Journal of Psychosomatic Research 137 (2020) 110211

Contents lists available at ScienceDirect

Journal of Psychosomatic Research

journal homepage: www.elsevier.com/locate/jpsychores

Time to get personal? The impact of researchers choices on the selection of treatment targets using the experience sampling methodology☆

Jojanneke A. Bastiaansen[a,b], Yoram K. Kunkels[a], Frank J. Blaauw[c,d], Steven M. Boker[e], Eva Ceulemans[f], Meng Chen[g], Sy-Miin Chow[g], Peter de Jonge[a,c], Ando C. Emerencia[c], Sacha Epskamp[h], Aaron J. Fisher[i], Ellen L. Hamaker[j], Peter Kuppens[f], Wolfgang Lutz[k], M. Joseph Meyer[e], Robert Moulder[e], Zita Oravecz[g], Harriëtte Riese[a], Julian Rubel[l], Oisín Ryan[j], Michelle N. Servaas[a], Gustav Sjobeck[e], Evelien Snippe[a], Timothy J. Trull[m], Wolfgang Tschacher[n], Date C. van der Veen[a], Marieke Wichers[a], Phillip K. Wood[m], William C. Woods[o], Aidan G.C. Wright[o], Casper J. Albers[c], Laura F. Bringmann[a,c,*]

[a] *Interdisciplinary Center Psychopathology and Emotion regulation, University of Groningen, University Medical Center Groningen, Department of Psychiatry, Groningen, the Netherlands*
[b] *Department of Education and Research, Friesland Mental Health Care Services, Leeuwarden, the Netherlands*
[c] *Department of Psychology, University of Groningen, Groningen, the Netherlands*
[d] *Distributed Systems group, Faculty of Science and Engineering, University of Groningen, Groningen, the Netherlands*
[e] *Department of Psychology, University of Virginia, Charlottesville, USA*
[f] *Faculty of Psychology and Educational Sciences, University of Leuven, Leuven, Belgium*
[g] *Department of Human Development and Family Studies, Pennsylvania State University, State College, USA*
[h] *Department of Psychology, University of Amsterdam, Amsterdam, Netherlands*
[i] *Department of Psychology, University of California Berkeley, Berkeley, USA*
[j] *Department of Methodology and Statistics, Faculty of Social and Behavioural Sciences, Utrecht University, Utrecht, the Netherlands*
[k] *Department of Psychology, University of Trier, Trier, Germany*
[l] *Department of Psychology, Justus-Liebig-University Giessen, Germany*
[m] *Department of Psychological Sciences, University of Missouri, Columbia, USA*
[n] *University Hospital of Psychiatry and Psychotherapy, University of Bern, Bern, Switzerland*
[o] *Department of Psychology, University of Pittsburgh, Pittsburgh, USA*

ARTICLE INFO

Keywords:
Time-series analysis
Electronic diary
Personalized medicine
Mental disorders
Psychological networks
Crowdsourcing science

ABSTRACT

Objective: One of the promises of the experience sampling methodology (ESM) is that a statistical analysis of an individual's emotions, cognitions and behaviors in everyday-life could be used to identify relevant treatment targets. A requisite for clinical implementation is that outcomes of such person-specific time-series analyses are not wholly contingent on the researcher performing them.

Methods: To evaluate this, we crowdsourced the analysis of one individual patient's ESM data to 12 prominent research teams, asking them what symptom(s) they would advise the treating clinician to target in subsequent treatment.

Results: Variation was evident at different stages of the analysis, from preprocessing steps (e.g., variable

☆ This project was initiated by the *iLab* of the Department of Psychiatry, University Medical Center Groningen, Groningen, the Netherlands (http://ilab-psychiatry.nl). Researchers were funded by a variety of sources, none of which had a role in the design of the study, data collection, analysis, or interpretation of data, nor in writing the manuscript. A. G. C. Wright: National Institute of Mental Health (L30 MH101760); E. Ceulemans and P. Kuppens: KU Leuven Research Council grant (GOA/15/003) and Fund for Scientific Research-Flanders grant (FWO G074319N, G066316N); F. J. Blaauw: The Netherlands Initiative for Education Research (NRO) grant (no.644405–16-401); H. Riese and M. Wichers: Innovatiefonds De Friesland (grant no. DS81); J. A. Bastiaansen, M. N. Servaas and H. Riese: charitable foundation Stichting tot Steun VCVGZ (grant no. 239); L. F. Bringmann: Netherlands Organization for Scientific Research Veni Grant (NWO-Veni 191G.037); M. Wichers: European Research Council (ERC) under the European Union's Horizon 2020 research and innovative programme (ERC-CoG-2015; No. 68146); O. Ryan: Netherlands Organization for Scientific Research Talent Grant (NWO Onderzoekstalent 406–15-128); P. K. Wood: National Institute on Alcohol Abuse and Alcoholism (AA024133); T. J. Trull: National Institute on Alcohol Abuse and Alcoholism (AA024133; AA019546); S.-M. Chow: National Institutes of Health (NIH U24AA027684) and National Science Foundation (NSF IGE-1806874).

* Corresponding author at: Faculty of Behavioural and Social Sciences, Department of Psychometrics and Statistics, University of Groningen, Grote Kruisstraat 2/1, 9712 TS Groningen, the Netherlands.
E-mail address: l.f.bringmann@rug.nl (L.F. Bringmann).

https://doi.org/10.1016/j.jpsychores.2020.110211
Received 11 February 2020; Received in revised form 15 July 2020; Accepted 31 July 2020

selection, clustering, handling of missing data) to the type of statistics and rationale for selecting targets. Most teams did include a type of vector autoregressive model, examining relations between symptoms over time. Although most teams were confident their selected targets would provide useful information to the clinician, not one recommendation was similar: both the number (0–16) and nature of selected targets varied widely.
Conclusion: This study makes transparent that the selection of treatment targets based on personalized models using ESM data is currently highly conditional on subjective analytical choices and highlights key conceptual and methodological issues that need to be addressed in moving towards clinical implementation.

1. Introduction

Clinicians rely on evidence-based guidelines for the assessment and treatment of psychiatric disorders such as major depressive disorder (MDD, [1],[2]). These guidelines are built on predominantly group-based (i.e., nomothetic) research. The outcome of nomothetic research represents knowledge that is true on average for the population under investigation [3]. Clinicians, however, rarely meet an average individual in their day-to-day practice. Even within the same diagnostic category, patients vary widely in the combination and intensity of symptoms as well as the development of symptoms over time. There are, for instance, 1030 unique symptom combinations that all qualify for a diagnosis of MDD and none of them is very common [4]. In addition, patients vary widely in their response to treatments [5].

By identifying individual characteristics that determine disease susceptibility as well as treatment response [6], personalized medicine promises to move from treatments that are effective on average towards identifying the best evidence-based treatment for any individual [7,8]. However, if we were to actually tailor treatments to the individual patient [9], we need to look beyond differences between individuals and additionally examine processes within the individual [10,11]. Thus, a more person-specific (i.e., idiographic) research approach is required to complement our current nomothetic focus [12,13,14,15,16], and as such facilitate personalized psychiatric treatments.

The experience sampling methodology (ESM) has been positioned as one of the best opportunities for personalized medicine in psychiatry [17,18]. ESM, which is also commonly referred to as ecological momentary assessment or ambulatory assessment [11], is a structured method that can capture intraindividual changes in psychological processes across time and context through multiple in-the-moment assessments within one person (e.g., through electronic diaries, [19,20]). ESM studies have shown that many symptoms of patients with severe psychiatric disorders show person-specific, meaningful and widespread variation over time [20,21]. Stavrakakis et al. [22], for instance, analyzed temporal relationships between variables at the individual level and showed that the dynamic relationship between affect and physical activity varies considerably between patients with MDD. Person-specific analyses based on ESM time-series data could have great potential for use in clinical practice, because they could provide personalized and contextualized feedback to patients and clinicians [23,24,25].

This idea has mainly been put into practice by experience sampling intervention (ESI) studies for, amongst others, individuals with depressive symptoms [26,27,28,29]. These interventions provide patients with personalized graphical feedback by showing summary statistics (e.g., a patient's average daily positive affect) or outcomes of individual statistical models on dynamic within-person or person-environment relationships (e.g., relationships between affect and physical activity). The aim of these ESM-based interventions is to help patients get insight in their daily emotions, activities, thoughts, and behaviors, to ultimately induce behavioral change and decrease symptoms [30].

There are many other conceivable clinical applications of ESM, including a more detailed monitoring of treatment response (e.g., [31]), but also a more precise assessment of treatment needs ('precision diagnosis', [24]), and hence a more personalized intervention selection and targeted treatment delivery [10]. Korotitsch and Nelson-Gray [32] already suggested two decades ago that self-monitoring data may be used specifically to "identify particular targets for treatment and help decide which aspects of treatment may be most beneficial to a particular patient" (1999, p. 416). Currently, such clinical applications are often data-driven. In a proof-of-principle study, Kroeze et al. [33] discussed ESM-based graphical feedback on the interplay between symptoms and behaviors with a patient suffering from treatment-resistant anxiety and depression. They report that the apparent central role of somatic symptoms convinced the patient to start a treatment that she had repeatedly refused before (i.e., interoceptive exposure). In a larger study by Fisher and colleagues [34,35], 40 patients with a primary diagnosis of MDD and/or generalized anxiety disorder (GAD) completed a 30-day ESM period prior to therapy. The ESM data was then used to inform the selection and sequencing of specific psychotherapeutic intervention modules based on the idea that "interventions for symptoms shown to drive the behavior of other symptoms are preferentially delivered earlier in therapy" ([10], p. 500). For patient "Peter", for example, treatment modules targeting depressive symptoms were delivered earlier, because his person-specific dynamic factor model showed that changes in levels of depression preceded changes in anxiety [10].

This personalized psychotherapy study focused on temporal relationships between symptoms. Different analytic approaches might, however, lead to different outcomes. This has recently been demonstrated by Silberzahn et al. [36] for a clearly-defined and relatively straightforward research question: whether soccer players with dark skin tone are more likely than those with light skin tone to receive red cards from referees. By crowdsourcing data analysis, a strategy in which multiple research teams simultaneously investigate the same research question, they disclosed diversity in analytical approaches and demonstrated how subjective choices influenced results. In theory, a patient's ESM data could be used to answer similarly specific research questions (e.g., in what context do somatic symptoms aggravate most), which would probably lead to a relatively high convergence in outcomes across teams. However, in clinical practice, ESM data have typically been used to answer broad questions (e.g., what treatment module should be delivered first, [10]). Silberzahn et al. [36] suggest that crowdsourcing data analysis could also add great value for research questions that are more complex or broad, not only by uncovering the extent of variability in analytical approaches and resulting outcomes, but also by disclosing different underlying conceptualizations of the research question. Both the issue of analysis-contingent results and conceptualizing the research question are especially pressing in clinical applications of person-specific analyses, because different outcomes can have different implications for patients.

In this study, we will use a crowdsourcing data analysis strategy, in which several expert research teams from around the world are invited to simultaneously investigate the same clinically relevant research question for one single dataset: "What symptom(s) would you advise the treating clinician to target subsequent treatment on, based on a person-centered (-specific) analysis of this particular patient's ESM data?" We will evaluate how much researchers vary in their analytical approach towards these individual time-series data and to what degree outcomes vary based on analytical choices. In addition, we will evaluate how much researchers value the outcomes of their analyses for use in clinical practice. This many-labs project will not only provide a

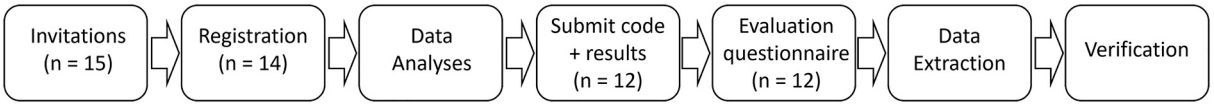

Fig. 1. Flowchart of the study. This figure illustrates the study procedure from inviting research teams to the project team verifying analytical approaches with the research teams.

window on what time-series methods are used in the field today, but also highlight important issues that need to be addressed before these methods can be taken from the realm of the researcher and presented as a tool to patients and clinicians.

2. Methods

2.1. Data analysts

The project group (J.A. Bastiaansen, Y. K. Kunkels, C. J. Albers, L. F. Bringmann) wrote a project description, which included an overview of the research question, a description of the dataset, the planned timeline, and rules for participation. This document was sent to 15 research teams (Fig. 1) selected by the project group for their expertise in ESM (demonstrated by at least one ESM publication, preferably on a clinical topic) and/or the statistical analysis of idiographic data (using any technique). Teams were not obliged to include a clinician. Thirteen groups registered for participation in the project and were sent an ESM dataset (described below) via e-mail. Data were sent to one additional research team, who applied for the project themselves and were accepted based on expertise. Of the initial 14 applications, 12 research teams submitted their code accompanied by a report describing analysis strategy and outcomes. Multiple co-authorships per team were allowed to accommodate the workload of the project. In total, the project involved 28 researchers, who each approved the manuscript and contributed to their team's analysis plan, data analysis, or the description of the procedure and (the interpretation of the) results.

2.2. Dataset

The data were drawn from a multiphase personalized psychotherapy study [34,35]. In brief, participants with a primary diagnosis of GAD and/or MDD completed measurements on their momentary experiences four times per day for at least 30 consecutive days, prior to therapy. Surveys were conducted during participant's self-reported waking hours, which were divided into 3 four-hour blocks that comprised 4 measurements at quasi-random times (with the additional constraint that surveys should be spaced at least 30 min apart). During each survey, subjects were prompted to think about the period of time since the last survey. Items were scored on a visual analogue scale ranging from 0 to 100 with the extremes labeled as 'not at all' and 'as much as possible'. Item order was randomized at each measurement. Each survey included 23 items related to depression and anxiety psychopathology (e.g., felt down or depressed, felt a loss of interest or pleasure, felt frightened or afraid). We use the term momentary for these items, because questions pertained to experiences over a short period of time (4-h blocks). In addition, three items pertaining to sleep were measured on a daily basis. We selected the multivariate times series of one participant based on the following criteria: primary diagnosis of MDD, more than 100 time points in the dataset, and some missingness (as this is typically present in ESM datasets). The selected subject (ID 3) was a white 25-year-old male with a primary diagnosis of MDD and a comorbid GAD. His scores on the Hamilton Rating Scale for Depression [37] and the Hamilton Rating Scale for Anxiety [38] were 16 and 15, respectively. His dataset comprised 122 time points (113 entries and 9 missings) spread across 30 days. The first measurement of the day was offered to the participant around 9 AM and the last measurement around 9 PM. The full item list and dataset are available in Appendix A and at our OSF page, respectively.

2.3. Procedure

For a flowchart of the study, see Fig. 1. After registration, research teams were sent the ESM data. Each team decided on the best strategy to investigate the research question: "What symptom(s) would you advise the treating clinician to target subsequent treatment on, based on a person-centered (-specific) analysis of this particular patient's ESM data?" Teams were requested to submit a report comprising a structured summary of their analytical approach (including information about e.g., data preprocessing, statistical techniques, and software packages) and their results (i.e., a list of target symptoms). After submission of a report, all team members were asked to fill out a short questionnaire (https://osf.io/t5289/) on their expertise and contributions to the project. Teams were additionally asked (https://osf.io/egdu6/) for qualitative feedback on the project and answered, on a 7-point scale with the extremes labeled 'not at all' and 'very', questions on the suitability of the dataset, suitability of their analysis, expected target similarity across teams, clinical usefulness of their selected targets, and readiness of ESM for use in clinical practice. Subsequently, the project group reran the submitted code and reached out to the teams via e-mail to fix bugs and check details. The project group compiled summary tables of the analytical approaches and selected targets for intervention and verified this with the teams (see documentation in team folders at our OSF page: https://osf.io/h3djy/). The project group wrote a first draft of the methods and results section for final verification by the research teams. Finally, the project group completed a full draft of this manuscript, which was sent to all analysts for commenting. The final version of the manuscript was approved by all authors.

3. Results

3.1. Data analysts

Twelve independent[1] teams of researchers submitted their analytical approaches and clarified these, if necessary. Teams worked in five different countries (Belgium, Germany, Switzerland, the Netherlands, the United States). Research teams varied in size from one to four individuals (Mo = 2). Characteristics of the researchers can be found in Fig. 2. Teams included as highest academic rank a Full Professor (n = 7), Associate Professor (n = 2), or Assistant Professor (n = 3). All teams had published at least one paper using ESM and/or at least one paper that was primarily focused on methodology or statistics regarding longitudinal or time-series data. In addition, ten out of twelve (83%) teams included a member that had taught at least one undergraduate or graduate statistics course. Furthermore, ten teams (83%) had published one or more papers on depression and/or anxiety disorders, and eight teams (67%) included a member who had worked in a clinical setting with patients with depression and/or anxiety disorders. Hence, teams generally were not only well-versed with relevant statistics, but also knowledgeable about mood and anxiety disorders.

[1] Two research teams were from different departments of the same university, but worked independently nonetheless.

3.2. Analysis software

Teams used one ($n = 7$) or two ($n = 5$) different standard software programs for their analyses, namely R ($n = 7$), Mplus ($n = 2$), SAS ($n = 2$), LISREL ($n = 1$), Matlab ($n = 1$), Stata ($n = 1$), and open source packages DyFa ($n = 1$, beta version 3.0 (unreleased)), and OpenMx (n = 1, version 2.9: [39]). In most cases, scripts ran errorless or errors were easily fixed, for instance by repeating the analysis with a set seed (i.e., initially randomly generated numbers are fixed to ensure that rerunning the analysis does not change the results). In two cases, there were bugs in the teams' code that needed to be fixed in order for the analysis to provide reproducible results.

3.3. Analytical approaches

There is no standardized approach towards analyzing ESM data; no guidelines outlining steps that need to be taken. Therefore, we used the teams' scripts to recreate the most relevant analysis stages, from preprocessing steps (e.g., variable selection, clustering, and handling of missing data) to the type of statistical analyses. Below, we describe similarities and differences between the approaches of the twelve teams. These sections are inevitably dense in details. For a summary of the variation in analytical approaches, see Text box 1.

Text box 1. *Variation in Analytical Approaches Summary.*

- **Variable selection**: Most teams discarded the (day-level) sleep variables and incorporated all available momentary items in their analyses without specific preselections.
- **Preprocessing** (Table 1)**:** The majority did not standardize the data beforehand, applied a form of detrending, and used an imputation technique to account for missing data.
- **Clustering**: Although many teams used related techniques (Table 1), there were major differences in the clustering of items (Fig. 3). Due to these differences, the input for subsequent inferential analyses varied across teams.
- **Statistical Analyses** (Table 2)**:**
  - Only a handful of teams analyzed mean levels of variables.
  - Most teams included at least one type of analysis focused on relations between variables over time. The exact model, however, varied.
  - Additional analyses examined how variables covaried at the same time point, which item(s) had the overall highest influence on other items in a network (i.e., centrality analysis), and how effects changed over time.

3.3.1. Variable selection

Teams were free to select variables from the provided dataset. One team reported that their first step was to examine the construct validity of items. This team (no. 5) examined whether items were unambiguously formulated and excluded the anhedonia item (I felt a loss of interest or pleasure), because they found it unclear what criterion the patient should use to determine whether “a loss” was present. The team additionally excluded the tension[2] item (I experienced muscle tension). They observed it initially correlated negatively with positive affect items (i.e., tension went down when positive affect items went up), but that the sign of the correlation coefficient became positive towards the end of the ESM study. Hence, the team concluded that the meaning of the tension item might have changed from a negative connotation (stress-related tension) to a more positive connotation (activity-related muscle tension). This team also examined whether variables fluctuated and excluded one item (I avoided activities) due to low within-person variability (i.e., the standard deviation (*SD*) was below 10% of the scale). Furthermore, this team excluded all items pertaining to positive affect for analyses assuming stationarity, because they not only found a change in the mean levels of these items over time (which time series analyses could correct for), but also a shift in the correlational structure between these items. Another team (no. 12) used an automated procedure to perform checks on variable distributions (z-skewness) and within-person variability (mean square of successive difference, *MSSD*), but did not discard any variables based on their criteria ($MSSD < 50$ and/or skewness > 4).

Most teams examined all available momentary items. Eight teams excluded the three sleep variables, because their statistical analysis of choice could not deal with day-level variables and/or the relatively few number of observations ($n = 30$). Team 12, however, included varying sets of six or less variables (including the sleep variables) in their model through an iterative process. Team 3 also included sleep items in their analyses. Team 6 purposefully selected two sleep items in combination

[2] One other team also excluded tension, but after clustering; tension did not clearly measure one thing, but loaded on different clusters.

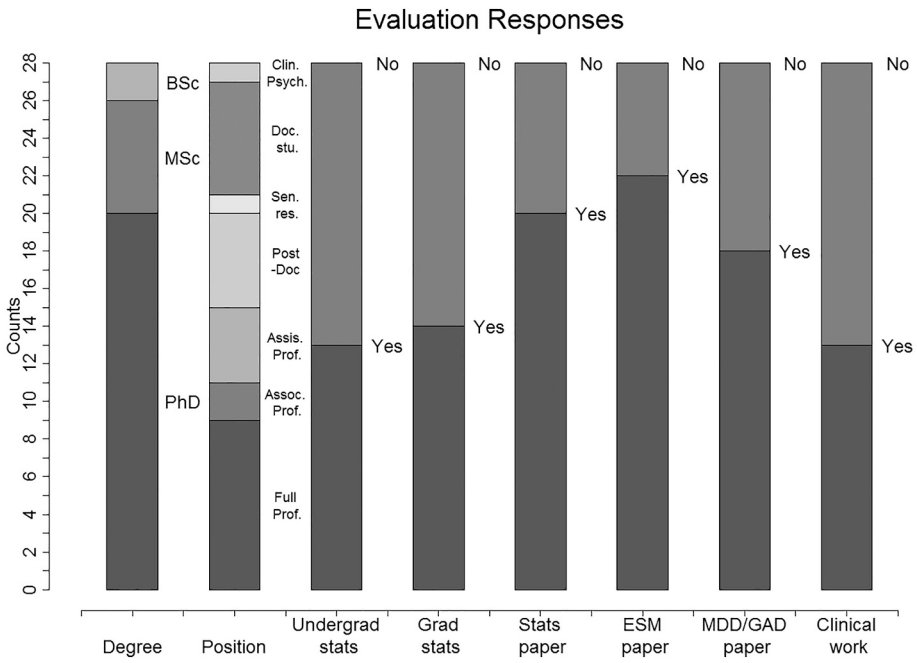

Fig. 2. Characteristics of the researchers. The bars summarize the responses of the 28 researchers to the eight questions in the expertise section of the evaluation questionnaire, regarding researchers' highest academic degree (bachelor, master, doctorate), current position (full professor, associate professor, senior researcher, assistant professor, clinical psychologist, post-doc, doctoral student), experience in teaching undergraduate-level and graduate-level statistics, publications on methodology or statistics concerning time-series data, publications using experience sampling methodology, publications focused on depression and/or anxiety disorders, and clinical experience with depression and/or anxiety.

Table 1
Data handling choices.

Team No.	Clustering technique	Clusters (*n*)	Detrending	Standardizing	Missing data handling
1	Orthogonal PCA	3	No	Yes	Listwise deletion
2	Exploratory and confirmatory dynamic FA [a]	3	Yes	No	Listwise deletion, Imputation by aggregating the four-daily measurements into twice-daily measurements
3	Time-series exploratory FA	9	Yes	Yes	Listwise deletion, Imputation (Maximum Likelihood estimation)
4	Theory-driven	4	Yes	No	Imputation (spline regression)
5	Oblique PCA	4	Yes	No	Listwise deletion
6	–	–	No	No	Imputation (Kalman filter; DSEM)
7	Exploratory and confirmatory FA	2	Yes	No	Listwise deletion, Imputation (Maximum Likelihood estimation)
8	–[b]	0	Yes	Yes	Listwise deletion
9	Oblique exploratory FA	1	Yes	No	Imputation (cubic spline interpolation)
10	Orthogonal PCA	5	No	No	Listwise deletion
11	Oblique exploratory FA	4	No	No	Imputation (Kalman filter; DSEM)
12	–	0	Yes	No	Imputation (Amelia II)

Note. PCA = Principal Component Analysis, FA = Factor Analysis, DSEM = Dynamic Structural Equation Model [a] In contrast to the other teams, who applied a clustering technique before moving on to statistical models, this team's clustering technique was contained within their statistical model. [b] This team did not cluster items prior to their statistical analyses, but created clusters after their analyses based on visual inspection and clinical theoretical reasoning. Note that three teams handled missing data differently in different analyses (nos. 2, 3 and 7).

with merely three momentary variables based on theory (i.e., the role of sleep in triggering core affective symptoms), and then chose their statistical analysis accordingly. Team 1 also used the sleep items in a separate analysis to examine relationships between sleep problems and affective symptoms.

3.3.2. Clustering

Three teams only used individual items in their statistical analyses. The other nine teams grouped the items (at least some of) into clusters[3] prior to at least one statistical analysis to reduce data dimensionality. One of these nine teams (no. 4) used theoretical reasoning to create clusters for positive affect, negative affect, depressive symptoms, and generalized anxiety symptoms. The other eight teams created clusters in a data-driven manner through six different but related techniques (i.e., variants of factor or principal component analysis, for details see Table 1). Nonetheless, no two teams had exactly the same clustering.

[3] We use the term cluster loosely to include the output of both PCA (components) and FA (factors).

In total, 35 clusters (range: 1–9, *Mdn* = 4) were created of which 29 had unique content (i.e., cluster compositions differed in at least one item). The remaining six clusters had an 'identical twin', that is, there were three pairs of clusters comprising the exact same items for two teams. Fig. 3 shows for each research team how items were clustered and illustrates the diversity in outcomes. We applied cluster numbering to align four types of clusters that were somewhat comparable across several teams.

Cluster 1 (green circles in Fig. 3): teams 9 and 11 both had a cluster labeled positive affect comprising the items enthusiastic, content, positive and accepted. Four additional teams had a cluster comprising positive affect items in (slightly different) combinations. One team (no. 1) had a cluster named feeling bad/good that included both positive and negative items.

Cluster 2 (blue circles in Fig. 3): teams 4 and 10 both had a cluster labeled depression comprising five items, namely guilty, anhedonia, hopeless, down, and fatigue. Five other teams had a cluster comprising at least three of these items (amongst other items) in a cluster that they labeled MDD, depressed, depression, or low-arousal negative affect.

Cluster 3 (red circles in Fig. 3): teams 10 and 11 both created a

cluster comprising the items irritable, restless, worried, and concentrate. Five additional teams had a cluster comprising at least three of these items in addition to other negative items. One team had a cluster comprising the item irritable with a mixture of positive and negative items. The variable composition of cluster 3 is reflected by the diversity in cluster names (nervous, anxiety, GAD, high-arousal negative affect, high arousal distress, mental unrest, negative affect).

Cluster 4 (yellow circles in Fig. 3): six teams had a cluster that comprised at least one of the items tension, threatened, or afraid. Here, the diversity in cluster names also reflected the variable composition of these clusters (bodily discomfort/threat/avoidance, defensive, GAD, anxiety, threatened, threat engagement). The remaining clusters were even less comparable across teams and are indicated in Fig. 3 by a grayscale.

In sum, there was a wide variety in cluster outcomes with no two teams having exactly the same clustering. However, six teams did have a cluster comprising predominantly positive affect items, and seven teams had a cluster that comprised items that most of them labeled as depression. Multiple teams also included at least one cluster comprising negative affect items, but the content and labeling of these clusters was rather variable.

We should note here that one of the teams (no. 8) that did not cluster items prior to their statistical analyses, did create clusters after their analyses to interpret the results. Based on visual inspection and clinical theoretical reasoning by a clinician, they created a 'depression' cluster and an 'irritable-distress' cluster, which partly overlap with cluster 2 and 3, respectively. These clusters are indicated by lighter shades of blue and red in Fig. 3.

3.3.3. Handling of data

Teams generally performed few preprocessing steps (Table 1). Nine teams used the raw data; the other three teams standardized the data beforehand. Many teams (8/12) applied some form of detrending (i.e., removing trends from the time series such as a change in the mean over time), either beforehand or within their model (e.g., by adding a linear trend to the model). Many teams (8/12) used an imputation technique to account for missing data in at least one of their analyses, for instance through smoothing (e.g., cubic splines: [40]) or Bayesian techniques (e.g., a Kalman filter: [41]). Other teams simply dealt with missing data through listwise deletion (i.e., if the value of a single variable was missing for a certain measurement the entire record for that measurement was excluded from analysis).

Three out of the twelve teams checked for the robustness of their outcomes across a couple of variations of their model (no. 2, 4 and 6). For instance, team 2 ran their model on the raw, non-equally spaced data (i.e., four measurements during the day and none at night), but also ran their model on data converted to approximately equidistant intervals (i.e., a morning and an evening measurement spaced 12 h apart). Furthermore, one team (no. 12) took robustness into account by selecting the associations that were most prevalent across multiple model configurations and/or those that replicated across imputation strategies. This team noticed that their imputation procedure did not adequately handle the relatively large number of missing values at the end of the ESM study and recomputed their models after removing the last part of the time series (which led to different results).

Five team reports provided descriptive statistics (i.e., basic summaries of the data through plots and/or measures such as means and variances) before moving on to cluster procedures or other more advanced inferential modelling techniques.

3.3.4. Statistical analyses

The teams performed various different statistical analyses, which are outlined in Table 2 (and summarized in Text box 1). A handful of teams, for instance, analyzed mean levels of items. The variety of analyses can be further broken down into three themes, which are described below.

3.3.4.1. Contemporaneous and lagged effects. Vector-autoregressive (VAR) modelling was part of the analyses of all teams except for one. VAR models are used to determine whether the time series of one variable (i.e., an item or cluster) is useful in predicting its own time series from one moment to the next (autoregressive associations) and the time series of another variable from one time point to another (cross-lagged associations, [42,43]). Most teams that used a VAR model examined autoregressive (11/11) and cross-lagged (10/11) associations between items or clusters from one measurement to the next (lag 1), which were on average spaced 3 h apart. Two teams (nos. 3, 12) did not only include autoregressive associations from one time point to the next, but also included the effect on the time point after that (i.e., autoregressive association lag 2). Team 3 did not only use a discrete VAR-based model, but also used a continuous time modelling approach. Whereas a discrete VAR model assumes equidistant measurements (which is often – and also in the current instance – not the case), a continuous-time VAR model can handle variables that are measured on different time scales (e.g., momentary variables combined with day-level variables such as sleep). Teams 6 and 12 used alternative approaches to analyze variables with different time scales. Team 12 applied an imputation technique, whereas team 6 changed the structure of the data to a combination of wide and long format.[4]

Some teams (6 out of 11) not only used VAR to estimate effects across time, but also used their VAR model to examine how variables covaried at the same time point (contemporaneous effects or lag 0). The one team (no. 7) that did not use a VAR model studied contemporaneous effects between items through a regression-based network. Another team (no. 4) studied contemporaneous effects through spline regression.

One team (no. 5) not only examined lagged associations between symptoms using a VAR-based model, but also examined unidirectional lagged associations between behavioral items and symptoms. That is, they selected behavioral items that predicted higher symptom levels at a later time point.

3.3.4.2. Networks and centrality analysis. Three teams (nos. 7, 8 and 9) stated they took a network approach, in which items are typically not clustered but individually related to each other [44–46,47]. To reduce data dimensionality these teams used data-driven techniques that reduce the number of parameters [48,49]. Two of these teams (nos. 7 and 9) additionally performed a centrality analysis [50], which aims to identify the item(s) that had the overall highest influence on other items in a network [51,52,53].

3.3.4.3. Changes across time. Most models assumed that the data were normally distributed and stationary (i.e., time series do not change over time) or corrected for non-stationarity (detrending, [54]). Some teams, however, were explicitly interested in how the effects in their regression or VAR models changed over time. For instance, team 3 relaxed the stationarity assumption in their time series factor analysis model [55,56]. Another team (no. 4) examined how associations between variables varied across time using a regression spline method. Rather than examining smooth changes across time, one team (no. 5) examined abrupt changes (i.e., how structural changes in clusters during the ESM period preceded structural changes in other clusters) by means of a change point analysis [57].

[4] In the approach of team 6, all the data for one day were given in wide format as a row, while the days were included in long format. This means that the two sleep variables and the six ESM measurements were represented as different columns, and each measurement was included only once in this data setup.

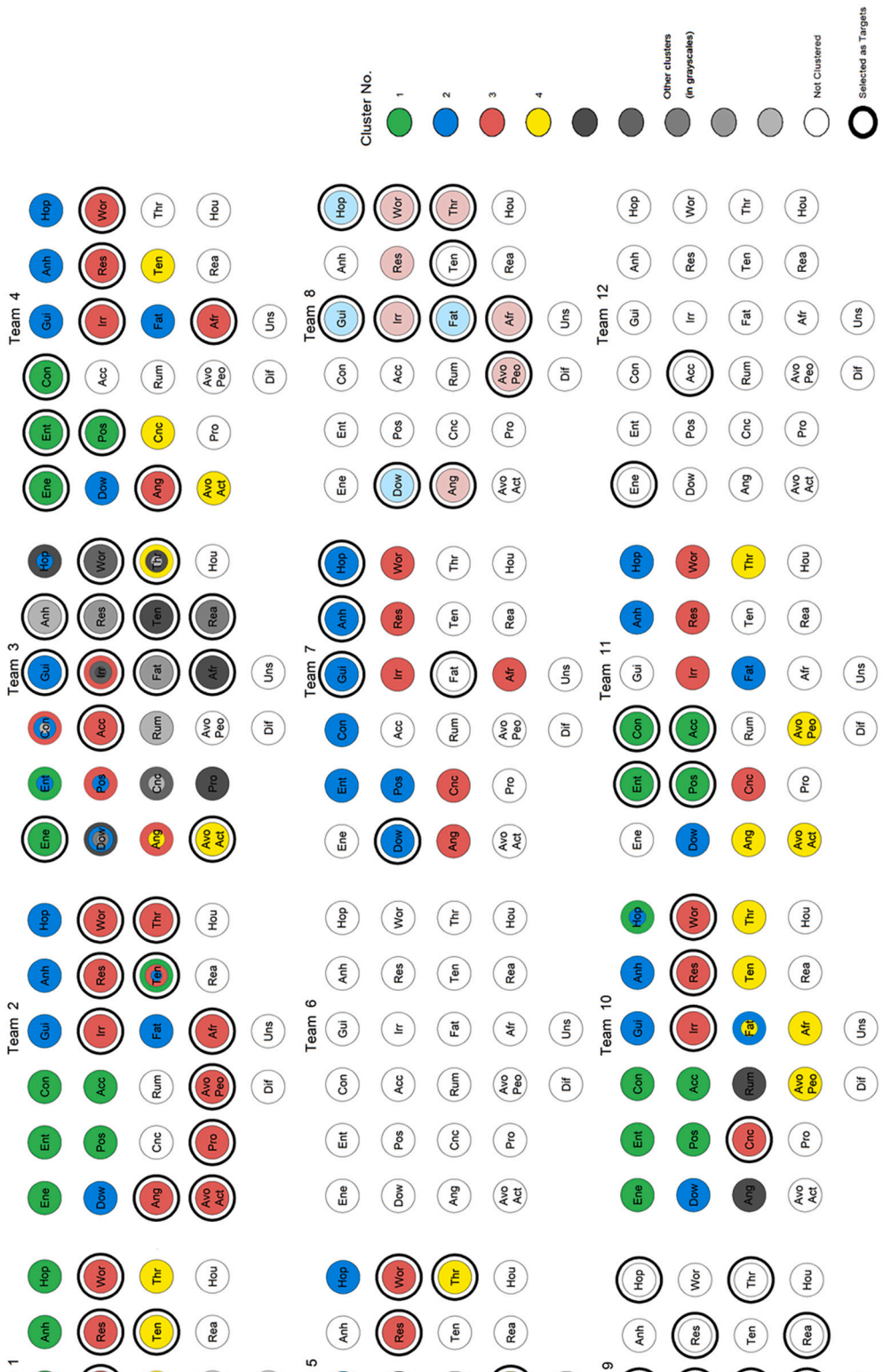

Fig. 3. Clustering and target selection per research team. Each figure part shows for a research team how items (represented by circles) were clustered and which items were eventually selected as targets (bold outline). Clusters that were somewhat comparable were aligned: cluster 1 (green) comprises predominantly positive affect items, cluster 2 (blue) comprises items that some teams labeled as depression, and cluster 3 (red) and cluster 4 (yellow) mainly comprise negative affect items. Team 8 created clusters after rather than prior to their statistical analyses; these clusters are indicated by lighter shades of blue and red. Additional clusters are represented by different shades of gray. A multi-colored circle indicates that this item was part of multiple clusters. Note that teams that included clusters in their analyses did not necessarily use them for target selection. See Table 3 and Table 4 for the target selection results. Ene = energetic, Ent = enthusiastic, Con = content, Gui = guilty, Anh = anhedonia, Hop = hopeless, Dow = down, Pos = positive, Acc = accepted, Irr = irritable, Res = restless, Wor = worried, Ang = angry, Cnc = concentrate, Rum = ruminate, Fat = fatigue, Ten = tension, Thr = threatened, Avo Act = avoid activities, Pro = procrastinate, Avo Peo = avoid people, Afr = afraid, Rea = reassure, Hou = hours of sleep, Dif = difficulty sleeping, Uns = unsatisfying sleep.

Table 2
Overview of statistical analyses including those used for target selection.

Team	Mean-level analysis	VAR-related analysis						Other analyses
	Yes/No	Yes/No	Clusters	Lag 0	Cross-Lag 1	Auto-Lag 1	Auto-Lag 2	
1	Yes	Yes	✓		✓	✓		Additional VAR analysis on sleep items and affective symptoms
2	No	Yes	✓	✓	✓	✓		
3	No	Yes				✓	✓	Additional VAR-analysis based on a continuous time model Time-series exploratory FA
4	Yes	Yes	✓		✓	✓		Spline regression analysis with only concurrent (no lagged) variables
5	Yes	Yes	✓		✓	✓		Regression analysis (10 items) Change point analysis (1 item)
6	No	Yes		✓[b]	✓	✓		
7	No	No						LASSO regression with concurrent (no lagged) variables Centrality analysis
8	Yes	Yes		✓	✓	✓		
9	No	Yes	✓[a]	✓	✓	✓		Centrality analysis
10	No	Yes	✓		✓	✓		
11	No	Yes	✓	✓	✓	✓		
12	No	Yes		✓	✓	✓	✓	

Note. Checkmarks indicate which analysis was executed. Analyses that were eventually used for target selection are indicated by light gray shading. VAR = vector-autoregressive model, Lag 0 = contemporaneous associations, Lag 1 = lagged associations from one time point to the next, Lag 2 = lagged associations across two time points, Auto = autoregressive effect (i.e., the effect of a variable on itself from one time point to the next). [a] Only one cluster amidst a series of individual variables. [b] This team considers their lag 0 model as lagged in nature; their variables have the same time stamp but actually refer to different times (i.e., sleep during preceding night and mood during the day).

3.4. Intervention targets

3.4.1. Target selection rationale

Not all performed statistical analyses were used to support the final selection of intervention targets. The shaded parts in Table 2 show what varying sources of information the teams based their target selection on. Only two teams (nos. 1, 8) used descriptive statistics for target selection. One team (no. 8) examined descriptives of "items related to the criteria of the established DSM-5 diagnoses", "items related to coping" (e.g., avoiding people), and other items such as the item angry, which was finally selected as one of the targets because of its multiple, relatively high peaks in its time series. Descriptive statistics were used (on top of information from cross-lagged and contemporaneous associations and clustering based on visual inspection and clinical theoretical reasoning) to formulate a clinical "working hypothesis" about the patient. Another team (no. 1) set out to determine (1) which symptoms caused the most suffering based on mean levels, (2) lagged associations between sleep problems and symptoms, and (3) lagged associations between different symptoms. In the absence of significant cross-lagged associations, this team selected their targets solely based on the highest mean self-reported rating for negative symptoms, and lowest mean self-reported ratings for positive symptoms. Rather than examining overall symptom levels, a third team (no. 5) analyzed whether there was a *shift* in the mean level of certain symptoms during the ESM period (in addition to examining lagged associations between symptoms and between behaviors and symptoms).

All teams that examined cross-lagged associations (n = 11) selected targets based on these effects or at least intended to do so. For instance, one team (no. 12) selected the item accepted, because it 'reduced' rumination at a later time point, and energetic because it 'reduced' muscle tension. In the absence of significant cross-lagged associations, one team (no. 1) reverted to variable mean scores to select targets (as mentioned above) and three teams (nos. 2, 3 and 11) selected their targets based on the autoregressive effects (i.e., the overspill of variables on themselves). In addition, team 3 selected items that showed cyclical patterns (rapid changes) or had the highest factor loadings in their time series factor analysis. One team (no. 6) did not select any targets, because they found little –if any– evidence for their theory-driven hypothesis. However, if results would have been convincing, they would have selected targets based on their analyses of cross-lagged associations between sleep problems and affective symptoms.

Three of the teams that used a VAR model for information on autoregressive (no. 11) or cross-lagged associations (nos. 8 and 9) to select their targets, also used that model for information on contemporaneous associations between variables. One team (no. 4) only used their VAR model for information on cross-lagged associations and relied on a separate regression analysis for information on contemporaneous associations. Another team (no. 7) solely used information on contemporaneous associations based on a regression analysis to select targets.

Whereas most teams based their targets on cross-lagged or contemporaneous associations between sets of variables, two teams (nos. 7 and 9, which both took a network approach) selected targets based on the average out-strength across all modeled associations, that is, they selected items that had the overall highest influence on other items (centrality measure). Team 9 additionally included items that were most strongly influenced by the most central items.

Table 3
Number and type of selected targets.

Team No.	Potential items	Selected items		Clustering of selected items
	n	*n*	%	
1	26	9	35	
2	23	10	43	Cluster 3
3	26	13	50	
4	17	9	53	Cluster 1 + Cluster 3
5	20	7	35	Cluster 3 + Cluster 4 + 2 individual items
6	5	0	–	
7	21	5	24	
8	23	11	48	a
9	23	16[b]	70	Cluster 1 + 12 individual items
10	23	4	17	Cluster 3
11	23	4	17	Cluster 1
12	26	2	8	

Note. Every item is a potential target if it has been included by a team in at least one statistical analysis including clustering (N.B.: teams could have included different subsets of items in different analyses). The percentage of selected items refers to the relative number of potential items that were selected by the team. Cluster 1 commonly comprised items related to positive affect. Clusters 3 and 4 comprised varying subsets of NA items. [a] This team did not perform statistical clustering but created two clusters based on visual inspection from a clinical theoretical viewpoint after their analyses to formulate a working hypothesis as a starting point in treatment. Eventually, individual items were selected as targets. [b] This team suggested to target symptoms and behaviors across 4 consecutive phases.

3.4.2. Selected targets

Teams varied in both the number (Table 3) and nature (Fig. 3, Table 4) of selected targets. Table 3 shows that teams selected between 2 and 16 (*Mdn* = 9) of the potential items (*Mdn* = 23) either as individual targets (5 teams), as part of a target cluster (4 teams) or as a combination of cluster(s) and individual items (2 teams). Selected targets per team are shown as circles with a bold outline in Fig. 3 and are listed in Table 4. Table 4 additionally shows per item how many teams selected it as a target (either as an individual item or as part of a cluster), which ranged from 0 to 7 teams (*Mdn* = 4). The most often selected items (by 7 teams) were irritable, restless, and worried. None of the teams selected the exact same set of items.

Table 4
Selected target items per research team.

	Team number												
Item	1	2	3	4	5	6	7	8	9	10	11	12	Sum
Irritable	✓	✓	✓	✓				✓	✓	✓			7
Restless	✓	✓	✓	✓	✓				✓	✓			7
Worried	✓	✓	✓	✓	✓			✓		✓			7
Afraid		✓	✓	✓	✓			✓	✓				6
Accepted	✓		✓						✓		✓	✓	5
Threatened		✓	✓		✓			✓	✓				5
Angry		✓		✓				✓	✓				4
Avoid people		✓			✓			✓	✓				4
Content	✓			✓					✓		✓		4
Enthusiastic	✓			✓					✓		✓		4
Fatigue			✓				✓	✓	✓				4
Guilty			✓				✓	✓	✓				4
Positive	✓			✓					✓		✓		4
Tension	✓	✓	✓					✓					4
Energetic			✓	✓								✓	3
Down							✓	✓	✓				3
Hopeless							✓	✓	✓				3
Procrastinate		✓			✓				✓				3
Anhedonia			✓				✓						2
Avoid activities		✓	✓										2
Concentrate					✓					✓			2
Reassure			✓						✓				2
Ruminate	✓												1
Difficulty sleeping													0
Hours of sleep													0
Unsatisfying sleep													0

Note. The outer right column shows the total number of times an item was reported by the (twelve) teams as a potential target for intervention. For information on which teams selected target items individually and/or as part of a cluster, see Table 3.

Of the seven teams that included clusters in (some of) their analyses, six eventually selected one or two clusters as targets (Table 3). Cluster diversity made it difficult to determine whether teams identified similar clusters as targets: only clusters 1 and 2 were reasonably comparable across six and seven teams, respectively. Three teams selected cluster 1 (comprising predominantly PA items), amongst other targets. In contrast, none of the teams with a cluster 2 (commonly labeled as depression) selected it as a target. Four teams selected one or two clusters with negative affect items, but - as mentioned above - the content of these clusters varied widely.

Importantly, teams using the same number of clusters or similar analysis techniques also varied in their selected targets. For instance, teams 1 and 10 both used clustering through orthogonal PCA followed by VAR modelling. Whereas team 1 found three clusters, no significant cross-lagged effects, and finally selected nine individual items, team 10 found five clusters, significant cross-lagged effects, and selected one cluster comprising four items (of which three were also selected by team 1).

3.4.3. Treatment selection

Teams were not asked to provide specific treatment recommendations, but simply to list what symptom(s) they would advise the treating clinician to target subsequent treatment on. In their reports, five teams (nos. 1, 2, 3, 7, and 10) listed their selected targets without specifying *how* these should be intervened on (e.g., team 7: "interventions targeting depressed mood are thus indicated").

In contrast, two teams specifically advised behavioral activation therapy to target positive affect (no. 11) or both positive and negative affect (no. 4: by "increasing behaviors and activities that are pleasurable"). Another team (no. 12) tentatively suggested acceptance and commitment therapy and mindfulness-based therapy to increase feelings of acceptance and improve feeling energetic. One team (no. 9) did not refer to existing treatments, but created a four-phase plan for the treating clinician that included specific recommendations (e.g., "In this phase it seems crucial to work with the patient on his management of his resources and the importance of making breaks. It seems as if he cannot accept his need to rest some times and reacts with feelings of guilt"). Another team (no. 8) also used their observations to formulate a clinical "working hypothesis". If their working hypothesis were to be confirmed by the patient, this team would suggest cognitive behavioral analysis system of psychotherapy and relaxation exercises to improve emotion regulation. This team emphasized that final decisions about which symptoms to target by which interventions "can only be made in dialogue with the patient". Similarly, team 5 suggested that their selected targets should only be used to start a dialogue between clinician and patient about the first target for intervention. Moreover, they point out that in this case the patient's own clinical question was unknown, but this should - in their opinion - be the starting point of any analyses.

In addition to teams 5 and 8, two other teams (nos. 1, 6) noted that in order to tailor interventions to the individual one should look beyond the ESM data and include clinical information. For instance, information on "the symptoms that the patient is most eager to change" (no. 6) or the aspects the clinician sees as most important such as those symptoms causing the most suffering (no. 1).

3.5. Team evaluations

The teams evaluated the project by responding to five closed questions (1–7 scale, Appendix B); 8 of the 12 teams also provided

additional comments in the open fields of the questionnaire. Together, these data show that teams varied widely in how suitable they found the dataset for answering the research question (range: 1–6, *Mdn* = 4.5). Some teams reported the availability of many observations as a strength (no. 8), although more might have been better (no. 5), while others advocated a longer time frame given the number of variables (nos. 2, 6, 11). Team 6 refrained from selecting targets, because they deemed the uncertainty of parameter estimates too large and the statistical power too low. The fact that there were multiple assessments per day was seen as a nice feature, but team 11 noted there was no justification for the timing of measurements; others noted that the lags between measurements might have been too large to catch relevant psychopathological processes (nos. 5, 7). Two teams stated that item selection could have been more strategic (nos. 2, 5). For instance, team 5 suggested that more items on external stressors, activities, social contexts, physical activity and possibly other behaviors would have been desirable, as "behavior is probably more effective as an advice for targeting than symptoms themselves".

Given any limitations the dataset might have had, research teams were moderately positive about the suitability of their own analytical approach (range: 3–7, *Mdn* = 5). In general, teams were only moderately confident that other teams would come up with the same targets for intervention (range: 1–6, *Mdn* = 4), but they were confident that the targets they selected could provide useful information for the clinician (range: 3–6, *Mdn* = 6). Some teams were very positive about the readiness for person-specific analyses based on ESM data for use in clinical practice, while others emphasized there are still many hurdles to be taken or that it depends on how ESM is used (range: 1–7, *Mdn* = 5).

4. Discussion

Twelve research teams simultaneously investigated the same clinically relevant research question: "What symptom(s) would you advise the treating clinician to target subsequent treatment on, based on a person-centered (-specific) analysis of this particular patient's ESM data?" We examined how much researchers varied in their analytical approach towards these individual time series data and to what degree outcomes varied based on analytical choices.

4.1. Variation in analytical approaches

We used the teams' scripts to recreate the most relevant analysis steps, and associated similarities and differences. We observed some differences in variable selection, but most teams discarded the (day-level) sleep variables and incorporated all available momentary items in their analyses without specific pre-selections. Teams made different choices in whether and how data were preprocessed (e.g., standardization, detrending, missing data). There were major differences in the clustering of items: although many teams used related techniques, no two teams ended up with exactly the same clusters. Due to these differences, the input for subsequent inferential analyses varied across teams. Interestingly, most teams included at least one type of VAR-based analysis, examining relations between variables (e.g., symptoms) over time. The exact model, however, varied (e.g., whether contemporaneous effects were incorporated or not).

4.2. Variation in target selection rationale

Statistical analyses were often the starting point, but some teams additionally used clinical arguments for the selection of targets. Few teams used descriptive statistics such as mean levels as target selection criterion. Most teams selected (or intended to select) intervention targets based on cross-lagged associations, which show what behaviors or symptoms are related to other symptoms at the next time point. For instance, if avoiding people related to feeling less positive at the next time point, avoiding people would have been selected as an intervention target. In the absence of significant cross-lagged associations, three teams selected their targets based on the autoregressive effects, that is, they selected variables that had an effect on itself from one time point to the next. For instance, if being enthusiastic at one time point strongly related to being enthusiastic at the next time point, enthusiastic would have been chosen as a target for intervention. Five teams (additionally) used information on contemporaneous associations between variables. For instance, if feeling irritable correlated with feeling worried at the same time point, those symptoms would have been chosen as targets. Two teams did not select targets based on specific associations between pairs of variables, but based on centrality: the variable with the highest average out-strength across *all* modeled associations was chosen as target.

4.3. Variation in selected targets

Both the number and nature of selected targets varied widely: teams selected between 0 and 16 variables, either as individual targets, as part of a target cluster, or as a combination of clusters and individual items. None of the teams had the exact same set of targets, not even teams using the same number of clusters or similar analysis techniques. Thus, depending on which of the 12 teams our hypothetical clinician would have consulted to analyze the ESM data of this patient with MDD and comorbid GAD, he/she would have received a different list of symptoms to target in subsequent treatment. There were, however, also some similarities: while most items were only selected by a minority of teams, the items irritable, restless, and worried were selected by seven teams (in combination with other targets). Furthermore, of the six teams with a reasonably comparable cluster comprising positive affect items (cluster 1), three selected this cluster as a target (either alone, in combination with another cluster, or in combination with individual items). Two of these teams specifically recommended behavioral activation, which is one of the standard recommendations for depression either as a component of cognitive behavioral therapy or as a stand-alone therapy ([1,2]).

4.4. Highlighted issues

Our project highlights several important issues that need to be addressed in moving ESM towards clinical implementation.

4.4.1. Different conceptualizations of important treatment targets

First, the variation in target selection rationale reveals underlying conceptual differences in what teams perceive as 'relevant targets for intervention'. Target selection based on the mean implies, for instance, that symptoms that are most severely affected are most important. Target selection purely based on VAR-based models implies, however, that symptoms are important targets if they either correlate with themselves across time (auto-lag), correlate with other symptoms across time (cross-lag), or correlate with other symptoms at the same time point (contemporaneous effect), on top of all other included effects [58]. Other analyses reveal that symptoms were deemed important if they were most representative of a cluster, rapidly changed, or shifted in mean level across time. These underlying ideas were rarely made explicit. This study shows that clinicians, patients and researchers need to discuss the most relevant information that can be obtained through ESM to support treatment target selection. These ideas should then be put to the test: what information from personalized models is most predictive of treatment change (e.g., are dynamic symptom-symptom relationships a better predictor of treatment response than mean symptom levels)? We should note that in this project, the research question concentrated on symptoms as potential treatment targets. This was partly prompted by the relative scarcity of items on behaviors and context in the available dataset. Naturally, a person's behaviors and daily life context could also make important targets for intervention.

4.4.2. The need for contextualizing person-specific analyses

A second issue, which was raised by several teams, is that ESM data might mean very little in isolation. In our set-up, teams were relatively 'agnostic', that is, they had little background knowledge about the patient's current context and personal history (e.g., previous episodes and interventions). This fueled mostly[5] data-driven approaches. In order to tailor interventions to the individual it might be more fruitful to look beyond ESM data and include clinical information at various stages, starting with the formulation of a clearly-defined, clinically and personally relevant research question. Ideally, the latter will not only guide design choices, such as the selection of variables that are deemed most relevant by the patient and clinician and most reliable by the researcher, but also set the stage for a specific analytical strategy. Several teams were hesitant to make any final decisions about which symptoms to target and advocated target selection should not be mainly data-driven, but done in a dialogue between clinician and patient (for an example see [33]). Other researchers have argued that person-specific analyses need not only be contextualized by personal information but also by comparing individuals to other similarly of differentially affected individuals; examining in what aspects an individual deviates from the norm is essential in targeting maladaptive processes [16].

4.4.3. The need for further scrutinizing person-specific analyses

Third, the variation in analytical approaches demonstrates that there is no standardized manner of analyzing individual ESM data yet. Our study uncovered many potential sources of variation in outcomes. However, we cannot pinpoint the specific impact of the diverging choices we observed. Extensive simulation studies could provide insight here: by generating data under various conditions (e.g., low, medium and high levels of missing data) and measuring the performance of different approaches (e.g., different imputation techniques). Because the true nature of the data generating process of our dataset is unknown, there is no objective way to judge which of the 12 approaches performed the best. Simulation studies could provide insight in which approaches are performing better, on average, or for which type of data (e.g., depending on the number of observations, number of variables, amount of missingness, amount of measurement error, etc., [59]). Furthermore, future research could investigate the impact of other choices by fixing those aspects, for instance, by fixing the clusters beforehand and investigating whether this decreases variation in outcomes.

Person-specific ESM models are still in their infancy. Rather than providing answers, our study shows there are many questions that need to be answered before the field can move towards a goldstandard. Although it is too early to settle on the 'best' approach (from a methodological point of view), teams used several practices that seem worthwhile to adopt on a larger scale: an examination of item validity and variability before advancing to inferential modelling techniques, an outcome robustness check (e.g., across various model configurations), and the inclusion of summary statistics (e.g., the mean) in addition to more complex statistics such as measures of relationships between variables. We would further like to emphasize here that there could never be a one-size-fits-all approach. As we discussed in the previous section, analyses will need to be tailored to the specific research question and individual patient's data.

4.4.4. The need for transparency in person-specific analyses

Fourth, our study underscores the need for transparency in science, particularly in this strongly data-driven and exploratory field, to avert "a lurking replicability crisis" ([60], p. 999). None of the analytic approaches were inherently invalid. Instead, the multiplicity of plausible processing steps implies that there could be several sensible statistical results based on the same original dataset [61]. Or, as one team put it: there may be "many right suggestions to extract from all these data". At many steps in the analysis process, choices between various reasonable (and unreasonable) options have to be made [62]. The route one takes in this 'garden of forking paths' [63] can have a considerable effect on the outcome of the analysis. Thus, researchers need to be transparent about their choices for a reader to be able to appraise the results. Furthermore, researchers should try to mitigate data-contingent analysis decisions, for instance by pre-registration of the analysis plan, prior to observing the data ([64]; for an ESM template see: [65]). In this project, pre-registration could have had an additional advantage: it could have made explicit that teams had different conceptualizations of the research question and therefore different analysis goals. The variability in analytical approaches in our study is, hence, due to a mixture of teams choosing different paths within the same garden and teams actually working in neighboring ones.

In the previous sections, we focused particularly on challenges with the use of individual ESM data for the selection of targets for treatment. The need for developing best practices in analyzing ESM data is, however, essential to a wide range of clinical applications (from a more precise assessment of treatment needs, to a better tailored treatment programme, and a more detailed monitoring of treatment response). The same holds true for transparency about how analyses are planned and executed. The conceptualization of "important treatment targets" is somewhat specific to our current purpose, although the formulation of a specific research question will be an important consideration for other clinical applications as well (e.g., what represents a "treatment need", how do you define "treatment response"). Furthermore, those applications will also have to find a way to best take into account a person's immediate context and past history. The coming years will likely see a surge of new ESM applications aimed at supporting various aspects of the care process. Evaluating not only the efficacy but also the reliability and validity of each of these applications will be key. This project showed that, for the latter purpose, crowdsourcing data analysis is a useful new tool.

This was the first study that assessed the diversity of analytical approaches for one individual time-series ESM dataset. We found that different research teams chose different analytical approaches and that outcomes – and hence, recommendations to the clinician on treatment targets- varied widely. This study highlights conceptual and methodological issues that need to be addressed in moving person-specific analyses based on ESM data towards clinical implementation. Developing best practices for formulating well-defined, clinically and personally relevant research questions and converting them into appropriate and acceptable study designs with matching analytical strategies is essential [66]. This will require a great collaborative effort between researchers, patients, and clinicians.

Author contributions

The project group (J. A. Bastiaansen, Y. K. Kunkels, C. J. Albers, L. F. Bringmann) designed and coordinated the study, analyzed the output by the research teams, and wrote the manuscript. All other authors contributed to their team's analysis plan, data analysis, or the description of the procedure and (the interpretation of the) results, and contributed to and approved the final manuscript.

[5] We use the phrase "mostly" here, because approaches were not solely based on information in the dataset: they were also grounded in teams' ideas on what entailed an important treatment target (see previous section) and their preferred statistical techniques. Moreover, some teams explicitly relied on clinical arguments in various stages. There was, for instance, one team (no. 4) that created clusters based on theories of MDD and GAD, and one team (no. 6) that explicitly tested the theory that sleep triggers affective symptoms. Some teams explicitly used clinical reasoning in combination with statistical information (e.g., no. 8) to select treatment targets, and several teams provided specific treatment recommendations based on clinical reasoning and/or current guidelines. Notwithstanding these theoretical components, we use the term data-driven, because in this project (in contrast to actual clinical practice) the dataset was inevitably the starting point.

Funding

Researchers were funded by a variety of sources, none of which had a role in the design of the study, data collection, analysis, or interpretation of data, nor in writing the manuscript. A. G. C. Wright: National Institute of Mental Health (L30 MH101760); E. Ceulemans and P. Kuppens: KU Leuven Research Council grant (GOA/15/003) and Fund for Scientific Research-Flanders grant (FWO G074319N, G066316N); F. J. Blaauw: The Netherlands Initiative for Education Research (NRO) grant (no.644405–16-401); H. Riese and M. Wichers: Innovatiefonds De Friesland (grant no. DS81); J. A. Bastiaansen, M. N. Servaas and H. Riese: charitable foundation Stichting tot Steun VCVGZ (grant no. 239); L. F. Bringmann: Netherlands Organization for Scientific Research Veni Grant (NWO-Veni 191G.037); M. Wichers: European Research Council (ERC) under the European Union's Horizon 2020 research and innovative programme (ERC-CoG-2015; No. 68146); O. Ryan: Netherlands Organization for Scientific Research Talent Grant (NWO Onderzoekstalent 406–15-128); P. K. Wood: National Institute on Alcohol Abuse and Alcoholism (AA024133); T. J. Trull: National Institute on Alcohol Abuse and Alcoholisma (AA024133; AA019546); S.-M. Chow: National Institutes of Health (NIH U24AA027684) and National Science Foundation (NSF IGE-1806874).

Disclosures

Further information on this study is available online as a project on the Open Science Framework (OSF): https://osf.io/h3djy/. This includes the project description, the dataset, materials (item list, evaluation questionnaires), a summary of each team's analytical approach, and – for the eleven out of twelve teams that agreed with our open science statement – their (anonymized) original reports and analytical code. This research was conducted using previously published, publicly available data and according to ethical standards. A preprint of the manuscript can be found here: https://doi.org/10.31234/osf.io/c8vp7.

Declaration of Competing Interest

The authors declared no conflicts of interest with respect to the authorship or the publication of this article.

Acknowledgements

This project was initiated by the *iLab* of the Department of Psychiatry, University Medical Center Groningen, Groningen, the Netherlands (http://ilab-psychiatry.nl).

Appendix A. ESM item list

Variable name	Abbreviation	Variable type (momentary/day)	Full item text *(to what degree have you)*	Scale
Down	Dow	Momentary	Felt down or depressed	0–100 (*not at all – as much as possible*)
Hopeless	Hop	Momentary	Felt hopeless	0–100 (*not at all – as much as possible*)
Angry	Ang	Momentary	Felt angry	0–100 (*not at all – as much as possible*)
Anhedonia	Anh	Momentary	Experienced loss of interest or pleasure	0–100 (*not at all – as much as possible*)
Afraid	Afr	Momentary	Felt frightened or afraid	0–100 (*not at all – as much as possible*)
Guilty	Gui	Momentary	Felt worthless or guilty	0–100 (*not at all – as much as possible*)
Worried	Wor	Momentary	Felt worried	0–100 (*not at all – as much as possible*)
Restless	Res	Momentary	Felt restless	0–100 (*not at all – as much as possible*)
Irritable	Irr	Momentary	Felt irritable	0–100 (*not at all – as much as possible*)
Concentrate	Cnc	Momentary	Had difficulty concentrating	0–100 (*not at all – as much as possible*)
Tension	Ten	Momentary	Experienced muscle tension	0–100 (*not at all – as much as possible*)
Fatigue	Fat	Momentary	Felt fatigued	0–100 (*not at all – as much as possible*)
Positive	Pos	Momentary	Felt positive	0–100 (*not at all – as much as possible*)
Content	Con	Momentary	Felt content	0–100 (*not at all – as much as possible*)
Enthusiastic	Ent	Momentary	Felt enthusiastic	0–100 (*not at all – as much as possible*)
Energetic	Ene	Momentary	Felt energetic	0–100 (*not at all – as much as possible*)
avoid_act(ivities)	Avo Act	Momentary	Avoided activities	0–100 (*not at all – as much as possible*)
avoid_people	Avo Peo	Momentary	Avoided people	0–100 (*not at all – as much as possible*)
procrast(inate)	Pro	Momentary	Procrastinated	0–100 (*not at all – as much as possible*)
Reassure	Rea	Momentary	Sought reassurance	0–100 (*not at all – as much as possible*)
Ruminate	Rum	Momentary	Dwelled on the past	0–100 (*not at all – as much as possible*)
Threatened	Thr	Momentary	Felt threatened, judged, or intimidated	0–100 (*not at all – as much as possible*)
Accepted	Acc	Momentary	Felt accepted or supported	0–100 (*not at all – as much as possible*)
Hours (of sleep)	Hou	Day	How many hours did you sleep last night?	0 to 24
Difficult(y sleeping)	Dif	Day	Experienced difficulty falling or staying asleep	0–100 (*not at all – as much as possible*)
Unsatisfy(ing sleep)	Uns	Day	Experienced restless or unsatisfying sleep	0–100 (*not at all – as much as possible*)

Note. Item order was randomized at each measurement.

Appendix B. Responses to the closed evaluation questions

Suitability of the dataset	Suitability own analysis approach	Expected target similarity across teams	Clinical usefulness of own selected targets	Readiness ESM for clinical practice
1	4	1	3	5
3	5	5	6	4
3	5	3	7	6
4	4	6	5	1
4	5	4	4	5
4	3	2	6	5
5	5	2	6	4
5	6	5	6	5
5	6	5	6	5

6	5	4	6	5
6	5	3	5	7
6	7	4	7	7

Note. Answers to the closed evaluation questions filled in by the teams on a 7-point scale with the endpoints 1 (“not at all”) and 7 (“very”). Each row represents a team's responses, sorted in ascending order to the first question.

References

[1] American Psychiatric Association, Practice guideline for the treatment of patients with major depressive disorder (third edition), Am. J. Psychiatry 167 (2010) 10), 1–152.

[2] National Institute for Health and Care Excellence, Depression in Adults: Recognition and Management (NICE Guideline 90), www.nice.org.uk/CG90, (2009).

[3] J.T. Lamiell, ‘Nomothetic’ and ‘idiographic’: contrasting Windelband’s understanding with contemporary usage, Theory Psychol. 8 (1) (1998) 23–38, https://doi.org/10.1177/0959354398081002.

[4] E.I. Fried, R.M. Nesse, Depression is not a consistent syndrome: an investigation of unique symptom patterns in the STAR*D study, J. Affect. Disord. 172 (2015) 96–102, https://doi.org/10.1016/j.jad.2014.10.010.

[5] R. Uher, Genes, environment, and individual differences in responding to treatment for depression, Harvard Rev. Psychiatry 19 (3) (2011) 109–124, https://doi.org/10.3109/10673229.2011.586551.

[6] U. Ozomaro, C. Wahlestedt, C.B. Nemeroff, Personalized medicine in psychiatry: problems and promises, BMC Med. 11 (1) (2013) 132, , https://doi.org/10.1186/1741-7015-11-132.

[7] T.R. Insel, Translating scientific opportunity into public health impact: a strategic plan for research on mental illness, Arch. Gen. Psychiatry 66 (2) (2009) 128–133, https://doi.org/10.1001/archgenpsychiatry.2008.540.

[8] G.E. Simon, R.H. Perlis, Personalized medicine for depression: can we match patients with treatments? Am. J. Psychiatr. 167 (12) (2010) 1445–1455, https://doi.org/10.1176/appi.ajp.2010.09111680.

[9] I. Elfeddali, C.M. van der Feltz-Cornelis, J. van Os, S. Knappe, E. Vieta, H.-U. Wittchen, C. Obradors-Tarragó, J.M. Haro, Horizon 2020 priorities in clinical mental health research: results of a consensus-based ROAMER expert survey, Int. J. Environ. Res. Public Health 11 (10) (2014) 10915–10939, https://doi.org/10.3390/ijerph111010915.

[10] A.J. Fisher, J.F. Boswell, Enhancing the personalization of psychotherapy with dynamic assessment and modeling, Assessment 23 (4) (2016) 496–506, https://doi.org/10.1177/1073191116638735.

[11] T.J. Trull, U. Ebner-Priemer, Ambulatory assessment, Annu. Rev. Clin. Psychol. 9 (1) (2013) 151–176, https://doi.org/10.1146/annurev-clinpsy-050212-185510.

[12] D.H. Barlow, M.K. Nock, Why can’t we be more idiographic in our research? Perspect. Psychol. Sci. 4 (1) (2009) 19–21, https://doi.org/10.1111/j.1745-6924.2009.01088.x.

[13] E.L. Hamaker, Why researchers should think ‘within-person’: A paradigmatic rationale, Handbook of Research Methods for Studying Daily Life, Guilford Press, 2012, pp. 43–61.

[14] P.C.M. Molenaar, A manifesto on psychology as idiographic science: bringing the person back into scientific psychology, this time forever, Measurement Interdiscip. Res. Perspect. 2 (4) (2004) 201–218, https://doi.org/10.1207/s15366359mea0204_1.

[15] W. Tschacher, F. Ramseyer, Modeling psychotherapy process by time-series panel analysis (TSPA), Psychother. Res. 19 (4–5) (2009) 469–481, https://doi.org/10.1080/10503300802654496.

[16] A.G.C. Wright, J. Zimmermann, Applied ambulatory assessment: integrating idiographic and nomothetic principles of measurement, Psychol. Assess. 31 (12) (2019) 1467–1480, https://doi.org/10.1037/pas0000685.

[17] I. Myin-Germeys, Z. Kasanova, T. Vaessen, H. Vachon, O. Kirtley, W. Viechtbauer, U. Reininghaus, Experience sampling methodology in mental health research: new insights and technical developments, World Psychiatry 17 (2) (2018) 123–132, https://doi.org/10.1002/wps.20513.

[18] A.G.C. Wright, W.C. Woods, Personalized models of psychopathology, Annu. Rev. Clin. Psychol. 16 (1) (2020) 49–74, https://doi.org/10.1146/annurev-clinpsy-102419-125032.

[19] M. Csikszentmihalyi, R. Larson, Validity and reliability of the experience-sampling method, J. Nerv. Ment. Dis. 175 (9) (1987) 526–536.

[20] I. Myin-Germeys, M. Oorschot, D. Collip, J. Lataster, P. Delespaul, J. van Os, Experience sampling research in psychopathology: opening the black box of daily life, Psychol. Med. 39 (9) (2009) 1533–1547, https://doi.org/10.1017/S0033291708004947.

[21] A.G.C. Wright, C.J. Hopwood, Advancing the assessment of dynamic psychological processes, Assessment 23 (4) (2016) 399–403, https://doi.org/10.1177/1073191116654760.

[22] N. Stavrakakis, S.H. Booij, A.M. Roest, P. de Jonge, A.J. Oldehinkel, E.H. Bos, Temporal dynamics of physical activity and affect in depressed and nondepressed individuals, Health Psychol. 34 (2015) 1268–1277, https://doi.org/10.1037/hea0000303.

[23] J.E. Palmier-Claus, I. Myin-Germeys, E. Barkus, L. Bentley, A. Udachina, P.A.E.G. Delespaul, S.W. Lewis, G. Dunn, Experience sampling research in individuals with mental illness: reflections and guidance, Acta Psychiatr. Scand. 123 (1) (2011) 12–20, https://doi.org/10.1111/j.1600-0447.2010.01596.x.

[24] J. van Os, P. Delespaul, J. Wigman, I. Myin-Germeys, M. Wichers, Beyond DSM and ICD: introducing ‘precision diagnosis’ for psychiatry using momentary assessment technology, World Psychiatry 12 (2) (2013) 113–117, https://doi.org/10.1002/wps.20046.

[25] M. Wichers, J.A. Hartmann, I.M.A. Kramer, C. Lothmann, F. Peeters, L. van Bemmel, I. Myin-Germeys, P. Delespaul, J. van Os, C.J.P. Simons, Translating assessments of the film of daily life into person-tailored feedback interventions in depression, Acta Psychiatr. Scand. 123 (5) (2011) 402–403, https://doi.org/10.1111/j.1600-0447.2011.01684.x.

[26] J.A. Bastiaansen, M. Meurs, R. Stelwagen, L. Wunderink, R.A. Schoevers, M. Wichers, A.J. Oldehinkel, Self-monitoring and personalized feedback based on the experiencing sampling method as a tool to boost depression treatment: A protocol of a pragmatic randomized controlled trial (ZELF-i), *BMC Psychiatry 18* (1) (2018) 276, https://doi.org/10.1186/s12888-018-1847-z.

[27] M.N. Burns, M. Begale, J. Duffecy, D. Gergle, C.J. Karr, E. Giangrande, D.C. Mohr, Harnessing context sensing to develop a mobile intervention for depression, J. Med. Internet Res. 13 (3) (2011) e55, , https://doi.org/10.2196/jmir.1838.

[28] S.D. Kauer, S.C. Reid, A.H.D. Crooke, A. Khor, S.J.C. Hearps, A.F. Jorm, L. Sanci, G. Patton, Self-monitoring using mobile phones in the early stages of adolescent depression: randomized controlled trial, J. Med. Internet Res. 14 (3) (2012) e67, https://doi.org/10.2196/jmir.1858.

[29] I. Kramer, C.J.P. Simons, J.A. Hartmann, C. Menne-Lothmann, W. Viechtbauer, F. Peeters, K. Schruers, A.L. van Bemmel, I. Myin-Germeys, P. Delespaul, J. van Os, M. Wichers, A therapeutic application of the experience sampling method in the treatment of depression: a randomized controlled trial, World Psychiatry 13 (1) (2014) 68–77, https://doi.org/10.1002/wps.20090.

[30] I. Myin-Germeys, A. Klippel, H. Steinhart, U. Reininghaus, Ecological momentary interventions in psychiatry, Curr. Opin. Psychiatry 29 (4) (2016) 258–263, https://doi.org/10.1097/YCO.0000000000000255.

[31] S.J.W. Verhagen, J.A. Berben, C. Leue, A. Marsman, P.A.E.G. Delespaul, J. van Os, R. Lousberg, Demonstrating the reliability of transdiagnostic mHealth routine outcome monitoring in mental health services using experience sampling technology, PLoS One 12 (10) (2017) e0186294, , https://doi.org/10.1371/journal.pone.0186294.

[32] W.J. Korotitsch, R.O. Nelson-Gray, An overview of self-monitoring research in assessment and treatment, Psychol. Assess. 11 (4) (1999) 415–425, https://doi.org/10.1037/1040-3590.11.4.415.

[33] R. Kroeze, D.C. van der Veen, M.N. Servaas, J.A. Bastiaansen, R.C. Oude Voshaar, D. Borsboom, H.G. Ruhe, R.A. Schoevers, H. Riese, Personalized feedback on symptom dynamics of psychopathology: a proof-of-principle study, J. Person-Orient. Res. 3 (1) (2017) 1–11, https://doi.org/10.17505/jpor.2017.01.

[34] A.J. Fisher, H.G. Bosley, K.C. Fernandez, J. Reeves, A. Diamond, P.D. Soyster, J. Barkin, Open Trial of a Personalized Modular Treatment for Mood and Anxiety, (2019), https://doi.org/10.31234/osf.io/8ezhm.

[35] A.J. Fisher, J.W. Reeves, G. Lawyer, J.D. Medaglia, J.A. Rubel, Exploring the idiographic dynamics of mood and anxiety via network analysis, J. Abnorm. Psychol. 126 (8) (2017) 1044–1056, https://doi.org/10.1037/abn0000311.

[36] R. Silberzahn, E.L. Uhlmann, D.P. Martin, P. Anselmi, F. Aust, E. Awtrey, Š. Bahník, F. Bai, C. Bannard, E. Bonnier, R. Carlsson, F. Cheung, G. Christensen, R. Clay, M.A. Craig, A.D. Rosa, L. Dam, M.H. Evans, I.F. Cervantes, ... M. Vianello, Many analysts, one data set: Making transparent how variations in analytic choices affect results, Advances in Methods and Practices in Psychological Science 1 (3) (2018) 337–356, https://doi.org/10.1177/2515245917747646.

[37] M. Hamilton, A rating scale for depression, J. Neurol. Neurosurg. Psychiatry 23 (1) (1960) 56–62.

[38] M. Hamilton, The assessment of anxiety states by rating, Br. J. Med. Psychol. 32 (1) (1959) 50–55, https://doi.org/10.1111/j.2044-8341.1959.tb00467.x.

[39] S. Boker, M. Neale, H. Maes, M. Wilde, M. Spiegel, T. Brick, J. Spies, R. Estabrook, S. Kenny, T. Bates, P. Mehta, J. Fox, OpenMx: an open source extended structural equation modeling framework, Psychometrika 76 (2) (2011) 306–317, https://doi.org/10.1007/s11336-010-9200-6.

[40] J.J. Faraway, Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models, vol. 24, Chapman & Hall/CRC, 2006.

[41] T. Asparouhov, E.L. Hamaker, B. Muthén, Dynamic structural equation models, Struct. Equ. Model. Multidiscip. J. 25 (3) (2018) 359–388, https://doi.org/10.1080/10705511.2017.1406803.

[42] C. Chatfield, The Analysis of Time Series: An Introduction (Fifth Edition), Chapman and Hall/CRC, 1996.

[43] H. Lütkepohl, New Introduction to Multiple Time Series Analysis, Springer-Verlag, 2005.

[44] D. Borsboom, A network theory of mental disorders, World Psychiatry 16 (1) (2017) 5–13, https://doi.org/10.1002/wps.20375.

[45] D. Borsboom, A.O.J. Cramer, Network analysis: an integrative approach to the structure of psychopathology, Annu. Rev. Clin. Psychol. 9 (1) (2013) 91–121, https://doi.org/10.1146/annurev-clinpsy-050212-185608.

[46] L.F. Bringmann, M.I. Eronen, Don’t blame the model: reconsidering the network approach to psychopathology, Psychol. Rev. 125 (4) (2018) 606–615, https://doi.org/10.1037/rev0000108.

[47] A.O.J. Cramer, L.J. Waldorp, H.L.J. van der Maas, D. Borsboom, Comorbidity: a network perspective, Behav. Brain Sci. 33 (2–3) (2010) 137–150 discussion 150-193 https://doi.org/10.1017/S0140525X09991567.
[48] G. Costantini, S. Epskamp, D. Borsboom, M. Perugini, R. Mõttus, L.J. Waldorp, A.O.J. Cramer, State of the aRt personality research: a tutorial on network analysis of personality data in R, J. Res. Pers. 54 (2015) 13–29, https://doi.org/10.1016/j.jrp.2014.07.003.
[49] R. Tibshirani, Regression shrinkage and selection via the lasso: a retrospective, J. Royal Stat. Soc. Ser. B (Stat. Methodol.) 73 (3) (2011) 273–282, https://doi.org/10.1111/j.1467-9868.2011.00771.x.
[50] L.F. Bringmann, T. Elmer, S. Epskamp, R.W. Krause, D. Schoch, M. Wichers, J.T.W. Wigman, E. Snippe, What do centrality measures measure in psychological networks? J. Abnomr. Psychol. 128 (8) (2019) 892–903, https://doi.org/10.1037/abn0000446.
[51] L.F. Bringmann, N. Vissers, M. Wichers, N. Geschwind, P. Kuppens, F. Peeters, D. Borsboom, F. Tuerlinckx, A network approach to psychopathology: new insights into clinical longitudinal data, PLoS One 8 (4) (2013) e60188, , https://doi.org/10.1371/journal.pone.0060188.
[52] A.O.J. Cramer, D. Borsboom, S.H. Aggen, K.S. Kendler, The pathoplasticity of dysphoric episodes: differential impact of stressful life events on the pattern of depressive symptom inter-correlations, Psychol. Med. 42 (5) (2012) 957–965, https://doi.org/10.1017/S003329171100211X.
[53] W. Lutz, B. Schwartz, S.G. Hofmann, A.J. Fisher, K. Husen, J.A. Rubel, Using network analysis for the prediction of treatment dropout in patients with mood and anxiety disorders: a methodological proof-of-concept study, Sci. Rep. 8 (1) (2018) 7819, , https://doi.org/10.1038/s41598-018-25953-0.
[54] T.A. Walls, J.L. Schafer, Models for Intensive Longitudinal Data, Oxford University Press, 2006.
[55] S. Boker, M. Neale, J. Rausch, Latent differential equation modeling with multivariate multi-occasion indicators, Recent Developments on Structural Equation Models, Springer, 2004, pp. 151–174.
[56] P.D. Gilbert, E. Meijer, Time Series Factor Analysis with an Application to Measuring Money (No. 05F10; p. 37), University of Groningen, SOM Research School, 2005, http://som.eldoc.ub.rug.nl/reports/themeF/2005/05F10/.
[57] M. Basseville, I.V. Nikiforov, Detection of Abrupt Changes: Theory and Application, Prentice-Hall, Inc, 1993.
[58] K. Bulteel, F. Tuerlinckx, A. Brose, E. Ceulemans, Using raw VAR regression coefficients to build networks can be misleading, Multivar. Behav. Res. 51 (2–3) (2016) 330–344, https://doi.org/10.1080/00273171.2016.1150151.
[59] L.L. Doove, T.F. Wilderjans, A. Calcagnì, I. Van Mechelen, Deriving optimal data-analytic regimes from benchmarking studies, Comput. Stat. Data Anal. 107 (2017) 81–91, https://doi.org/10.1016/j.csda.2016.10.016.
[60] E.I. Fried, A.O.J. Cramer, Moving forward: challenges and directions for psychopathological network theory and methodology, Perspect. Psychol. Sci. 12 (6) (2017) 999–1020, https://doi.org/10.1177/1745691617705892.
[61] S. Steegen, F. Tuerlinckx, A. Gelman, W. Vanpaemel, Increasing transparency through a multiverse analysis, Perspect. Psychol. Sci. 11 (5) (2016) 702–712 (10.1177%2F1745691616658637).
[62] J.P. Simmons, L.D. Nelson, U. Simonsohn, False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant, Psychol. Sci. 22 (11) (2011) 1359–1366, https://doi.org/10.1177/0956797611417632.
[63] A. Gelman, E. Loken, The Garden of Forking Paths: Why Multiple Comparisons can be a Problem, Even when there is no “Fishing Expedition” or “p-Hacking” and the Research Hypothesis was Posited ahead of Time, (2013), pp. 1–17.
[64] M.R. Munafò, B.A. Nosek, D.V.M. Bishop, K.S. Button, C.D. Chambers, N. Percie du Sert, U. Simonsohn, E.-J. Wagenmakers, J.J. Ware, J.P.A. Ioannidis, A manifesto for reproducible science, Nat. Hum. Behav. 1 (1) (2017) 0021, https://doi.org/10.1038/s41562-016-0021.
[65] O.J. Kirtley, G. Lafit, R. Achterhof, A.P. Hiekkaranta, I. Myin-Germeys, Making the Black Box Transparent: A Pre-Registration Template for Studies Using Experience Sampling Methods (ESM), Advances in Methods and Practices in Psychological Science (2019), https://doi.org/10.31234/osf.io/seyq7.
[66] T.J. Trull, U.W. Ebner-Priemer, Ambulatory assessment in psychopathology research: a review of recommended reporting guidelines and current practices, J. Abnorm. Psychol. 129 (1) (2020) 56–63, https://doi.org/10.1037/abn0000473.