis forms an in-depth investigation of the theoretically expected increases in sensitivity and specificity of diagnostic approaches in clinical trials through the use of multiple baseline and follow-up assessments, as practically implemented in ecological momentary assessment.

Furthermore, this study is concerned with the comparison of currently used techniques for determining meaningful change in longitudinal clinical trials, which follow either a single-point approach or an intense-assessment approach to measuring psychopathology. An individualized adaptation of the Reliable Change Index, the $\text{RCI}_{\text{ind,pre-SD}}$, which implements within-subjects rather than between-subjects variability, is proposed and tested. Agreement and differences between these estimates for the clinical significance of symptom changes will be examined in order to evaluate the psychometric quality of their classifications.

For the purpose of investigating the above-mentioned research questions in combination, an exploratory simulation study will be conducted, in which the included clinical change methods will be compared for both a classical questionnaire format and an EMA format of a short scale for depressive symptoms.

# Chapter 3

# Method

The present study constitutes an investigation of classification outcomes from different methods in simulated scenarios of a psychiatric trial with depressive patients.

## 3.1 Study Design

The methodology follows a repeated-measures design with simulated patient data. Data were generated on the basis of various empirical data sets, ensuring the validity of findings. Simulation studies are characterized by relying on data that is generated and sampled pseudo-randomly on the basis of known distributions of the respective variables of interest. They can be used to empirically investigate the performance of statistical methods under specific conditions and allow to draw conclusions about them independently of context-specific influences that would otherwise be present in studies with real-world data (Morris et al., 2019).

## 3.2 Data Simulation Procedure

All following analyses are based on mathematically simulated data sets that were generated for a previous study by Schuster et al. (2020). A detailed description of the simulation process can be found in the supplementary material of their article online.[1] Estimated parameters and the data-generation process will be described in the following sections.

Data were simulated for two essential scenarios, questionnaire and EMA, on the basis of empirical trials that were conducted with clinical samples of patients with

[1] https://doi.org/10.1016/j.invent.2020.100313

diagnosed depression. As described in sections 2 and 3 of the appendix of the original article, the dependence structure between subsequent assessments was simulated using copulas (i.e. a Frank copula for control conditions and a Clayton copula for treatment conditions). While Schuster et al. (2020) generated and analyzed both "actual trial data" (scenario 2), which were informed by the EVIDENT trial (Klein et al., 2016), and realistic, synthetic data without a specific basis (scenarios 1 and 3), only the latter category was used in this thesis. The difference between them is that the so-called "actual trial data" scenario has differing pairwise correlations between subsequent assessments ($r_{tt}$ between .46 and .69), while the dependence structure of the other scenarios was set constantly as the average empirically found inter-correlations between assessments (i.e. $r_{tt} = .4$ between subsequent EMA assessments and $r_{tt} = .7$ between subsequent questionnaire assessments). As a special form, one simulation also included varying parameters accounting for potential systematic bias, such as a learning effect in EMA data (e.g., participant reactivity to repeated measurements.

All simulated scenarios were based on data sets implementing the Patient Health Questionnaire-9 (PHQ-9, K. Kroenke et al., 2001), which is commonly used to evaluate changes in the degree of self-reported depressive symptoms. In this short scale, all 9 items are scored on a 4-point Likert scale (0-3), resulting in a total score between 0 and 27, with higher scores indicating more severe depressive mood. In practice, the PHQ-9 is being used in both assessment modalities, i.e. for questionnaire assessments and EMA.

### 3.2.1 Simulated Scenarios

To give an overview, the characteristics of all investigated trial scenarios are summarized in Table 3.1. The simulation process leading to this data is described in Appendix .2.

**Questionnaire Scenarios** include treatment conditions within intense 30-fold assessment intervals ($PP_{30.30}$), intense 5-fold random-window assessment intervals ($PP_{5.5\text{-Window}}$), and single pre-post assessments ($PP_{1.1}$), as well as no-treatment (control) conditions within intense 5-fold assessment intervals ($PP_{5.5}$) and single pre-post assessments ($PP_{1.1}$).

**EMA Scenarios** include treatment conditions within intense 30-fold assessment intervals ($EMA_{30.30}$), intense 5-fold random-window assessment intervals ($EMA_{5.5\text{-Window}}$), and intense 5-fold random-day assessment intervals ($EMA_{5.5\text{-Days}}$), as well as no-

treatment (control) conditions within intense 5-fold assessment intervals ($EMA_{5.5}$) and single pre-post assessments ($EMA_{1.1}$).

**Table 3.1**
*Final Data Structure of the Simulated Scenarios, Including Standard-Questionnaire and EMA Scenarios With Three Different Sampling Frequencies Each Under Treatment Conditions, and Standard-Questionnaire and EMA Scenarios With Two Different Sampling Frequencies Each Under No-Treatment (Control) Conditions*

| Condition | Scenario | $N$ | $d$ | Frequency | $r_{tt}$ | $\alpha$ |
|---|---|---|---|---|---|---|
| Treat | Standard Questionnaire | 8180 | 0.89 | 30-30 | 0.65 | 0.98 |
| | | | | 5-5 Window | 0.66 | **0.87** |
| | | | | 1-1 | 0.10 | 0.17 |
| | EMA | 8040 | 0.88 | 30-30 | 0.31 | 0.93 |
| | | | | 5-5 Window | 0.31 | **0.62** |
| | | | | 5-5 Days | 0.30 | 0.70 |
| No-Treat | Standard Questionnaire | 99810 | 0.00 | 5-5 | 0.66 | **0.83** |
| | | | | 1-1 | 0.36 | 0.53 |
| | EMA | 99964 | 0.00 | 5-5 | 0.33 | **0.51** |
| | | | | 1-1 | 0.17 | 0.29 |

*Note.* Treat = treatment condition, No-Treat = no-treatment (control) condition, $d$ = effect size Cohen´s $d$ for the mean difference between the first pre- and the first post-assessment, $r_{tt}$ = average correlation between subsequent assessments (i.e. test-retest reliability), $\alpha$ = internal consistency Cronbach´s $\alpha$ between assessments, highlighting in bold font indicates which $\alpha$ estimate was implemented in Reliable Change Indices for each scenario respectively

Treatment-condition data sets for both diagnostic methods showed an overall effect size of Cohen´s $d$ between 0.88 (EMA) and 0.89 (standard questionnaire) for the symptom change from pre- to post-treatment assessments. Their overall treatment effect would therefore be considered large (Cohen, 2013), while lying within the range of real empiric effect sizes reported in research on psychiatric outcomes. No-treatment scenarios with $d = 0$, on the other hand, were used in separate analyses to investigate how precise the included clinical significance methods recognized cases without meaningful changes (i.e. specificity), and in turn, how many cases were falsely classified as clinically meaningful changes (i.e. false positives).

The correlation matrices of pre- and post-treatment assessments are given in Appendix .1. The pairwise correlation coefficients between subsequent assessments,

$r_{tt}$, were roughly equal both within and between pre- and post-treatment intervals for all included scenarios (pooled estimates of $r_{tt}$ are given in Table 3.1). In detail, in treatment scenarios, the average correlation between subsequent assessments was $r_{tt}$ = .64-.65 in $PP_{30.30}$, $r_{tt}$ = .65-.66 in $PP_{5.5\text{-Window}}$, $r_{tt}$ = .10 in $PP_{1.1}$, $r_{tt}$ = .31 in $EMA_{30.30}$, $r_{tt}$ = .31 in $EMA_{5.5\text{-Window}}$, and $r_{tt}$ = .29-.30 in $EMA_{5.5\text{-Days}}$; and in no-treatment scenarios: $r_{tt}$ = .66-.67 in $PP_{5.5}$, $r_{tt}$ = .36 in $PP_{1.1}$, $r_{tt}$ = .33 in $EMA_{5.5}$, and $r_{tt}$ = .17 in $EMA_{1.1}$.

As indicated in Table 3.1, the internal consistency Cronbach´s $\alpha$, was implemented to calculate some of the clinical significance methods (see Chapter 3.4.2). It varied strongly between different assessment frequencies, which is expected, as $\alpha$ typically increases with the number of assessments. To control its differential effects on the classification accuracy of methods in these scenarios, $\alpha$ was taken from the 5-fold random-window scenario for use in calculations within each standard-questionnaire and EMA modality, e.g., $\alpha$ = .87 in $PP_{5.5\text{-Window}}$ for all three treatment-condition questionnaire scenarios and $\alpha$ = .62 in $EMA_{5.5\text{-Window}}$ for all three treatment-condition EMA scenarios.

For a comparison to previous studies using the PHQ-9, for instance, Titov et al. (2011) reported a roughly similar internal consistency of $\alpha$ = .74 (before treatment) and $\alpha$ = .81 (after treatment), while other studies reported higher empirical estimates of $\alpha$, e.g., $\alpha$ = .87 in a study by Hepner et al. (2009; see also Adewuya et al., 2006; K. Kroenke et al., 2001; Kurt Kroenke et al., 2010; Lamers et al., 2008).

## 3.3 Clinical Interpretation of PHQ-9 Scores

Within the PHQ-9 (K. Kroenke et al., 2001), depressive mood is evaluated as a sum score of its 9 items. The items of the PHQ-9 correspond to the diagnostic criteria of major depressive disorder in the DSM-IV (*Diagnostic and Statistical Manual of Mental Disorders: DSM-IV*, 1995). The severity of depressive symptoms is operationalized by the items retrospectively asking for the patient´s self-report on the frequency of experienced symptoms over the past 2 weeks. Each item scores on a Likert scale between 0 and 3, where 0 corresponds to *"Not at all"*, 1 to *"Several days"*, 2 to *"More than half the days"*, and 3 to *"Nearly every day"*. The total score of these subjective ratings therefore represents the severity of depressive mood based on the number of experienced symptoms and the frequency with which they were perceived over a period of 2 weeks.

**Table 3.2**
*Clinical Interpretation of Depressive Symptom Levels on the PHQ-9 Scale*

| PHQ-9 Score | Classification | Depression Severity |
|---|---|---|
| 0–4 | 0 | Minimal or none |
| 5–9 | 1 | Mild |
| 10–14 | 2 | Moderate |
| 15–19 | 3 | Moderately severe |
| 20–27 | 4 | Severe |

It is important to note that, although the items are originally formulated retrospectively, they are also commonly adapted to a daily EMA strategy by instructing participants to rate the specified symptoms on their intensity on the present day, instead of on their frequency over the past 2 weeks. Because of the intended daily-assessment structure underlying the simulated data sets, precisely this item variation of the PHQ-9 is assumed for the present study.

The clinical interpretation categories of PHQ-9 Total scores are displayed in Table 3.2 (see Karin, Dear, Heller, Crane, et al., 2018; K. Kroenke et al., 2001). The documentation of the PHQ-9 explicitly discourages diagnosing depressive disorders solely on the basis of the questionnaire. The interpretation categories given in Table 3.2 therefore do not exactly correspond to diagnostic levels of depression (with regards to classification systems, such as ICD-10 or DSM-5), but to scale-specific recommendations for interpretation.

## 3.4 Classification Methods for Clinically Significant Change

The following classification methods will be evaluated regarding their accuracy and agreement with each other:

- the Percentage Change Method (PC),
- the Reliable Change Index (RCI);
  - Reliable Change Index by Jacobson & Truax ($RCI_{JT}$);
  - Repeated-Assessment Individualized Reliable Change Index with individual pre-treatment standard deviation ($RCI_{ind,pre\text{-}SD}$).

Their respective criteria for clinically significant change are listed in Table 3.3.

**Table 3.3**
*Clinically Significant Change Interpretation of PHQ-9 Scores, According to Percentage Change and Reliable Change Criteria*

| Method | PHQ-9 (Post) | Index Value | Classification | Interpretation |
|---|---|---|---|---|
| $CSI_{PC}$ | $\leq 9$ | PC $\geq$ 50 | -1 | Significant Improvement |
| | $> 9$ | -50 $<$ PC $<$ 50 | 0 | No Significant Change |
| | $> 9$ | PC $\leq$ -50 | 1 | Significant Deterioration |
| $CSI_{RCI}$ | $\leq 9$ | RCI $<$ -1.96 | -1 | Significant Improvement |
| | $> 9$ | -1.96 $\leq$ RCI $\leq$ 1.96 | 0 | No Significant Change |
| | $> 9$ | RCI $>$ 1.96 | 1 | Significant Deterioration |

### 3.4.1 Percentage Change

The Percentage Change method ($PC$), also known as Percentage Improvement method ($PI$), describes longitudinal changes in test scores proportionally on an individual level. It results in an index that describes a subject´s post-treatment score as a proportion of his or her pre-treatment score. A positive result indicates that the post-treatment score is smaller than the pre-treatment score (i.e. improvement), while a negative result indicates a post-treatment score higher than the pre-treatment score (i.e. deterioration). The formula to calculate a PC index is given in Equation (3.1). Note that it can be applied equally as simple to single assessments as to average scores calculated from assessment intervals.

$$PC = \left(1 - \frac{\overline{x_2}}{\overline{x_1}}\right) \cdot 100 \tag{3.1}$$

*Note.* $\overline{x_2}$ = mean of subject´s posttest scores, $\overline{x_1}$ = mean of subject´s pretest scores

When applied on a common scoring system for a psychological construct, i.e. including only non-negative scores, the resulting index can assume values smaller than or equal to 100. This is a consequence of the fact that a person can not reduce his or her scores by more than 100%, as the lower bound of the scale itself is non-negative (most typically 0). But depending on the specific scale, it may well be possible that a subject can increase scores from pre- to post-treatment by more than 100%, indicated by a post-treatment score greater than two times the size of the pre-treatment score. Hence,

the negative limit of the index (i.e. the most extreme expression of deterioration) is not defined a priori, but rather scale- and data-specific, as it is determined by the maximum of the empirical distribution of pre-treatment scores in relation to the highest achievable score on the scale.

Percentage Change (Percentage Improvement) rates are commonly reported in psycho-pharmacological studies, mostly involving cutoff criteria of 25 or 50% (Hiller et al., 2012). Furthermore, particularly for clinical trials on depressive disorders, a large body of research generally recommends using a criterion of $\geq 50\%$ improvement to indicate significant treatment response (e.g., see Bandelow et al., 2006; Frank et al., 1991; Hiller et al., 2012; Lecrubier, 2002; McMillan et al., 2010; Nierenberg & DeCecco, 2001; Rush et al., 2006).

Following the Clinical Significance method with a two-fold criterion, consisting of (1) proportional change significance in terms of Percentage Change and (2) clinical significance on a scale-specific cutoff score, subjects in the present study are evaluated according to the criteria given in Table 3.3. The listed criteria were adopted from the original validation study of the PHQ-9, which defined clinically significant improvement as (1) percentage improvement of PC $\geq 50\%$, combined with (2) a post-treatment score $\leq 9$ (McMillan et al., 2010).

### 3.4.2 Reliable Change Index

The Reliable Change Index ($RCI$) was first introduced by Jacobson et al. (1984) and Jacobson & Truax (1991). It is defined as a standardized difference score that determines whether a score change is statistically significant, i.e. substantially exceeds the error variance of the assessment method. Hence, it determines if the observed score difference can be attributed to treatment effects rather than to naturally occurring variance in the sample.

#### Reliable Change (Jacobson et al., 1984; Jacobson & Truax, 1991)

Contemplating the sole reliance on statistical significance of tests, Jacobson & Truax (1991) criticized widespread research approaches for the following problems: (1) comparisons on a group level ignore intra-individual variability and change, and (2) statistically significant group differences are therefore not synonymous with clinical relevance. To address these issues, the authors formulated a two-fold approach to evaluating clinically significant change, consisting of both statistical reliability (i.e. the RC index) and clinical significance in terms of symptom severity scores (Jacobson et

al., 1984).

The RC Index is a standardized measure of the raw score difference between two assessments. It quantifies the extent by which the score difference exceeds the error variance of the assessment method. A significant RCI therefore indicates that the observed change exceeds the measurement error by an extent upon which it can be confidently assumed that it is not caused by error variance, but rather by other factors, such as an applied clinical treatment. The conventionally applied significance cutoff is $|RCI| > 1.96$, derived from the z score for 95% confidence, i.e. a two-sided $\alpha$ probability <.05. An $RCI > 1.96$ indicates statistically reliable deterioration, an $RCI < -1.96$ indicates reliable improvement, and $-1.96 \geq RCI \leq 1.96$ indicates no reliable change.

$$RCI = \frac{x_2 - x_1}{s_{diff}} \tag{3.2}$$

$$s_{diff} = \sqrt{2 \cdot (SE)^2} \tag{3.3}$$

$$SE = s_1 \cdot \sqrt{1 - r_{xx'}} \tag{3.4}$$

*Note.* $x_2$ = subject´s posttest score, $x_1$ = subject´s pretest score, $s_{diff}$ = standard error of difference between test scores, $SE$ = standard error of measurement, $s_1$ = standard deviation of test scores at pretest, $r_{xx'}$ = reliability of the measure

For instance, a significant RC index of ±2 would show that the score difference was equal to two standard deviations, weighted by the reliability of the method. Furthermore, Jacobson et al. (1984) and Jacobson & Truax (1991) offer an additional formula for the calculation of a significance cutoff given in raw scores:

$$\textit{significance cutoff} = 1.96 \cdot s_{diff} = 1.96 \cdot \sqrt{2 \cdot (s_1 \cdot \sqrt{1 - r_{xx'}})^2} \tag{3.5}$$

*Note.* *significance cutoff* = (absolute) cutoff score for reliable change (95%-criterion)

This formula defines the raw score that an individual would have to gain or lose in the given test to be recognized as reliably changed. It is also based on the whole sample´s characteristics. The estimates should be calculated using the standard deviation of either a control group, a normal population, or an experimental group at the baseline assessment (for an adaptation using within-subjects variability, see the proposed $RCI_{ind,pre\text{-}SD}$ in the following section). It also includes the test-retest reliability, which is oftentimes available in the test documentation or in published validation studies.

Following from the assumption of normally distributed change scores, an individual RCI score could also be interpreted in the sense of percentage ranks, i.e.: Assuming normality, it is expected that $X\%$ of participants getting the same treatment under the same conditions would show an improvement or deterioration of at most the same extent.

Regarding the recommended use cases of the $\mathrm{RCI_{JT}}$, there is a general consensus between the original authors and following studies. For instance, Hinton-Bayre (2000) argues that the $\mathrm{RCI_{JT}}$ is appropriate when only pretest data and the test reliability are available and a true change in the construct, independently of treatment effects, is not expected. If normative retest data are available, he argues for the inclusion of the post-treatment variance. The absence of true change in the construct is a critical precondition, because in many assessment contexts there are practice effects, regression toward the mean or divergence from the mean, and natural fluctuations in the construct. If these changes are expected independently of an intervention, they need to be taken into account as error variance (Busch et al., 2015). When true changes are expected additionally to the effects of experimental interventions, e.g., in the form of practice effects or spontaneous remission, especially regression-based calculation approaches can be used to correct the obtained scores (e.g., hierarchical linear models).

### Defining an Individualized Reliable Change Index

The Individualized Reliable Change Index, $RCI_{ind}$, is proposed as an adaptation of the originally defined RCI to repeated-measurement data including more than two timepoints, such as data from EMA procedures. $\mathrm{RCI_{ind}}$ scores are standardized estimates for the reliable change of a single person, as they are based on each individual´s variance, instead of relying on group-level variance estimates.

The proposed individualized formula, the $\mathrm{RCI_{ind,pre\text{-}SD}}$, is adapted to include more than two single assessments in its numerator: The originally included score difference between two single scores is replaced by the average difference between a pre-treatment and a post-treatment assessment interval. Through averaging, the number of assessments in each interval can vary between subjects and missing data could be imputed or, after careful consideration, even ignored. It also includes a subject-level standard error, $SE_{D,pre}$. This standard error is calculated using the reliability of the assessment method and the standard deviation of each individual´s scores throughout the pre-treatment interval. In this way, the individual´s pre-post difference is relativized by their own measurement error, which includes both within-

subject fluctuation of scores and the reliability (i.e. consistency) of the method. Similar to the $\text{RCI}_{\text{JT}}$, an individual cutoff score for significant change can also be calculated easily.

$$RCI_{ind,pre\text{-}SD} = \frac{\overline{x_2} - \overline{x_1}}{SE_{D,pre}} \tag{3.6}$$

$$SE_{D,pre} = \sqrt{2 \cdot (s_x \cdot (1 - r_{xy})^2)} \tag{3.7}$$

$$\textit{significance cutoff} = 1.96 \cdot SE_{D,pre} = 1.96 \cdot \sqrt{2 \cdot (s_x \cdot (1 - r_{xy})^2)} \tag{3.8}$$

*Note.* $\overline{x_2}$ = mean of subject´s posttest scores, $\overline{x_1}$ = mean of subject´s pretest scores, $SE_{D,pre}$ = standard error of difference between the test scores in the individual´s pre interval, $s_x$ = individual standard deviation of pretest time points, $r_{xy}$ = reliability (internal consistency Cronbach´s $\alpha$) of the measure, *significance cutoff* = (absolute) cutoff score for reliable change (95%-criterion)

Significance cutoff scores for the individualized RCI give the absolute scale points that an individual would have to gain or lose on the respective scale to be classified as reliably changed. However, contrary to the $\text{RCI}_{\text{JT}}$, the cutoff score in Equation (3.7) is calculated individually on each subject. Thus, it is not assumed that there exists a universal cutoff score to decide clinically significant change for all participants in a sample. Rather, each subject would need to pass a personally defined range of points in either direction to be considered reliably deteriorated or improved.

The $\text{RCI}_{\text{ind,pre-SD}}$ formula is a simple adaptation of the original RC approach by Jacobson et al. (1984) and Jacobson & Truax (1991), but, resulting from a few changes, the proposed estimate is interpreted differently, specifically because (1) it is calculated over assessment intervals rather than single assessments and (2) it is no longer based on a group-level estimate of variance in a single assessment, but on the individual´s score fluctuation over whole intervals. The inclusion of a subject's individual standard deviation(s), rather than group- or population-level estimates of variability, neither understates the individual error term nor inflates it through the influence of test scores of subjects other than the individual of interest. Nonetheless, the variability of a sample or population's responses to the given test is still considered informative for the individual, and is nevertheless a part of the test's reliability, i.e. included in the measurement error.

For the estimate of reliability in the standard error, similar to the $\text{RCI}_{\text{JT}}$, it is recommended to use the internal consistency Cronbach´s $\alpha$ instead of the test-retest reliability $r_{tt}$. While $r_{tt}$ is used in some common RCI approaches, in contexts where

unstable psychological constructs are measured over time, the internal consistency Cronbach's $\alpha$ is more appropriate for the following reasons: As Cronbach (1947) described in his review of reliability coefficients, the test-retest reliability can be an accurate estimate of measurement accuracy only if the measured construct is expected to be stable over time. By definition, measurement variance in a single test score can only be distinguished from real construct variance over time if the construct does not vary between assessments in reality (Maassen et al., 2009; Wyrwich, 2004). Since constructs examined in clinical research and practice are not expected to be stable, but are examined specifically for changes over time, the test-retest reliability is not considered suitable for calculating a reliable index of change. This issue has been addressed in early studies, but is still often not taken into account in clinical research. For instance, Martinovich et al. (1996) and Tingey et al. (1996) similarly recommend using the internal consistency instead of the retest reliability.

A person´s average difference in test scores between the two assessment intervals, however large or small it may be, is relativized not only by the reliability (i.e. consistency) of the measurement instrument, but also by the person's own variability in responses to the instrument. Consequently, a person with a relatively large mean difference in a measured construct, but also with much variability in individual test scores over time, will be assigned a smaller RCI than another person with the same mean difference and less variability in test scores. This is because there is a reasonably higher confidence in the accuracy of the resulting difference in test scores for the latter person than for the former.

Following the Clinical Significance method with a two-fold criterion, consisting of (1) statistical significance on a Reliable Change Index and (2) clinical significance on a scale-specific cutoff score, subjects in the present study are evaluated according to the criteria given in Table 3.3. It should be noted that the originally introduced 3-class interpretation of RCIs can also be extended to better specify different levels of improvement, as suggested by Lambert & Ogles (2009): Patients could be classified as *recovered* (if the passed both criteria), *improved* (if they passed only the RCI criterion in the positive direction), *unchanged* (if they did not pass the RCI criterion), or *deteriorated* (if they passed the RCI criterion in the negative direction).

## 3.5 Analyses of Sensitivity and Specificity

After defining the statistical terms of interest in Chapter 2.4, their precise implementations regarding the research question and data ought to be clarified shortly.

In the present study, positive cases are equivalent to true cases of meaningful change, and therefore include both *true* improvement and deterioration. Negative cases are equivalent to true cases of no meaningful change. In order to apply the definition of sensitivity (see Equation (2.1)) to the three classes of change interpretation, a class-weighted average sensitivity can be calculated in the following way:

$$Sensitivity_{cwa} = \frac{1}{n}\sum_{k=1}^{c} tp^{(k)} = \frac{tp^{(det)}}{tp^{(det)} + fn^{(det)}} + \frac{tp^{(nc)}}{tp^{(nc)} + fn^{(nc)}} + \frac{tp^{(imp)}}{tp^{(imp)} + fn^{(imp)}} \tag{3.9}$$

Equally, the class-weighted average specificity is calculated in the following way:

$$Specificity_{cwa} = \frac{1}{n}\sum_{k=1}^{c} tn^{(k)} = \frac{tn^{(det)}}{tn^{(det)} + fp^{(det)}} + \frac{tn^{(nc)}}{tn^{(nc)} + fp^{(nc)}} + \frac{tn^{(imp)}}{tn^{(imp)} + fp^{(imp)}} \tag{3.10}$$

*Note.* $c$ = number of classes $k$ (i.e. 3: deteriorated; not changed; improved); $n$ = total number of cases; $tp^k$ = proportion of true positive cases in class k; $fn^k$ = proportion of false negative cases in class k; $tn^k$ = proportion of true negative cases in class k; $fp^k$ = proportion of false positive cases in class k

## 3.6 False-Positive Rate and Specificity in a Control Group

False-positive rates and the specificity of clinical change methods were investigated in questionnaire and EMA scenarios with overall within-subjects effect sizes of Cohen´s $d \approx 0$, representing the scores of a control group in a clinical trial. The characteristics of these data sets are summarized in Table 3.1. Although some participants in these simulated scenarios showed a substantial symptom improvement or deterioration, the overall pre-post symptom changes were closely distributed around 0, with the vast majority of cases showing no meaningful changes in absolute scores. The main advantage of using zero-effect data sets for this analysis is that the absence of a general treatment effect, along with equally distributed random positive and negative effects, enables the a-priori assumption that proportions of cases identified as changed should be minimal in the most specific calculation methods. The respective cases would only constitute false-positive classifications (i.e. both classifications of improvement and of deterioration), as the number of truly changed participants would

be $P = TP + FN = 0$, implying that cases of true change could neither be detected (i.e. $TP = 0$) nor overlooked (i.e. $FN = 0$) in these scenarios. Following from their definitions, classification sensitivity (see Equation (2.1)) could not be calculated under these conditions, while the classification specificity (see Equation (2.2)) could be appropriately estimated with regard to the known *ground truth* of the whole sample consisting of only negative (i.e. non-changed) cases.

The *false-positive rate (FPR)* is given by:

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN} \tag{3.11}$$

Hence, classification methods can be compared regarding their false-positive rates and their specificity (i.e. probability of true-positive classifications) on the basis of this data.

# Chapter 4

# Results

All steps of data preparation and statistical analyses were performed using the statistical programming language R (R Core Team, 2020). A complete list of additionally loaded packages is provided in Appendix .4.1.

## 4.1 Clinical Significance Under Treatment Conditions

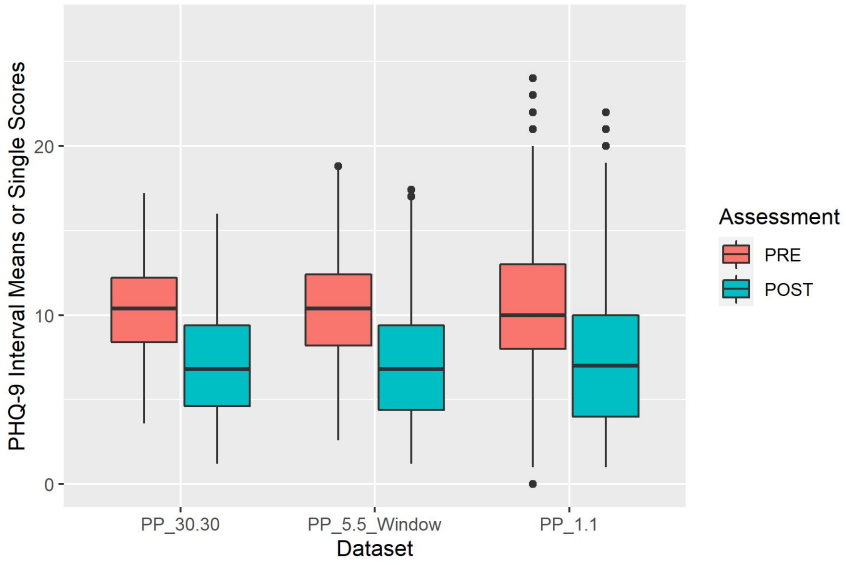
### 4.1.1 Pre–Post Differences in Symptom Scores

The first step of analyses was concerned with the question if the respective variants of questionnaire and EMA scenarios were sufficiently comparable between each other. Their similarity was necessary for the purpose of comparing the course of clinical symptoms of the same participants between different assessment frequencies. The simulation and pre-processing methods described in the previous chapter and in Appendix .2 ensured that scenarios with different questionnaire and EMA frequencies were linked to comprise the same participants. Although all questionnaire data sets had an identical sample ($N = 8180$), and equally all EMA data sets had an identical sample ($N = 8040$), it was furthermore necessary that they had respectively similar overall effect sizes between pre- and post-treatment intervals of symptom scores. The data sets were judged comparable if they showed similar overall pre- and post-treatment levels of depression with similar average standard deviations.

### Questionnaire Scenarios

The within-subjects pre-post treatment effect was equal among all questionnaire scenarios ($PP_{5.5\text{-Window}}$, $PP_{30.30}$, and $PP_{1.1}$), Cohen´s $d = 0.89$.[1] As shown in Figure 4.1, the data set $PP_{5.5\text{-Window}}$ had average pre-treatment interval depression levels of $\overline{x_1} = 10.33$ ($s_{x_1} = 1.90$) and post-treatment levels of $\overline{x_2} = 7.10$ ($s_{x_2} = 2.33$); $PP_{30.30}$ had average pre-treatment interval depression levels of $\overline{x_1} = 10.33$ ($s_{x_1} = 1.87$) and post-treatment levels of $\overline{x_2} = 7.09$ ($s_{x_2} = 2.31$); and $PP_{1.1}$ had average pre-treatment single-assessment depression levels of $\overline{x_1} = 10.37$ ($s_{x_1} = 3.25$) and post-treatment levels of $\overline{x_2} = 7.12$ ($s_{x_2} = 4.00$). For a more detailed overview of within-subjects treatment effects within the three questionnaire scenarios, see Figure 1 in Appendix .3.

**Figure 4.1**
*Box Plots for PHQ-9 Score Distributions of 5-Fold ($PP_{5.5\text{-Window}}$) and 30-Fold ($PP_{30.30}$) Pre- and Post-Interval Mean Scores and Single Pre-Post Scores ($PP_{1.1}$) in a Simulated Standard-Questionnaire Scenario*



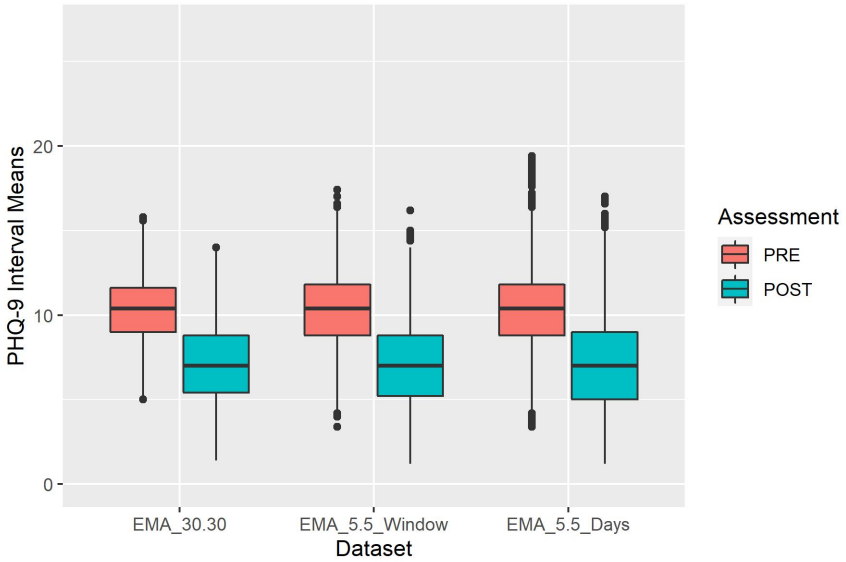

### EMA Scenarios

The within-subjects pre-post treatment effect was equal for all EMA scenarios ($EMA_{30.30}$, $EMA_{5.5\text{-Window}}$, and $EMA_{5.5\text{-Days}}$), Cohen´s $d = 0.88$. As shown in Figure 4.2, the data set $EMA_{30.30}$ had average pre-treatment interval depression levels

[1] The sample-level effect size was calculated between the first pre-treatment and the first post-treatment PHQ-9 assessment.

of $\overline{x_1} = 10.31$ ($s_{x_1} = 2.51$) and post-treatment levels of $\overline{x_2} = 7.12$ ($s_{x_2} = 3.12$); $\text{EMA}_{\text{5.5-Window}}$ had average pre-treatment interval depression levels of $\overline{x_1} = 10.32$ ($s_{x_1} = 2.58$) and post-treatment levels of $\overline{x_2} = 7.10$ ($s_{x_2} = 3.18$); and $\text{EMA}_{\text{5.5-Days}}$ had average pre-treatment interval depression levels of $\overline{x_1} = 10.31$ ($s_{x_1} = 2.42$) and post-treatment levels of $\overline{x_2} = 7.11$ ($s_{x_2} = 2.98$). Compared to the 30-fold questionnaire scenario $\text{PP}_{30.30}$, the standard deviations of pre- and post-treatment interval averages, $s_{x_1}$ and $s_{x_2}$, were higher in the 30-fold EMA scenario $\text{EMA}_{30.30}$. This difference was also present between both 5-fold EMA scenarios ($\text{EMA}_{\text{5.5-Window}}$ and $\text{EMA}_{\text{5.5-Days}}$) and the 5-fold questionnaire scenario ($\text{PP}_{\text{5.5-Window}}$). Hence, there was generally more variation in within-subject interval-wise average depression levels in EMA scenarios than in questionnaire scenarios with corresponding assessment frequencies. For a more detailed overview of within-subjects treatment effects within the three EMA scenarios, see Figure 2 in Appendix .3.

**Figure 4.2**
*Box Plots for PHQ-9 Score Distributions of 30-Fold ($EMA_{30.30}$), 5-Fold Random Window ($EMA_{5.5\text{-}Window}$), and 5-Fold Random Days ($EMA_{5.5\text{-}Days}$) Pre-Treatment and Post-Treatment Interval Mean Scores in a Simulated EMA Scenario*




## 4.1.2 Comparison of Classification Methods

In this section, results of applying the investigated classification methods under treatment conditions are summarized and compared with each other in terms of their sensitivity and specificity. Summary and evaluation tables are given for questionnaire and EMA scenarios.

**Table 4.1**

*Evaluation of Performances Within Classification Methods Between Different Assessment Frequencies of Questionnaire and EMA Scenarios in Reference to their Respective 30-Fold Scenarios*

| | Questionnaire | | | | | EMA | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Method | Freq. | Sens. | Spec. | Acc. | $\kappa$ | Freq. | Sens. | Spec. | Acc. | $\kappa$ |
| PC | 30.30 (BL) | 1 | 1 | 1 | 1 | 30.30 (BL) | 1 | 1 | 1 | 1 |
| | 5.5 Window | 0.89 | 0.95 | 0.92 | 0.83 | 5.5 Window | 0.82 | 0.92 | 0.90 | 0.74 |
| | 1.1 | 0.71 | 0.84 | 0.72 | 0.47 | 5.5 Days | 0.79 | 0.90 | 0.85 | 0.64 |
| $CSI_{PC}$ | 30.30 (BL) | 1 | 1 | 1 | 1 | 30.30 (BL) | 1 | 1 | 1 | 1 |
| | 5.5 Window | 0.90 | 0.95 | 0.94 | 0.83 | 5.5 Window | 0.80 | 0.91 | 0.91 | 0.71 |
| | 1.1 | 0.69 | 0.84 | 0.77 | 0.47 | 5.5 Days | 0.80 | 0.88 | 0.87 | 0.61 |
| RCI | 30.30 (BL) | 1 | 1 | 1 | 1 | 30.30 (BL) | 1 | 1 | 1 | 1 |
| | 5.5 Window | 0.84 | 0.93 | 0.88 | 0.77 | 5.5 Window | 0.79 | 0.90 | 0.87 | 0.71 |
| | 1.1 | $0.59^{JT}$ | $0.81^{JT}$ | $0.66^{JT}$ | $0.39^{JT}$ | 5.5 Days | 0.75 | 0.88 | 0.81 | 0.60 |
| $CSI_{RCI}$ | 30.30 (BL) | 1 | 1 | 1 | 1 | 30.30 (BL) | 1 | 1 | 1 | 1 |
| | 5.5 Window | 0.87 | 0.94 | 0.91 | 0.83 | 5.5 Window | 0.82 | 0.90 | 0.89 | 0.72 |
| | 1.1 | $0.63^{JT}$ | $0.82^{JT}$ | $0.73^{JT}$ | $0.49^{JT}$ | 5.5 Days | 0.75 | 0.88 | 0.84 | 0.61 |

*Note.* $N = 8.180$, (BL) = baseline for performance evaluation, PC and $CSI_{PC}$ refer to mean percentage change in multiple-assessment and to percentage change in single-assessment scenarios; $CSI_{RCI}$ refers to the individualized $CSI_{RCI_{ind,pre\text{-}SD}}$ in multiple-assessment and to the $CSI_{RCI_{JT}}$ (highlighted with $^{JT}$) in single-assessment scenarios; Freq. = assessment frequency; Sens. = sensitivity; Spec. = specificity; Acc. = accuracy (percentage agreement with reference method); $\kappa$ = Cohen´s $\kappa$

Table 4.1 shows the results of within-method/between-frequencies comparisons of classification outcomes for both questionnaire and EMA scenarios. With the respective 30-fold scenario as a reference for each method, classification performances were compared between different 5-fold and single-point assessment scenarios.

Within questionnaire assessments, accuracy levels across all four methods were consistently higher in 5-fold Random Window than in Single-Point scenarios. The increases in accuracy ranged between 17–22% (largest increase for RCI), while increases in sensitivity ranged between 18–25% (largest increase for RCI) and increases in specificity ranged between 11–12%. All four methods reached only moderate levels of accuracy (.66–.77) when applied in a Single-Point scenario, but high levels of accuracy (.88–.94) in a 5-fold Random Window scenario.

Similarly, within EMA assessments, accuracy levels across all four methods were also consistently higher in 5-fold Random Window than in 5-fold Random Days scenarios. The increases in accuracy ranged between 4–6%, while increases in sensitivity ranged

between 0–7% (no increase for $CSI_{PC}$ and largest increase for $CSI_{RCI}$) and increases in specificity ranged between 2–3%. Overall, despite small advantages of $EMA_{5.5\ Window}$ over $EMA_{5.5\ Days}$, all methods resulted in high levels of agreement (accuracy between .81–.91) in comparison to their 30-fold reference scenario.

### Questionnaire Scenarios

From the perspective of Clinical Significance criteria, i.e. based partially on the interpretation of PHQ-9 pre- and post-treatment levels of depression (see Table 3.3), the baseline distribution of severity levels among participants – i.e. symptomatic pre-treatment levels of depression (PHQ-9 $\geq$ 9) in 55–60% of participants in questionnaire scenarios – implies that only these respective proportions of each sample would be able to show clinically significant improvement after the treatment. Simultaneously, the baseline distributions imply that between 40–45% of participants in questionnaire scenarios would theoretically be able to deteriorate significantly over the course of the treatment.

For a summary of classification results in standard- and intense-questionnaire scenarios, see Table 4.2. The summary table displays the absolute and relative frequencies of classification categories (i.e., sig. deteriorated, not sig. changed, and sig. improved) which resulted from applying the investigated clinical significance methods within questionnaire scenarios. To interpret the classification performances, it is important to note that the (Mean) PC and the RCI methods did not include an external, symptom-level criterion, while the $CSI_{PC}$ and the $CSI_{RCI}$ methods did include an external cutoff score.

Consistently across all assessment frequencies and methods, except for the $RCI_{ind,pre\text{-}SD}$ method, the biggest proportion of participants was classified as not changed and the second biggest proportion as improved, whereas the $RCI_{ind,pre\text{-}SD}$ method classified the biggest proportion of the sample as improved and the second biggest proportion as not changed. As the $CSI_{PC}$ method in the 30-fold assessment scenario was defined as the reference for classifications, the other methods needed to yield as similar outcomes as possible to be considered precise. The ground truth for questionnaire scenarios was therefore given as a distribution of 181 (2.2 %) deteriorated cases, 6275 (76.7 %) cases with no significant change, and 1724 (21.1 %) improved cases.

Results from the performance evaluation of classification methods in single- and intense-assessment questionnaire scenarios are summarized in Table 4.3. Note that specificity levels will also be examined separately under no-treatment conditions in

**Table 4.2**
*Classifications Resulting From Clinical Change Methods in Questionnaire Scenarios*

| Frequency | Method | deteriorated (%) | no sig. change (%) | improved (%) |
|---|---|---|---|---|
| 30.30 | $CSI_{PC}$ (BL) | **181 (2.2 %)** | **6275 (76.7 %)** | **1724 (21.1 %)** |
| | Mean PC | 224 (2.7 %) | 5417 (66.2 %) | 2539 (31.0 %) |
| | $CSI_{RCI_{ind,pre\text{-}SD}}$ | 352 (4.3 %) | 4839 (59.2 %) | 2989 (36.5 %) |
| | $RCI_{ind,pre\text{-}SD}$ | 737 (9.0 %) | 2182 (26.7 %) | 5261 (64.3 %) |
| 5.5 Window | $CSI_{PC}$ | 226 (2.8 %) | 6157 (75.3 %) | 1797 (22.0 %) |
| | Mean PC | 302 (3.7 %) | 5226 (63.9 %) | 2652 (32.4 %) |
| | $CSI_{RCI_{ind,pre\text{-}SD}}$ | 378 (4.6 %) | 4845 (59.2 %) | 2957 (36.1 %) |
| | $RCI_{ind,pre\text{-}SD}$ | 828 (10.1 %) | 2204 (26.9 %) | 5148 (62.9 %) |
| 1.1 | $CSI_{PC}$ | 457 (5.6 %) | 5463 (66.8 %) | 2260 (27.6 %) |
| | PC | 729 (8.9 %) | 4327 (52.9 %) | 3124 (38.2 %) |
| | $CSI_{RCI_{JT}}$ | 487 (6.0 %) | 4613 (56.4 %) | 3080 (37.7 %) |
| | $RCI_{JT}$ | 749 (9.2 %) | 3360 (41.1 %) | 4071 (49.8 %) |

*Note.* $N = 8.180$; (BL) = baseline for performance evaluation (highlighted in bold font)

chapter 4.2.

Considering the $CSI_{PC}$ among different sampling frequencies, classifications were highly similar between the 5-fold Random Window and the 30-fold scenario (accuracy = .94), while the classification accuracy was considerably lower in the single-point scenario (.77). Similarly high sensitivity levels were generally achieved by all methods in the 5-fold Random Window and the 30-fold scenario, except for the $RCI_{ind,pre\text{-}SD}$ method, which resulted in considerably lower sensitivity levels.

For both 5-fold Random Window and 30-fold questionnaire scenarios, sensitivity and specificity levels >.90 were achieved by the Mean PC, the $CSI_{RCI_{ind,pre\text{-}SD}}$, and the $CSI_{PC}$ methods, indicating good classification performances with accuracy levels (i.e. percentage agreement) between 81–94 %. Hence, these methods were able to correctly identify both significantly changed and not changed symptom trajectories considerably well.

Within the single-point questionnaire scenario, however, all computed methods resulted in moderate performances at best. Specificity levels were consistently higher than sensitivity levels, with the highest sensitivity of .70 resulting from the PC method. Thus, none of the $CSI_{PC}$, PC, $CSI_{RCI_{JT}}$, and $RCI_{JT}$ methods were able to identify deteriorated, not changed, and improved cases reasonably well within the single-point questionnaire scenario.

**Table 4.3**
*Performance of Classification Methods in Questionnaire Scenarios in Reference to the Clinical Significance Method*

| Frequency | Method | Sensitivity | Specificity | Accuracy | Kappa |
|---|---|---|---|---|---|
| 30.30 | $CSI_{PC}$ (BL) | 1 | 1 | 1 | 1 |
| | Mean PC | 0.95 | 0.95 | 0.90 | 0.75 |
| | $CSI_{RCI_{ind,pre\text{-}SD}}$ | 0.92 | 0.92 | 0.82 | 0.62 |
| | $RCI_{ind,pre\text{-}SD}$ | 0.70 | 0.75 | 0.50 | 0.24 |
| 5.5 Window | $CSI_{PC}$ | 0.90 | 0.95 | 0.94 | 0.83 |
| | Mean PC | 0.87 | 0.91 | 0.84 | 0.63 |
| | $CSI_{RCI_{ind,pre\text{-}SD}}$ | 0.87 | 0.91 | 0.81 | 0.58 |
| | $RCI_{ind,pre\text{-}SD}$ | 0.69 | 0.75 | 0.50 | 0.24 |
| 1.1 | $CSI_{PC}$ | 0.69 | 0.84 | 0.77 | 0.47 |
| | PC | 0.70 | 0.81 | 0.67 | 0.35 |
| | $CSI_{RCI_{JT}}$ | 0.68 | 0.83 | 0.71 | 0.40 |
| | $RCI_{JT}$ | 0.67 | 0.80 | 0.59 | 0.30 |

*Note.* (BL) = baseline for performance evaluation, Accuracy = percentage agreement with reference method; Kappa = Cohen´s $\kappa$

There was a consistent advantage of the $CSI_{PC}$ over the PC method (increase in accuracy of 10%), as well as an advantage of the $CSI_{RCI_{pre\text{-}SD}}$ over the $RCI_{ind,pre\text{-}SD}$ (increase in accuracy of 31–32%), and of the $CSI_{RCI_{JT}}$ over the $RCI_{JT}$ (increase in accuracy of 12%), indicating that the inclusion of a cutoff score for clinical significance strongly increased the accuracy of each of the methods.

### EMA Scenarios

From the perspective of Clinical Significance criteria, i.e. based partially on the interpretation of PHQ-9 pre- and post-treatment levels of depression (see Table 3.3), the baseline distribution of severity levels among participants – i.e. symptomatic pre-treatment levels of depression (PHQ-9 $\geq$ 9) in 58–60% of participants in EMA scenarios – implies that only these respective proportions of the sample would be able to show clinically significant improvement after the treatment. Simultaneously, the baseline distributions imply that between 40–42% of participants in EMA scenarios would theoretically be able to deteriorate significantly over the course of the treatment.

For a summary of classification results in EMA scenarios, see Table 4.4. The summary table displays the absolute and relative frequencies of classification categories

(i.e., sig. deteriorated, not sig. changed, and sig. improved) which resulted from applying the investigated clinical significance methods within EMA scenarios. To interpret the classification performances, it is important to note that the (Mean) PC and the RCI methods did not include an external, symptom-level criterion, while the $CSI_{PC}$ and the $CSI_{RCI}$ methods did include an external cutoff score.

**Table 4.4**
*Classifications Resulting From Clinical Change Methods in EMA Scenarios*

| Frequency | Method | deteriorated (%) | no sig. change (%) | improved (%) |
|---|---|---|---|---|
| 30.30 | $CSI_{PC}$ (BL) | **59 (0.7 %)** | **6523 (81.1 %)** | **1458 (18.1 %)** |
| | Mean PC | 73 (0.9 %) | 6068 (75.5 %) | 1899 (23.6 %) |
| | $CSI_{RCI_{ind,pre\text{-}SD}}$ | 54 (0.7 %) | 5945 (73.9 %) | 2041 (25.4 %) |
| | $RCI_{ind,pre\text{-}SD}$ | 82 (1.0 %) | 5382 (66.9 %) | 2576 (32.0 %) |
| 5.5 Window | $CSI_{PC}$ | 99 (1.2 %) | 6359 (79.1 %) | 1582 (19.7 %) |
| | Mean PC | 123 (1.5 %) | 5858 (72.9 %) | 2059 (25.6 %) |
| | $CSI_{RCI_{ind,pre\text{-}SD}}$ | 87 (1.1 %) | 5899 (73.4 %) | 2054 (25.5 %) |
| | $RCI_{ind,pre\text{-}SD}$ | 127 (1.6 %) | 5283 (65.7 %) | 2630 (32.7 %) |
| 5.5 Days | $CSI_{PC}$ | 183 (2.3 %) | 6169 (76.7 %) | 1688 (21.0 %) |
| | Mean PC | 225 (2.8 %) | 5559 (69.1 %) | 2256 (28.1 %) |
| | $CSI_{RCI_{ind,pre\text{-}SD}}$ | 152 (1.9 %) | 5550 (69.0 %) | 2338 (29.1 %) |
| | $RCI_{ind,pre\text{-}SD}$ | 224 (2.8 %) | 4783 (59.5 %) | 3033 (37.7 %) |

*Note.* $N = 8.040$; (BL) = baseline for performance evaluation (highlighted in bold font)

Consistently across all assessment frequencies and methods, the biggest proportion of participants was classified as not changed and the second biggest proportion as improved. As the $CSI_{PC}$ method in the 30-fold assessment scenario was defined as the reference for classifications, the other methods needed to yield as similar outcomes as possible to be considered precise. The ground truth for EMA scenarios was therefore given as a distribution of 59 (0.7 %) deteriorated cases, 6523 (81.1 %) cases with no significant change, and 1458 (18.1 %) improved cases. Overall, the $CSI_{PC}$ method yielded fairly similar distributions among the three sampling frequencies.

Results from the performance evaluation of classification methods in different EMA scenarios are summarized in Table 4.5. Note that specificity levels will also be examined separately under no-treatment conditions in chapter 4.2.

Considering the $CSI_{PC}$ among different sampling frequencies, classifications were relatively similar between the 30-fold and both 5-fold scenarios. Overall, a comparison of performance metrics within each method between the three sampling frequencies

**Table 4.5**
*Performance of Classification Methods in EMA Scenarios in Reference to the Clinical Significance Method*

| Frequency | Method | Sensitivity | Specificity | Accuracy | Kappa |
|---|---|---|---|---|---|
| 30.30 | $CSI_{PC}$ (BL) | 1 | 1 | 1 | 1 |
| | Mean PC | 0.98 | 0.98 | 0.94 | 0.84 |
| | $CSI_{RCI_{ind,pre\text{-}SD}}$ | 0.70 | 0.90 | 0.86 | 0.61 |
| | $RCI_{ind,pre\text{-}SD}$ | 0.68 | 0.87 | 0.79 | 0.48 |
| 5.5 Window | $CSI_{PC}$ | 0.80 | 0.91 | 0.91 | 0.71 |
| | Mean PC | 0.80 | 0.90 | 0.86 | 0.62 |
| | $CSI_{RCI_{ind,pre\text{-}SD}}$ | 0.67 | 0.85 | 0.82 | 0.50 |
| | $RCI_{ind,pre\text{-}SD}$ | 0.66 | 0.84 | 0.76 | 0.41 |
| 5.5 Days | $CSI_{PC}$ | 0.80 | 0.88 | 0.87 | 0.61 |
| | Mean PC | 0.82 | 0.88 | 0.82 | 0.53 |
| | $CSI_{RCI_{ind,pre\text{-}SD}}$ | 0.70 | 0.84 | 0.78 | 0.44 |
| | $RCI_{ind,pre\text{-}SD}$ | 0.69 | 0.82 | 0.71 | 0.35 |

*Note.* Accuracy = percentage agreement with reference method; Kappa = Cohen´s $\kappa$

shows that for every method, its specificity, accuracy, and Cohen´s $\kappa$ indicated that the agreement of classifications with the 30-fold reference scenario was the highest within the 30-fold scenario, followed by the 5-fold Random Window scenario and the 5-fold Random Days scenario. In contrast, the sensitivity levels of the Mean PC, the $CSI_{RCI_{ind,pre\text{-}SD}}$, and the $RCI_{ind,pre\text{-}SD}$ methods did not consistently follow this rank order, although the differences were relatively small. However, considering all performance metrics, the examined clinical significance methods calculated for random assessment windows were generally more representative of participants´ *true* change classifications given by the $CSI_{PC}$ (calculated on their individual 30-fold sample of assessments). This appears to indicate an accuracy advantage of implementing intervals of subsequent days over random days for assessing depressive symptoms.

Specificity levels were consistently higher than sensitivity levels. The highest sensitivity levels were achieved by the Mean PC method in the 30-fold scenario (sensitivity = .98, specificity = .98, $\kappa$ = .84), followed by almost similarly high levels from the $CSI_{PC}$ method (sensitivity = .80, specificity = .91, $\kappa$ = .71) and the Mean PC method (sensitivity = .80, specificity = .90, $\kappa$ = .62) in the 5-fold Random Window scenario and in the 5-fold Random Days scenario. Within the 5-fold Random Days scenario, considerably good performances were also only achieved by the $CSI_{PC}$

method (sensitivity = .80, specificity = .88, $\kappa = .61$) and the related Mean PC method (sensitivity = .82, specificity = .88, $\kappa = .53$).

In both 5-fold scenarios, the $\text{CSI}_{\text{RCI}_{\text{ind,pre-SD}}}$ and the $\text{RCI}_{\text{ind,pre-SD}}$ methods achieved specificity levels >.80, but much lower sensitivity levels between .66 and .70. Regarding the performance of the proposed individualized RCI formula, the results showed a consistently small advantage of the $\text{CSI}_{\text{RCI}_{\text{ind,pre-SD}}}$ over the $\text{RCI}_{\text{ind,pre-SD}}$, throughout all EMA scenarios, indicating that the inclusion of a cutoff score for clinical significance increased the accuracy of the $\text{RCI}_{\text{ind,pre-SD}}$ by 6–7%. For the PC method, the inclusion of a cutoff score for clinical significance (i.e. $\text{CSI}_{\text{PC}}$) increased its accuracy by 5%.

## 4.2 Clinical Significance Under No-Treatment Conditions

### 4.2.1 False–Positive Rate and Specificity in a Control Group

#### Questionnaire Scenarios

The within-subjects pre-post treatment effect was equal among both no-treatment questionnaire scenarios ($\text{PP}_{5.5}$ and $\text{PP}_{1.1}$), Cohen´s $d = 0.00$.[2] The data set $\text{PP}_{5.5}$ had average pre-treatment interval depression levels of $\overline{x_1} = 10.40$ ($s_{x_1} = 2.66$) and post-treatment levels of $\overline{x_2} = 10.39$ ($s_{x_2} = 2.65$), and $\text{PP}_{1.1}$ had average pre-treatment single-assessment depression levels of $\overline{x_1} = 10.38$ ($s_{x_1} = 3.45$) and post-treatment levels of $\overline{x_2} = 10.39$ ($s_{x_2} = 3.46$).

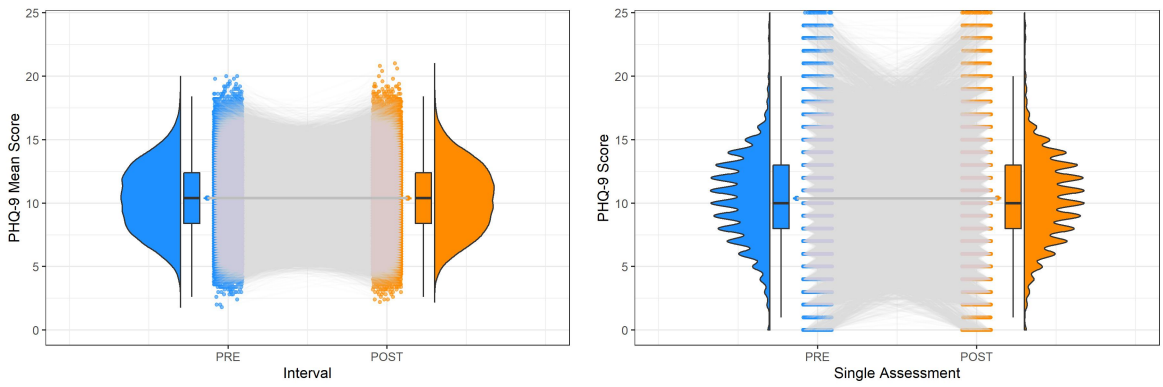
Figure 4.3 gives a more complete overview over within-subjects treatment effects observed in both questionnaire scenarios, with individual score changes depicted by thin gray lines, the overall average pre-post effect given by the bold black line, and pre- and post-treatment score distributions depicted as density and box plots. The left plot in Figure 4.3 displays individual changes for participants in the 5-fold questionnaire scenario between pre- and post-treatment intervals. The right plot in Figure 4.3 displays individual changes for participants in the single-assessment questionnaire scenario between their pre- and post-treatment assessments.

Results of the analyses of specificity across all clinical significance methods in both no-treatment questionnaire scenarios are given in Table 4.6. It should be noted that, in contrast to the simulated treatment scenarios in chapter 4.1, the present simulation

[2]The sample-level effect size was calculated between the first pre-treatment and the first post-treatment PHQ-9 assessment.

**Figure 4.3**

*PHQ-9 Score Distributions of (1) 5-Fold Individual Pre- and Post-Treatment Interval Mean Scores and (2) Single Individual Pre- and Post-Treatment Scores of a No-Treatment Control Group in a Simulated Standard-Questionnaire Scenario*




of no-treatment conditions did not require selecting a reference method to serve as the ground truth for classifications. Instead, specificity levels were calculated from false-positive and true-negative rates, based on the fact that no treatment effects were present in these scenarios and the assumption that therefore all cases classified as *deteriorated* or *improved* were false-positive judgments.

In both scenarios, the highest specificity levels were achieved by the $CSI_{PC}$ method (.86–.88) and the Mean PC method (.79–.86), while the lowest specificity levels were achieved by the $RCI_{ind,pre\text{-}SD}$ method. Surprisingly, the $RCI_{JT}$ and the $CSI_{RCI_{JT}}$ methods resulted in higher specificity levels within the single-point scenario (.67 and .79, respectively) than the $RCI_{ind,pre\text{-}SD}$ and the $CSI_{RCI_{ind,pre\text{-}SD}}$ methods within the 5-fold scenario (.52 and .74, respectively).

Especially considering that the $PP_{1.1}$ scenario only included two single assessments, the specificity of the $CSI_{PC}$ method can be considered high (.86), closely followed by the PC method and the $CSI_{RCI_{JT}}$ method with specificity = .79. In conclusion, particularly the $CSI_{PC}$ and the Mean PC methods were able to correctly identify symptom trajectories without clinically significant changes considerably well.

### EMA Scenarios

The within-subjects pre-post treatment effect was equal among both no-treatment EMA scenarios ($EMA_{5.5}$ and $EMA_{1.1}$), Cohen´s $d = 0.00$.[3] The data set $EMA_{5.5}$ had

[3]The sample-level effect size was calculated between the first pre-treatment and the first post-treatment PHQ-9 assessment.

**Table 4.6**
*Specificity of Classification Methods for a No-Treatment Control Group in Simulated Questionnaire Scenarios*

| Frequency | Method | False Positives | True Negatives | Specificity |
|---|---|---|---|---|
| 5.5 | Mean PC | 14291 | 85519 | 0.86 |
| | $CSI_{PC}$ | 11811 | 87999 | 0.88 |
| | $RCI_{ind,pre\text{-}SD}$ | 48316 | 51494 | 0.52 |
| | $CSI_{RCI_{ind,pre\text{-}SD}}$ | 26020 | 73790 | 0.74 |
| 1.1 | PC | 21029 | 78777 | 0.79 |
| | $CSI_{PC}$ | 13953 | 85857 | 0.86 |
| | $RCI_{JT}$ | 33364 | 66446 | 0.67 |
| | $CSI_{RCI_{JT}}$ | 21376 | 78434 | 0.79 |

*Note.* $N = 99.810$

average pre-treatment interval depression levels of $\overline{x_1} = 10.40$ ($s_{x_1} = 2.01$) and post-treatment levels of $\overline{x_2} = 10.40$ ($s_{x_2} = 2.01$), and $EMA_{1.1}$ had average pre-treatment single-assessment depression levels of $\overline{x_1} = 10.40$ ($s_{x_1} = 3.45$) and post-treatment levels of $\overline{x_2} = 10.39$ ($s_{x_2} = 3.45$).

**Figure 4.4**
*PHQ-9 Score Distributions of (1) 5-Fold Individual Pre- and Post-Treatment Interval Mean Scores and (2) Single Individual Pre- and Post-Treatment Scores of a No-Treatment Control Group in a Simulated EMA Scenario*

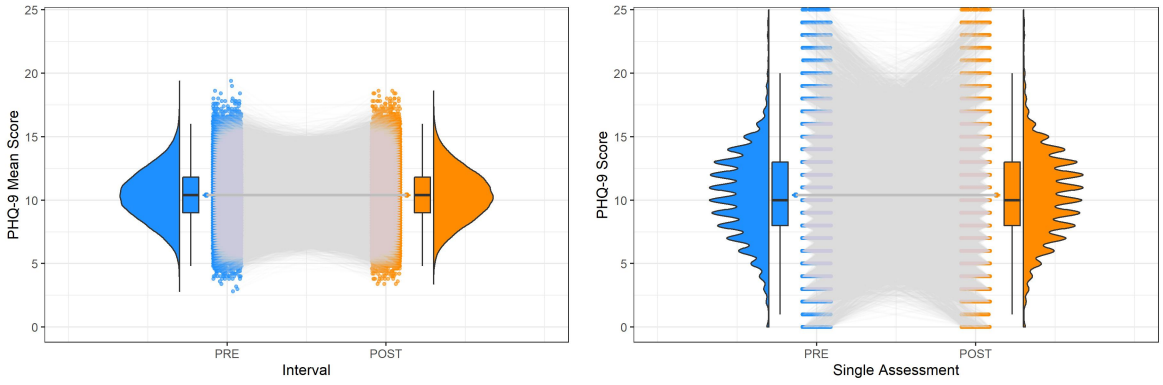



Figure 4.4 gives a more complete overview over within-subjects treatment effects observed in both EMA scenarios, with individual score changes depicted by thin gray lines, the overall average pre-post effect given by the bold black line, and pre- and post-treatment score distributions depicted as density and box plots. The left plot

in Figure 4.4 displays individual changes for participants in the 5-fold EMA scenario between pre- and post-treatment intervals. The right plot in Figure 4.4 displays individual changes for participants in the single-assessment EMA scenario between their pre- and post-treatment assessments.

**Table 4.7**
*Specificity of Classification Methods for a No-Treatment Control Group in Simulated EMA Scenarios*

| Frequency | Method | False Positives | True Negatives | Specificity |
|---|---|---|---|---|
| 5.5 | Mean PC | 7896 | 92068 | 0.92 |
| | $CSI_{PC}$ | 6975 | 92989 | 0.93 |
| | $RCI_{ind,pre\text{-}SD}$ | 10921 | 89043 | 0.89 |
| | $CSI_{RCI_{ind,pre\text{-}SD}}$ | 7816 | 92148 | 0.92 |
| 1.1 | PC | 26644 | 73315 | 0.73 |
| | $CSI_{PC}$ | 6975 | 92989 | 0.93 |
| | $RCI_{JT}$ | 13600 | 86364 | 0.86 |
| | $CSI_{RCI_{JT}}$ | 11236 | 88728 | 0.88 |

*Note.* $N = 99.964$

Results of the analyses of specificity across all clinical significance methods in both no-treatment EMA scenarios are given in Table 4.7. It should be noted that, in contrast to the simulated treatment scenarios in chapter 4.1, the present simulation of no-treatment conditions did not require selecting a reference method to serve as the ground truth for classifications. Instead, specificity levels were calculated from false-positive and true-negative rates, based on the fact that no treatment effects were present in these scenarios and the assumption that therefore all cases classified as *deteriorated* or *improved* were false-positive judgments.

In the $EMA_{5.5}$ scenario, all methods achieved similarly high specificity levels between .89–.93. The $CSI_{PC}$ method achieved the highest specificity level in the $EMA_{1.1}$ scenario (.93), followed by the $CSI_{RCI_{JT}}$ method with specificity = .88, while the PC method resulted in the lowest specificity level of .73.

In conclusion, all examined methods achieved high specificity levels in the 5-fold no-treatment EMA scenario and, in the single-point EMA scenario, particularly the $CSI_{PC}$ method was similarly able to correctly identify symptom trajectories without clinically significant changes with a high precision.

# Chapter 5

# Discussion

The present thesis aimed to investigate possible increases in precision of research designs in clinical trials through the use of ecological momentary assessment instead of single-point questionnaire assessments, as well through the use of psychometrically valid classification methods for determining meaningful symptom changes. An exploratory simulation study was conducted, in which a selection of clinical significance methods was compared for both a classical questionnaire format and an EMA format of the PHQ-9 scale for depressive symptoms. The following methods were evaluated: Percentage Change (PC), the Reliable Change Index $RCI_{JT}$ (Jacobson et al., 1984; Jacobson & Truax, 1991), and the Individualized Reliable Change Index $RCI_{ind,pre\text{-}SD}$ introduced in this thesis, as well as their Clinical Significance variants CSI, which included an additional cutoff criterion for determining significant changes.

In this chapter, the results presented in the previous section will be summarized and interpreted. Furthermore, strengths and limitations of this study will be discussed and a final conclusion will be given.

## 5.1 Discussion of Results

The investigated increases in classification precision resulting from the implementation of multiple daily rather than single-point assessments were consistently found for both questionnaire and EMA formats, but were most pronounced in standard-questionnaire scenarios: By choosing 5-fold assessment intervals rather than two single questionnaire assessments, the accuracy of clinical significance methods was increased by between 17–22%. In particular, the sensitivity was increased by between 18–25% and the specificity by between 11–12%. All of the investigated methods especially benefited from assessment intervals in questionnaire scenarios, where they improved

from moderate to high levels of accuracy in determining significantly improved and deteriorated symptom changes. For EMA scenarios, the classification accuracy was increased by between 4–6%, the sensitivity by between 0–7%, and the specificity by between 2–3%, through applying 5-fold intervals of subsequent days rather than 5-fold random-day assessments. Overall, these considerable increases in precision could justify the additional effort of implementing 5-fold assessment intervals over single-point assessments in clinical research and practice.

In 30-fold, 5-fold Random Window, and single-point questionnaire scenarios under treatment conditions, the methods with the highest sensitivity and specificity were the $CSI_{PC}$, the Mean PC, the $CSI_{RCI_{ind,pre\text{-}SD}}$, and the $RCI_{ind,pre\text{-}SD}$ methods. Comparing only 5-fold and 30-fold scenarios, all methods (i.e. Mean PC, $RCI_{ind,pre\text{-}SD}$, and $CSI_{RCI_{ind,pre\text{-}SD}}$) reached similarly high sensitivity and specificity levels >.90.

In a single-point standard-questionnaire scenario, specificity levels were consistently higher than sensitivity levels, indicating that all examined methods were better able to identify negative cases (i.e., correctly identifying when each of the clinical change categories *is not* present) than positive cases (i.e., correctly identifying when each of the clinical change categories *is* present).

In 30-fold, 5-fold Random Window, and 5-fold Random Days EMA scenarios under treatment conditions, the methods with the highest sensitivity and specificity were the $CSI_{PC}$, the Mean PC, and the $RCI_{ind,pre\text{-}SD}$ methods, respectively. Contrary to expectations, there were no large differences between classifications calculated from 30 vs. 5 pre- and post-treatment assessments, except in the Mean PC method, which dropped in sensitivity and specificity from the 30-fold to both 5-fold scenarios. Across all methods and sampling frequencies, specificity levels were consistently higher than sensitivity levels. In both 5-fold Random Window and Random Days scenarios, considerably good performances were only achieved by the $CSI_{PC}$ and the Mean PC methods.

Classifications in an EMA scenario with randomly selected windows (i.e., comprised of 5 subsequent days pre- and post-treatment) were generally more similar to the "true" classifications from their full 30-fold assessment interval than the classifications calculated over 5-fold randomly selected sets of days. This result indicates an advantage of study designs with regular daily assessments over designs with randomly selected assessment days. This suggestion seems especially reasonable when a kind of treatment is provided, because its expected effects could presumably be captured most reliably when other systematic, time-sensitive influences are minimized. Similarly, when studying the effects of a therapy over time, random and treatment-independent

influences could be detected and statistically controlled more easily in time series with evenly spaced assessments than with random intervals between them.

Regarding the examined clinical significance methods under treatment conditions in general, sensitivity and specificity analyses revealed the strongest performances throughout different sampling frequencies resulting from the Clinical Significance approach with a Percentage Change criterion ($CSI_{PC}$), followed by the Clinical Significance approach with an Individualized Reliable Change Index ($RCI_{ind,pre\text{-}SD}$).

When considering the performance of classification methods, the category of *no significant change* is equally as important as the directed change categories, because in some contexts, in which deterioration in symptoms is expected over time, the *no change* category can serve as evidence of therapy effects (Estrada et al., 2018). In order to specifically investigate the ability of different approaches to identify these cases, they were also implemented in "no-treatment" data sets, which simulated waitlist control groups.

Within no-treatment questionnaire scenarios, the highest specificity levels were achieved by the $CSI_{PC}$ (in $PP_{1.1\ No\text{-}Treat}$ even higher than in $PP_{1.1\ Treat}$), followed by the PC method and the $RCI_{JT}$ method. With the lowest specificity in no-treatment questionnaire scenarios, and therefore the highest number of patients falsely determined as significantly changed, the $RCI_{ind,pre\text{-}SD}$ method generally appeared to be too imprecise for standard-questionnaire studies.

Under no-treatment conditions in EMA scenarios, the overall ability of methods to identify true-negative cases could be considered very good. In the 5-fold scenario, all methods achieved specificity levels $>.90$, while the $CSI_{PC}$ method also did in the single-point scenario. Calculated for only two assessments, the $RCI_{JT}$ method also reached acceptable specificity levels $>.80$. There was an interesting difference between specificity levels in questionnaire and EMA scenarios: Although both were separately investigated in 5-fold and single-point assessment frequencies, every method, except for the PC method, achieved higher specificity levels in EMA scenarios than in standard-questionnaire scenarios. This difference suggests that the PC method performed better (i.e. yielded less false-positive classifications) when calculated with the lower between-assessment intercorrelations present in EMA data.

The (Mean) PC and $RCI_{ind,pre\text{-}SD}$ methods are proportional and therefore individualized estimates, which are characterized by not making implicit assumptions about the nature of individual change. The $RCI_{JT}$ method, on the other hand, has an inherent assumption of linear change in the sense that it judges on the basis of a

fixed, sample-level score difference that has to be achieved by participants in order to be regarded as meaningfully improved or deteriorated. This approach treats all individuals equally in that it does not take into account the individual symptom severity expressed at their baseline assessment. Subjects with a low symptom severity at baseline may not be able to show a score reduction $\geq$ the pre-defined meaningful difference, and hence could not be regarded as meaningfully improved, while subjects with a high symptom severity at baseline could pass the required score improvement and be regarded as meaningfully improved, even though their post-treatment score could still be within the clinical range of scores. In contrast, if a method inherently assumes proportional change, it defines the absolute score difference to be regarded as meaningfully changed proportionally, in order to account for the influence of baseline severity. By setting proportional differences as cutoff criteria for classification categories, observed changes are evaluated individually in relation to baseline severity (Karin, Dear, Heller, Gandy, et al., 2018).

These characteristics may explain why individualized methods generally resulted in higher sensitivity and specificity levels than the linearly calculated (i.e. more baseline-dependent) $RCI_{JT}$, especially in repeated-assessment scenarios. This is in line with previous research showing that proportional estimates were better apt to model treatment effects, with much higher sensitivity and specificity levels and a lower baseline dependency than linear estimates (Karin, Dear, Heller, Gandy, et al., 2018).

Regarding the achieved precision, results of this study suggest a clear superiority of repeated-assessment over single-point assessment approaches, as well as an advantage of random-window over random-day assessment intervals in EMA scenarios. This resulting advantage of EMA over retrospective standard-questionnaire formats is in line with previous research (e.g., Vork et al., 2019). Furthermore, specificity ratings were altogether better in EMA scenarios than in standard-questionnaire scenarios, suggesting that true-negative cases (i.e. individuals with no meaningful symptom changes) are generally better detected in EMA data than in questionnaire data.

## 5.2 Strengths and Limitations

The biggest strength of the present study is the empirical basis of the simulated trial data. The generation of questionnaire and EMA data sets in scenarios with different effect sizes and sampling frequencies, which were prepared to comprise respectively identical samples of participants, allowed for comparing individual changes in depressive symptoms across these assessment designs. Questionnaire and EMA scenarios also

showed sufficiently similar reliabilities, pre- and post-treatment levels of depressive symptoms, and effect sizes between different scenarios to be comparable without introducing significant systematic biases. A particular emphasis was put on within-subjects comparisons in order to maximize the validity and practicability of findings regarding the sensitivity and specificity of classification methods. In combination with the validated and externally anchored reference method for clinically significant change (i.e. Clinical Significance with Percentage Change criteria and external cutoff scores), these preconditions were created to ensure a high external validity of the results reported in this study.

One notable limitation of the study regards other characteristics of the simulated data sets. Although based on empirically gathered data from clinical samples, on average, the simulated baseline-interval scores were arguably low and mainly corresponded to mild and moderate levels of depression. The PHQ-9 scale has a maximum score of 27 points, but the data used for the analyses in this study did only reach a maximum of 25 points. Therefore, for instance, it would have been possible to add a constant score of 2 points to every single assessment, if the intention would have been to correct the data sets to represent more severe levels of depression. This overall correction would not have had any impact on the effect size (Cohen´s $d$), the underlying covariance matrix of assessments, test-retest reliabilities, the internal consistency Cronbach´s $\alpha$, or the proportional clinical change methods (Percentage Change and Individualized Reliable Change Index). It would only have affected the proportions of cases that were identified by the Clinical Significance method as moved from the clinical to the non-clinical population, or vice versa. This is because the standard definition of clinically significant change in PHQ-9 scores includes the 50 % change criterion, as well as passing the cutoff score of 9 points defining the border between the clinical and the non-clinical distribution (see McMillan et al., 2010). However, following from the comparative approach in this study, a constant-value correction would not have altered the measures of interest in this methodological comparison, and neither the conclusions that are drawn from its results. This example is intended to emphasize the generalizability of conclusions regarding the sensitivity and specificity of the analyzed methods in comparison to each other: Assuming that the simulated data sets realistically represent empirically observed clinical trial data, the resulting differences in agreement between methods would be the same, regardless of the symptom-severity levels.

Another notable limitation concerns the formula for the newly introduced $\mathrm{RCI}_{\mathrm{ind,pre\text{-}SD}}$, which implements the within-subjects standard deviation to determine

the standard error of measurement individually, rather than the between-subjects standard deviation. Although this adaptation is intended, it makes comparisons between both approaches difficult, because the within-subjects variability is often smaller than between-subjects variability. In this case, the $RCI_{ind,pre\text{-}SD}$ results in more liberal estimates than the $RCI_{JT}$ and therefore more subjects being classified as significantly changed. This should be noted when applying the $RCI_{ind,pre\text{-}SD}$, especially when comparing it with other methods.

A strength concerning the empirical basis of this study is the use of the PHQ-9 for simulated scenarios, as it can be considered a highly reliable and valid (i.e. its items correspond to the DSM-IV criteria of Major Depressive Disorder) scale, which is extensively studied and applied under both standard-questionnaire and repeated-measurement conditions. However, the use of a special variant of the PHQ-9, which was adapted to daily assessments in EMA scenarios, may confound the originally assessed frequency of depressive symptoms in the past two weeks with the now assessed severity of symptoms on the respective day. This limitation may affect the generalizability of the results in this study.

Another important limitation concerns the reference standard that was used in parts of the analyses to evaluate the included calculation methods. The clinical significance criteria defined by McMillan et al. (2010), namely a percentage change of $\geq 50$ % and passing a given cutoff score, were selected for their empirical basis and adaptation to the simulated questionnaire. Besides this recommended definition of the CS method for the PHQ-9, it would also have been reasonable to implement the Jacobson & Truax (1991) definition, which would consist of the same cutoff score and a Reliable Change Index in the place of percentage change. In this study, the preference for the method including a PC criterion automatically led to higher sensitivity and specificity scores of the PC method, as it was, by definition, very closely related to the reference method of Clinically Significant Change. More importantly, despite the limitations that follow from this specific choice, imperfect reference classifications, unfortunately, are a general problem in almost every context where new assessment methods are evaluated against established ones.

A potential next step for future research may be comparing the included classification methods for varying levels of reliability, as was done, e.g., by Atkins et al. (2005). In a simulation study with a similar data structure, pairwise agreement results between clinical significance methods could be computed for a range of empirically relevant reliabilities. Results from the study by Atkins et al. (2005) showed that for all pairwise comparisons, the agreement between methods increased with higher

reliabilities, with Cohen´s Kappa >.90 for a theoretical reliability of $r = .95$. This extreme result indicates a high dependence of methods on reliability scores: By using very reliable assessment methods, it seems to become less relevant which classification method is applied, as they produce increasingly similar results for high reliabilities. Although reliability estimates as high as ≥.90 are not common in many fields of research and also depend on which measure of reliability is used (e.g., Cronbach´s $\alpha$ or $r_{tt}$), high reliabilities are certainly often reported for outcome measures in psychotherapy research.

Analogously, the agreement between classification methods could be examined for a range of different overall effect sizes. This simulation would then give deeper insights into important questions such as whether methods also converge in their agreement for increasing effect sizes, or if there are methods which show high sensitivity and specificity over a wide range of effect sizes.

## 5.3 Conclusion

The present thesis addressed two important methodological issues which have been prevalent in psychiatric trials for decades and arguably may have contributed to the replication crisis in subdivisions of psychology. The first problem is that the single-point assessment paradigm, which is predominantly used for measuring psychological symptoms over time, has a limited statistical precision to accurately measure the often highly dynamic expression of these symptoms. The second problem of interest is a lack of individual-level methods for evaluating clinically significant symptom changes in repeated-measurement studies. Various useful and widely used methods were introduced in the literature, but most of them were only formulated for either group-level comparisons or individual single-point pre-post assessment scenarios. The purpose of this explorative study was to describe and test ways to improve the psychometric quality of RCTs by conducting ecological momentary assessment studies and implementing individualized criteria for clinically meaningful change.

On the basis of this research, characteristics like the intensity, frequency, and duration of psychological interventions could be closely tailored to each patient individually, according to their symptom severity and fluctuation over time. Digital mental health tools allow for the precise evaluation of treatment effects within subjects and between subjects in a variety of easily applicable ways (e.g., see Bauer et al., 2004). For instance, within-patient effect sizes and cutoff scores for significant changes may be reported in clinical studies and meta-analyses, in order to increase the

transparency and practical benefits of methods. It would also be possible to monitor and analyze symptom changes in EMA apps by incorporating clinical significance methods in their algorithms, which could identify or even predict meaningful changes on the basis of markers for remission or deterioration. Some already widely used routine monitoring tools already offer patient feedback, therapy evaluation, and easy visualization of symptom changes for patients and practitioners. But in addition, individual symptom changes could also be relativized with internal and external factors (e.g., sleep, medication, occupational concerns), to distinguish random noise from treatment effects.

There are many approaches to increase the sensitivity and specificity of assessment strategies. In particular, the results of the present study lead to the recommendation for treatment outcome research to implement repeated-measurement designs with short and reliable symptom scales via EMA and to evaluate treatment effects following the Clinical Significance approach with a validated external cutoff criterion and either (1) a Percentage Change criterion of PC $\geq 50\%$ or (2) the proposed Individualized Reliable Change Index $RCI_{ind,pre\text{-}SD}$ (including the individual pre-treatment standard deviation).

# References


Adewuya, A. O., Ola, B. A., & Afolabi, O. O. (2006). Validity of the patient health questionnaire (PHQ-9) as a screening tool for depression amongst Nigerian university students. *Journal of Affective Disorders*, *96*(1-2), 89–93. https://doi.org/10.1016/j.jad.2006.05.021

Anderson, S. R., Tambling, R. B., Huff, S. C., Heafner, J., Johnson, L. N., & Ketring, S. A. (2014). The Development of a Reliable Change Index and Cutoff for the Revised Dyadic Adjustment Scale. *Journal of Marital and Family Therapy*, *40*(4), 525–534. https://doi.org/10.1111/jmft.12095

Armey, M. F., Schatten, H. T., Haradhvala, N., & Miller, I. W. (2015). Ecological Momentary Assessment (EMA) of Depression-Related Phenomena. *Current Opinion in Psychology*, *4*, 21–25. https://doi.org/10.1016/j.copsyc.2015.01.002

Atkins, D. C., Bedics, J. D., McGlinchey, J. B., & Beauchaine, T. P. (2005). Assessing Clinical Significance: Does It Matter Which Method We Use? *Journal of Consulting and Clinical Psychology*, *73*(5), 982–989. https://doi.org/10.1037/0022-006X.73.5.982

Bandelow, B., Baldwin, D. S., Dolberg, O. T., Andersen, H. F., & Stein, D. J. (2006). What is the threshold for symptomatic response and remission for major depressive disorder, panic disorder, social anxiety disorder, and generalized anxiety disorder? *The Journal of Clinical Psychiatry*, *67*(9), 1428–1434.

Bauer, S., Lambert, M. J., & Nielsen, S. L. (2004). Clinical Significance Methods: A Comparison of Statistical Techniques. *Journal of Personality Assessment*, *82*(1), 60–70. https://doi.org/10.1207/s15327752jpa820111

Berthold, M. R., Borgelt, C., Höppner, F., Klawonn, F., & Silipo, R. (2020). *Guide to Intelligent Data Science*. Springer International Publishing. https://doi.org/10.1007/978-3-030-45574-3

Beygelzimer, A., Kakadet, S., Langford, J., Arya, S., Mount, D., & Li Shengqiao. (2019). *FNN: Fast Nearest Neighbor Search Algorithms and Applications: R package version 1.1.3*. https://CRAN.R-project.org/package=FNN

Busch, R. M., Lineweaver, T. T., Ferguson, L., & Haut, J. S. (2015). Reliable Change Indices and Standardized Regression-Based Change Score Norms for Evaluating Neuropsychological Change in Children With Epilepsy. *Epilepsy & Behavior*, *47*, 45–54. https://doi.org/10.1016/j.yebeh.2015.04.052

Cohen, J. (2013). *Statistical Power Analysis for the Behavioral Sciences*. Academic press.

Cronbach, L. J. (1947). Test "Reliability": Its Meaning and Determination. *Psychometrika*, *12*(1), 1–16.

Csikszentmihalyi, M., & Larson, R. (2014). Validity and Reliability of the Experience-Sampling Method. In M. Csikszentmihalyi (Ed.), *Flow and the Foundations of Positive Psychology: The Collected Works of Mihaly Csikszentmihalyi* (pp. 35–54). Springer Netherlands. https://doi.org/10.1007/978-94-017-9088-83

*Diagnostic and Statistical Manual of Mental Disorders: DSM-IV* (4th ed.). (1995). American Psychiatric Association; American Psychiatric Assoc.

Ebner-Priemer, U. W., & Trull, T. J. (2009). Ecological Momentary Assessment of Mood Disorders and Mood Dysregulation. *Psychological Assessment*, *21*(4), 463–475. https://doi.org/10.1037/a0017075

Estrada, E., Ferrer, E., & Pardo, A. (2018). Statistics for Evaluating Pre-post Change: Relation Between Change in the Distribution Center and Change in the Individual Scores. *Frontiers in Psychology*, *9*, 2696. https://doi.org/10.3389/fpsyg.2018.02696

Fahrenberg, J., Myrtek, M., Pawlik, K., & Perrez, M. (2007). Ambulatory Assessment - Monitoring Behavior in Daily Life Settings. *European Journal of Psychological Assessment*, *23*(4), 206–213. https://doi.org/10.1027/1015-5759.23.4.206

Ferrer, R., & Pardo, A. (2014). Clinically Meaningful Change: False Positives in the Estimation of Individual Change. *Psychological Assessment*, *26*(2), 370–383. https://doi.org/10.1037/a0035419

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 of the United States of America*, *115*(27), E6106–E6115. https://doi.org/10.1073/pnas.1711978115

Frank, E., Prien, R. F., Jarrett, R. B., Keller, M. B., Kupfer, D. J., Lavori, P. W., Rush, A. J., & Weissman, M. M. (1991). Conceptualization and rationale for consensus definitions of terms in major depressive disorder. Remission, recovery, relapse, and recurrence. *Archives of General Psychiatry*, *48*(9), 851–855. https://doi.org/10.1001/archpsyc.1991.01810330075011

Haley, S. M., & Fragala-Pinkham, M. A. (2006). Interpreting Change Scores of Tests and Measures Used in Physical Therapy. *Physical Therapy*, *86*(5), 735–743. https://doi.org/10.1093/ptj/86.5.735

Hepner, K. A., Hunter, S. B., Edelen, M. O., Zhou, A. J., & Watkins, K. (2009). A comparison of two depressive symptomatology measures in residential substance abuse treatment clients. *Journal of Substance Abuse Treatment*, *37*(3), 318–325. https://doi.org/10.1016/j.jsat.2009.03.005

Hiller, W., Schindler, A. C., & Lambert, M. J. (2012). Defining Response and Remission in Psychotherapy Research: A Comparison of the RCI and the Method of Percent Improvement. *Psychotherapy Research*, *22*(1), 1–11. https://doi.org/10.1080/10503307.2011.616237

Hinton-Bayre, A. (2000). Reliable Change formula query. *Journal of the International Neuropsychological Society*, *6*(3), 362–363.

Holmes, E. A., Bonsall, M. B., Hales, S. A., Mitchell, H., Renner, F., Blackwell, S. E., Watson, P., Goodwin, G. M., & 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*, e720. https://doi.org/10.1038/tp.2015.207

Ismay, C., & Solomon, N. (2020). *thesisdown: An updated R Markdown thesis template using the bookdown package: R package version 0.1.0*. https://github.com/ismayc/thesisdown

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy Outcome Research: Methods for Reporting Variability and Evaluating Clinical Significance. *Behavior Therapy*, *15*(4), 336–352.

Jacobson, N. S., & Truax, P. (1991). Clinical Significance: A Statistical Approach to

Defining Meaningful Change in Psychotherapy Research. *Journal of Consulting and Clinical Psychology*, *59*(1), 12–19. https://doi.org/10.1037/0022-006X.59.1.12

Karin, E., Dear, B. F., Heller, G. Z., Crane, M. F., & Titov, N. (2018). Wish You Were Here: Examining Characteristics, Outcomes, and Statistical Solutions for Missing Cases in Web-Based Psychotherapeutic Trials. *JMIR Mental Health*, *5*(2), e22. https://doi.org/10.2196/mental.8363

Karin, E., Dear, B. F., Heller, G. Z., Gandy, M., & Titov, N. (2018). Measurement of Symptom Change Following Web-Based Psychotherapy: Statistical Characteristics and Analytical Methods for Measuring and Interpreting Change. *JMIR Mental Health*, *5*(3), e10200. https://doi.org/10.2196/10200

Klein, J. P., Berger, T., Schröder, J., Späth, C., Meyer, B., Caspar, F., Lutz, W., Arndt, A., Greiner, W., Gräfe, V., Hautzinger, M., Fuhr, K., Rose, M., Nolte, S., Löwe, B., Anderssoni, G., Vettorazzi, E., Moritz, S., & Hohagen, F. (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. https://doi.org/10.1159/000445355

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: Validity of a Brief Depression Severity Measure. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Kroenke, Kurt, Spitzer, R. L., Williams, J. B. W., & Löwe, B. (2010). The Patient Health Questionnaire Somatic, Anxiety, and Depressive Symptom Scales: a systematic review. *General Hospital Psychiatry*, *32*(4), 345–359. https://doi.org/10.1016/j.genhosppsych.2010.03.006

Lambert, M. J. (2013). *Bergin and Garfield's handbook of psychotherapy and behavior change* (6. Aufl.). Wiley. http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=566385

Lambert, M. J., & Ogles, B. M. (2009). Using Clinical Significance in Psychotherapy Outcome Research: The Need for a Common Procedure and Validity Data. *Psychotherapy Research*, *19*(4-5), 493–501. https://doi.org/10.1080/10503300902849483

Lamers, F., Jonkers, C. C. M., Bosma, H., Penninx, B. W. J. H., Knottnerus, J. A., & van Eijk, J. T. M. (2008). Summed Score of the Patient Health Questionnaire-9 was a Reliable and Valid Method for Depression Screening in Chronically ill Elderly Patients. *Journal of Clinical Epidemiology*, *61*(7), 679–687. https://doi.org/10.1016/j.jclinepi.2007.07.018

Lecrubier, Y. (2002). How do you define remission? *Acta Psychiatrica Scandinavica*, *106*, 7–11.

Maassen, G. H., Bossema, E., & Brand, N. (2009). Reliable Change and Practice Effects: Outcomes of Various Indices Compared. *Journal of Clinical and Experimental Neuropsychology*, *31*(3), 339–352. https://doi.org/10.1080/13803390802169059

Martinovich, Z., Saunders, S., & Howard, K. (1996). Some comments on assessing clinical significance. *Psychotherapy Research : Journal of the Society for Psychotherapy Research*, *6*(2), 124–132. https://doi.org/10.1080/10503309612331331648

McMillan, D., Gilbody, S., & Richards, D. (2010). Defining successful treatment outcome in depression using the PHQ-9: a comparison of methods. *Journal of Affective Disorders*, *127*(1-3), 122–129. https://doi.org/10.1016/j.jad.2010.04.030

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

Nierenberg, A. A., & DeCecco, L. M. (2001). Definitions of antidepressant treatment response, remission, nonresponse, partial response, and other relevant outcomes: a focus on treatment-resistant depression. *Journal of Clinical Psychiatry*, *62*, 5–9.

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. https://doi.org/10.1016/j.jad.2014.11.040

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org

Rot, M. aan het, 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. https://doi.org/10.1016/j.cpr.2012.05.007

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

Schuster, R., Schreyer, M. L., Kaiser, T., Berger, T., Klein, J. P., Moritz, S., Laireiter, A.-R., & Trutschnig, W. (2020). Effects of Intense Assessment on Statistical Power in Randomized Controlled Trials: Simulation Study on Depression. *Internet Interventions*, *20*. https://doi.org/10.1016/j.invent.2020.100313

Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology*, *4*, 1–32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

Silk, J. S., Forbes, E. E., Whalen, D. J., Jakubcak, J. L., Thompson, W. K., Ryan, N. D., Axelson, D. A., Birmaher, B., & Dahl, R. E. (2011). Daily Emotional Dynamics in Depressed Youth: A Cell Phone Ecological Momentary Assessment Study. *Journal of Experimental Child Psychology*, *110*(2), 241–257. https://doi.org/10.1016/j.jecp.2010.10.007

Stone, A. A., & Broderick, J. E. (2007). Real-time data collection for pain: appraisal and current status. *Pain Medicine (Malden, Mass.)*, *8 Suppl 3*, S85–93. https://doi.org/10.1111/j.1526-4637.2007.00372.x

Tingey, R., Lambert, M., Burlingame, G., & Hansen, N. (1996). Assessing Clinical Significance: Proposed Extensions to Method. *Psychotherapy Research*, *6*(2), 109–123. https://doi.org/10.1080/10503309612331331638

Titov, N., Dear, B. F., McMillan, D., Anderson, T., Zou, J., & Sunderland, M. (2011). Psychometric comparison of the PHQ-9 and BDI-II for measuring response during treatment of depression. *Cognitive Behaviour Therapy*, *40*(2), 126–136. https://doi.org/10.1080/16506073.2010.550059

Vork, L., Mujagic, Z., Drukker, M., Keszthelyi, D., Conchillo, J. M., Hesselink, M. A. M., van Os, J., Masclee, A. A. M., Leue, C., & Kruimel, J. W. (2019). The

Experience Sampling Method: Evaluation of Treatment Effect of Escitalopram in IBS With Comorbid Panic Disorder. *Neurogastroenterology and Motility*, *31*(1). https://doi.org/10.1111/nmo.13515

Wyrwich, K. W. (2004). Minimal Important Difference Thresholds and the Standard Error of Measurement: Is There a Connection? *Journal of Biopharmaceutical Statistics*, *14*(1), 97–110. https://doi.org/10.1081/BIP-120028508

# Appendix

## .1 Appendix A: Pairwise Correlations Between Assessments

The first appendix includes tables displaying correlation coefficients for pairwise comparisons between single PHQ-9 assessments. Thus, the following correlation matrices give estimates of the retest-reliability $r_{tt}$ for all possible combinations of assessments in the respective dataset. For the sake of readability, they are only included for 5-fold questionnaire and EMA scenarios, but not for 30-fold scenarios. Pairwise correlations between consecutive assessments are displayed in bold font on the diagonals.

### .1.1 Questionnaire Scenarios

**Table 1**
*Matrix of Correlations Between Single Pre- and Post-Assessments in the 5-Fold Random-Window Standard-Questionnaire Scenario*

| | PRE_1 | PRE_2 | PRE_3 | PRE_4 | PRE_5 | POST_1 | POST_2 | POST_3 | POST_4 | POST_5 |
|---|---|---|---|---|---|---|---|---|---|---|
| PRE_1 | 1 | **0.65** | 0.53 | 0.50 | 0.55 | 0.11 | 0.11 | 0.12 | 0.12 | 0.12 |
| PRE_2 | 0.65 | 1 | **0.65** | 0.52 | 0.51 | 0.10 | 0.12 | 0.11 | 0.12 | 0.12 |
| PRE_3 | 0.53 | 0.65 | 1 | **0.64** | 0.54 | 0.10 | 0.10 | 0.12 | 0.12 | 0.11 |
| PRE_4 | 0.50 | 0.52 | 0.64 | 1 | **0.64** | 0.09 | 0.09 | 0.11 | 0.10 | 0.10 |
| PRE_5 | 0.55 | 0.51 | 0.54 | 0.64 | 1 | **0.09** | 0.08 | 0.11 | 0.11 | 0.10 |
| POST_1 | 0.11 | 0.10 | 0.10 | 0.09 | 0.09 | 1 | **0.65** | 0.53 | 0.51 | 0.53 |
| POST_2 | 0.11 | 0.12 | 0.10 | 0.09 | 0.08 | 0.65 | 1 | **0.65** | 0.52 | 0.52 |
| POST_3 | 0.12 | 0.11 | 0.12 | 0.11 | 0.11 | 0.53 | 0.65 | 1 | **0.63** | 0.52 |
| POST_4 | 0.12 | 0.12 | 0.12 | 0.10 | 0.11 | 0.51 | 0.52 | 0.63 | 1 | **0.65** |
| POST_5 | 0.12 | 0.12 | 0.11 | 0.10 | 0.10 | 0.53 | 0.52 | 0.52 | 0.65 | 1 |

### .1.2 EMA Scenarios

**Table 2**
*Matrix of Correlations Between Single Pre- and Post-Assessments in the 5-Fold Random-Window EMA Scenario*

| | PRE_1 | PRE_2 | PRE_3 | PRE_4 | PRE_5 | POST_1 | POST_2 | POST_3 | POST_4 | POST_5 |
|---|---|---|---|---|---|---|---|---|---|---|
| PRE_1 | 1.00 | **0.32** | 0.18 | 0.24 | 0.30 | 0.02 | 0.00 | 0.02 | 0.03 | 0.01 |
| PRE_2 | 0.32 | 1.00 | **0.30** | 0.19 | 0.22 | 0.03 | 0.02 | 0.01 | 0.02 | 0.02 |
| PRE_3 | 0.18 | 0.30 | 1.00 | **0.31** | 0.17 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 |
| PRE_4 | 0.24 | 0.19 | 0.31 | 1.00 | **0.31** | 0.01 | 0.03 | 0.03 | 0.01 | 0.02 |
| PRE_5 | 0.30 | 0.22 | 0.17 | 0.31 | 1.00 | **0.03** | 0.01 | 0.02 | 0.01 | 0.02 |
| POST_1 | 0.02 | 0.03 | 0.02 | 0.01 | 0.03 | 1.00 | **0.30** | 0.19 | 0.21 | 0.26 |
| POST_2 | 0.00 | 0.02 | 0.02 | 0.03 | 0.01 | 0.30 | 1.00 | **0.30** | 0.19 | 0.19 |
| POST_3 | 0.02 | 0.01 | 0.02 | 0.03 | 0.02 | 0.19 | 0.30 | 1.00 | **0.31** | 0.17 |
| POST_4 | 0.03 | 0.02 | 0.02 | 0.01 | 0.01 | 0.21 | 0.19 | 0.31 | 1.00 | **0.32** |
| POST_5 | 0.01 | 0.02 | 0.02 | 0.02 | 0.02 | 0.26 | 0.19 | 0.17 | 0.32 | 1.00 |

**Table 3**
*Matrix of Correlations Between Single Pre- and Post-Assessments in the 5-Fold Random-Days EMA Scenario*

| | PRE_1 | PRE_2 | PRE_3 | PRE_4 | PRE_5 | POST_1 | POST_2 | POST_3 | POST_4 | POST_5 |
|---|---|---|---|---|---|---|---|---|---|---|
| PRE_1 | 1.00 | **0.28** | 0.34 | 0.36 | 0.33 | 0.00 | 0.02 | 0.01 | 0.00 | 0.02 |
| PRE_2 | 0.28 | 1.00 | **0.30** | 0.35 | 0.35 | 0.00 | 0.02 | 0.01 | 0.00 | 0.02 |
| PRE_3 | 0.34 | 0.30 | 1.00 | **0.31** | 0.36 | 0.00 | 0.02 | 0.01 | 0.01 | 0.02 |
| PRE_4 | 0.36 | 0.35 | 0.31 | 1.00 | **0.28** | -0.01 | 0.01 | 0.00 | -0.02 | -0.01 |
| PRE_5 | 0.33 | 0.35 | 0.36 | 0.28 | 1.00 | **0.01** | 0.01 | 0.02 | 0.02 | 0.01 |
| POST_1 | 0.00 | 0.00 | 0.00 | -0.01 | 0.01 | 1.00 | **0.28** | 0.34 | 0.35 | 0.33 |
| POST_2 | 0.02 | 0.02 | 0.02 | 0.01 | 0.01 | 0.28 | 1.00 | **0.30** | 0.34 | 0.32 |
| POST_3 | 0.01 | 0.01 | 0.01 | 0.00 | 0.02 | 0.34 | 0.30 | 1.00 | **0.31** | 0.32 |
| POST_4 | 0.00 | 0.00 | 0.01 | -0.02 | 0.02 | 0.35 | 0.34 | 0.31 | 1.00 | **0.29** |
| POST_5 | 0.02 | 0.02 | 0.02 | -0.01 | 0.01 | 0.33 | 0.32 | 0.32 | 0.29 | 1.00 |

## .2 Appendix B: Data Pre-Processing

The data sets originally generated for the study of Schuster et al. (2020) were further prepared to be apt for the particular questions of the present study. The strategy and process will be thoroughly described step by step in the following subsections, satisfying the required level of transparency needed for future reproduction and replication attempts. In the same regard, the R code utilized for these crucial steps of data preparation is included in Appendix .4).

## .2.1 Extension of Individual Assessments

### K-Nearest-Neighbor Search

In order to investigate the sensitivity and specificity of estimates obtained through single and short-interval assessment formats in comparison to each subject´s respective *true symptom levels* – defined by the score fluctuation in their underlying structure of daily assessments – it was necessary to extend the originally simulated assessment intervals. As both the questionnaire and EMA scenarios were first modeled for 5-fold intervals, they were extended for further analyses to obtain 30-fold pre- and post assessment intervals. This was achieved with the following approach.
In both simulated data sets comprising $N = 100.000$ participants each, subjects with equal interval means and standard deviations were matched using a k-nearest-neighbor (KNN) search algorithm. In particular, this was done using the k-dimensional tree algorithm within the function `get.knn()` from the R package *FNN* (Beygelzimer et al., 2019). This KNN-search method compares all cases to one another on one or more dimensions of interest by computing the Euclidian distances between them. For instance, to compare two participants $p = (p_1, p_2)$ and $q = (q_1, q_2)$ regarding the symptom severity and variability within their baseline assessments, with $p_1$ and $q_1$ denoting the mean scores and $p_2$ and $q_2$ denoting the standard deviations of their respective baseline intervals, the Euclidian distance $d$ between them is given by Equation (1):

$$d(p, q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2} \tag{1}$$

Cases were matched separately by pre- and post-treatment intervals to ensure an appropriate balance between (1) within-interval similarity and (2) individual between-interval changes:

1. within-interval similarity between matched cases: cases need to have an exactly equal mean score and fluctuation within the respective interval
2. individual between-interval changes: intervals are matched and concatenated separately: matching pre-treatment intervals are concatenated case-wise and matching post-treatment intervals are concatenated case-wise in another step.

By specifying the KNN-search function with $k = 5$, it calculates the similarity of all cases to each other and matches each case with its 5 nearest neighbors, resulting in lists of 6 matched case IDs for each specific (observed) combination of interval means and standard deviations. In this way, cases with similar average symptom scores and similar pre- and post standard deviations were matched inside each data

set (questionnaire and EMA). Thereby, only participants with both similar average score changes from pre to post and similar intra-individual variability were matched together.

**Generation of 30-fold Individual Assessment Intervals**

The individual assessment intervals of these similar cases were then concatenated after one another in order to extend the number of simulated assessments from 5-fold to 30-fold intervals for each participant. In detail, for each combination of 6 perfect neighbors regarding the pre-treatment interval, the IDs of these neighbors were used to bind their 5-fold pre-treatment intervals together to obtain a table of cases with 30-fold pre-treatment intervals. Within this data set of matched pre-case IDs, cases were first sorted within each set of 6 matched IDs (i.e., within rows), then sorted by rows (i.e., by the lowest ID in each row), and then filtered to contain only unique combinations of matched case IDs. This was also done for all cases that were matched by their post-treatment intervals.
Finally, pre- and post-KNN lists were joined by the first, and therefore lowest, ID in each case row. Hence, the number of cases was further filtered to comprise only cases which contained both 6-fold pre- and 6-fold post-case IDs, i.e. only cases with 6 pre-nearest neighbors and 6 post-nearest neighbors, which could be linked together by their lowest ID. Using this KNN-search information, the final 30-fold assessment intervals were created by concatenating the assessments of matched cases from the originally simulated questionnaire- and EMA-like data sets. Both the questionnaire-like and EMA-like samples were reduced by the extension process by about 92 %, resulting in sample sizes of $N = 8.240$ (questionnaire) and $N = 8.087$ (EMA). R code for this procedure is provided in the Appendices .4.2 and .4.3.
It should be noted that this strategy to extend assessment intervals, i.e. by stringing together 5-fold intervals from multiple different cases, was only considered appropriate because the originally simulated data presented no signs of autoregressive effects within individual intervals, i.e. neither systematic longitudinal effects between consecutive assessments (i.e., overall improvement or deterioration of symptoms within an interval) nor systematic variability (i.e., regression towards the mean or regression towards the tail). These assumptions can be confirmed, for instance, from the correlation matrices given in Appendix .1.

## .2.2 Random Sampling of Assessments From the Intense-Assessment Intervals

In order to realistically simulate drawing arbitrary 5-fold (EMA-like) samples of assessments from each subject´s 30-fold intervals, the following approach was taken within both questionnaire and EMA data sets (see the R code in Appendix .4.4).
For each subject individually, 5-fold windows of pre-treatment and post-treatment assessments were randomly drawn from their respective 30-fold intervals in order to create the scenarios $PP_{5.5\text{-Window}}$ and $EMA_{5.5\text{-Window}}$. This scenario simulates a study design in which participants are monitored via questionnaires or EMA on 5 consecutive days before and after receiving a treatment. Furthermore, only within EMA data, for each subject individually, 5 single pre-treatment and post-treatment assessments were randomly drawn from their respective 30-fold intervals in order to create the scenario $EMA_{5.5\text{-Days}}$. This scenario simulates a study design in which participants are monitored via EMA on 5 arbitrary and not necessarily consecutive days before and after receiving a treatment. This was included to analyze potential systematic differences between implementing a daily vs. a non-daily EMA routine.

## .2.3 Exclusion of Cases Without Variance

A small number of cases with no symptom variability (i.e. with perfectly constant scores) throughout one or both of their assessment intervals were excluded from all analyses. This criterion for exclusion was formulated because it was deemed improbable for participants to show no fluctuation in PHQ-9 scores over 5-fold or, even more improbable, over 30-fold assessments. Including these cases would also have affected the outcome of the $RCI_{\text{ind, pre SD}}$, which incorporates individual standard deviations as estimates of within-subject fluctuations Calculating this estimate of reliable change (see Equation (3.7)) with an individual interval standard deviation of 0 would result in an infinite value for the $CSI_{RCI_{\text{ind,pre-SD}}}$.
The exclusion of these cases was the last step of pre-processing and resulted in the final structure of data sets, as displayed in Table 3.1, with the following sample sizes: Among **treatment condition** trials, $n = 60$ cases were excluded from questionnaire scenarios and $n = 47$ cases were excluded from EMA scenarios, resulting in final samples comprising $N = 8.180$ participants with questionnaire assessments and $N = 8.040$ participants with EMA assessments. Among **no-treatment condition** trials, $n = 190$ cases were excluded from questionnaire scenarios and $n = 36$ cases were excluded from EMA scenarios, resulting in final samples comprising $N = 99.810$ participants

with questionnaire assessments and $N = 99.964$ participants with EMA assessments. No-treatment scenarios had larger final sample sizes than treatment scenarios, as the pre-processing steps described above (knn-search, interval extension, and random sampling of assessments) were not applied on them. Within those scenarios, analyses of false-positive rates and specificity levels were only conducted in 5-fold and single-assessment frequencies, therefore not requiring the generation of 30-fold assessment intervals.

# .3 Appendix C: Distributions of Individual Symptom Changes

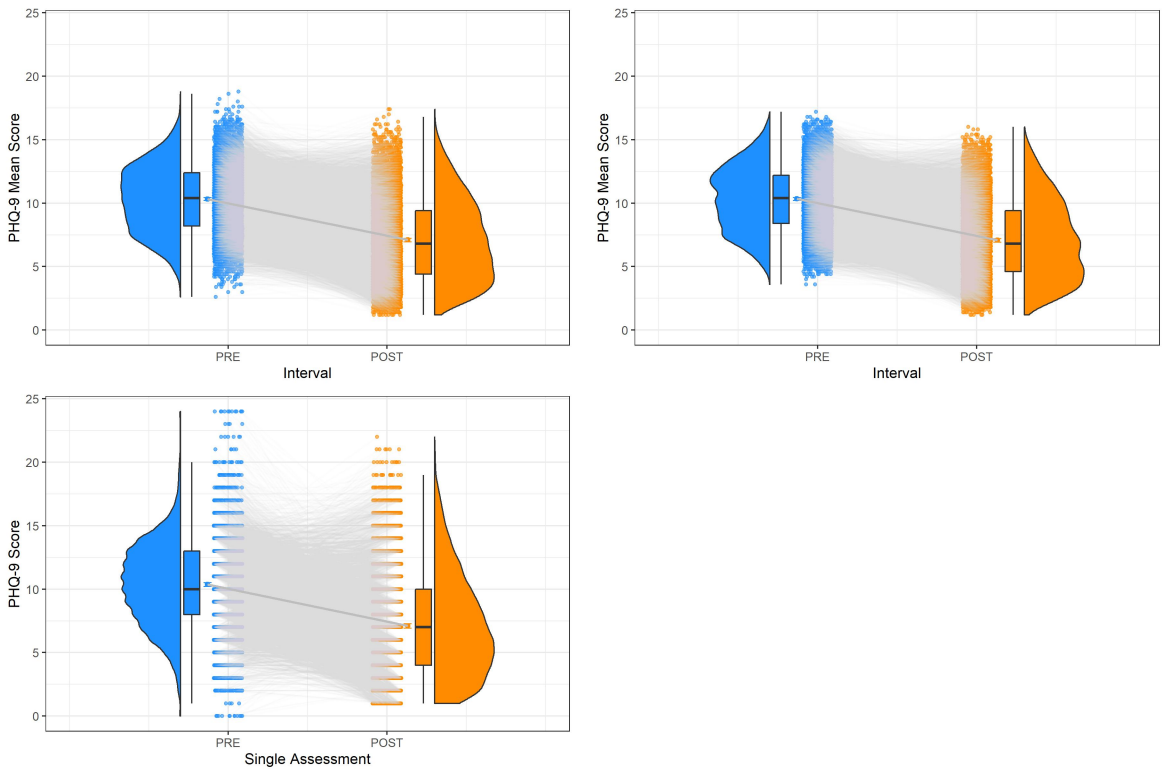
## .3.1 Questionnaire Scenarios

Figure 1 gives an complete overview over within-subjects treatment effects observed in the three questionnaire scenarios, with individual score changes depicted by thin gray lines, the overall average pre-post effect given by the bold black line, and pre- and post-treatment score distributions depicted as density and box plots. The top-left plot in Figure 1 displays individual changes for participants in the 5-fold questionnaire scenario between pre- and post-treatment intervals.[1] The top-right plot in Figure 1 displays individual changes for participants in the 30-fold questionnaire scenario between pre- and post-treatment intervals. The bottom-left plot in Figure 1 displays individual changes for participants in the single-assessment questionnaire scenario between their pre- and post-treatment assessments.

[1]Repeated-measures box- and violin plots were created following an open-visualizations tutorial available on `https://jorvlan.github.io/publications/repmes_tutorial_R.pdf`, which is now also available in the R package *raincloudplots*.

**Figure 1**

*PHQ-9 Score Distributions of (1) 5-Fold and (2) 30-Fold Individual Pre- and Post-Treatment Interval Mean Scores and (3) Single Individual Pre- and Post-Treatment Scores in a Simulated Standard-Questionnaire Scenario*

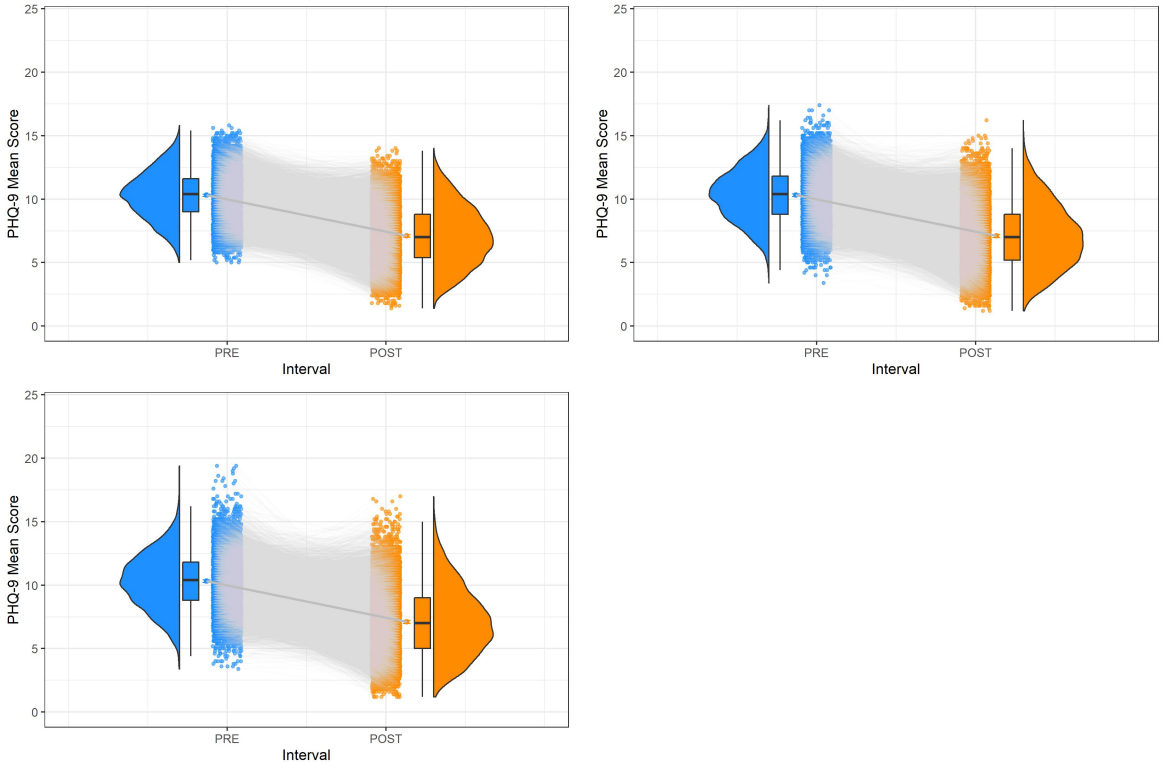



## .3.2 EMA Scenarios

Figure 2 gives an overview over within-subjects treatment effects observed in the three EMA scenarios, with individual score changes depicted by thin gray lines, the overall average pre-post effect given by the bold black line, and pre- and post-treatment score distributions depicted as density and box plots. The top-left plot in Figure 2 displays individual changes for participants in the 30-fold EMA scenario between pre- and post-treatment intervals. The top-right plot in Figure 2 displays individual changes for participants in the 5-fold Random Window EMA scenario between pre- and post-treatment intervals. The bottom-left plot in Figure 2 displays individual changes for participants in the 5-fold Random Days EMA scenario between pre- and post-treatment intervals.

**Figure 2**

*Individual Mean Differences in PHQ-9 Scores Between (1) 30-Fold, (2) 5-Fold Random Window, and (3) 5-Fold Random Days Pre-Treatment and Post-Treatment Intervals in a Simulated EMA Scenario*




# .4 Appendix D: R Code

The third appendix includes information about the R version and packages that were used to prepare and process data, as well as R code for the most important pre-processing steps and the computation of clinical change methods.

## .4.1 R Session Information and Used Packages

```
toLatex(sessionInfo())
```

- R version 4.0.2 (2020-06-22), `x86_64-w64-mingw32`
- Locale: `LC_COLLATE=German_Germany.1252`, `LC_CTYPE=German_Germany.1252`, `LC_MONETARY=German_Germany.1252`, `LC_NUMERIC=C`, `LC_TIME=German_Germany.1252`

- Running under: `Windows 10 x64 (build 19042)`

- Matrix products: default

- Base packages: base, datasets, graphics, grDevices, methods, stats, utils

- Other packages: bootstrap 2019.6, citr 0.3.2, DescTools 0.99.39, devtools 2.3.2, dplyr 1.0.3, FNN 1.1.3, forcats 0.5.1, foreign 0.8-81, gghalves 0.1.1, ggplot2 3.3.5, haven 2.3.1, kableExtra 1.3.1, knitr 1.31, lattice 0.20-41, lubridate 1.7.9.2, overlapping 1.6, papaja 0.1.0.9997, plot.matrix 1.5.2, plyr 1.8.6, psych 2.0.12, purrr 0.3.4, raincloudplots 0.1.0, readr 1.4.0, rmarkdown 2.6, Rmisc 1.5, sjmisc 2.8.6, stringr 1.4.0, summarytools 0.9.8, testthat 3.0.1, thesisdown 0.1.0, tibble 3.0.5, tidyr 1.1.2, tidyverse 1.3.0, timetk 2.6.1, usethis 2.0.0

- Loaded via a namespace (and not attached): assertthat 0.2.1, backports 1.2.0, base64enc 0.1-3, bookdown 0.21, boot 1.3-26, broom 0.7.4, cachem 1.0.3, callr 3.5.1, cellranger 1.1.0, checkmate 2.0.0, class 7.3-18, cli 2.2.0, codetools 0.2-18, colorspace 2.0-0, compiler 4.0.2, crayon 1.3.4, DBI 1.1.1, dbplyr 2.0.0, desc 1.2.0, digest 0.6.27, e1071 1.7-4, ellipsis 0.3.1, evaluate 0.14, Exact 2.1, expm 0.999-6, fansi 0.4.2, fastmap 1.1.0, fs 1.5.0, furrr 0.2.2, future 1.21.0, generics 0.1.0, gld 2.6.2, globals 0.14.0, glue 1.4.2, gower 0.2.2, grid 4.0.2, gtable 0.3.0, hms 1.0.0, htmltools 0.5.1.1, httpuv 1.5.5, httr 1.4.2, insight 0.13.0.1, ipred 0.9-9, jsonlite 1.7.2, later 1.1.0.1, lava 1.6.8.1, lifecycle 0.2.0, listenv 0.8.0, lmom 2.8, magick 2.6.0, magrittr 2.0.1, MASS 7.3-53, Matrix 1.2-18, matrixStats 0.57.0, memoise 2.0.0, mime 0.9, miniUI 0.1.1.1, mnormt 2.0.2, modelr 0.1.8, munsell 0.5.0, mvtnorm 1.1-1, nlme 3.1-151, nnet 7.3-15, pacman 0.5.1, pander 0.6.3, parallel 4.0.2, parallelly 1.23.0, pillar 1.4.7, pkgbuild 1.2.0, pkgconfig 2.0.3, pkgload 1.1.0, prettyunits 1.1.1, processx 3.4.5, prodlim 2019.11.13, promises 1.1.1, pryr 0.1.4, ps 1.5.0, R6 2.5.0, rapportools 1.0, Rcpp 1.0.6, readxl 1.3.1, recipes 0.1.15, remotes 2.2.0, reprex 1.0.0, rlang 0.4.10, rootSolve 1.8.2.1, rpart 4.1-15, rprojroot 2.0.2, rsample 0.0.8, rstudioapi 0.13, rvest 0.3.6, scales 1.1.1, sessioninfo 1.1.1, shiny 1.6.0, sjlabelled 1.1.7, splines 4.0.2, stringi 1.5.3, survival 3.2-7, tcltk 4.0.2, tidyselect 1.1.0, timeDate 3043.102, tmvnsim 1.0-2, tools 4.0.2, vctrs 0.3.6, viridisLite 0.3.0, webshot 0.5.2, withr 2.4.1, xfun 0.22, xml2 1.3.2, xtable 1.8-4, xts 0.12.1, yaml 2.2.1, zoo 1.8-8

### .4.2 K-Nearest-Neighbor Search

K-Nearest-Neighbor Search (using `get.knn()` from the package *FNN*) for the questionnaire data set $PP_{5.5}$ as an example (similar procedure for both the EMA and the questionnaire data set).

```
pacman::p_load(dplyr, FNN)

# opening the originally simulated data set (N = 100.000) and
# calculating interval means and standard deviations
PP_5.5 = read.delim("cor_07_k20/cor_07_dataset_k20.txt",
                    row.names=NULL) %>%
  select(PRE1_1:POST1_5) %>%
  add_column(., .before = "PRE1_1", ID = 1:nrow(.)) %>%
  as_tibble()

pre_5mzp = c("PRE1_1","PRE1_2","PRE1_3","PRE1_4","PRE1_5")
post_5mzp = c("POST1_1","POST1_2","POST1_3","POST1_4","POST1_5")

PP_5.5$PRE_Mean = apply(PP_5.5[pre_5mzp], 1, mean)
PP_5.5$POST_Mean = apply(PP_5.5[post_5mzp], 1, mean)
PP_5.5$MeanDiff = PP_5.5$PRE_Mean - PP_5.5$POST_Mean
PP_5.5$ind.pretestSD = apply(PP_5.5[pre_5mzp], 1, sd)
PP_5.5$ind.posttestSD = apply(PP_5.5[post_5mzp], 1, sd)
save(PP_5.5, file = "cor_07_k20/PP_5.5.RData")

# PRE interval: finding the k=5 nearest neighbors regarding their
# mean score and standard deviation (with distance == 0)
pre_data = PP_5.5 %>% select(PRE_Mean, ind.pretestSD)
PP_PRE_KNN_df = FNN::get.knn(pre_data, k=5, algorithm = "kd_tree")

x = as_tibble(PP_PRE_KNN_df[[1]], .name_repair = "minimal")
colnames(x) = c("neighbor1", "neighbor2", "neighbor3",
                "neighbor4", "neighbor5")
y = as_tibble(PP_PRE_KNN_df[[2]], .name_repair = "minimal")
colnames(y) = c("distance1", "distance2", "distance3",
                "distance4", "distance5")
```

```
PP_PRE_KNN_df = bind_cols(x, y) %>%
  add_column(., .before = "neighbor1", ID = 1:nrow(.)) %>%
  filter(distance1 == 0 & distance2 == 0 & distance3 == 0 &
           distance4 == 0 & distance5 == 0)

# POST interval: finding the k=5 nearest neighbors regarding their
# mean score and standard deviation (with distance == 0)
post_data = PP_5.5 %>% select(POST_Mean, ind.posttestSD)
PP_POST_KNN_df = FNN::get.knn(post_data, k=5, algorithm = "kd_tree")

x = as_tibble(PP_POST_KNN_df[[1]], .name_repair = "minimal")
colnames(x) = c("neighbor1", "neighbor2", "neighbor3",
                "neighbor4", "neighbor5")
y = as_tibble(PP_POST_KNN_df[[2]], .name_repair = "minimal")
colnames(y) = c("distance1", "distance2", "distance3",
                "distance4", "distance5")

PP_POST_KNN_df = bind_cols(x, y) %>%
  add_column(., .before = "neighbor1", ID = 1:nrow(.)) %>%
  filter(distance1 == 0 & distance2 == 0 & distance3 == 0 &
           distance4 == 0 & distance5 == 0)

# filtering the resulting knn combinations to keep only unique
# rows of 6 perfectly matching neighbors
# PRE interval
PP_PRE_KNN_df = PP_PRE_KNN_df %>%
  select(ID, neighbor1, neighbor2, neighbor3, neighbor4, neighbor5) %>%
  apply(., 1, sort) %>%
  t() %>%
  as_tibble() %>%
  arrange(., V1, V2, V3, V4, V5, V6) %>%
  distinct() %>%
  filter(V1 != V2 & V2 != V3 & V3 != V4 & V4 != V5 & V5 != V6) %>%
  group_by(V1) %>%
  filter(row_number() == 1) %>%
```

```
  ungroup()

colnames(PP_PRE_KNN_df) = c("ID1_PRE", "ID2_PRE", "ID3_PRE",
                            "ID4_PRE", "ID5_PRE", "ID6_PRE")

# POST interval
PP_POST_KNN_df = PP_POST_KNN_df %>%
  select(ID, neighbor1, neighbor2, neighbor3, neighbor4, neighbor5) %>%
  apply(., 1, sort) %>%
  t() %>%
  as_tibble() %>%
  arrange(., V1, V2, V3, V4, V5, V6) %>%
  distinct() %>%
  filter(V1 != V2 & V2 != V3 & V3 != V4 & V4 != V5 & V5 != V6) %>%
  group_by(V1) %>%
  filter(row_number() == 1) %>%
  ungroup()

colnames(PP_POST_KNN_df) = c("ID1_POST", "ID2_POST", "ID3_POST",
                             "ID4_POST", "ID5_POST", "ID6_POST")

# joining the matched IDs of pre- and post-neighbors in a data frame
PP_KNNs = inner_join(PP_PRE_KNN_df, PP_POST_KNN_df,
              by = c("ID1_PRE" = "ID1_POST"))
PP_KNNs = PP_KNNs %>%
  add_column(., .before = "ID2_POST", ID1_POST = PP_KNNs$ID1_PRE)
save(PP_KNNs, file = "cor_07_k20/PP_KNNs.RData")
```

### .4.3 Extension of Assessment Intervals

Extension of assessment intervals for the questionnaire data set $PP_{5.5}$ as an example (similar procedure for both the EMA and the questionnaire data set).

```
pacman::p_load(dplyr)
load("cor_07_k20/PP_KNNs.RData")
load("cor_07_k20/PP_5.5.RData")
```

```
PP_KNNs = PP_KNNs %>% as.data.frame()

PP_30.30 = data.frame(
  ID1_PRE = c(), ID2_PRE = c(), ID3_PRE = c(),
  ID4_PRE = c(), ID5_PRE = c(), ID6_PRE = c(),
  ID1_POST = c(), ID2_POST = c(), ID3_POST = c(),
  ID4_POST = c(), ID5_POST = c(), ID6_POST = c(),

  PRE1_1 = c(), PRE1_2 = c(), PRE1_3 = c(), PRE1_4 = c(),
  PRE1_5 = c(), PRE1_6 = c(), PRE1_7 = c(), PRE1_8 = c(),
  PRE1_9 = c(), PRE1_10 = c(), PRE1_11 = c(), PRE1_12 = c(),
  PRE1_13 = c(), PRE1_14 = c(), PRE1_15 = c(), PRE1_16 = c(),
  PRE1_17 = c(), PRE1_18 = c(), PRE1_19 = c(), PRE1_20 = c(),
  PRE1_21 = c(), PRE1_22 = c(), PRE1_23 = c(), PRE1_24 = c(),
  PRE1_25 = c(), PRE1_26 = c(), PRE1_27 = c(), PRE1_28 = c(),
  PRE1_29 = c(), PRE1_30 = c(),

  POST1_1 = c(), POST1_2 = c(), POST1_3 = c(), POST1_4 = c(),
  POST1_5 = c(), POST1_6 = c(), POST1_7 = c(), POST1_8 = c(),
  POST1_9 = c(), POST1_10 = c(), POST1_11 = c(), POST1_12 = c(),
  POST1_13 = c(), POST1_14 = c(), POST1_15 = c(), POST1_16 = c(),
  POST1_17 = c(), POST1_18 = c(), POST1_19 = c(), POST1_20 = c(),
  POST1_21 = c(), POST1_22 = c(), POST1_23 = c(), POST1_24 = c(),
  POST1_25 = c(), POST1_26 = c(), POST1_27 = c(), POST1_28 = c(),
  POST1_29 = c(), POST1_30 = c())

for (i in 1:length(PP_KNNs$ID1_PRE)) {
  PP_30.30[i,"ID1_PRE"] = PP_KNNs[i,"ID1_PRE"]
  PP_30.30[i,"ID2_PRE"] = PP_KNNs[i,"ID2_PRE"]
  PP_30.30[i,"ID3_PRE"] = PP_KNNs[i,"ID3_PRE"]
  PP_30.30[i,"ID4_PRE"] = PP_KNNs[i,"ID4_PRE"]
  PP_30.30[i,"ID5_PRE"] = PP_KNNs[i,"ID5_PRE"]
  PP_30.30[i,"ID6_PRE"] = PP_KNNs[i,"ID6_PRE"]
  PP_30.30[i,"ID1_POST"] = PP_KNNs[i,"ID1_POST"]
  PP_30.30[i,"ID2_POST"] = PP_KNNs[i,"ID2_POST"]
  PP_30.30[i,"ID3_POST"] = PP_KNNs[i,"ID3_POST"]
```

```
PP_30.30[i,"ID4_POST"] = PP_KNNs[i,"ID4_POST"]
PP_30.30[i,"ID5_POST"] = PP_KNNs[i,"ID5_POST"]
PP_30.30[i,"ID6_POST"] = PP_KNNs[i,"ID6_POST"]

PP_30.30[i,"PRE1_1"] = PP_5.5[PP_KNNs[i,"ID1_PRE"],"PRE1_1"]
PP_30.30[i,"PRE1_2"] = PP_5.5[PP_KNNs[i,"ID1_PRE"],"PRE1_2"]
PP_30.30[i,"PRE1_3"] = PP_5.5[PP_KNNs[i,"ID1_PRE"],"PRE1_3"]
PP_30.30[i,"PRE1_4"] = PP_5.5[PP_KNNs[i,"ID1_PRE"],"PRE1_4"]
PP_30.30[i,"PRE1_5"] = PP_5.5[PP_KNNs[i,"ID1_PRE"],"PRE1_5"]
PP_30.30[i,"PRE1_6"] = PP_5.5[PP_KNNs[i,"ID2_PRE"],"PRE1_1"]
PP_30.30[i,"PRE1_7"] = PP_5.5[PP_KNNs[i,"ID2_PRE"],"PRE1_2"]
PP_30.30[i,"PRE1_8"] = PP_5.5[PP_KNNs[i,"ID2_PRE"],"PRE1_3"]
PP_30.30[i,"PRE1_9"] = PP_5.5[PP_KNNs[i,"ID2_PRE"],"PRE1_4"]
PP_30.30[i,"PRE1_10"] = PP_5.5[PP_KNNs[i,"ID2_PRE"],"PRE1_5"]

PP_30.30[i,"PRE1_11"] = PP_5.5[PP_KNNs[i,"ID3_PRE"],"PRE1_1"]
PP_30.30[i,"PRE1_12"] = PP_5.5[PP_KNNs[i,"ID3_PRE"],"PRE1_2"]
PP_30.30[i,"PRE1_13"] = PP_5.5[PP_KNNs[i,"ID3_PRE"],"PRE1_3"]
PP_30.30[i,"PRE1_14"] = PP_5.5[PP_KNNs[i,"ID3_PRE"],"PRE1_4"]
PP_30.30[i,"PRE1_15"] = PP_5.5[PP_KNNs[i,"ID3_PRE"],"PRE1_5"]
PP_30.30[i,"PRE1_16"] = PP_5.5[PP_KNNs[i,"ID4_PRE"],"PRE1_1"]
PP_30.30[i,"PRE1_17"] = PP_5.5[PP_KNNs[i,"ID4_PRE"],"PRE1_2"]
PP_30.30[i,"PRE1_18"] = PP_5.5[PP_KNNs[i,"ID4_PRE"],"PRE1_3"]
PP_30.30[i,"PRE1_19"] = PP_5.5[PP_KNNs[i,"ID4_PRE"],"PRE1_4"]
PP_30.30[i,"PRE1_20"] = PP_5.5[PP_KNNs[i,"ID4_PRE"],"PRE1_5"]

PP_30.30[i,"PRE1_21"] = PP_5.5[PP_KNNs[i,"ID5_PRE"],"PRE1_1"]
PP_30.30[i,"PRE1_22"] = PP_5.5[PP_KNNs[i,"ID5_PRE"],"PRE1_2"]
PP_30.30[i,"PRE1_23"] = PP_5.5[PP_KNNs[i,"ID5_PRE"],"PRE1_3"]
PP_30.30[i,"PRE1_24"] = PP_5.5[PP_KNNs[i,"ID5_PRE"],"PRE1_4"]
PP_30.30[i,"PRE1_25"] = PP_5.5[PP_KNNs[i,"ID5_PRE"],"PRE1_5"]
PP_30.30[i,"PRE1_26"] = PP_5.5[PP_KNNs[i,"ID6_PRE"],"PRE1_1"]
PP_30.30[i,"PRE1_27"] = PP_5.5[PP_KNNs[i,"ID6_PRE"],"PRE1_2"]
PP_30.30[i,"PRE1_28"] = PP_5.5[PP_KNNs[i,"ID6_PRE"],"PRE1_3"]
PP_30.30[i,"PRE1_29"] = PP_5.5[PP_KNNs[i,"ID6_PRE"],"PRE1_4"]
PP_30.30[i,"PRE1_30"] = PP_5.5[PP_KNNs[i,"ID6_PRE"],"PRE1_5"]
```

```
PP_30.30[i,"POST1_1"] = PP_5.5[PP_KNNs[i,"ID1_POST"],"POST1_1"]
PP_30.30[i,"POST1_2"] = PP_5.5[PP_KNNs[i,"ID1_POST"],"POST1_2"]
PP_30.30[i,"POST1_3"] = PP_5.5[PP_KNNs[i,"ID1_POST"],"POST1_3"]
PP_30.30[i,"POST1_4"] = PP_5.5[PP_KNNs[i,"ID1_POST"],"POST1_4"]
PP_30.30[i,"POST1_5"] = PP_5.5[PP_KNNs[i,"ID1_POST"],"POST1_5"]
PP_30.30[i,"POST1_6"] = PP_5.5[PP_KNNs[i,"ID2_POST"],"POST1_1"]
PP_30.30[i,"POST1_7"] = PP_5.5[PP_KNNs[i,"ID2_POST"],"POST1_2"]
PP_30.30[i,"POST1_8"] = PP_5.5[PP_KNNs[i,"ID2_POST"],"POST1_3"]
PP_30.30[i,"POST1_9"] = PP_5.5[PP_KNNs[i,"ID2_POST"],"POST1_4"]
PP_30.30[i,"POST1_10"] = PP_5.5[PP_KNNs[i,"ID2_POST"],"POST1_5"]

PP_30.30[i,"POST1_11"] = PP_5.5[PP_KNNs[i,"ID3_POST"],"POST1_1"]
PP_30.30[i,"POST1_12"] = PP_5.5[PP_KNNs[i,"ID3_POST"],"POST1_2"]
PP_30.30[i,"POST1_13"] = PP_5.5[PP_KNNs[i,"ID3_POST"],"POST1_3"]
PP_30.30[i,"POST1_14"] = PP_5.5[PP_KNNs[i,"ID3_POST"],"POST1_4"]
PP_30.30[i,"POST1_15"] = PP_5.5[PP_KNNs[i,"ID3_POST"],"POST1_5"]
PP_30.30[i,"POST1_16"] = PP_5.5[PP_KNNs[i,"ID4_POST"],"POST1_1"]
PP_30.30[i,"POST1_17"] = PP_5.5[PP_KNNs[i,"ID4_POST"],"POST1_2"]
PP_30.30[i,"POST1_18"] = PP_5.5[PP_KNNs[i,"ID4_POST"],"POST1_3"]
PP_30.30[i,"POST1_19"] = PP_5.5[PP_KNNs[i,"ID4_POST"],"POST1_4"]
PP_30.30[i,"POST1_20"] = PP_5.5[PP_KNNs[i,"ID4_POST"],"POST1_5"]

PP_30.30[i,"POST1_21"] = PP_5.5[PP_KNNs[i,"ID5_POST"],"POST1_1"]
PP_30.30[i,"POST1_22"] = PP_5.5[PP_KNNs[i,"ID5_POST"],"POST1_2"]
PP_30.30[i,"POST1_23"] = PP_5.5[PP_KNNs[i,"ID5_POST"],"POST1_3"]
PP_30.30[i,"POST1_24"] = PP_5.5[PP_KNNs[i,"ID5_POST"],"POST1_4"]
PP_30.30[i,"POST1_25"] = PP_5.5[PP_KNNs[i,"ID5_POST"],"POST1_5"]
PP_30.30[i,"POST1_26"] = PP_5.5[PP_KNNs[i,"ID6_POST"],"POST1_1"]
PP_30.30[i,"POST1_27"] = PP_5.5[PP_KNNs[i,"ID6_POST"],"POST1_2"]
PP_30.30[i,"POST1_28"] = PP_5.5[PP_KNNs[i,"ID6_POST"],"POST1_3"]
PP_30.30[i,"POST1_29"] = PP_5.5[PP_KNNs[i,"ID6_POST"],"POST1_4"]
PP_30.30[i,"POST1_30"] = PP_5.5[PP_KNNs[i,"ID6_POST"],"POST1_5"]}
```

### .4.4 Random Sampling of 5-fold EMA Windows and Days

Random sampling of 5-fold EMA assessments from 30-fold intervals for the generation of individual (1) 5-fold windows (similar process for the 5-fold Random-Window standard-questionnaire data set) and (2) 5-fold single assessment days.

```
pacman::p_load(dplyr)
set.seed(42)

pre_5mzp = c("PRE1_1","PRE1_2","PRE1_3","PRE1_4","PRE1_5")
post_5mzp = c("POST1_1","POST1_2","POST1_3","POST1_4","POST1_5")

pre_30mzp = c("PRE1_1","PRE1_2","PRE1_3","PRE1_4","PRE1_5",
              "PRE1_6","PRE1_7","PRE1_8","PRE1_9","PRE1_10",
              "PRE1_11","PRE1_12","PRE1_13","PRE1_14","PRE1_15",
              "PRE1_16","PRE1_17","PRE1_18","PRE1_19","PRE1_20",
              "PRE1_21","PRE1_22","PRE1_23","PRE1_24","PRE1_25",
              "PRE1_26","PRE1_27","PRE1_28","PRE1_29","PRE1_30")
post_30mzp = c("POST1_1","POST1_2","POST1_3","POST1_4","POST1_5",
               "POST1_6","POST1_7","POST1_8","POST1_9","POST1_10",
               "POST1_11","POST1_12","POST1_13","POST1_14","POST1_15",
               "POST1_16","POST1_17","POST1_18","POST1_19","POST1_20",
               "POST1_21","POST1_22","POST1_23","POST1_24","POST1_25",
               "POST1_26","POST1_27","POST1_28","POST1_29","POST1_30")

# (1) 5-fold windows (EMA_5.5_Window)
# random sampling of 5-fold pre- and post-assessment windows (5 days
# in a row) from individual 30-fold intervals (EMA_30.30)
EMA_5.5_Window = data.frame(ID = c(),
  Pre_MZP1 = c(), Pre_MZP2 = c(), Pre_MZP3 = c(), Pre_MZP4 = c(),
  Pre_MZP5 = c(), Post_MZP1 = c(), Post_MZP2 = c(), Post_MZP3 = c(),
  Post_MZP4 = c(), Post_MZP5 = c(), PRE1_1 = c(), PRE1_2 = c(),
  PRE1_3 = c(), PRE1_4 = c(), PRE1_5 = c(), POST1_1 = c(),
  POST1_2 = c(), POST1_3 = c(), POST1_4 = c(), POST1_5 = c())

for (i in EMA_30.30$ID) {
  a = sample(1:26, 1)
```

```
  EMA_5.5_pre_Window = pre_30mzp[seq(from = a, to = a+4)]
  b = sample(1:26, 1)
  EMA_5.5_post_Window = post_30mzp[seq(from = b, to = b+4)]

  EMA_5.5_Window[i,"ID"] = i
  EMA_5.5_Window[i,"Pre_MZP1"] = EMA_5.5_pre_Window[1]
  EMA_5.5_Window[i,"Pre_MZP2"] = EMA_5.5_pre_Window[2]
  EMA_5.5_Window[i,"Pre_MZP3"] = EMA_5.5_pre_Window[3]
  EMA_5.5_Window[i,"Pre_MZP4"] = EMA_5.5_pre_Window[4]
  EMA_5.5_Window[i,"Pre_MZP5"] = EMA_5.5_pre_Window[5]
  EMA_5.5_Window[i,"Post_MZP1"] = EMA_5.5_post_Window[1]
  EMA_5.5_Window[i,"Post_MZP2"] = EMA_5.5_post_Window[2]
  EMA_5.5_Window[i,"Post_MZP3"] = EMA_5.5_post_Window[3]
  EMA_5.5_Window[i,"Post_MZP4"] = EMA_5.5_post_Window[4]
  EMA_5.5_Window[i,"Post_MZP5"] = EMA_5.5_post_Window[5]

  EMA_5.5_Window[i,"PRE1_1"] = EMA_30.30[i,EMA_5.5_pre_Window[1]]
  EMA_5.5_Window[i,"PRE1_2"] = EMA_30.30[i,EMA_5.5_pre_Window[2]]
  EMA_5.5_Window[i,"PRE1_3"] = EMA_30.30[i,EMA_5.5_pre_Window[3]]
  EMA_5.5_Window[i,"PRE1_4"] = EMA_30.30[i,EMA_5.5_pre_Window[4]]
  EMA_5.5_Window[i,"PRE1_5"] = EMA_30.30[i,EMA_5.5_pre_Window[5]]
  EMA_5.5_Window[i,"POST1_1"] = EMA_30.30[i,EMA_5.5_post_Window[1]]
  EMA_5.5_Window[i,"POST1_2"] = EMA_30.30[i,EMA_5.5_post_Window[2]]
  EMA_5.5_Window[i,"POST1_3"] = EMA_30.30[i,EMA_5.5_post_Window[3]]
  EMA_5.5_Window[i,"POST1_4"] = EMA_30.30[i,EMA_5.5_post_Window[4]]
  EMA_5.5_Window[i,"POST1_5"] = EMA_30.30[i,EMA_5.5_post_Window[5]]}

# (2) 5-fold single assessment days (EMA_5.5_Days)
# random sampling of 5-fold pre- and post-assessments (not necessarily
# days in a row) from individual 30-fold intervals (EMA_30.30)
EMA_5.5_Days = data.frame(ID = c(),
  Pre_MZP1 = c(), Pre_MZP2 = c(), Pre_MZP3 = c(), Pre_MZP4 = c(),
  Pre_MZP5 = c(), Post_MZP1 = c(), Post_MZP2 = c(), Post_MZP3 = c(),
  Post_MZP4 = c(), Post_MZP5 = c(), PRE1_1 = c(), PRE1_2 = c(),
  PRE1_3 = c(), PRE1_4 = c(), PRE1_5 = c(), POST1_1 = c(),
  POST1_2 = c(), POST1_3 = c(), POST1_4 = c(), POST1_5 = c())
```

```
for (i in EMA_30.30$ID) {
  EMA_5.5_pre_Days = pre_30mzp[sort(sample(1:30, 5))]
  EMA_5.5_post_Days = post_30mzp[sort(sample(1:30, 5))]

  EMA_5.5_Days[i,"ID"] = i
  EMA_5.5_Days[i,"Pre_MZP1"] = EMA_5.5_pre_Days[1]
  EMA_5.5_Days[i,"Pre_MZP2"] = EMA_5.5_pre_Days[2]
  EMA_5.5_Days[i,"Pre_MZP3"] = EMA_5.5_pre_Days[3]
  EMA_5.5_Days[i,"Pre_MZP4"] = EMA_5.5_pre_Days[4]
  EMA_5.5_Days[i,"Pre_MZP5"] = EMA_5.5_pre_Days[5]
  EMA_5.5_Days[i,"Post_MZP1"] = EMA_5.5_post_Days[1]
  EMA_5.5_Days[i,"Post_MZP2"] = EMA_5.5_post_Days[2]
  EMA_5.5_Days[i,"Post_MZP3"] = EMA_5.5_post_Days[3]
  EMA_5.5_Days[i,"Post_MZP4"] = EMA_5.5_post_Days[4]
  EMA_5.5_Days[i,"Post_MZP5"] = EMA_5.5_post_Days[5]

  EMA_5.5_Days[i,"PRE1_1"] = EMA_30.30[i,EMA_5.5_pre_Days[1]]
  EMA_5.5_Days[i,"PRE1_2"] = EMA_30.30[i,EMA_5.5_pre_Days[2]]
  EMA_5.5_Days[i,"PRE1_3"] = EMA_30.30[i,EMA_5.5_pre_Days[3]]
  EMA_5.5_Days[i,"PRE1_4"] = EMA_30.30[i,EMA_5.5_pre_Days[4]]
  EMA_5.5_Days[i,"PRE1_5"] = EMA_30.30[i,EMA_5.5_pre_Days[5]]
  EMA_5.5_Days[i,"POST1_1"] = EMA_30.30[i,EMA_5.5_post_Days[1]]
  EMA_5.5_Days[i,"POST1_2"] = EMA_30.30[i,EMA_5.5_post_Days[2]]
  EMA_5.5_Days[i,"POST1_3"] = EMA_30.30[i,EMA_5.5_post_Days[3]]
  EMA_5.5_Days[i,"POST1_4"] = EMA_30.30[i,EMA_5.5_post_Days[4]]
  EMA_5.5_Days[i,"POST1_5"] = EMA_30.30[i,EMA_5.5_post_Days[5]]}
```

## .4.5 R Code for the Calculation of Clinical Change Methods

### Percentage Change

Calculation of the Percentage Change method PC for interpreting the score difference between two assessment intervals (i.e. Mean Percentage Change), as well as between two single assessments for the questionnaire data set as an example (similar process for both the EMA and the questionnaire data set).

```
pacman::p_load(dplyr)

### PP_5.5:
PP_5.5$Mean_PC = (1-(PP_5.5$POST_Mean / PP_5.5$PRE_Mean)) * 100

# creating the interpretation categories for Percentage Change,
# ranging from -2 (strong deterioration) to 2 (strong improvement):
PP_5.5 = PP_5.5 %>%
  mutate(Mean_PC_klass = case_when(
    Mean_PC <= -50 ~ -2,
    Mean_PC > -50 & Mean_PC <= -25 ~ -1,
    Mean_PC > -25 & Mean_PC < 25 ~ 0,
    Mean_PC >= 25 & Mean_PC < 50 ~ 1,
    Mean_PC >= 50 ~ 2,
    TRUE ~ Mean_PC))

### PP_30.30:
PP_30.30$Mean_PC = (1-(PP_30.30$POST_Mean / PP_30.30$PRE_Mean)) * 100

# creating the interpretation categories for Percentage Change,
# ranging from -2 (strong deterioration) to 2 (strong improvement):
PP_30.30 = PP_30.30 %>%
  mutate(Mean_PC_klass = case_when(
    Mean_PC <= -50 ~ -2,
    Mean_PC > -50 & Mean_PC <= -25 ~ -1,
    Mean_PC > -25 & Mean_PC < 25 ~ 0,
    Mean_PC >= 25 & Mean_PC < 50 ~ 1,
    Mean_PC >= 50 ~ 2,
    TRUE ~ Mean_PC))

### PP_1.1:
PP_1.1$PC = (1 - (PP_1.1$POST / PP_1.1$PRE)) * 100

# creating the interpretation categories for Percentage Change,
# ranging from -2 (strong deterioration) to 2 (strong improvement):
PP_1.1 = PP_1.1 %>%
```

```
  mutate(PC_klass = case_when(
    PC <= -50 ~ -2,
    PC > -50 & PC <= -25 ~ -1,
    PC > -25 & PC < 25 ~ 0,
    PC >= 25 & PC < 50 ~ 1,
    PC >= 50 ~ 2,
    TRUE ~ as.numeric(PC)))
```

### Clinical Significance

Implementation of the Clinical Significance method CSI (see McMillan et al., 2010) for interpreting the score difference between two assessment intervals, as well as between two single assessments for the questionnaire data set as an example (similar process for both the EMA and the questionnaire data set).

```
# creating the interpretation categories for Clinically Sig. Change,
# ranging from -1 (improvement) to 1 (deterioration):
pacman::p_load(dplyr)

### PP_5.5:
PP_5.5 = PP_5.5 %>%
   mutate(CSI_klass = case_when(
     PRE_Mean >= 10 & POST_Mean <= 9 & Mean_PC >= 50 ~ -1,
     PRE_Mean <= 9 & POST_Mean >= 10 & Mean_PC <= -50 ~ 1,
     TRUE ~ 0))

### PP_30.30:
PP_30.30 = PP_30.30 %>%
   mutate(CSI_klass = case_when(
     PRE_Mean >= 10 & POST_Mean <= 9 & Mean_PC >= 50 ~ -1,
     PRE_Mean <= 9 & POST_Mean >= 10 & Mean_PC <= -50 ~ 1,
     TRUE ~ 0))

### PP_1.1:
PP_1.1 = PP_1.1 %>%
   mutate(CSI_klass = case_when(
```

```
    PRE >= 10 & POST <= 9 & PC >= 50 ~ -1,
    PRE <= 9 & POST >= 10 & PC <= -50 ~ 1,
    TRUE ~ 0))
```

### Average Internal Consistency

Calculation of the average internal consistency Cronbach´s $\alpha$ as the estimate of reliability to be used to compute Reliable Change Indices. The population´s internal consistency of PHQ-9 assessments was first calculated within each 5-fold interval (pre and post). Then, $\alpha_{pre}$ and $\alpha_{post}$ were Fisher-Z transformed to take the average of both estimates, and finally this value was transformed back to obtain a pooled Cronbach´s $\alpha$. The same calculation method was used for questionnaire and EMA data sets.

```
pacman::p_load(DescTools)
PRE_alpha = CronbachAlpha(PP_5.5[pre_5mzp])
POST_alpha = CronbachAlpha(PP_5.5[post_5mzp])
PP_5.5_Alpha = FisherZInv(mean(c(FisherZ(PRE_alpha),
                                 FisherZ(POST_alpha))))
```

### Reliable Change Index (Jacobson et al., 1984; Jacobson & Truax, 1991)

Calculation of the Reliable Change Index $RCI_{JT}$ and its population-level significance cutoff sensu Jacobson et al. (1984) and Jacobson & Truax (1991) for the difference between two single assessments for the questionnaire data set as an example (similar process for both the EMA and the questionnaire data set).

```
pacman::p_load(dplyr)

PP_1.1$RCI_JT = (PP_1.1$POST - PP_1.1$PRE) /
    sqrt(2 * (sd(PP_1.1$PRE) * sqrt(1 - PP_5.5_Alpha)) ^ 2)
RCI_JT_Cutoff = 1.96 * sqrt(2 * (sd(PP_1.1$PRE) *
    sqrt(1 - PP_5.5_Alpha)) ^ 2)

# creating the interpretation categories for the RCI(JT), ranging
# from -1 (reliable improvement) to 1 (reliable deterioration):
PP_1.1 = PP_1.1 %>%
  mutate(RCI_JT_klass = case_when(
```

```
    PRE >= 10 & POST <= 9 & RCI_JT < -1.96 ~ -1,
    PRE <= 9 & POST >= 10 & RCI_JT > 1.96 ~ 1,
    TRUE ~ 0))
```

### Individualized Reliable Change Index (Pre-SD)

Calculation of a proposed Individualized Reliable Change Index $RCI_{ind,pre\text{-}SD}$ and its corresponding individual significance cutoff for the difference between two assessment intervals, including the subject´s standard deviation from the baseline interval as a measure of individual variability. The same calculation method was used for both questionnaire and EMA data sets.

```
pacman::p_load(dplyr)

### PP_5.5:
PP_5.5$SEd_pre = sqrt(2 * (PP_5.5$ind.pretestSD *
                           sqrt(1 - PP_5.5_Alpha)) ^ 2)
PP_5.5$RCI_ind_preSD = (PP_5.5$POST_Mean - PP_5.5$PRE_Mean) /
                           PP_5.5$SEd_pre
PP_5.5$RCI_ind_preSD_Cutoff =  1.96 * PP_5.5$SEd_pre

# creating the interpretation categories for the RCI(JT), ranging
# from -1 (reliable improvement) to 1 (reliable deterioration):
PP_5.5 = PP_5.5 %>%
  mutate(RCI_ind_preSD_klass = case_when(
    PRE_Mean >= 10 & POST_Mean <= 9 & RCI_ind_preSD < -1.96 ~ -1,
    PRE_Mean <= 9 & POST_Mean >= 10 & RCI_ind_preSD > 1.96 ~ 1,
    TRUE ~ 0))

### PP_30.30:
PP_30.30$SEd_pre = sqrt(2 * (PP_30.30$ind.pretestSD *
                           sqrt(1 - PP_5.5_Alpha)) ^ 2)
PP_30.30$RCI_ind_preSD = (PP_30.30$POST_Mean - PP_30.30$PRE_Mean) /
                           PP_30.30$SEd_pre
PP_30.30$RCI_ind_preSD_Cutoff =  1.96 * PP_30.30$SEd_pre
```

```
# creating the interpretation categories for the RCI(JT), ranging
# from -1 (reliable improvement) to 1 (reliable deterioration):
PP_30.30 = PP_30.30 %>%
  mutate(RCI_ind_preSD_klass = case_when(
    PRE_Mean >= 10 & POST_Mean <= 9 & RCI_ind_preSD < -1.96 ~ -1,
    PRE_Mean <= 9 & POST_Mean >= 10 & RCI_ind_preSD > 1.96 ~ 1,
    TRUE ~ 0))
```