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

Preprint · March 2019

DOI: 10.31234/osf.io/c8vp7

32 authors, including:

Jojanneke A Bastiaansen
University of Groningen
42 PUBLICATIONS **896** CITATIONS
SEE PROFILE

Yoram K. Kunkels
University of Groningen
3 PUBLICATIONS **1,524** CITATIONS
SEE PROFILE

Frank Blaauw
University of Groningen
17 PUBLICATIONS **101** CITATIONS
SEE PROFILE

Eva Ceulemans
KU Leuven
143 PUBLICATIONS **1,642** CITATIONS
SEE PROFILE

Time to get personal?

The impact of researchers' choices on the selection of treatment targets using the experience sampling methodology

Authors

Jojanneke A. Bastiaansen [1,2], Yoram K. Kunkels [1], Frank J. Blaauw [3,4], Steven M. Boker [5], Eva Ceulemans [6], Meng Chen [7], Sy-Miin Chow [7], Peter de Jonge [3], Ando C. Emerencia [3], Sacha Epskamp [8], Aaron J. Fisher [9], Ellen L. Hamaker [10], Peter Kuppens [6], Wolfgang Lutz [11], M. Joseph Meyer [5], Robert Moulder [5], Zita Oravecz [7], Harriëtte Riese [1], Julian Rubel [11], Oisín Ryan [10], Michelle N. Servaas [1], Gustav Sjobeck [5], Evelien Snippe [1], Timothy J. Trull [12], Wolfgang Tschacher [13], Date C. van der Veen [1], Marieke Wichers [1], Phillip K. Wood [12], William C. Woods [14], Aidan G.C. Wright [14], Casper J. Albers [3], Laura F. Bringmann [1,3]

Affiliations

[1] Interdisciplinary Center Psychopathology and Emotion regulation, Department of Psychiatry, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands; [2] Department of Education and Research, Friesland Mental Health Care Services, Leeuwarden, The Netherlands; [3] Department of Psychology, University of Groningen, Groningen, The Netherlands; [4] Distributed Systems group, Faculty of Science and Engineering, University of Groningen, Groningen, The Netherlands; [5] Department of Psychology, University of Virginia, Charlottesville, USA; [6] Faculty of Psychology and Educational Sciences, University of Leuven, Leuven, Belgium; [7] Department of Human Development and Family Studies, Pennsylvania State University, State College, USA; [8] Department of Social and Behavioural Sciences, University of Amsterdam, Amsterdam, Netherlands; [9] Department of Psychology, University of California Berkeley, Berkeley, USA; [10] Department of Methodology and Statistics, Faculty of Social and Behavioural Sciences, Utrecht University, Utrecht, The Netherlands; [11] Department of Psychology, University of Trier, Trier, Germany; [12] Department of Psychological Sciences, University of Missouri, Columbia, USA; [13] University Hospital of Psychiatry and Psychotherapy, University of Bern, Bern, Switzerland; [14] Department of Psychology, University of Pittsburgh, Pittsburgh, USA

Corresponding author

L.F. Bringmann, 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: l.f.bringmann@rug.nl

Abstract

One of the promises of the experience sampling methodology (ESM) is that it could be used to identify relevant targets for treatment, based on a statistical analysis of an individual's emotions, cognitions and behaviors in everyday-life. A requisite for clinical implementation is that outcomes of person-centered analyses are not wholly contingent on the researcher performing them. To evaluate how much researchers vary in their analytical approach and to what degree outcomes vary based on analytical choices, 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. The dataset was from a 25-year-old male with a primary diagnosis of major depressive disorder and comorbid generalized anxiety disorder, who completed momentary assessments related to depression and anxiety psychopathology prior to psychotherapy. Variation was evident at different stages of the analysis, from preprocessing steps (e.g., variable selection, clustering, handling of missing data) to the type of statistics. Most teams did include a type of vector autoregressive model, which examines relations between variables (e.g., symptoms) over time. Although most teams were confident their selected targets would provide useful information to the clinician, not one advice was similar: both the number (0-16) and nature of selected targets varied widely. 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 methodological issues that need to be addressed in moving toward clinical implementation.

Research proposal, data and materials: osf.io/h3djy/

Keywords

experience sampling methodology; ecological momentary assessment; depression; anxiety; psychopathology; personalized medicine; intervention selection; crowdsourcing; time-series analysis; many labs; psychological networks

Introduction

Clinicians rely on evidence-based guidelines for the assessment and treatment of psychiatric disorders such as major depressive disorder (MDD; American Psychiatric Association, 2010; National Institute for Health and Care Excellence, 2009). 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 (Lamiell, 1998). 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 (Fried & Nesse, 2015). In addition, patients vary widely in their response to treatments (Uher, 2011).

By identifying individual characteristics that determine disease susceptibility as well as treatment response (Ozomaro, Wahlestedt, & Nemeroff, 2013), personalized medicine promises to move from treatments that are effective on average towards identifying the best treatment for any individual (Insel, 2009; Simon & Perlis, 2010). However, if we were to actually tailor treatments to the individual patient (Elfedalli et al., 2014), we need to look beyond differences between individuals and additionally examine processes within the individual (Fisher & Boswell, 2016; Trull & Ebner-Priemer, 2013). Thus, a more person-centered (i.e., idiographic) research approach is required to complement our current nomothetic focus (Barlow & Nock, 2009; Hamaker, 2012; Molenaar, 2004; Ramseyer et al., 2014; Wright & Zimmermann, in press), and as such facilitate personalized medicine.

The experience sampling methodology (ESM) has been positioned as one of the best opportunities for personalized medicine in psychiatry (Myin-Germeys et al., 2018; Wright & Zimmermann, in press). ESM 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 (Larson & Csikszentmihalyi, 1983). ESM studies have shown that many symptoms of patients with severe psychiatric disorders show person-specific, meaningful and widespread variation over time (Myin-Germeys et al., 2009; Wright & Hopwood, 2016). Stavrakakis and colleagues (2015), 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-centered analyses based on ESM data could have great potential for use in clinical practice, because they could provide personalized and contextualized feedback to patients and clinicians (Van Os et al., 2013; Palmier-Claus et al., 2011; Wichers et al., 2011).

This idea has been put into practice by experience sampling intervention (ESI) studies for, amongst others, individuals with depressive symptoms (e.g., Burns et al., 2011; Kauer et al., 2012; Kramer et al., 2014; Bastiaansen et al., 2018). 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 (Myin-Germeys et al., 2016).

Self-monitoring data may also be used more specifically to "identify particular targets for treatment and help decide which aspects of treatment may be most beneficial to a particular patient" (Korotitsch & Nelson-Gray, 1999). In a proof-of-principle study, Kroeze and colleagues (2017) 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 (Fisher & Boswell, 2016; Fisher, Reeves, Lawyer, Medaglia, & Rubel, 2017; Fisher et al., in press), 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" (Fisher & Boswell, 2016). To this end, they examined temporal relationships between symptoms.

Different analytic approaches might, however, lead to different outcomes, and reveal different conceptualizations of 'the most important symptom that needs to be targeted first'. Data preprocessing and analysis typically comprise many steps that involve choices between often several reasonable (and unreasonable) options, which can induce many researcher degrees of freedom (Simmons, Nelson, & Simonsohn, 2011). It is unclear how much diversity in analytical approaches there is in the ESM field and to what degree outcomes vary based on analytical choices. Knowledge on the robustness of outcomes over individual, subjective, choices is vital if we want to be able to take these methods from the realm of the researcher and present them as a tool to patients and clinicians.

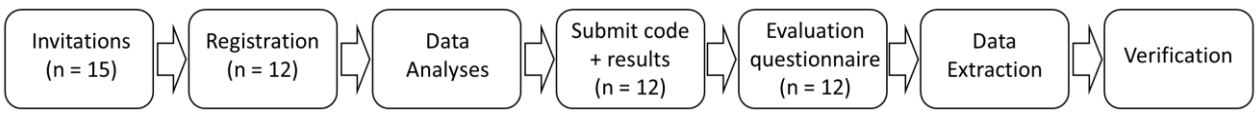
In this study, we will use a crowdsourcing data analysis strategy (Silberzahn et al., 2018), in which several expert research teams from around the world are invited to simultaneously investigate the same clinically relevant research question for one single dataset: "What symptom(s) would you advise the treating clinician to target subsequent treatment on, based on a person-centered 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.

Methods

Data analysts

The project group (JAB, YKK, CJA, LFB) wrote a project description (available at our Open Science Framework (OSF) page: https://osf.io/h3djy/), 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 selected by the project group for their expertise in ESM and/or the statistical analysis of idiographic data (for a flowchart of the study, see Figure 1). 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. Our aim was to work in a transparent manner and make the (anonymized) analytical approaches per research team publicly available through the OSF. Teams were asked to indicate whether they objected to this procedure.

Figure 1 Flowchart of the study



Dataset

The data were drawn from a multiphase personalized psychotherapy study (Fisher & Boswell, 2016; Fisher, Reeves, Lawyer, Medaglia, & Rubel, 2017; Fisher et al., in press). 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 at a random time within each of the four 3-hour blocks (but note that invitations for the surveys were spaced at least 30 minutes 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'. Each survey included 23 momentary[1] items related to depression and anxiety psychopathology (e.g., felt down or depressed, felt a loss of interest or pleasure, felt frightened or afraid). 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 (Hamilton, 1960) and the Hamilton Rating Scale for Anxiety (Hamilton, 1959) were 16 and 15, respectively. The full item list and dataset are available at our OSF page.

Procedure

For a flowchart of the study, see Figure 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 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.

Results

Data analysts

Twelve independent[2] 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 =

[1] We use the term momentary, because questions pertained to experiences over a short period of time (3-hour blocks).
[2] Two research teams were from different departments of the same university, but worked independently nonetheless.

2). Teams included as highest academic rank a Full Professor (n = 7), Associate Professor (n = 2), or Assistant Professor (n = 3). All teams 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%) 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. Characteristics of the analysts can be found in Supplementary Figure 1.

Analytical approaches

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; Boker et al., 2011). 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 replicable results. Eleven out of twelve teams agreed with our open science statement. The code for the analyses of these research teams is available at our OSF page. We examined variation between the research teams at different stages of the analysis, from preprocessing steps (e.g., variable selection, clustering, handling of missing data) to the type of statistical analyses.

Variable selection

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[3] item (I experienced muscle tension), because they thought it might have changed meaning from a negative connotation (stress-related tension) to a more positive connotation (activity-related muscle tension) during the ESM study: at first, tension correlated negatively with positive affect items (i.e., tension went down when positive affect items went up), but at the end, tension correlated positively with positive affect items. 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, because they found that the mean levels of these items changed significantly over time (i.e., a violation of the stationarity assumption for time series analysis). 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 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

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

also used the sleep items in a separate analysis to examine relationships between sleep problems and affective symptoms.

Clustering

Three teams only used individual items in their statistical analyses. The other nine teams grouped the items (at least some of) into clusters[4] 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.

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. Figure 2 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 Figure 2): 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 Figure 2): 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 Figure 2): 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 Figure 2): 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 Figure 2 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 labelled 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 Figure 2.

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

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[1]	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	- [2]	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)

PCA = Principal Component Analysis, FA = Factor Analysis, DSEM = Dynamic Structural Equation Model, [1] 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. [2] 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).

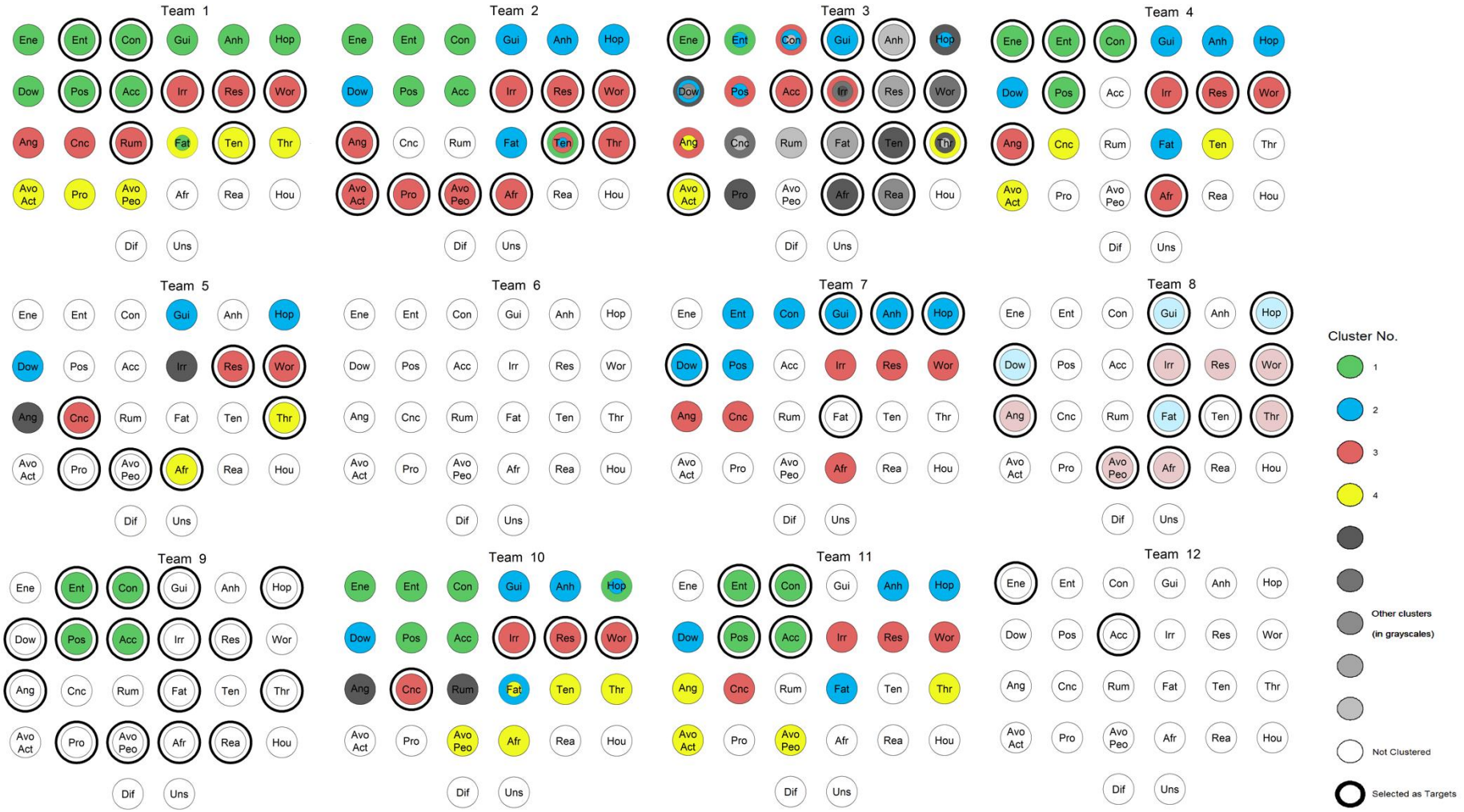


Figure 2 Clustering and target selection per research team

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 labelled 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. 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 = procrast, Avo Peo = avoid people, Afr = afraid, Rea = reassure, Hou = hours, Dif = difficult, Uns = unsatisfy.

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; Faraway, 2006) or Bayesian techniques (e.g., a Kalman filter; Asparouhov, Hamaker, & Muthén, 2018). 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 hours 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. The latter are outlined in Table 2 and discussed below.

Statistical analyses

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; Chatfield, 2003, Lütkepohl, 2005). 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 hours 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 modeling 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 based on imputation techniques.

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[5]). The one team (no. 7) that did not use a VAR model studied contemporaneous

[5] Note that a lag 0 regression does not have to represent a contemporaneous effect. Team 6, for instance, argued that some variables with the same time stamp actually refer to different times (i.e., sleep during preceding night and mood during the day) and associations should hence be seen as lagged in nature.

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.

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 (Borsboom, 2017; Borsboom & Cramer, 2013; Bringmann & Eronen, 2018; Cramer, Waldorp, van der Maas & Borsboom, 2010). To reduce data dimensionality these teams used data-driven techniques that reduce the number of parameters (Costantini et al. 2015; Tibshirani, 2011). Two of these teams (nos. 7 and 9) additionally performed a centrality analysis, which aims to identify the item(s) that had the overall highest influence on other items in a network (Cramer, Borsboom, Aggen, & Kendler, 2012; Bringmann et al., 2013; Lutz et al., 2018).

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; Walls & Schafer, 2006). 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 (Boker, Neale, & Rausch, 2004; Gilbert & Meijer, 2005). 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 (Basseville & Nikiforov, 1993).

Table 2 Statistical analyses

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

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)

[1] information eventually used for target selection. ^ only one cluster amidst a series of individual variables † 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).

Intervention targets

Target selection rationale

Table 2 shows that teams based their target selection on varying sources of information.

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.

Selected targets

We considered an item as a potential target if it had been included by the team in at least one statistical analysis (including clustering). Note, however, that teams could have included different subsets of items in different analyses. 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 Figure 2. Table 4 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.

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, among other targets. In contrast, none of the teams with a cluster 2 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 modeling. 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).

Table 3 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	†
9	23	16 ‡	70	Cluster 1 + 12 individual items
10	23	4	17	Cluster 3
11	23	4	17	Cluster 1
12	26	2	8	

Notes: Every item is a potential target if it has been included by a team in at least one statistical analysis (including clustering). 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. † 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. ‡ This team suggested to target symptoms and behaviors across 4 consecutive phases.

Table 4 Target selection frequency per item

Irritable	7	Angry	4	Positive	4	Anhedonia	2	Hours	0
Restless	7	Avoid people	4	Tension	4	Avoid activities	2	Unsatisfy	0
Worried	7	Content	4	Energetic	3	Concentrate	2		
Afraid	6	Enthusiastic	4	Down	3	Reassure	2		
Accepted	5	Fatigue	4	Hopeless	3	Ruminate	1		
Threatened	5	Guilty	4	Procrast	3	Difficult	0		

Note: This table shows the number of times an item was reported by a team as a potential target for intervention.

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).

Team evaluations

Responses to the closed evaluation questions are provided in Supplementary Table 1. Eight of the 12 teams also provided additional comments in the open fields of the questionnaire. 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-centered 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).

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 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.

Variation in analytical approaches: There were 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 some 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).

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.

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 (American Psychiatric Association, 2010; National Institute for Health and Care Excellence, 2009).

Our project highlights several important issues that need to be addressed in moving ESM toward clinical implementation. 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 (Bulteel et al., 2016). 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 what the most relevant information is 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?

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 data-driven approaches. In order to tailor interventions to the individual it might be more fruitful to look beyond 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 purely data-driven, but done in a dialogue between clinician and patient (for an example see Kroeze et al., 2017). Other researchers have argued that person-centered 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 (Wright & Zimmermann, in press).

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.; Doove et al., 2017). 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.

Fourth, our study underscores the need for transparency in science. 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 (Steegen et al., 2016). 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 (Simmons, Nelson, & Simonsohn, 2011). The route one takes in this 'garden of forking paths' (Gelman & Loken, 2013) 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 (Munafò et al., 2017).

This was the first study that assessed the robustness of outcomes over subjective analytical choices for one individual time-series ESM dataset. We found that different research teams chose different analytical approaches and that outcomes – and hence, advice to the clinician on treatment targets- varied widely. This study highlights conceptual and methodological issues that need to be addressed in moving person-centered analyses based on ESM data toward clinical implementation. The translation to clinical practice requires a collaborative effort between researchers, patients, and clinicians.

References

- American Psychiatric Association. (2010). Practice guideline for the treatment of patients with major depressive disorder (third edition). *The American Journal of Psychiatry, 167*(10), 1–152.
- Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic structural equation models. *Structural Equation Modeling, 25*, 359–388. https://doi.org/10.1080/10705511.2017.1406803
- Barlow, D. H., & Nock, M. K. (2009). Why can't we be more idiographic in our research? *Perspectives on Psychological Science*, 4(1), 19-21. https://doi.org/10.1111/j.1745-6924.2009.01088.x
- Basseville, M., & Nikiforov, I. V. (1993). *Detection of abrupt changes: theory and application*. Englewood Cliffs: Prentice Hall. Retrieved from http://www.irisa.fr/sisthem/kniga/
- Bastiaansen, J. A., Meurs, M., Stelwagen, R., Wunderink, L., Schoevers, R. A., Wichers, M., & Oldehinkel, A. J. (2018). 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), 276. https://doi.org/10.1186/s12888-018-1847-z
- Boker, S., Neale, M., Maes, H., Wilde, M., Spiegel, M., Brick, T., … Fox, J. (2011). OpenMx: An open source extended structural equation modeling framework. *Psychometrika, 76*(2), 306–317. https://doi.org/10.1007/s11336-010-9200-6

- Boker, S., Neale, M., & Rausch, J. (2004). Latent differential equation modeling with multivariate multi-occasion indicators. In *Recent Developments on Structural Equation Models* (pp. 151–174). Dordrecht: Springer. Retrieved from https://link.springer.com/chapter/10.1007/978-1-4020-1958-6_9
- Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry*, *16*(1), 5–13. https://doi.org/10.1002/wps.20375
- Borsboom, D., & Cramer, A. (2013). Network analysis: an integrative approach to the structure of psychopathology. *Annual Review Of Clinical Psychology*, *9*(1), 91-121. https://doi.org/10.1146/annurev-clinpsy-050212-185608
- Bringmann, L., & Eronen, M. (2018). Don't blame the model: reconsidering the network approach to psychopathology. *Psychological Review*, *125*(4), 606-615. https://doi.org/10.1037/rev0000108
- Bringmann, L. F., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., … Tuerlinckx, F. (2013). A network approach to psychopathology: new insights into clinical longitudinal data. *PloS One*, *8*(4), e60188. https://doi.org/10.1371/journal.pone.0060188
- Bulteel, K., Tuerlinckx, F., Brose, A., & Ceulemans, E. (2016). Using raw VAR regression coefficients to build networks can be misleading. *Multivariate Behavioral Research*, *51*(2–3), 330–344. https://doi.org/10.1080/00273171.2016.1150151
- Burns, M. N., Begale, M., Duffecy, J., Gergle, D., Karr, C. J., Giangrande, E., & Mohr, D. C. (2011). Harnessing context sensing to develop a mobile intervention for depression. *Journal of Medical Internet Research*, *13*(3), e55. https://doi.org/10.2196/jmir.1838
- Chatfield, C. (1996). *The analysis of time series: An introduction* (5th ed.). Boca Raton, FL: Chapman and Hall/CRC.
- Costantini, G., Epskamp, S., Borsboom, D., Perugini, M., Mõttus, R., Waldorp, L. J., & Cramer, A. O. J. (2015). State of the aRt personality research: A tutorial on network analysis of personality data in R. *Journal of Research in Personality*, *54*, 13–29. https://doi.org/10.1016/j.jrp.2014.07.003
- Cramer, A. O., Borsboom, D., Aggen, S. H., & Kendler, K. S. (2012). The pathoplasticity of dysphoric episodes: differential impact of stressful life events on the pattern of depressive symptom inter-correlations. *Psychological medicine, 42(5)*, 957-965.
- Cramer, A. O., Waldorp, L. J., van der Maas, H. L., & Borsboom, D. (2010). Comorbidity: A network perspective. *Behavioral and Brain Sciences, 33*, 137-150.
- Doove, L. L., Wilderjans, T. F., Calcagnì, A., & Van Mechelen, I. (2017). Deriving optimal data-analytic regimes from benchmarking studies. Computational Statistics & Data Analysis, 107, 81–91. https://doi.org/10.1016/j.csda.2016.10.016
- Elfeddali, I., van der Feltz-Cornelis, C. M., van Os, J., Knappe, S., Vieta, E., Wittchen, H.-U., … Haro, J. M. (2014). Horizon 2020 priorities in clinical mental health research: results of a consensus-based ROAMER expert survey. *International Journal of Environmental Research and Public Health*, *11*(10), 10915–10939. https://doi.org/10.3390/ijerph111010915
- Faraway, J. J. (2006). *Extending the linear model with R: Generalized linear, mixed and nonparametric regression models.* Boca Raton, FL: Chapman and Hall/CRC.
- Fisher, A. J., Bosley, H. G., Fernandez, K. C., Reeves, J., Diamond, A., Soyster, P. D., & Barkin, J. (in press). Open trial of a personalized modular treatment for mood and anxiety. *Behaviour Research and Therapy*. https://doi.org/10.31234/osf.io/8ezhm
- Fisher, A. J., & Boswell, J. F. (2016). Enhancing the personalization of psychotherapy with dynamic assessment and modeling. *Assessment*, *23*(4), 496–506. https://doi.org/10.1177/1073191116638735
- Fisher, A. J., Reeves, J. W., Lawyer, G., Medaglia, J. D., & Rubel, J. A. (2017). Exploring the idiographic dynamics of mood and anxiety via network analysis. *Journal of*

Abnormal Psychology, *126*(8), 1044–1056. https://doi.org/10.1037/abn0000311
- Fried, E. I., & Nesse, R. M. (2015). Depression is not a consistent syndrome: An investigation of unique symptom patterns in the STAR*D study. *Journal of Affective Disorders*, *172*, 96–102. https://doi.org/10.1016/j.jad.2014.10.010
- Gelman, A., & Loken, E. (2013). 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. Retrieved from http://www.stat.columbia.edu/~gelman/research/unpublished/
- Gilbert, P. D., & Meijer, E. (2005). Time series factor analysis with an application to measuring money (Tech. Rep. No. 05F10). University of Groningen, SOM Research School. Retrieved from http://som.eldoc.ub.rug.nl/reports/themeF/2005/05F10/
- Hamaker, E. L. (2012). Why researchers should think 'within-person': A paradigmatic rationale. In *Handbook of research methods for studying daily life* (pp. 43–61). New York, NY: Guilford Press.
- Hamilton, M. (1959). The assessment of anxiety states by rating. *British Journal of Medical Psychology*, *32*(1), 50–55. https://doi.org/10.1111/j.2044-8341.1959.tb00467.x
- Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery & Psychiatry, 23(1), 56-62.* https://doi.org/10.1136/jnnp.23.1.56
- Insel, T. R. (2009). Translating scientific opportunity into public health impact: a strategic plan for research on mental illness. *Archives of General Psychiatry*, *66*(2), 128–133. https://doi.org/10.1001/archgenpsychiatry.2008.540
- Kauer, S. D., Reid, S. C., Crooke, A. H. D., Khor, A., Hearps, S. J. C., Jorm, A. F., … Patton, G. (2012). Self-monitoring using mobile phones in the early stages of adolescent depression: randomized controlled trial. *Journal of Medical Internet Research*, *14*(3), e67. https://doi.org/10.2196/jmir.1858
- Korotitsch, W. J., & Nelson-Gray, R. O. (1999). An overview of self-monitoring research in assessment and treatment. *Psychological Assessment*, *11*(4), 415–425. https://doi.org/10.1037/1040-3590.11.4.415
- Kramer, I., Simons, C. J. P., Hartmann, J. A., Menne-Lothmann, C., Viechtbauer, W., Peeters, F., … Wichers, M. (2014). A therapeutic application of the experience sampling method in the treatment of depression: a randomized controlled trial. *World Psychiatry*, *13*(1), 68–77. https://doi.org/10.1002/wps.20090
- Kroeze, R., van der Veen, D. C., Servaas, M. N., Bastiaansen, J. A., Oude Voshaar, R. C., Borsboom, D., … Riese, H. (2017). Personalized feedback on symptom dynamics of psychopathology: A proof-of-principle study. *Journal for Person-Oriented Research*, *3(1)*, 1–11. https://doi.org/10.17505/jpor.2017.01
- Lamiell, J. T. (1998). 'Nomothetic' and 'idiographic': contrasting Windelband's understanding with contemporary usage. *Theory & Psychology, 8,* 23-38. https://doi.org/10.1177%2F0959354398081002
- Larson, R., & Csikszentmihalyi, M. (1983). "The experience sampling method". *New Directions for Methodology of Social and Behavioral Science, 15*, 41-56.
- Lütkepohl, H. (2005*). New introduction to multiple time series analysis*. New York, NY: Springer.
- Lutz, W., Schwartz, B., Hofmann, S. G., Fisher, A. J., Husen, K., & Rubel, J. A. (2018). Using network analysis for the prediction of treatment dropout in patients with mood and anxiety disorders: A methodological proof-of-concept study. *Scientific Reports*, *8*(1), 7819. https://doi.org/10.1038/s41598-018-25953-0
- Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research & Perspective*, *2*(4), 201–218.

https://doi.org/10.1207/s15366359mea0204_1
- Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., … Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 0021. https://doi.org/10.1038/s41562-016-0021
- Myin-Germeys, I., Kasanova, Z., Vaessen, T., Vachon, H., Kirtley, O., Viechtbauer, W., & Reininghaus, U. (2018). Experience sampling methodology in mental health research: new insights and technical developments. *World Psychiatry*, *17*(2), 123–132. https://doi.org/10.1002/wps.20513
- Myin-Germeys, I., Klippel, A., Steinhart, H., & Reininghaus, U. (2016). Ecological momentary interventions in psychiatry. *Current Opinion in Psychiatry*, *29*(4), 258–263. https://doi.org/10.1097/YCO.0000000000000255
- Myin-Germeys, I., Oorschot, M., Collip, D., Lataster, J., Delespaul, P., & van Os, J. (2009). Experience sampling research in psychopathology: opening the black box of daily life. *Psychological Medicine*, *39*(9), 1533–1547. https://doi.org/10.1017/S0033291708004947
- National Institute for Health and Care Excellence (2009). Depression in adults: recognition and management (NICE Guideline 90). Available at: www.nice.org.uk/CG90. [Accessed 20 Dec 2018].
- Ozomaro, U., Wahlestedt, C., & Nemeroff, C. B. (2013). Personalized medicine in psychiatry: problems and promises. BMC Medicine, 11(1), 132. https://doi.org/10.1186/1741-7015-11-132
- Palmier-Claus, J. E., Myin-Germeys, I., Barkus, E., Bentley, L., Udachina, A., Delespaul, P. A. E. G., … Dunn, G. (2011). Experience sampling research in individuals with mental illness: reflections and guidance. *Acta Psychiatrica Scandinavica*, *123*(1), 12–20. https://doi.org/10.1111/j.1600-0447.2010.01596.x
- Ramseyer, F., Kupper, Z., Caspar, F., Znoj, H., & Tschacher, W. (2014). Time-series panel analysis (TSPA): Multivariate modeling of temporal associations in psychotherapy process. *Journal of Consulting and Clinical Psychology*, *82*(5), 828–838. https://doi.org/10.1037/a0037168
- Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., … Vianello, M. (2018). Many analysts, one data set: making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632
- Simon, G. E., & Perlis, R. H. (2010). Personalized medicine for depression: can we match patients with treatments? *American Journal of Psychiatry*, *167*(12), 1445–1455. https://doi.org/10.1176/appi.ajp.2010.09111680
- Stavrakakis, N., Booij, S. H., Roest, A. M., de Jonge, P., Oldehinkel, A. J., & Bos, E. H. (2015). Temporal dynamics of physical activity and affect in depressed and nondepressed individuals. *Health Psychology*, *34*, 1268–1277. https://doi.org/10.1037/hea0000303
- Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177%2F1745691616658637
- Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(3), 273–282. https://doi.org/10.1111/j.1467-9868.2011.00771.x
- Trull, T. J., & Ebner-Priemer, U. (2013). Ambulatory assessment. *Annual review of*

clinical psychology, *9*, 151-176. https://doi.org/10.1146/annurev-clinpsy-050212-185510
- Uher, R. (2011). Genes, Environment, and Individual Differences in Responding to Treatment for Depression: Harvard Review of Psychiatry, 19(3), 109–124. https://doi.org/10.3109/10673229.2011.586551
- van Os, J., Delespaul, P., Wigman, J., Myin-Germeys, I., & Wichers, M. (2013). Beyond DSM and ICD: introducing "precision diagnosis" for psychiatry using momentary assessment technology. *World Psychiatry*, *12*(2), 113–117. https://doi.org/10.1002/wps.20046
- Walls, T. A., & Schafer, J. L. (2006). *Models for intensive longitudinal data.* Oxford, England: Oxford University Press.
- Wichers, M., Hartmann, J. A., Kramer, I. M. A., Lothmann, C., Peeters, F., van Bemmel, L., … Simons, C. J. P. (2011). Translating assessments of the film of daily life into person-tailored feedback interventions in depression. *Acta Psychiatrica Scandinavica*, *123*(5), 402–403. https://doi.org/10.1111/j.1600-0447.2011.01684.x
- Wright, A. G., & Hopwood, C. J. (2016). Advancing the assessment of dynamic psychological processes. *Assessment, 23*, 399-403. https://doi.org/10.1177/1073191116654760
- Wright, A. G. C., & Zimmermann, J. (in press). Applied ambulatory assessment: integrating idiographic and nomothetic principles of measurement. *Psychological Assessment.* https://doi.org/10.31234/osf.io/6qc5x

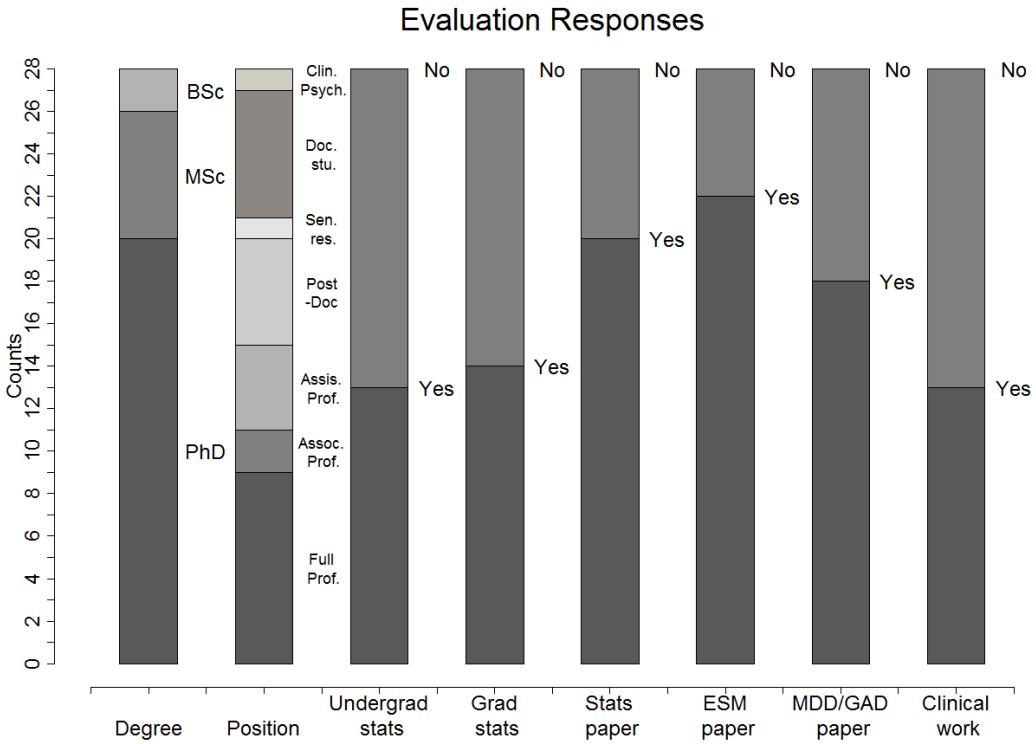
Author contributions

The project group (JAB, YKK, CJA, LFB) 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.

Funding

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. AGCW: National Institute of Mental Health (L30 MH101760); EC and PK: KU Leuven Research Council grant (GOA/15/003) and Fund for Scientific Research-Flanders grant (FWO G074319N, G066316N); FJB: The Netherlands Initiative for Education Research (NRO) grant (no.644405-16-401); JAB, MNS and HR: charitable foundation Stichting tot Steun VCVGZ (grant no. 239); MW: European Research Council (ERC) under the European Union's Horizon 2020 research and innovative programme (ERC-CoG-2015; No. 68146); OR: Netherlands Organization for Scientific Research Talent Grant (NWO Onderzoekstalent 406-15-128); PKW: National Institute on Alcohol Abuse and Alcoholism (AA024133); TJT: National Institute on Alcohol Abuse and Alcoholism (AA024133; AA019546).

Supplementary Materials



Supplementary Figure 1 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.

Supplementary Table 1 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

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.