AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Assessment

© 2024 American Psychological Association
ISSN: 1040-3590

2024, Vol. 36, Nos. 6–7, 379–394
https://doi.org/10.1037/pas0001314

Forecasting the Onset of Depression With Limited Baseline Data Only: A Comparison of a Person-Specific and a Multilevel Modeling Based Exponentially Weighted Moving Average Approach

Evelien Schat[1], Francis Tuerlinckx[1], Marieke J. Schreuder[1], Bart De Ketelaere[2], and Eva Ceulemans[1]
[1] Research Group of Quantitative Psychology and Individual Differences, Faculty of Psychology and Educational Sciences, KU Leuven
[2] Department of Biosystems, Mechatronics, Biostatistics and Sensors, KU Leuven

The onset of depressive episodes is preceded by changes in mean levels of affective experiences, which can be detected using the exponentially weighted moving average procedure on experience sampling method (ESM) data. Applying the exponentially weighted moving average procedure requires sufficient baseline data from the person under study in healthy times, which is needed to calculate a control limit for monitoring incoming ESM data. It is, however, not trivial to obtain sufficient baseline data from a single person. We therefore investigate whether historical ESM data from healthy individuals can help establish an adequate control limit for the person under study via multilevel modeling. Specifically, we focus on the case in which there is very little baseline data available of the person under study (i.e., up to 7 days). This multilevel approach is compared with the traditional, person-specific approach, where estimates are obtained using the person's available baseline data. Predictive performance in terms of Matthews correlation coefficient did not differ much between the approaches; however, the multilevel approach was more sensitive at detecting mean changes. This implies that for low-cost and nonharmful interventions, the multilevel approach may prove particularly beneficial.

Public Significance Statement
This study investigates early depression detection using the exponentially weighted moving average procedure with limited baseline data from the person under study in healthy times. The findings suggest that using historical data from other healthy individuals is useful and enhances sensitivity in detecting mean changes for the person under study.

Keywords: exponentially weighted moving average procedure, real-time monitoring, experience sampling method, multilevel models, detection of mean changes

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

A major challenge in clinical practice is identifying in real time when a person is at risk of developing a (recurrent) depressive episode, which would allow for timely intervention. So far, research into the prediction of depression has largely been performed at group level and has focused on time-invariant characteristics such as genetics (Bigdeli et al., 2017; Craddock & Forty, 2006; Levinson, 2006), personality (Klein et al., 2011), and early-life experiences (Agid et al., 1999; Heim et al., 2004). These studies, however, only inform us about who has an increased risk of developing depression. To identify when a person is at risk of developing (recurrent)

Evelien Schat https://orcid.org/0000-0003-1169-3984

Evelien Schat, Eva Ceulemans, and Francis Tuerlinckx were supported by a research grant from the Research Council of KU Leuven (C14/19/054). The computational resources and services used in this work were provided by the Vlaams Supercomputer Centrum, funded by the Hercules Foundation and the Flemish Government Department Economie, Wetenschap, and Innovatie. The authors thank Kristof Meers for his help in using the supercomputer and Heininga, V. E., Pe, M. L., Dejonckheere, E., Provenzano, J., Sels, L., and Cloos, L. J. R. for sharing the Experience Sampling Method data.

R code and additional simulation results can be found at https://osf.io/dar5v/. See https://osf.io/dar5v/ for an overview of which data sets are publicly available and where they can be accessed. The results appearing in the article were presented at the Society for Ambulatory Assessment Conference 2023.

Evelien Schat played a lead role in formal analysis, software, visualization, writing–original draft, and writing–review and editing and an equal role in conceptualization and methodology. Francis Tuerlinckx played a lead role in funding acquisition and supervision and an equal role in conceptualization, methodology, and writing–review and editing. Marieke J. Schreuder played a lead role in writing–review and editing. Bart De Ketelaere played a lead role in supervision, a supporting role in conceptualization and methodology, and an equal role in writing–review and editing. Eva Ceulemans played a lead role in funding acquisition and supervision, a supporting role in writing–original draft, and an equal role in conceptualization, methodology, and writing–review and editing.

Correspondence concerning this article should be addressed to Evelien Schat, Research Group of Quantitative Psychology and Individual Differences, Faculty of Psychology and Educational Sciences, KU Leuven, Tiensestraat 102, Box 3713, B-3000 Leuven, Belgium. Email: evelien.schat@kuleuven.be

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

depression, we need to monitor time-varying characteristics that may function as early warning signals. Monitoring emotional experiences seems promising for this purpose, as these experiences are inherently dynamic over time (Davidson et al., 2000; Frijda, 2007; Kuppens, 2015; Larsen, 2000). Moreover, deviations from typical patterns of emotional experiences are considered to be a central manifestation of depression (Houben et al., 2015).

To capture the dynamic nature of emotional experiences and how they unfold over time, researchers frequently use experience sampling methods (ESM; Myin-Germeys et al., 2009, 2018). In ESM studies, participants report on their affective states several times a day, for multiple days, during the course of their daily routines. Retrospective within-person analyses of ESM data suggest that the dynamics of affective experiences indeed often change prior to a depressive episode (Curtiss et al., 2023; Ludwig et al., 2023; Smit et al., n.d.; Wichers & Groot, 2016). However, from a prevention perspective, there is a need for methods that can detect when these affective experiences change in real time.

Statistical process control (SPC) procedures are deemed particularly useful for this purpose (Montgomery, 2009; Shewhart, 1931). SPC procedures originate from industry, where they are used to monitor manufacturing process in real time. Recently, SPC has been introduced in the field of psychopathology as a tool to monitor for early warning signals of depression in real time (Schat, Tuerlinckx, De Ketelaere, & Ceulemans, 2023; Schat, Tuerlinckx, Smit, et al., 2023; Smit et al., 2019; Smit & Snippe, 2023; Snippe et al., 2023). These studies have shown that the exponentially weighted moving average (EWMA) procedure (Roberts, 1959) performs especially well for this purpose and that there are benefits to monitoring day averages rather than the scores at the individual measurement occasions (Schat, Tuerlinckx, De Ketelaere, & Ceulemans, 2023; Schat, Tuerlinckx, Smit, et al., 2023). A potential drawback of the EWMA procedure is its probability of signaling false positives—that is, predicting that someone will have a depressive episode while this is not the case. However, this drawback characterizes all statistical tests. Moreover, the significant association between the obtained early warnings and the occurrence of a depressive episode indicates a favorable balance of true positives versus false positives.

To apply the EWMA procedure in practice, two distinct phases are required. In Phase I, the in-control distribution of the monitored process is captured. For psychological purposes, this involves estimating the mean (μ_{1i}) and standard deviation (σ_{1i}) of repeatedly assessed emotions for a particular individual i (e.g., Alice). These estimates are then used to compute a control limit,[1] which describes the boundary of Alice's normal emotions. In Phase II, incoming emotions reported by Alice are transformed into EWMA scores and monitored over time. As long as these scores remain below the control limit, Alice's emotions can be considered "normal" (or in control). When a score falls above the control limit, however, emotions are deviating from what is normal (or out of control), which may in turn call for a preventive intervention. Because the control limit is person-specific, namely based on Alice's Phase I data, the EWMA method is in line with the trend toward "personalized psychopathology" (Wright & Woods, 2020).

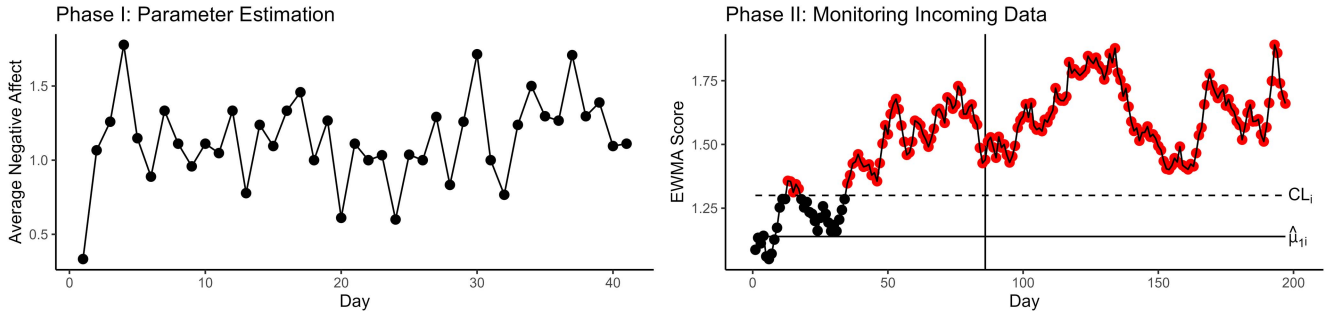
Figure 1 shows an example of the EWMA procedure applied to the ESM data of a person with a history of major depressive disorder who had been using antidepressants for the previous 8.5 years (Groot, 2010; Wichers & Groot, 2016). The participant took part in a 239-day ESM study, in which the antidepressant dose was reduced. Importantly, the study had two distinct phases, namely a relatively stable period during which the participant was in remission (i.e., Days 1–41) as well as a period during which he gradually tapered his antidepressant medication (i.e., Days 42–239). Correspondingly, we used the first 41 days as Phase I (i.e., to define the control limit) and the remaining data as Phase II (i.e., to monitor changes in emotions). Around Day 127 (i.e., Day 86 in Phase II), the participant relapsed into depression. Applying the EWMA procedure to the participant's day averages of negative affect (i.e., mean of the items "restless," "irritated," and "down") resulted in many out-of-control scores (indicated by the red dots), with the first occurring on Day 13 in Phase II. The detection of a change in the monitored process occurred well in advance of the relapse on Day 86 in Phase II and thus would have allowed for early intervention. This was later replicated in the Transitions in Depression Tapering study (Smit et al., 2020), in which 41 formerly depressed patients (gradually) discontinued their antidepressant medication while participating in a 4-month ESM study. Twenty-two participants experienced recurrent depressive symptoms. There was a distinct Phase I period of 28 days during which the participants were in remission, after which (gradual) discontinuation of their antidepressants started. The remaining days (i.e., Days 29–122) were used for Phase II monitoring using EWMA, and this confirmed that the EWMA procedure may be useful for forecasting the onset of depression by monitoring mental states (Schreuder et al., 2024; Snippe et al., 2023).

Despite the promising initial results, the implementation of SPC for psychological data is faced with a major challenge, namely obtaining a sufficient amount of Phase I data (i.e., baseline data) for the person we are monitoring. Ideally, these data should (a) cover multiple weeks (or even months) of frequent mood assessments and (b) be collected during a stable, healthy period. With respect to the first requirement, simulation results suggest that a minimum of 50 Phase I days are recommended for the EWMA procedure, while multivariate procedures require even more data (Schat, Tuerlinckx, Smit, et al., 2023). Although a number of studies have shown that collecting ESM data across many months is feasible (Bos et al., 2020; Dejonckheere, Houben, et al., 2021; Helmich et al., 2020; Ludwig et al., 2023; Myin-Germeys et al., 2018; Olthof et al., 2020; Schreuder et al., 2020; Smit & Snippe, 2023; Snippe et al., 2023; Wichers & Groot, 2016), it is not trivial to obtain 50 days of Phase I data. Indeed, the Transitions in Depression Tapering Phase I of 28 days would generally be considered long when compared to the more traditional baseline period in ESM research, which usually consists of 7 days at maximum (e.g., study on exam results, Dejonckheere, Mestdagh, et al., 2021; study on anticipatory social stress, Koval & Kuppens, 2012). In this study, we therefore focus on this more traditional scenario and investigate what we can best do in this case. In addition, it is difficult to guarantee that individuals do not experience major changes in their mental health during this period, which violates the second requirement. Thus, establishing a reliable control limit that reflects the boundary of "normal" emotions

[1] It is possible to implement two-sided monitoring, where one has an upper and a lower control limit, which describe the range of normal emotions. Regarding one-sided monitoring, one can use either the upper or lower control limit. Here, we choose for one-sided monitoring using the upper control limit, as warning signals are only expected to occur in one direction.

Figure 1

EWMA Procedure Applied to Day Averages of Negative Affect



Note. In the right panel, the dashed horizontal line indicates the control limit (CL_i). The solid horizontal line indicates the estimated Phase I mean ($\hat{\mu}_{1i}$). The solid vertical line indicates the day of the relapse into depression (Day 86 in Phase II). The red dots indicate the out-of-control scores that exceed the CL_i. EWMA = exponentially weighted moving average. See the online article for the color version of this figure.

for a particular person may be difficult, which severely limits the potential utility of SPC.

The issue of insufficient Phase I data has also surfaced in other fields. To illustrate, in agriculture research, alternative ways to compute control limits have been proposed, for example, to monitor the feeding patterns of pigs (Maselyne et al., 2018) or the daily average egg weight (Mertens et al., 2008). These studies have tried to solve the issue by using historical data to derive SPC control limits for the objects under study. Correspondingly, in our case, extensive historical data drawn from standard 1- or 2-week ESM studies may be a rich source of insights regarding the in-control distribution of healthy individuals (e.g., Cloos et al., 2023; Dejonckheere et al., 2018; Dejonckheere, Kalokerinos, et al., 2019; Eisele et al., 2022; Houben et al., 2017). Nevertheless, the challenge remains of how to combine ESM data from other individuals with little Phase I data (i.e., up to 7 days) of the person under study to establish a useful control limit in the light of the ubiquitous individual differences in the level and variability of people's emotions (Dejonckheere, Mestdagh, et al., 2019; Nelson et al., 2020; Thompson et al., 2017).

Drawing upon established literature, a good way to combine historical data with information from the person under study is through multilevel modeling. This allows us to use the historical information while still molding it toward the person under study. Specifically, for someone who has a few repeated assessments of mood (e.g., 1–7 days of ESM data), we could estimate their mean and standard deviation by augmenting the limited Phase I data with the multilevel model for the historical data.[2] An alternative, person-specific approach would be to solely use the limited available ESM data as Phase I data. We will compare these two approaches with respect to (a) deriving the control limit in Phase I and (b) accurately predicting the transition into depression in Phase II.

The remainder of this article is structured as follows: First, we describe the key ideas of the EWMA procedure. Next, we introduce the multilevel modeling approach of estimating the Phase I parameters and evaluate these estimates by comparing them to the person-specific estimates (i.e., the within-person sample mean and standard deviation) and the known parameters (i.e., the true within-person mean and standard deviations). We then further evaluate the estimates by performing a simulation study to investigate predictive Phase II performance of the EWMA procedure using a control limit based on the estimates. Finally, we present a discussion of the results and directions for future research.

EWMA Procedure

In this section, we first explain how we preprocess the ESM data and provide a rationale for this. Next, we discuss how to compute the EWMA scores that are monitored in Phase II, followed by a description on how the Phase I data are used to calculate the control limit needed for Phase II monitoring. Finally, we elaborate on how EWMA performance is measured and its desired behavior.

Data Preprocessing

The EWMA procedure (Roberts, 1959) is based on a number of assumptions, and its expected performance is contingent on these assumptions being met. Scores at the individual measurement occasions are assumed to be normally distributed, independent across time (i.e., serially independent), and there should be no missing values (i.e., equal spacing between measurement occasions). It is unlikely that these assumptions are met when using ESM data, but certain preprocessing steps can minimize the impact thereof. Specifically, simulation studies have shown that it is beneficial to apply the EWMA procedure to day averages (i.e., aggregated measurement occasions within days) rather than the scores at the individual measurement occasions (Schat, Tuerlinckx, De Ketelaere, & Ceulemans, 2023; Schat, Tuerlinckx, Smit, et al., 2023). The averaging step renders data distributions less skewed, decreases serial dependence, and minimizes issues associated with missing data (Schat, Tuerlinckx, Smit, et al., 2023). Importantly, this step also improves the EWMA performance by increasing the size of the mean change in the monitored process due to the reduced variance of the day statistics as compared to the original observed scores.

[2] Note that in case no Phase I data are available of the person under study, the historical data can function as a replacement. Alternatively, one can use covariates of the person under study to obtain an estimate of the person's mean and standard deviation (Schreuder et al., 2023).

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Phase II: Monitored Score

The monitored EWMA scores are based on both past and current information, as they are a weighted sum of the self-reported ratings obtained thus far. The highest weight is assigned to the current rating, and the weights of the past ratings decrease exponentially over time. Specifically, the EWMA procedure computes a score z_{it} for individual i on every day $t(t = 1, \ldots, T)$:

$$z_{it} = \lambda x_{it} + (1 - \lambda) z_{i,t-1}, \tag{1}$$

where x_{it} denotes the observed day average on day t, $z_{i,t-1}$ represents the EWMA score on the previous day, and the starting value z_{i0} is equal to the estimated Phase I mean $\hat{\mu}_{1i}$ for individual i. The constant λ $(0 < \lambda \leq 1)$ is the weight given to x_{it}; the remaining weight is given to $z_{i,t-1}$. When interested in detecting smaller mean changes, lower values of λ are recommended. For ESM data, λ values between .05 and .15 have shown to work well (Schat, Tuerlinckx, Smit, et al., 2023; Smit et al., 2023; Snippe et al., 2023). Here, we set λ to .10.

Phase I: Parameter Estimation and Calculation of Control Limit

Using the day averages in Phase I, we obtain estimates of the in-control mean and standard deviation (i.e., $\hat{\mu}_{1i}$ and $\hat{\sigma}_{1i}$). These estimates are then used to compute the control limit, which is used in Phase II to determine whether the process is in control or not. As long as the monitored EWMA scores in Phase II fall below this control limit, the process is considered to be in control. When an EWMA score goes beyond the control limit, the process is flagged as out of control. The control limit (CL_i) is calculated as follows:

$$CL_i = \hat{\mu}_{1i} + L\hat{\sigma}_{1i}\sqrt{\frac{2}{(2 - \lambda)}}. \tag{2}$$

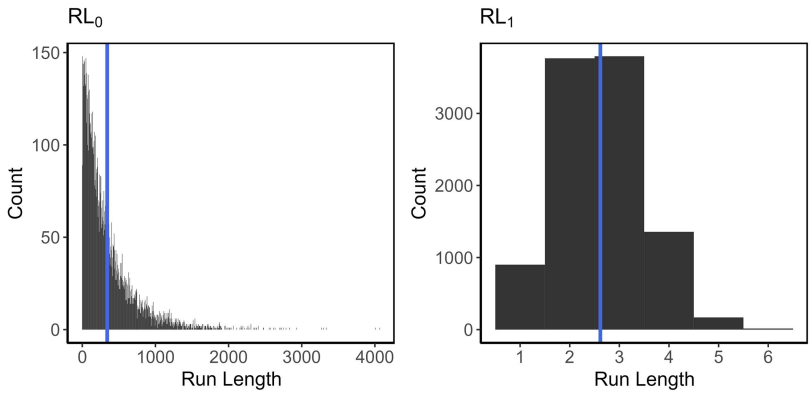
The L parameter determines the height of the control limit. We will explain in the next section how to obtain this L value for a specified EWMA performance.

Performance and Desired Behavior

In SPC literature, the run length (RL) indicates the day in Phase II at which the first out-of-control score is observed (Montgomery, 2009). If the process remains in control in Phase II (i.e., no change occurred; no depressive episode developed), the RL_0 is ideally high, as an out-of-control score is a false positive. On the other hand, if the process goes out of control (i.e., a change occurred; depression developed), the RL_1 is ideally short, as this allows for timely interventions. When multiple replicates from the same process (i.e., monitoring the same person again and again, which is obviously impossible in reality) could be obtained, the RL can largely vary. For example, in the event of a change (e.g., the onset of depression), the exact day at which the process is flagged as being out of control may vary. Consequently, in order to evaluate the performance of the EWMA procedure in simulations, one usually generates a large number of replicates from the same process, resulting in a distribution of RL values. To summarize this RL distribution, the average of the run length (ARL) distribution is usually reported. The in-control ARL_0 denotes the average run length given that the process remains in control in Phase II, and the out-of-control ARL_1 denotes the average run length given that the process goes out of control at the start of Phase II. Figure 2 shows an example of RL_0 and RL_1 distributions, where the former follows the typical positively skewed distribution of an in-control process. The ARL_0 and ARL_1 of the two distributions equal 340.5 and 2.6, respectively.

The L value in the formula for the control limit is related to these ARL values. Specifically, when the λ value is kept constant, increasing the value of L results in a higher control limit, which in turn leads to higher expected ARL_0 and ARL_1 values. Given this interrelationship, in practice L is often estimated from λ and the desired ARL_0, which users choose themselves. This can be done using the xewma.crit function of the R package spc (Knoth, 2020). We set ARL_0 to 370, and with a λ of .10, this yields an L of 2.40. An ARL_0 of 370 means that, in case the process stays in control (i.e., a person stays healthy), a false alarm is expected to occur on average after 370 days. Although the choice of 370 is common in the SPC literature, a higher or lower ARL_0 value may be used, depending on the cost and impact of an intervention (i.e., whether false positives are problematic; Smit et al., 2023).

Figure 2
Example of a RL_0 and RL_1 Distribution



Note. The horizontal blue lines indicate the ARL_0 (left) and ARL_1 (right) of the distributions. RL = run length; ARL = average of the run length. See the online article for the color version of this figure.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Study 1: Evaluating the Quality of the Obtained Phase I Estimates and Corresponding Control Limit

The challenge in the Phase I period is to obtain sufficient in-control data of the person we aim to monitor. In this section, we showcase the person-specific approach to obtaining estimates for μ_{1i} and σ_{1i} for the person under study, based on the person's available Phase I data, by means of an ESM data set that we will call the target data. For individuals in this target data set, we will compute control limits for monitoring changes in their negative affect, even though few Phase I data are available. Next, we elaborate on the multilevel modeling approach, which fits a multilevel model to historical data from a group of healthy individuals. This model is then used to generate estimates for μ_{1i} and σ_{1i} for the person under study, based on the person's available Phase I data (i.e., target data). As historical data, we use either the target data, excluding the person who is monitored (leave-one-out case), or five other ESM data sets (historical case). We evaluate the multilevel estimates resulting from both the leave-one-out and the historical cases by comparing them to the person-specific estimates and the true parameters of the person under study. Comparing the two multilevel cases will reveal whether it matters if the historical data differ from the data of the person under study in terms of design choices and settings (e.g., items). This study was not preregistered.

ESM Data Sets

The details of the six data sets are provided in Table 1 (Cloos et al., 2023; Dejonckheere, Kalokerinos, et al., 2019; Heininga et al., 2019; Pe et al., 2016; Provenzano et al., 2021; Sels et al., 2017). We consider the sixth data set of Cloos et al. (2023) as our target data, as this data set contained the most assessments per individual. In the leave-one-out case, these data also yield the historical data. In the historical case, the historical data consist of the first five data sets. In all studies, ESM data were collected, and participants rated their emotions multiple times a day (i.e., 7–10 times) for multiple days (i.e., 7–14 days). Negative affect items were measured on a continuous sliding scale, ranging from 0 (*not at all*) to 100 (*very much*). The negative affect items themselves and the amount of items that were measured differed between studies, which is a point we revisit in the Discussion section. Using the relevant items in each study, we calculated the day averages of negative affect for each person. All participants provided informed consent, and the studies were ethically approved. See https://osf.io/dar5v/ for an overview of which data sets are publicly available and where they can be accessed.

For all six data sets, we applied the following two inclusion criteria: First, we only included healthy individuals (i.e., not diagnosed with depression or other mood disorders), as we want the control limit to reflect the upper boundary of negative affect when people are healthy. Second, we only included individuals with a compliance of at least 75%.[3] This yielded a sample of 762 individuals in Data Sets 1–5 and a sample of 145 in Data Set 6.

In each data set, we computed the day averages of negative affect for each participant. The mean of negative affect in Data Sets 1–5 (calculated overall day averages of all participants within the study) falls between 9.58 and 15.44. The standard deviation of negative affect in these data sets ranges from 6.08 to 11.99. Both the mean and standard deviation values in Data Set 6 are higher, amounting to 19.21 and 12.56, respectively.

Obtaining Phase I Estimates Using the Person-Specific Approach

Ideally, the individual Phase I parameters (i.e., μ_{1i} and σ_{1i}) for the monitored process (e.g., negative affect) are known, such that they can be used to calculate rather than estimate the control limit. This is, in practice, never the case, also not for the 145 individuals in our target data. To still have a benchmark to compare the person-specific and multilevel estimates with, we consider the known parameters for each target person to be the mean and standard deviation of the 14 available day averages of negative affect.

When the Phase I parameters are not known, one usually takes the traditional, person-specific approach, where the mean ($\bar{y}_i$) and standard deviation (s_i) are estimated based on the person's available Phase I day averages of negative affect. These person-specific estimates are then used to compute the control limit. In this study, we vary the number of available Phase I days from 2 to 7 days, which are used to obtain the person-specific estimates.

To illustrate how these person-specific estimates and the corresponding control limit may fluctuate depending on the number of available Phase I days,[4] we present an example in Table 2 of a target person from Data Set 6. The person's known mean and standard deviation equal 22.72 and 4.60, respectively. The day averages y_{it} fluctuate over the days, which is reflected in the person-specific estimates and corresponding control limit. As more Phase I data become available, the estimates approach the known parameters.

Obtaining Phase I Estimates Using the Multilevel Modeling Approach

Modeling Historical Data

To obtain the multilevel estimates, we start by pooling the ESM data of our historical data. For each individual i, we compute a day average y_{it} of the item y on day t, where $i = [1, \ldots, I]$ and $t = [1, \ldots, T]$. In our case, y is negative affect. Next, we fit a multilevel model to the historical data. Specifically, we fit a random-intercept multilevel model on y_{it} and allow the Level 1 error variance to differ between persons (Nestler & Humberg, 2022):

$$y_{it} = \gamma_{0i} + \varepsilon_{it},\ \varepsilon_{it} \sim N(0,\ \sigma^2_{\varepsilon i}). \tag{3}$$

$$\gamma_{0i} = \beta_{00} + \nu_{0i},\ \nu_{0i} \sim N(0,\ \sigma^2_{\nu}). \tag{4}$$

$$\sigma^2_{\varepsilon i} = \exp(s_{00} + \omega_{0i}),\ \omega_{0i} \sim N(0,\ \sigma^2_{\omega}). \tag{5}$$

[3] We also ran the analyses in Studies 1 and 2 with a more relaxed inclusion criterion. Specifically, we included all participants in Data Sets 1–5 (i.e., 841 participants). For Data Set 6, we only excluded participants (i.e., $N = 5$) who did not have any data—and thus did not have a day average—in at least one of the first 7 days. Using this more relaxed inclusion criterion did not change our conclusions; the results can be found at https://osf.io/dar5v/. We advise to use a stringent threshold for the target person under study, as having (a lot of) missing data in a short Phase I period will render the estimates of the person-specific mean and standard deviation less reliable.

[4] Note that we do not update the estimates (and thus the control limits) every day after obtaining a new day average for the person under study but rather use the limited amount of Phase I data to obtain the person-specific and multilevel estimates.

Table 1
Overview of the ESM Studies

Data set	Protocol characteristic			Measurement characteristic		Descriptive	
	N participants	Occasions/day	*N* days	NA items	Scale	*M* NA	*SD* NA
1. Heininga et al. (2019)	38	10	7	Depressed, stressed, anxious, angry, restless	0–100	9.58	6.08
2. Pe et al. (2016)	516	10	7	Depressed, stressed, sad, anxious, angry	0–100	12.97	10.23
3. Dejonckheere, Kalokerinos, et al. (2019)	90	7	14	Stressed, sad, angry	0–100	15.44	11.44
4. Provenzano et al. (2021)	28	10	7	Stressed, sad, anxious, angry	0–100	10.05	9.48
5. Sels et al. (2017)	90	10	8	Depressed, stressed, sad, anxious, angry	0–100	12.64	11.99
6. Cloos et al. (2023)	145	10	14	Depressed, stressed, anxious, angry, exhausted, ashamed, guilty, gloomy, tired, shy, regretful, irritated, concerned, nervous	0–100	19.21	12.56

Note. *N* participants indicate the number of individuals we have after applying the exclusion criteria. The descriptives for each study are based on the day averages of NA of all participants, calculated using all NA items. ESM = experience sampling method; NA = negative affect; *M* = mean; *SD* = standard deviation.

The group-mean level of y equals β_{00}. Interindividual differences in y means are captured by the random intercept γ_{0i}, which expresses the deviation of the person-specific y mean from the overall (group) mean. The random effects v_{0i} are normally distributed, with mean zero and variance σ_v^2 (i.e., a large value implies big differences between individuals). ε_{it} denotes the Level 1 errors, which express daily variations in y level, and are normally distributed with variance $\sigma_{\varepsilon i}^2$. Interindividual differences in $\sigma_{\varepsilon i}^2$ are modeled by summing and exponentially transforming s_{00} and ω_{0i}, with s_{00} denoting the average of the logarithm of the individual-specific Level 1 error variances. ω_{0i} is the person-specific random effect, where ω_{0i} is normally distributed with a variance σ_ω^2. The exponential is taken to ensure that the Level 1 error variance remains positive.

To fit this multilevel model, we use the umels function of the mels package in R (Nestler & Humberg, 2022). Fitting this multilevel model to Data Sets 1–5, for instance, yields the following estimates: group-mean level of NA $\hat{\beta}_{00} = 12.57$, standard deviation of the random effect $\hat{\sigma}_\nu = 7.93$, group-level logarithm of the Level 1 error variance $\hat{s}_{00} = 3.02$ (for a typical person, this results in a Level 1 standard deviation of 4.53), and the standard deviation of the random effect $\hat{\sigma}_\omega = 1.27$. Applying the multilevel model to Data Set 6 yielded the estimates: $\hat{\beta}_{00} = 19.10$, $\hat{\sigma}_\nu = 10.76$, $\hat{s}_{00} = 3.29$, and $\hat{\sigma}_\omega = .98$.

Predicting Multilevel Estimates for a New Person

Once we fit the multilevel model, we compute a day average y_{it} for each available Phase I day for the person under study (i.e., a person from Data Set 6). Using the predict function, we obtain estimates of the random intercept $\hat{\gamma}_{0i}$ and Level 1 error variance $\hat{\sigma}_{\varepsilon i}^2$ for the person under study using the person's available Phase I day averages. These two estimates can then be used as our Phase I estimates: $\hat{\gamma}_{0i}$ for the Phase I mean $\hat{\mu}_{1i}$ and $\hat{\sigma}_{\varepsilon i}$ for the Phase I standard deviation $\hat{\sigma}_{1i}$. Of note, the parameter estimates obtained from the multilevel model may be subject to shrinkage, where the estimates are shrunk toward the group-level values (i.e., $\hat{\beta}_{00}$ and $\hat{s}_{00}$).

The multilevel estimates for our example person are shown in Table 2 for the historical approach, which fluctuate over the number of Phase I days. The estimates for the mean (i.e., $\hat{\gamma}_{0i}$) are lower than the person-specific estimates (i.e., $\bar{y}_i$), whereas the estimates for the standard deviation (i.e., $\hat{\sigma}_{\varepsilon i}$) are higher than the person-specific estimates (i.e., s_i).

Evaluation of the Person-Specific and Multilevel Estimates

To assess the accuracy of the person-specific and multilevel estimates obtained using the approaches described above, we

Table 2
Day Averages, Known Parameters, Person-Specific and Multilevel Estimates, and Corresponding Control Limits of a New Person

N Phase I days	y_{it}	Known parameter			Person-specific estimate			Multilevel estimate		
		μ_{1i}	σ_{1i}	CL_i	$\bar{y}_i$	s_i	CL_i	$\hat{\gamma}_{0i}$	$\hat{\sigma}_{\varepsilon i}$	CL_i
1	13.05	22.72	4.60	25.26						
2	17.91	22.72	4.60	25.26	15.48	3.44	23.96	15.03	4.29	25.60
3	21.89	22.72	4.60	25.26	17.62	4.43	28.53	16.97	4.83	28.88
4	24.89	22.72	4.60	25.26	19.41	5.09	31.96	18.61	5.32	31.72
5	21.15	22.72	4.60	25.26	19.76	4.48	30.79	19.18	4.83	31.09
6	25.66	22.72	4.60	25.26	20.74	4.67	32.26	20.16	4.97	32.41
7	26.23	22.72	4.60	25.26	21.53	4.74	33.22	20.97	5.00	33.30

Note. y_{it} denotes the day average of negative affect. μ_{1i} and σ_{1i} denote the known mean and standard deviation, respectively. $\bar{y}_i$ and s_i are the sample mean and standard deviation, respectively, based on the available day averages. $\hat{\gamma}_{0i}$ is the random intercept, $\hat{\sigma}_{\varepsilon i}$ is the standard deviation of the Level 1 errors, and CL_i is the control limit.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

compared these estimates to the known parameters. To this end, we determined the mean squared error (MSE) for the different estimates (i.e., $\bar{y}_i$, s_i, $\hat{\gamma}_{0i}$, $\hat{\sigma}_{\varepsilon i}$) across the 145 target persons. The MSE is the average squared difference between the estimate and the known parameter value (i.e., μ_{1i} for the mean and σ_{1i} for the standard deviation). Thus, the lower the MSE value, the more accurate the estimate is. We also calculated the correlation between the person-specific/multilevel estimates and known parameter values.

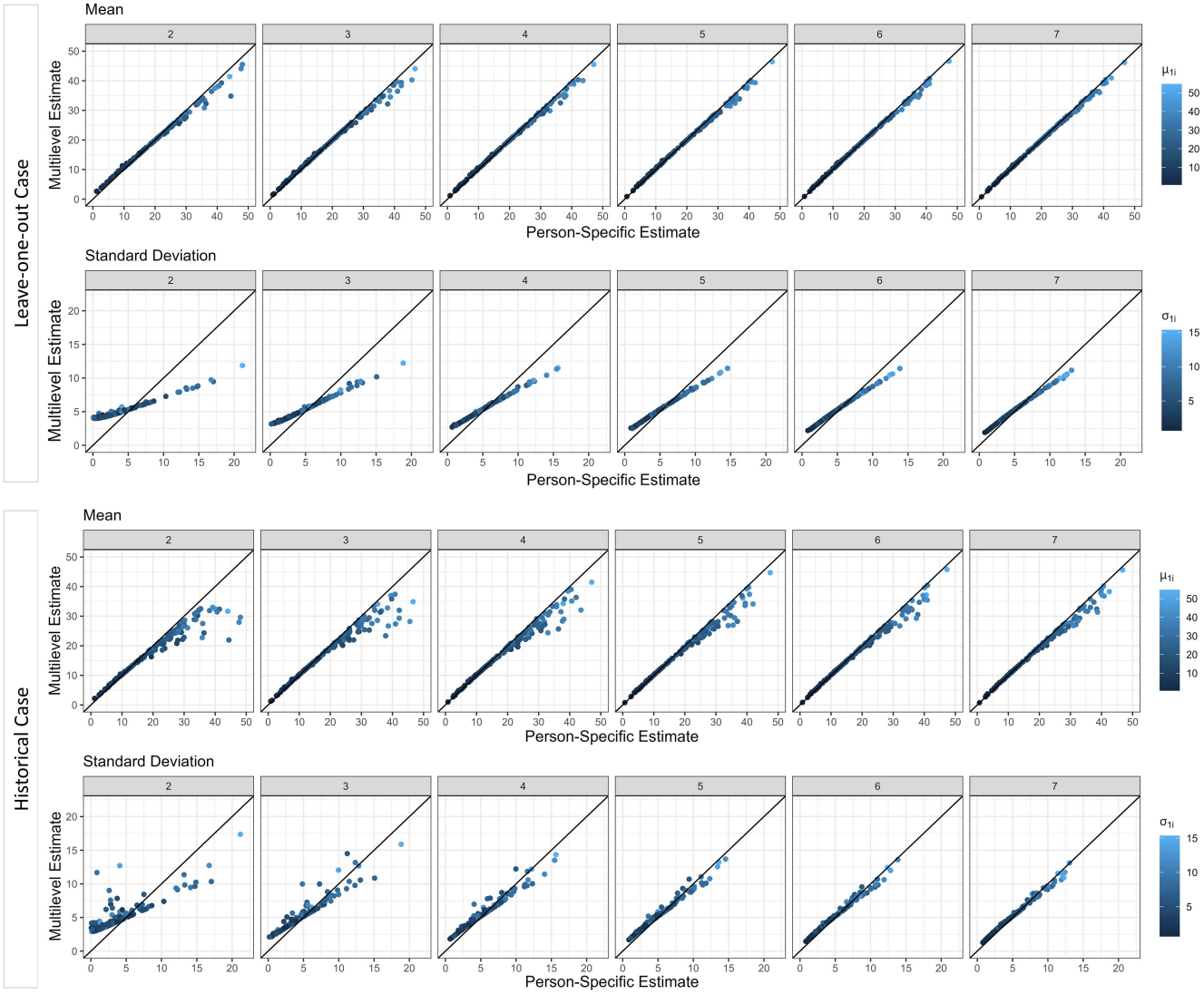
Results

Figure 3 shows the scatter plots of the person-specific and multilevel estimates of the mean and standard deviation for the leave-one-out case and the historical case. The columns indicate the number of Phase I days, and the color of the dots indicates the value of the known parameter (i.e., μ_{1i} and σ_{1i}). Table 3 shows the average MSE values of the person-specific and multilevel estimates (of the 145 persons) and the correlations between the estimates and known parameter values for the different number of Phase I days.

Mean. In the leave-one-out case, the person-specific (i.e., $\bar{y}_i$) and multilevel (i.e., $\hat{\gamma}_{0i}$) mean estimates are very similar. In Figure 3, we do observe some shrinkage, where the more extreme multilevel estimates are pulled toward the group-mean level. As such, the higher multilevel estimates are lower than the corresponding person-specific estimates, and the lower multilevel estimates are higher than the corresponding person-specific estimates. However, this effect decreases as the number of Phase I days increases and is minimal for seven Phase I days. In the historical case, we observe more shrinkage for the mean,

Figure 3

Scatter Plots of the Person-Specific and the Multilevel Estimates for the Leave-One-Out Case and Historical Case



Note. The first two and the last two rows show the results for the leave-one-out case and the historical case, respectively. Within a case, the top row shows the results for the mean and the bottom row for the standard deviation. The columns indicate the number of Phase I days. The color of the dots indicates the value of the known parameter (i.e.,μ_{1i} and σ_{1i}). μ_{1i} = known mean; σ_{1i} = standard deviation. See the online article for the color version of this figure.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 3
Average Mean Squared Error Values and Correlations

Case	*N* Phase I days	MSE				ρ			
		$\bar{y}_i$	$\hat{\gamma}_{0i}$	s_i	$\hat{\sigma}_{\varepsilon i}$	$\rho_{\bar{y}_i,\mu_{1i}}$	$\rho_{\hat{\gamma}_{0i},\mu_{1i}}$	$\rho_{s_i,\sigma_{1i}}$	$\rho_{\hat{\sigma}_{\varepsilon i},\sigma_{1i}}$
Leave-one-out	2	39.48	36.17	17.35	7.75	.84	.84	.36	.38
	3	26.65	25.37	9.72	6.47	.89	.89	.53	.53
	4	19.13	19.06	6.49	5.11	.92	.92	.65	.65
	5	16.14	16.28	5.48	4.73	.93	.93	.69	.68
	6	12.04	12.32	4.36	4.09	.95	.95	.74	.74
	7	9.54	9.72	3.30	3.23	.96	.96	.81	.80
Historical	2	39.48	38.53	17.35	7.40	.84	.84	.36	.50
	3	26.65	29.35	9.72	6.88	.89	.88	.53	.55
	4	19.13	21.79	6.49	5.11	.92	.91	.65	.67
	5	16.14	17.91	5.48	4.67	.93	.93	.69	.69
	6	12.04	13.60	4.36	3.90	.95	.95	.74	.75
	7	9.54	10.76	3.30	3.01	.96	.96	.81	.81

Note. $\hat{\gamma}_{0i}$ denotes the random intercept, $\hat{\sigma}_{\varepsilon i}$ the standard deviation of the Level 1 errors, $\bar{y}_i$ the sample mean, and s_i the sample standard deviation. μ_{1i} and σ_{1i} denote the known mean and standard deviation, respectively. Correlations between person-specific/multilevel estimates and known parameters are given by ρ. MSE = mean squared error.

where especially the higher multilevel estimates are lower than the corresponding person-specific estimates. This makes sense, as in our case the group-mean level of the historical data is lower than the mean level of the target data (see Table 1 and Obtaining Phase I Estimates Using the Multilevel Modeling Approach section). Again, this effect diminishes as the number of Phase I days increases.

In Table 3, we see that overall, the MSE values of all estimates decrease as the number of Phase I days increases, indicating that the estimates become more accurate (i.e., approach the known values) with more data. Similarly, the correlations increase as the number of Phase I days increases. In the leave-one-out case, the MSE of $\bar{y}_i$ is higher than the MSE of $\hat{\gamma}_{0i}$ for two to four Phase I days. From five Phase I days onward, the MSE of $\bar{y}_i$ is lower than that of $\hat{\gamma}_{0i}$. Thus, from five Phase I days onward, the person-specific estimate is closer to the known mean than the multilevel estimate. In the historical case, the intersection already occurs after two Phase I days. Thus, from three Phase I days onward, the MSE of $\bar{y}_i$ is lower than the MSE of $\hat{\gamma}_{0i}$. This corresponds with the larger amount of shrinkage observed in the historical case as compared to the leave-one-out case (Figure 3). This means that, aside from two Phase I days, the person-specific estimate outperforms the multilevel estimate. In both the leave-one-out and historical cases, the correlation between $\bar{y}_i$ and μ_{1i} initiates at .84 for two Phase I days and increases to .96 for seven Phase I days. These same values were observed for the correlation between $\hat{\gamma}_{0i}$ and μ_{1i} for two and seven Phase I days.

Standard Deviation. For the standard deviation, we also observe shrinkage in both the leave-one-out case and the historical case (Figure 3). The higher multilevel estimates (i.e., $\hat{\sigma}_{\varepsilon i}$) are lower than the corresponding person-specific estimates (i.e., s_i), and the lower multilevel estimates are higher than the corresponding person-specific estimates. Moreover, more Phase I days (as compared to the mean) are required to get a strong correlation between the person-specific and multilevel estimates and the known standard deviation. This is reasonable, considering that more observations are needed to get an accurate estimate for the standard deviation than for the mean. The correlation between s_i and σ_{1i} (in the leave-one-out case and the historical case) is quite low at .36 for two Phase I days and increases to .81 for seven Phase I days (Table 3). The correlation between $\hat{\sigma}_{\varepsilon i}$ and σ_{1i} also starts low in the leave-one-out case, at .38 for two Phase I days and increases to .80 for seven Phase I days. In the historical case, however, the correlation is .50 for two Phase I days and increases to .81 for seven Phase I days.

The MSE of $\hat{\sigma}_{\varepsilon i}$ is consistently lower than the MSE of s_i, implying that the multilevel estimates are more accurate than the person-specific estimates. This again reflects the fact that more data are needed to obtain an accurate estimate of the standard deviation, and thus, here, we benefit from having more data than merely the person's available Phase I data.

Study 2: Evaluating the Quality of Obtained Control Limits for Detecting Mean Changes

The next step is to investigate how well the estimates and corresponding control limit work for monitoring and detecting mean changes in the Phase II data of the person under study. Specifically, we aim to evaluate the predictive performance of the EWMA procedure with different control limits. As such, we conducted a simulation study in which we applied the EWMA procedure with a control limit based on the person-specific estimates, multilevel estimates, and known parameters. Specifically, we used the person-specific estimates, multilevel estimates (leave-one-out case and historical case), and the known parameters of the 145 individuals in the target data (i.e., Data Set 6), as obtained in the previous section. Prior to explaining our simulation study, we describe the statistical procedure used in two recent articles applying EWMA to forecast recurrence of depression (Schreuder et al., 2024; Snippe et al., 2023), which we used as a basis for our simulation design.

Statistical Procedure Used by Snippe et al. (2023) and Schreuder et al. (2024)

As mentioned in the introduction, Snippe et al. (2023) and Schreuder et al. (2024) investigated whether the recurrence of depression can be forecasted by monitoring mental states using the EWMA procedure. Specifically, these studies analyzed data of 41

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

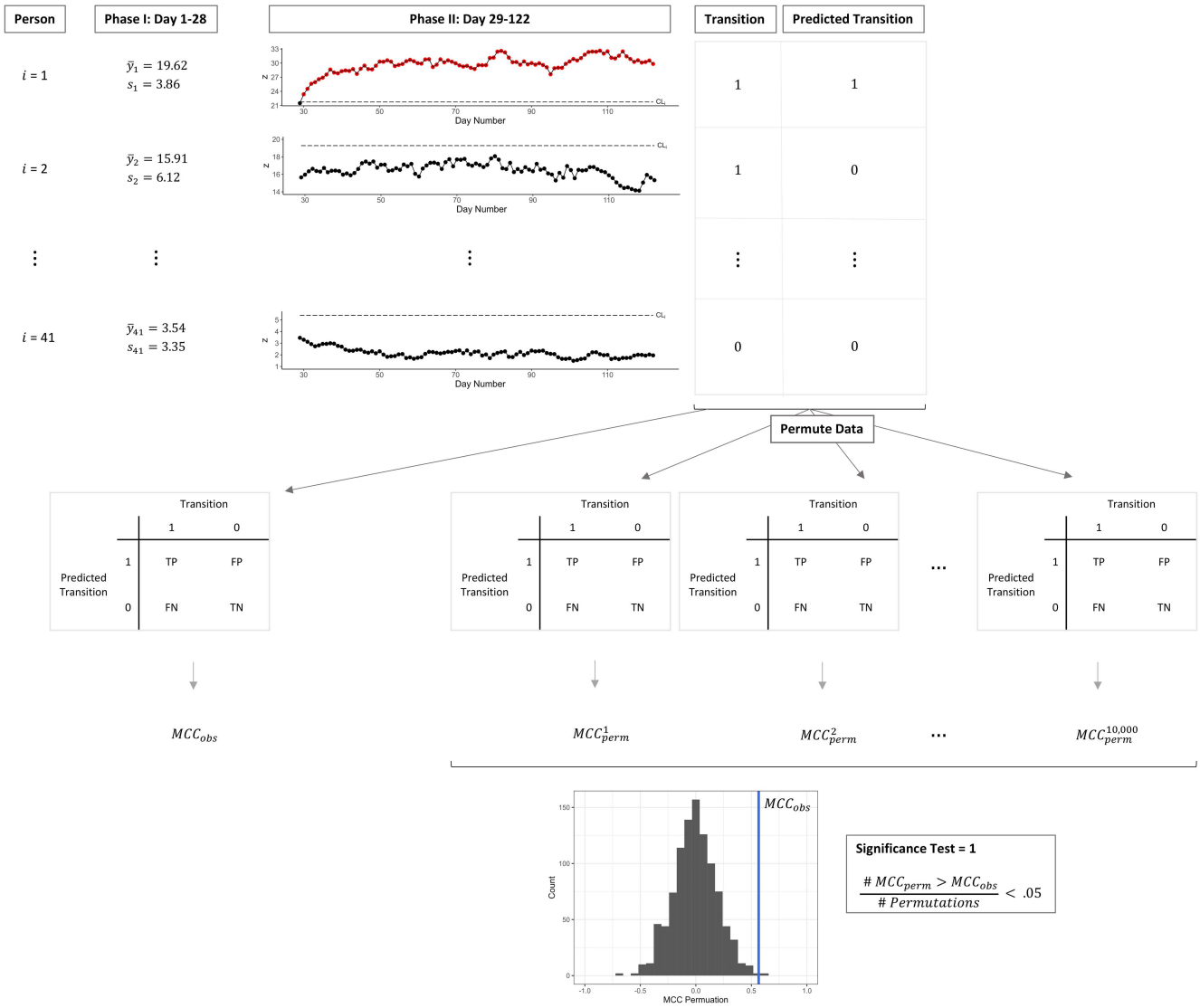
formerly depressed patients who participated in a 4-month ESM study while discontinuing their antidepressant medication. Twenty-two participants experienced a relapse into depression. There were two distinct phases: Phase I (i.e., days 1–28), in which participants were in remission, after which (gradual) continuation of their antidepressant started, followed by Phase II (days 29–122), in which participants were monitored using EWMA. The statistical procedure of the two studies is shown in Figure 4. First, in Phase I, the day averages of days 1–28 were used to obtain the person-specific estimates of the mean ($\bar{y}_i$) and standard deviation (s_i). Next, in Phase II, the day averages of days 29–122 were monitored using the EWMA procedure. In case there was an out-of-control EWMA score, the person was predicted to experience recurrent depression (hereinafter, transition).

Performance Measure

Predictive performance of the EWMA procedure was measured using Matthews correlation coefficient (MCC; Matthews, 1975). This measure ranges from −1 to 1, where an MCC of 1 means perfect prediction and an MCC of 0 means prediction at the chance-level. MCC essentially reflects a correlation between two binary variables (e.g., predicted transition vs. no transition and observed transition vs. no transition). In this case, the observed transition versus no transition is whether a person relapsed into depression or not. A predicted transition is operationalized as the detection of at least one out-of-control EWMA score, whereas a predicted no transition is no out-of-control EWMA score.

Figure 4

Statistical Procedure Used by Snippe et al. (2023) and Schreuder et al. (2024)



Note. TP = true positives; FP = false positives; FN = false negatives; TN = true negatives; MCC = Matthews correlation coefficient. See the online article for the color version of this figure.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Permutation Testing. To evaluate whether the observed MCC values were significantly larger than 0, permutation testing was used in Schreuder et al. (2024). Specifically, the predicted transition values were randomly shuffled and compared to the observed transition, leading to a new MCC value. For 1,000 permutations, a new MCC value was obtained, resulting in a sampling distribution. The p value of the observed MCC value was equal to the number of times the permutation MCC values were larger than the observed MCC value, divided by the total number of permutations.

Simulation Design

In our simulation study, we followed a similar procedure as described above. However, instead of using empirical Phase II data and the presence/absence of a transition, we simulated Phase II data for the 145 persons in Data Set 6. In addition, we simulated the transition itself and its magnitude in terms of a mean change.

First, for each of the 145 individuals, we sampled Phase II day averages of negative affect from a normal distribution, based on the person's known mean and standard deviation. Next, we randomly allocated the 145 individuals to the transition ($N = 73$) or no-transition group ($N = 72$). We set the transition to depression to 1 month (i.e., 31 days) after the start of Phase II,[5] as warning signals can be found up to a month prior to the transition (Smit et al., 2019; Snippe et al., 2023). Given that out of controls after the transition are not of interest, we limited the Phase II period to 31 days.

For those in the transition group, we introduced a mean change (i.e., proxy for depression) with varying sizes of 0, .79, 1.58, or 3.16 σ_{1i} at the day average level (based on 10 measurement occasions per day), which was introduced at the start of Phase II.[6] In other words, in three settings, participants did have a mean change in negative affect (i.e., .79, 1.58, or $3.16\sigma_{1i}$), and in one setting, they did not (i.e., $0\sigma_{1i}$). This resulted in a mean change in negative affect, with sizes of 0, .25, .50, and $1\sigma_{1i}$ in terms of Cohen's d, at the level of individual measurement occasions (Schat, Tuerlinckx, Smit, et al., 2023). The size of the change at day level is larger, as the day averages have a lower variance than the scores on the individual measurement occasions. We used 10 measurement occasions per day, as it is similar to most empirical ESM data sets that we used (see Table 1). For ease of interpretation, we refer to the mean changes in terms of Cohen's d in the remainder of this article. For those in the no-transition group, we did not introduce a mean change, as these individuals reflect a group of healthy individuals (i.e., they all had an effect size of $0\sigma_{1i}$). This procedure (i.e., data simulation, group allocation, introducing the four mean changes to the transition group) was replicated 10,000 times.

In each replication, we applied the EWMA procedure 19 times to the Phase II data of each person. We applied the EWMA procedure with the control limit based on the known parameters (i.e., μ_{1i} and σ_{1i}), multilevel estimates of the leave-one-out case (i.e., $\hat{\gamma}_{0i}$ and $\hat{\sigma}_{\varepsilon i}$), multilevel estimates of the historical case, and person-specific estimates (i.e., $\bar{y}_i$ and s_i). Whereas the first EWMA procedure was applied once, the latter three EWMA procedures were applied six times each, with estimates based on two to seven Phase I days. Finally, for each EWMA procedure, we noted down whether at least one EWMA score was flagged as out of control (1) or not (0). Thus, for each of the 76 combinations of settings (i.e., four mean change sizes × 19 EWMA procedures) and 10,000 replicates, we obtained a 145×2 matrix containing (a) the allocated transition group of the person (i.e., 1 transition, 0 no transition) and (b) whether there was an out-of-control EWMA score (i.e., 1 presence, 0 absence). The R code for the simulation study can be found at https://osf.io/dar5v/.

Performance Measures

We also measured predictive performance of the EWMA procedures using MCC. In our case, the observed transition versus no transition is whether a person is allocated to the transition or no-transition group. A predicted transition is operationalized as the detection of at least one out-of-control EWMA score, whereas a predicted no transition is no out-of-control EWMA score. In total, we obtained 76,000 MCC values (for each replicate of the 76 combinations of settings). Furthermore, we calculated the sensitivity and specificity of the procedures, as well as the true and false positives and negatives.

Permutation Testing. Permutation testing was again used to evaluate whether the obtained MCC values were significantly larger than 0. For each of the 76 combinations of settings, we did the following: First, for each of the 10,000 replicates, we obtained the number of detections that were found (i.e., the number of true positives + false positives). For each unique number of detections (to limit computational burden), we ran 1,000 permutations, in which we randomly allocated the presence (1) or absence (0) of a detection to 145 persons, of whom 73 belonged to the transition group (1) and 72 to the no-transition group (0). For each of the 1,000 permutations, we obtained a new MCC value, resulting in a sampling distribution. Next, we compared the observed MCC value to the sampling distribution based on the number of detections. Finally, we noted the result of the significance test with $\alpha = .05$: 1 if <5% of the permuted MCC values were larger than the observed MCC value, and 0 if ≥5% of the permuted MCC values were larger than the observed MCC value. For a mean change of $0\sigma_{1i}$, the percentage of significant MCC values is ideally low (i.e., 5%),

[5] We conducted additional simulations in which the timing of the transition differed between individuals. Specifically, we divided the individuals with a transition ($N = 73$) into three subgroups with different transition timings: 14 days after the start of Phase II for 24 individuals, 31 days for 24 individuals, and 45 days for the remaining individuals. Similarly, for the no-transition group ($N = 72$), the duration of Phase II varied: 24 individuals had 14 Phase II days, 24 individuals had 31 Phase II days, and 25 individuals had 45 Phase II days. Results are available at https://osf.io/dar5v/. A slight decrease in sensitivity and MCC was observed for mean changes .79 and $1.58\sigma_{1i}$, whereas specificity remained unaffected. The overall patterns across the different approaches (i.e., person-specific and multilevel) remained the same.

[6] We conducted additional simulations in which mean changes (i.e., up to 79, 1.58, or $3.16\sigma_{1i}$) were introduced gradually across the 31-day span of Phase II, as changes in affect can be more continuous. Specifically, the mean change was introduced by incrementally increasing it in 30 equal steps, resulting in a cumulative effect size of .79, 1.58, or $3.16\sigma_{1i}$. Results are available at https://osf.io/dar5v/. Compared to the results for abrupt changes, we observed lower sensitivity, unchanged specificity, and a lower MCC (i.e., fewer TP, more FN, unchanged FP and TN) for a mean change of .79 and $1.58\sigma_{1i}$. For $3.16\sigma_{1i}$, results were the same. This is in line with Smit et al. (2023), who stated that pinpointing the start of the change is often not possible, as the new process is still very similar to the in-control process. It therefore takes longer before changes are detected (i.e., lower sensitivity). However, if the gradual change is large enough or persists over an extended period, it can still be detected. Overall, the patterns across the different approaches (i.e., person-specific and multilevel) also remained the same.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

reflecting the Type I error rate. For the other mean changes, the percentage of significant tests is ideally high, indicating that the method is good at predicting the transition/no-transition groups in terms of MCC.

Results

We first discuss the Type I error rate and power of the EWMA procedures in terms of the percentage of significant MCC values. Next, we evaluate the predictive performance of the EWMA procedures in terms of the MCC values themselves, as well as sensitivity, specificity, and true and false positives and negatives. The results for the multilevel estimates are only shown for the historical case, as they were overall better than the results for the leave-one-out case (i.e., higher sensitivity and MCC values, slightly lower specificity; see https://osf.io/dar5v/).

Type I Error and Power

Figure 5 shows the average percentage of significant MCC values for the EWMA procedure with the control limit based on known parameters, person-specific estimates, and multilevel estimates. Specifically, the values indicate the average over all 10,000 replications. For the known parameters, the average percentage of significant MCC values lies at 5% for the mean change of $0\sigma_{1i}$, reflecting the significance level $\alpha = .05$. The Type I error rate of the other two procedures is higher but decreases as the number of Phase I days increases. The percentage of significant MCC values for the other mean changes (i.e., .25, .50 and $1\sigma_{1i}$) reflects power, which increases as the size of the mean change increases and as the number of Phase I days increases. For the smaller mean change of $.25\sigma_{1i}$, the percentage of significant MCC values is higher for the multilevel estimates than the person-specific estimates. For the mean change of $.50\sigma_{1i}$, the percentages of significant MCC values for the person-specific estimates and multilevel estimates become approximately equal after four Phase I days. At seven Phase I days, the MCC values of the person-specific and multilevel estimates approximate the MCC values of the known parameters. This is also the case for the larger mean change of $1\sigma_{1i}$, where the percentages of significant MCC values of the person-specific and multilevel estimates are similar to those of the known parameters.

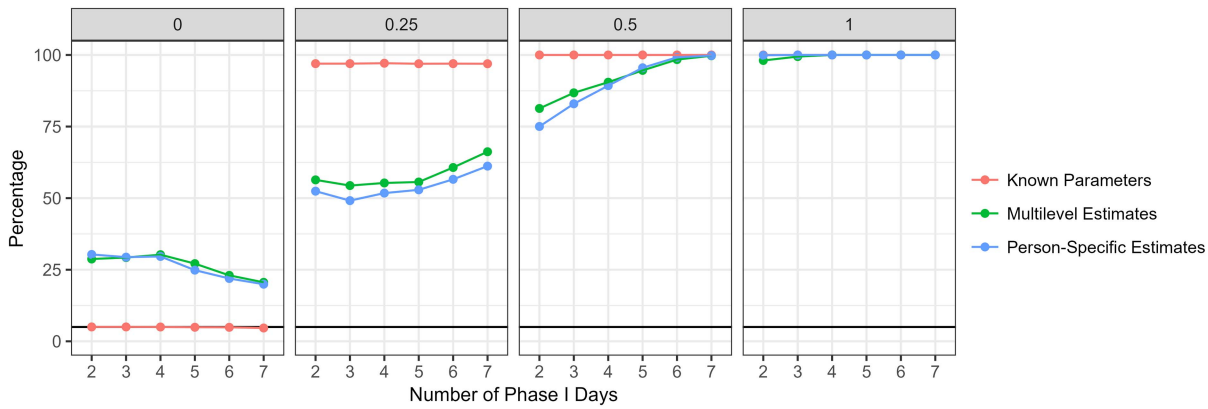
Performance Indices

To further understand the differences in Type I error rate and power between the known parameters, person-specific estimates, and the multilevel estimates, we investigate the different indices for predictive performance. Table 4 shows the average percentage of true and false positives and negatives, and the average MCC, sensitivity, and specificity. The values indicate the average over all 10,000 replications and are shown for mean changes .25 and $.50\sigma_{1i}$. The results for the mean change of $1\sigma_{1i}$ are not shown here as the results do not improve much, but can be found at https://osf.io/dar5v/.

Overall, four things stand out in Table 4. First, the values for the known parameters do not vary over the number of Phase I days because they are based on all 14 days (and thus the variation of the number of Phase I days plays no role). Second, for the person-specific and multilevel estimates, the average MCC and sensitivity values increase as the size of the mean change and the number of Phase I days increase, even though specificity remained relatively constant. Correspondingly, the percentage of true positives increases (without increasing false positives), while the percentage of false negatives decreases (without decreasing true negatives). It follows that, with a mean change of $.50\sigma_{1i}$ and seven Phase I days, one reaches similar performance with the person-specific and multilevel estimates, as one would with the known parameters. Third, EWMA with multilevel estimates is slightly more sensitive (but less specific) than EWMA with person-specific estimates. For the former, the percentage of true positives and false positives is higher, while for the latter, the percentage of true negatives and false negatives is higher. Finally, we observe that the MCC values of the person-specific estimates never clearly surpass those of the multilevel estimates, but are in many conditions worse: Especially for the smaller mean change of $.25\sigma_{1i}$, the MCC values of the multilevel approach were consistently higher. This was also the case for the medium mean change of $.50\sigma_{1i}$ in

Figure 5

Average Percentages of Significant MCC Values



Note. The columns indicate the size of the mean change (σ_{1i}) in terms of Cohen's *d* in the transition group. The solid horizontal line indicates the significance level of α at 5%. MCC = Matthews correlation coefficient. See the online article for the color version of this figure.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 4
Average Percentage of True and False Positives and Negatives, Sensitivity, Specificity, and MCC

Mean change	Phase I days	Known parameter						Person-specific estimate						Multilevel estimate—historical case					
		TP	FP	FN	TN	Sens	Spec	MCC	TP	FP	FN	TN	Sens	Spec	MCC	TP	FP	FN	TN	Sens	Spec	MCC
$.25\sigma_{1i}$	2	49.80	22.81	.55	26.84	.99	.54	.59	29.27	22.02	21.07	27.64	.58	.56	.14	31.67	23.09	18.68	26.56	.63	.53	.16
	3	49.80	22.81	.55	26.84	.99	.54	.59	28.01	22.03	22.34	27.62	.56	.56	.12	31.19	22.93	19.16	26.73	.62	.54	.16
	4	49.80	22.81	.55	26.84	.99	.54	.59	29.12	22.08	21.22	27.58	.58	.56	.14	31.63	22.9	18.72	26.75	.63	.54	.16
	5	49.80	22.81	.55	26.84	.99	.54	.59	29.67	22.03	20.67	27.63	.59	.56	.15	31.43	22.68	18.91	26.98	.62	.54	.17
	6	49.80	22.81	.55	26.84	.99	.54	.59	31.66	22.23	18.69	27.42	.63	.55	.18	33.99	22.79	16.36	26.86	.68	.54	.21
	7	49.80	22.81	.55	26.84	.99	.54	.59	33.97	22.26	16.38	27.40	.67	.55	.22	36.22	22.77	14.13	26.88	.72	.54	.26
$.50\sigma_{1i}$	2	50.34	22.81	0	26.84	1	.54	.61	39.47	22.02	10.87	27.64	.78	.56	.34	40.74	23.09	9.61	26.56	.81	.53	.35
	3	50.34	22.81	0	26.84	1	.54	.61	42.27	22.03	8.07	27.62	.84	.56	.40	43.69	22.93	6.65	26.73	.87	.54	.42
	4	50.34	22.81	0	26.84	1	.54	.61	46.34	22.08	4.01	27.58	.92	.56	.50	46.35	22.9	4.00	26.75	.92	.54	.49
	5	50.34	22.81	0	26.84	1	.54	.61	48.86	22.03	1.49	27.63	.97	.56	.57	48.26	22.68	2.09	26.98	.96	.54	.54
	6	50.34	22.81	0	26.84	1	.54	.61	50.09	22.23	.25	27.42	.99	.55	.61	49.78	22.79	.57	26.86	.99	.54	.59
	7	50.34	22.81	0	26.84	1	.54	.61	50.30	22.26	.04	27.40	1	.55	.61	50.24	22.77	.11	26.88	1	.54	.60

Note. TP = true positives; FP = false positives; FN = false negatives; TN = true negatives; Sens = sensitivity; Spec = specificity; MCC = Matthews correlation coefficient.

combination with less than four Phase I days. From four Phase I days onward, for the mean change of $.50\sigma_{1i}$, the MCC values of the person-specific approach were slightly higher.

Discussion

The EWMA procedure shows great potential for detecting early warning signals of depression in real time and, thus, identifying when someone is at risk of developing depression. However, a primary challenge of applying EWMA in practice lies in the Phase I period, namely obtaining sufficient Phase I data of the person under study. This is needed to obtain accurate estimates of the person's mean and standard deviation in healthy times, such that an accurate control limit can be obtained for monitoring in Phase II. While simulation studies have suggested that at least 50 Phase I days are needed, the practical implementation of such a long Phase I period poses challenges. Based on usual implementations in ESM research (Dejonckheere, Mestdagh, et al., 2021; Koval & Kuppens, 2012), we focused on the scenario in which very little Phase I data are available (i.e., up to 7 days). Nonetheless, the recommendation remains: If you are able to acquire a large amount of Phase I data, it is advisable to do so.

In other fields, historical data have been used to calculate control limits for the subject of interest (e.g., Maselyne et al., 2018; Mertens et al., 2008). Our study aimed to adopt a similar strategy by using ESM data of a large group of healthy individuals. Specifically, we proposed to fit a multilevel model with random intercepts and random Level 1 error variances to historical ESM data (i.e., day averages of negative affect). Next, we obtained estimates of the mean and standard deviation for the person under study by predicting the person's random intercept and Level 1 error variance using a small number of available observations for this person. We used empirical ESM data of six studies to compare the multilevel estimates with person-specific estimates with respect to deriving the control limit in Phase I. Moreover, we compared the accuracy with which these approaches predicted simulated transitions into depression in Phase II. In the remainder of this section, we first discuss the obtained results and provide recommendations. Next, we discuss applications for the multilevel EWMA approach and address some limitations and remaining challenges that warrant further research.

Phase I Parameter Estimates

For estimating within-person means based on varying lengths of Phase I, we saw that multilevel approaches initially outperformed the person-specific approach. After two (historical case) and five (leave-one-out case) Phase I days, however, this reversed. Regarding the standard deviation, the multilevel approaches consistently provided more accurate estimates than the person-specific approach. This illustrates that more data are needed to obtain an accurate estimate of the standard deviation, and thus approaches to estimate standard deviations benefit more from having historical data. Although MSE values decreased as the number of Phase I days increased, there is still room for improvement, and the values would decrease even further by including more Phase I days.

Phase II Predictive Performance

The multilevel and person-specific approaches were further investigated in the simulation study, which focused on predictive

performance in Phase II. Here, we found that both multilevel approaches (historical and leave-one-out) essentially yielded a similar performance. Both multilevel approaches generally outperformed the person-specific approach for detecting small changes (i.e., $.25\sigma_{1i}$), implying that small changes in mental health are more accurately detected when applying control limits that are based on historical data than with control limits that are based on a person's limited available Phase I data. For larger changes in mental health, this only held when fewer than four Phase I days were available.

Applications for the Multilevel EWMA Approach

The multilevel EWMA procedure is particularly useful in mental health applications since individual norms (i.e., a person's control limit) are typically unknown in this domain. In situations where low-cost interventions are applicable and having somewhat more false positives can be dealt with, for instance, by asking a "triage" follow-up question, we recommend using the multilevel approach. In this case, one has a higher chance of targeting the right person at the right time (e.g., finding an out-of-control EWMA score for someone who experiences a transition in mental health) as compared to the person-specific approach.

To illustrate, one type of application is Just-In-Time Adaptive Interventions (Spruijt-Metz & Nilsen, 2014). Such interventions are delivered online (e.g., via smartphone), right when individuals are likely to benefit from them. Ideally, Just-In-Time Adaptive Interventions should not be delivered upon false alarms, as this limits user engagement (Nahum-Shani et al., 2018). One way of limiting the rate of false alarms might be to take the reliability of momentary assessments into account when applying the EWMA procedure, for instance, by repeating certain items (Dejonckheere et al., 2022). One could also check whether there is a different reason (than impendent depression) for the extreme score based on context information. Another way of limiting the impact of false alarms could be to incorporate an opt-out option, giving individuals the choice to respond to or ignore the prompt (Nahum-Shani et al., 2015, 2018).

A second potential low-cost application is patient engagement in therapy. Patients often find it insightful to visualize their data in a time series format, which can be used in therapeutic settings. For example, reviewing specific days when they felt out of control can provide both the patient and their therapist with insights into underlying triggers or patterns. Having one or a few false positives will probably not hinder the detection of such patterns, as long as there are sufficient true positives.

In this study, we focused on mental health applications, specifically depression. However, utilizing historical data to estimate control limits may be useful in a broader scope of applications, such as in behavioral health (e.g., smoking cessation; Soyster & Fisher, 2019), physical health monitoring (e.g., heart pump device; Moazeni et al., 2023), or organizations (e.g., social network dynamics; Perry, 2020).

Directions for Future Research

Although the present study provided valuable insights and recommendations regarding ways to derive the EWMA control limit, there remain some challenges that merit future research. Regarding the empirical data sets that we used, we saw that the mean of the day averages was higher in Data Set 6 (i.e., the target data) compared to Data Sets 1–5 (Table 1). A clear distinction between Data Set 6 and the other data sets is the larger amount and wider range of negative affect items, which may contribute to a higher mean. Indeed, computing the mean per item shows that there is quite some variability between the items (see https://osf.io/dar5v/). Items that are found in most data sets (i.e., "depression," "stress," "anxiety," and "anger") exhibit a slightly higher mean value in Data Set 6. Nevertheless, other items such as "exhaustion" and "fatigue" contribute more to the higher overall mean of negative affect in Data Set 6. The observed differences in the multilevel estimates of the leave-one-out case and historical case were also reflected in the predictive performance of the EWMA procedure. Therefore, it appears that the specific subselection of items does have an impact on predictive performance, a factor that future research should consider.

The simulation results assessing Phase II predictive performance showed that, regardless of the size of the mean change and the number of Phase I days, the percentage of false positives stays constant. The presence of false positives can be linked to the skewed shape of the RL_0 distribution (see Figure 2), where there is a high probability of early out-of-control EWMA scores even though the process is in control (Gan, 1993). This is an inevitable outcome when using the EWMA procedure, and modifications of the parameters (i.e., λ and L) provide only a small workaround for this problem (Snippe et al., 2023). Hence, if the occurrence of false positives does not present significant concerns, using EWMA can be valuable for detecting early warning signals of depression. However, if the intervention carries potential risks, it may be worthwhile to explore alternative methods that are less prone to false positive detections. One could, for instance, adapt the EWMA procedure such that only multiple consecutive out-of-control scores are considered a warning signal rather than a single out-of-control score.

Although the multilevel approach worked best in most cases, the multilevel estimates were still very quickly governed by the person's own observed Phase I day averages. For example, we observed some shrinkage for two Phase I days (Figure 3); however, this quickly decreased as the number of Phase I days increased to 7. This may not be desirable because (a) the limited number of observations may not accurately reflect the person's own distribution in healthy times, and (b) we are not fully utilizing the knowledge we have regarding the range of scores that are considered healthy based on historical data. This could potentially be improved by including an additional type of information, namely person-specific covariates. For example, this may be information regarding demographics, personality traits, emotion regulation, life satisfaction, or self-esteem, which is often also collected in ESM studies (e.g., Dejonckheere, Kalokerinos, et al., 2019; Grommisch et al., 2020; Heininga et al., 2019; Houben et al., 2016; Kuppens et al., 2010; Zetsche et al., 2021). By including covariates in the multilevel model, the between-person variance of the random effects will decrease, and consequently, the ratio of the between-person and within-person variances will change. The multilevel estimates that we obtain for a new person depend on this ratio. Specifically, the speed at which the multilevel estimates are governed by the observed Phase I data depends on this ratio, together with the number of Phase I days. Thus, by including covariates in the model and thus changing the ratio of the variances, we may be able to make fuller use of information in the historical data.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

The multilevel approach is based on the assumption that the sample of individuals is a random sample of the underlying population. Indeed, in the present study, both the to-be-monitored individuals and the historical sample can be considered healthy. Nevertheless, if one aims to monitor individuals with a history of major depressive disorder, it is important to consider that they may inherently possess a higher susceptibility to depression and could potentially have a higher Phase I average compared to individuals without a history of depression. In future research, this can be addressed by including covariates in the multilevel model or by working with historical data from a sample of individuals with a history of major depressive disorder.

In the present study, EWMA was used to detect mean changes in negative affect, as such changes can serve as early warning signals of the onset of depression (Ludwig et al., 2023; Smit & Snippe, 2023; Smit et al., 2019; Snippe et al., 2023). However, the onset of depression can also manifest in other aspects of the data beyond just mean changes, including changes in variability and trends (Cabrieto et al., 2018, 2019; Wichers & Groot, 2016, 2020). Detecting changes in trends is challenging, as SPC is built on the assumption that a process fluctuates around a mean level with a constant variance and that observations are independent. If there is a trend in the data, leading to dependence in the data, it is typically recommended in SPC literature to detrend the data. This can be done, for example, by fitting an autoregressive model (e.g., first-order autoregressive or AR[1] model) to the data and monitoring the residuals (Schat, Tuerlinckx, Smit, et al., 2023). It is also possible to monitor changes in other statistics than the mean by applying EWMA to multiple relevant day statistics, such as day standard deviations (Schat, Tuerlinckx, De Ketelaere, & Ceulemans, 2023). Research has shown that monitoring day standard deviations on top of day averages is beneficial for predicting depression (Schreuder et al., 2024). Future research can expand on the present study by monitoring additional types of day statistics on top of day averages.

Individuals fluctuate in mood and may thus experience a mean change in negative affect that does not necessarily signal impending depression. This change could, for example, be due to events such as exam stress or the loss of a loved one. Integrating such contextual information into the analysis could allow us to more effectively distinguish between changes that are indicative of impendent depression and those that arise from other factors. This can, in principle, be done by employing regression analysis with contextual covariates as predictors and monitoring the residuals. Future research should therefore consider utilizing the contextual information that is typically also collected in ESM studies.

Finally, since we focused on monitoring the day averages of negative affect rather than the individual measurement occasions, we did not account for temporal dependency in the multilevel model. This is because the aggregation of scores within a day reduces the serial dependence considerably (Schat, Tuerlinckx, Smit, et al., 2023). However, when monitoring affect scores at the individual measurement occasion level, this approach may lead to inaccurate estimates. In such cases, it is essential to consider and capture the temporal dependency of the monitored scores, by, for example, assuming the Level 1 errors to follow an autoregressive (e.g., AR[1]) process (Goldstein et al., 1994) or by detrending the data.

Conclusion

The present study demonstrated the potential of using historical ESM data of healthy individuals to help establish an EWMA control limit for a new person under study, for whom only a few days of data are already available, through multilevel modeling. In terms of MCC values, the difference between the person-specific approach and the multilevel approach is small. However, when weighing the trade-off between sensitivity and specificity, the multilevel approach still offers advantages given that it is more sensitive at detecting mean changes. Especially for low-cost and nonharmful interventions, the multilevel approach may therefore prove particularly beneficial. Furthermore, there is potential to extract even more information from the historical data by using person-level covariates, which may further improve the prediction of depression.

References

Agid, O., Shapira, B., Zislin, J., Ritsner, M., Hanin, B., Murad, H., Troudart, T., Bloch, M., Heresco-Levy, U., & Lerer, B. (1999). Environment and vulnerability to major psychiatric illness: A case control study of early parental loss in major depression, bipolar disorder and schizophrenia. *Molecular Psychiatry*, *4*(2), 163–172. https://doi.org/10.1038/sj.mp.4000473

Bigdeli, T. B., Ripke, S., Peterson, R. E., Trzaskowski, M., Bacanu, S. A., Abdellaoui, A., Andlauer, T. F. M., Beekman, A. T. F., Berger, K., Blackwood, D. H. R., Boomsma, D. I., Breen, G., Buttenschøn, H. N., Byrne, E. M., Cichon, S., Clarke, T. K., Couvy-Duchesne, B., Craddock, N., de Geus, E. J. C., … Kendler, K. S. (2017). Genetic effects influencing risk for major depressive disorder in China and Europe. *Translational Psychiatry*, *7*(3), Article e1074. https://doi.org/10.1038/tp.2016.292

Bos, F. M., Snippe, E., Bruggeman, R., Doornbos, B., Wichers, M., & van der Krieke, L. (2020). Recommendations for the use of long-term experience sampling in bipolar disorder care: A qualitative study of patient and clinician experiences. *International Journal of Bipolar Disorders*, *8*(1), Article 38. https://doi.org/10.1186/s40345-020-00201-5

Cabrieto, J., Adolf, J., Tuerlinckx, F., Kuppens, P., & Ceulemans, E. (2018). Detecting long-lived autodependency changes in a multivariate system via change point detection and regime switching models. *Scientific Reports*, *8*(1), Article 15637. https://doi.org/10.1038/s41598-018-33819-8

Cabrieto, J., Adolf, J., Tuerlinckx, F., Kuppens, P., & Ceulemans, E. (2019). An objective, comprehensive and flexible statistical framework for detecting early warning signs of mental health problems. *Psychotherapy and Psychosomatics*, *88*(3), 184–186. https://doi.org/10.1159/000494356

Cloos, L., Ceulemans, E., & Kuppens, P. (2023). Development, validation, and comparison of self-report measures for positive and negative affect in intensive longitudinal research. *Psychological Assessment*, *35*(3), 189–204. https://doi.org/10.1037/pas0001200

Craddock, N., & Forty, L. (2006). Genetics of affective (mood) disorders. *European Journal of Human Genetics*, *14*(6), 660–668. https://doi.org/10.1038/sj.ejhg.5201549

Curtiss, J. E., Mischoulon, D., Fisher, L. B., Cusin, C., Fedor, S., Picard, R. W., & Pedrelli, P. (2023). Rising early warning signals in affect associated with future changes in depression: A dynamical systems approach. *Psychological Medicine*, *53*(7), 3124–3132. https://doi.org/10.1017/S0033291721005183

Davidson, R. J., Jackson, D. C., & Kalin, N. H. (2000). Emotion, plasticity, context, and regulation: Perspectives from affective neuroscience. *Psychological Bulletin*, *126*(6), 890–909. https://doi.org/10.1037/0033-2909.126.6.890

Dejonckheere, E., Demeyer, F., Geusens, B., Piot, M., Tuerlinckx, F., Verdonck, S., & Mestdagh, M. (2022). Assessing the reliability of single-item momentary affective measurements in experience sampling. *Psychological Assessment*, *34*(12), 1138–1154. https://doi.org/10.1037/pas0001178

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Dejonckheere, E., Houben, M., Schat, E., Ceulemans, E., & Kuppens, P. (2021). The short-term psychological impact of the COVID-19 pandemic in psychiatric patients: Evidence for differential emotion and symptom trajectories in Belgium. *Psychologica Belgica*, *61*(1), 163–172. https://doi.org/10.5334/pb.1028

Dejonckheere, E., Kalokerinos, E. K., Bastian, B., & Kuppens, P. (2019). Poor emotion regulation ability mediates the link between depressive symptoms and affective bipolarity. *Cognition and Emotion*, *33*(5), 1076–1083. https://doi.org/10.1080/02699931.2018.1524747

Dejonckheere, E., Mestdagh, M., Houben, M., Erbas, Y., Pe, M., Koval, P., Brose, A., Bastian, B., & Kuppens, P. (2018). The bipolarity of affect and depressive symptoms. *Journal of Personality and Social Psychology*, *114*(2), 323–341. https://doi.org/10.1037/pspp0000186

Dejonckheere, E., Mestdagh, M., Houben, M., Rutten, I., Sels, L., Kuppens, P., & Tuerlinckx, F. (2019). Complex affect dynamics add limited information to the prediction of psychological well-being. *Nature Human Behaviour*, *3*(5), 478–491. https://doi.org/10.1038/s41562-019-0555-0

Dejonckheere, E., Mestdagh, M., Verdonck, S., Lafit, G., Ceulemans, E., Bastian, B., & Kalokerinos, E. K. (2021). The relation between positive and negative affect becomes more negative in response to personally relevant events. *Emotion*, *21*(2), 326–336. https://doi.org/10.1037/emo0000697

Eisele, G., Vachon, H., Lafit, G., Kuppens, P., Houben, M., Myin-Germeys, I., & Viechtbauer, W. (2022). The effects of sampling frequency and questionnaire length on perceived burden, compliance, and careless responding in experience sampling data in a student population. *Assessment*, *29*(2), 136–151. https://doi.org/10.1177/1073191120957102

Frijda, N. H. (2007). *The laws of emotion* (1st ed.). Psychology Press. https://doi.org/10.4324/9781315086071

Gan, F. F. (1993). An optimal design of EWMA control charts based on median run length. *Journal of Statistical Computation and Simulation*, *45*(3–4), 169–184. https://doi.org/10.1080/00949659308811479

Goldstein, H., Healy, M. J., & Rasbash, J. (1994). Multilevel time series models with applications to repeated measures data. *Statistics in Medicine*, *13*(16), 1643–1655. https://doi.org/10.1002/sim.4780131605

Grommisch, G., Koval, P., Hinton, J. D. X., Gleeson, J., Hollenstein, T., Kuppens, P., & Lischetzke, T. (2020). Modeling individual differences in emotion regulation repertoire in daily life with multilevel latent profile analysis. *Emotion*, *20*(8), 1462–1474. https://doi.org/10.1037/emo0000669

Groot, P. C. (2010). Patients can diagnose too: How continuous self-assessment aids diagnosis of, and recovery from, depression. *Journal of Mental Health*, *19*(4), 352–362. https://doi.org/10.3109/09638237.2010.494188

Heim, C., Plotsky, P. M., & Nemeroff, C. B. (2004). Importance of studying the contributions of early adverse experience to neurobiological findings in depression. *Neuropsychopharmacology*, *29*(4), 641–648. https://doi.org/10.1038/sj.npp.1300397

Heininga, V. E., Dejonckheere, E., Houben, M., Obbels, J., Sienaert, P., Leroy, B., van Roy, J., & Kuppens, P. (2019). The dynamical signature of anhedonia in major depressive disorder: Positive emotion dynamics, reactivity, and recovery. *BMC Psychiatry*, *19*(1), Article 59. https://doi.org/10.1186/s12888-018-1983-5

Helmich, M. A., Wichers, M., Olthof, M., Strunk, G., Aas, B., Aichhorn, W., Schiepek, G., & Snippe, E. (2020). Sudden gains in day-to-day change: Revealing nonlinear patterns of individual improvement in depression. *Journal of Consulting and Clinical Psychology*, *88*(2), 119–127. https://doi.org/10.1037/ccp0000469

Houben, M., Claes, L., Vansteelandt, K., Berens, A., Sleuwaegen, E., & Kuppens, P. (2017). The emotion regulation function of nonsuicidal self-injury: A momentary assessment study in inpatients with borderline personality disorder features. *Journal of Abnormal Psychology*, *126*(1), 89–95. https://doi.org/10.1037/abn0000229

Houben, M., Van Den Noortgate, W., & Kuppens, P. (2015). The relation between short-term emotion dynamics and psychological well-being: A meta-analysis. *Psychological Bulletin*, *141*(4), 901–930. https://doi.org/10.1037/a0038822

Houben, M., Vansteelandt, K., Claes, L., Sienaert, P., Berens, A., Sleuwaegen, E., & Kuppens, P. (2016). Emotional switching in borderline personality disorder: A daily life study. *Personality Disorders*, *7*(1), 50–60. https://doi.org/10.1037/per0000126

Klein, D. N., Kotov, R., & Bufferd, S. J. (2011). Personality and depression: Explanatory models and review of the evidence. *Annual Review of Clinical Psychology*, *7*(1), 269–295. https://doi.org/10.1146/annurev-clinpsy-032210-104540

Knoth, S. (2020). *spc: Statistical process control—Calculation of ARL and other control chart performance measures* (0.6.4). https://cran.r-project.org/package=spc

Koval, P., & Kuppens, P. (2012). Changing emotion dynamics: Individual differences in the effect of anticipatory social stress on emotional inertia. *Emotion*, *12*(2), 256–267. https://doi.org/10.1037/a0024756

Kuppens, P. (2015). Its about time: A special section on affect dynamics. *Emotion Review*, *7*(4), 297–300. https://doi.org/10.1177/1754073915590947

Kuppens, P., Oravecz, Z., & Tuerlinckx, F. (2010). Feelings change: Accounting for individual differences in the temporal dynamics of affect. *Journal of Personality and Social Psychology*, *99*(6), 1042–1060. https://doi.org/10.1037/a0020962

Larsen, R. J. (2000). Toward a science of mood regulation. *Psychological Inquiry*, *11*(3), 129–141. https://doi.org/10.1207/S15327965PLI1103_01

Levinson, D. F. (2006). The genetics of depression: A review. *Biological Psychiatry*, *60*(2), 84–92. https://doi.org/10.1016/j.biopsych.2005.08.024

Ludwig, V. M., Reinhard, I., Mühlbauer, E., Hill, H., Severus, E., Bauer, M., Ritter, P., & Ebner-Priemer, U. W. (2023). *Limited evidence of autocorrelation signaling upcoming affective episodes: A 12-month e-diary study in patients with bipolar disorder*. PsyArXiv. https://doi.org/10.31234/osf.io/8ebmc

Maselyne, J., Van Nuffel, A., Briene, P., Vangeyte, J., De Ketelaere, B., Millet, S., Van den Hof, J., Maes, D., & Saeys, W. (2018). Online warning systems for individual fattening pigs based on their feeding pattern. *Biosystems Engineering*, *173*, 143–156. https://doi.org/10.1016/j.biosystemseng.2017.08.006

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *BBA—Protein Structure*, *405*(2), 442–451. https://doi.org/10.1016/0005-2795(75)90109-9

Mertens, K., Vaesen, I., Löffel, J., Ostyn, B., Kemps, B., Kamers, B., Bamelis, F., Zoons, J., Darius, P., Decuypere, E., De Baerdemaeker, J., & De Ketelaere, B. (2008). Data-based design of an intelligent control chart for the daily monitoring of the average egg weight. *Computers and Electronics in Agriculture*, *61*(2), 222–232. https://doi.org/10.1016/j.compag.2007.11.010

Moazeni, M., Numan, L., Brons, M., Houtgraaf, J., Rutten, F. H., Oberski, D. L., van Laake, L. W., Asselbergs, F. W., & Aarts, E. (2023). Developing a personalized remote patient monitoring algorithm: A proof-of-concept in heart failure. *Digital Health*, *4*(6), 455–463. https://doi.org/10.1093/ehjdh/ztad049

Montgomery, D. C. (2009). *Introduction to statistical quality control* (8th ed.). Wiley.

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., 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

Nahum-Shani, I., Hekler, E. B., & Spruijt-Metz, D. (2015). Building health behavior models to guide the development of just-in-time adaptive

interventions: A pragmatic framework. *Health Psychology*, *34S*(Suppl.), 1209–1219. https://doi.org/10.1037/hea0000306

Nahum-Shani, I., Smith, S. N., Spring, B. J., Collins, L. M., Witkiewitz, K., Tewari, A., & Murphy, S. A. (2018). Just-in-time adaptive interventions (JITAIs) in mobile health: Key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine*, *52*(6), 446–462. https://doi.org/10.1007/s12160-016-9830-8

Nelson, J., Klumparendt, A., Doebler, P., & Ehring, T. (2020). Everyday emotional dynamics in major depression. *Emotion*, *20*(2), 179–191. https://doi.org/10.1037/emo0000541

Nestler, S., & Humberg, S. (2022). A lasso and a regression tree mixed-effect model with random effects for the level, the residual variance, and the autocorrelation. *Psychometrika*, *87*(2), 506–532. https://doi.org/10.1007/s11336-021-09787-w

Olthof, M., Hasselman, F., Strunk, G., van Rooij, M., Aas, B., Helmich, M. A., Schiepek, G., & Lichtwarck-Aschoff, A. (2020). Critical fluctuations as an early-warning signal for sudden gains and losses in patients receiving psychotherapy for mood disorders. *Clinical Psychological Science*, *8*(1), 25–35. https://doi.org/10.1177/2167702619865969

Pe, M. L., Brose, A., Gotlib, I. H., & Kuppens, P. (2016). Affective updating ability and stressful events interact to prospectively predict increases in depressive symptoms over time. *Emotion*, *16*(1), 73–82. https://doi.org/10.1037/emo0000097

Perry, M. B. (2020). An EWMA control chart for categorical processes with applications to social network monitoring. *Journal of Quality Technology*, *52*(2), 182–197. https://doi.org/10.1080/00224065.2019.1571343

Provenzano, J., Fossati, P., Dejonckheere, E., Verduyn, P., & Kuppens, P. (2021). Inflexibly sustained negative affect and rumination independently link default mode network efficiency to subclinical depressive symptoms. *Journal of Affective Disorders*, *293*(June), 347–354. https://doi.org/10.1016/j.jad.2021.06.051

Roberts, S. W. (1959). Control chart tests based on geometric moving averages. *Technometrics*, *1*(3), 239–250. https://doi.org/10.1080/00401706.1959.10489860

Schat, E., Tuerlinckx, F., De Ketelaere, B., & Ceulemans, E. (2023). Real-time detection of mean and variance changes in experience sampling data: A comparison of existing and novel statistical process control approaches. *Behavior Research Methods*. Advance online publication. https://doi.org/10.3758/s13428-023-02103-7

Schat, E., Tuerlinckx, F., Smit, A. C., de Ketelaere, B., & Ceulemans, E. (2023). Detecting mean changes in experience sampling data in real time: A comparison of univariate and multivariate statistical process control methods. *Psychological Methods*, *28*(6), 1335–1357. https://doi.org/10.1037/met0000447

Schreuder, M. J., Groen, R. N., Wigman, J. T. W., Hartman, C. A., & Wichers, M. (2020). Measuring psychopathology as it unfolds in daily life: Addressing key assumptions of intensive longitudinal methods in the TRAILS TRANS-ID study. *BMC Psychiatry*, *20*(1), Article 351. https://doi.org/10.1186/s12888-020-02674-1

Schreuder, M. J., Kuppens, P., Schat, E., de Jonge, P., Hartman, C. A., & Ceulemans, E. (2023). *Can warning signals for mental health problems in at-risk young adults be informed by momentary emotions reported by the general population? A novel application of the principles of statistical process control* [Manuscript submitted for publication].

Schreuder, M. J., Schat, E., Smit, A. C., Snippe, E., & Ceulemans, E. (2024). Monitoring emotional intensity and variability to forecast depression recurrence in real time in remitted adults. *Journal of Consulting and Clinical Psychology*. Advance online publication. https://doi.org/10.1037/ccp0000871

Sels, L., Ceulemans, E., & Kuppens, P. (2017). Partner-expected affect: How you feel now is predicted by how your partner thought you felt before. *Emotion*, *17*(7), 1066–1077. https://doi.org/10.1037/emo0000304

Shewhart, W. A. (1931). *Economic control of quality of manufactured product*. Macmillan.

Smit, A. C., Helmich, M. A., Bringmann, L. F., Oldehinkel, A. J., Wichers, M., & Snippe, E. (n.d.). *Personalized detection of impending symptom transitions in depression using early warning signals during and shortly after antidepressant discontinuation* [Manuscript submitted for publication].

Smit, A. C., Schat, E., & Ceulemans, E. (2023). The exponentially weighted moving average procedure for detecting changes in intensive longitudinal data in psychological research in real-time: A tutorial showcasing potential applications. *Assessment*, *30*(5), 1354–1368. https://doi.org/10.1177/10731911221086985

Smit, A. C., & Snippe, E. (2023). Real-time monitoring of increases in restlessness to assess idiographic risk of recurrence of depressive symptoms. *Psychological Medicine*, *53*(11), 5060–5069. https://doi.org/10.1017/S0033291722002069

Smit, A. C., Snippe, E., Kunkels, Y. K., Riese, H., Helmich, M. A., & Wichers, M. (2020). *Transitions in depression (TRANS-ID) tapering*. https://www.osf.io/h75p9

Smit, A. C., Snippe, E., & Wichers, M. (2019). Increasing restlessness signals impending increase in depressive symptoms more than 2 months before it happens in individual patients. *Psychotherapy and Psychosomatics*, *88*(4), 249–251. https://doi.org/10.1159/000500594

Snippe, E., Smit, A. C., Kuppens, P., Burger, H., & Ceulemans, E. (2023). Recurrence of depression can be foreseen by monitoring mental states with statistical process control. *Journal of Psychopathology and Clinical Science*, *132*(2), 145–155. https://doi.org/10.1037/abn0000812

Soyster, P. D., & Fisher, A. J. (2019). Involving stakeholders in the design of ecological momentary assessment research: An example from smoking cessation. *PLOS ONE*, *14*(5), Article e0217150. https://doi.org/10.1371/journal.pone.0217150

Spruijt-Metz, D., & Nilsen, W. (2014). Dynamic models of behavior for just-in-time adaptive interventions. *IEEE Pervasive Computing*, *13*(3), 13–17. https://doi.org/10.1109/MPRV.2014.46

Thompson, R. J., Boden, M. T., & Gotlib, I. H. (2017). Emotional variability and clarity in depression and social anxiety. *Cognition and Emotion*, *31*(1), 98–108. https://doi.org/10.1080/02699931.2015.1084908

Wichers, M., Groot, P. C., Psychosystems, ESM Group, & EWS Group. (2016). Critical slowing down as a personalized early warning signal for depression. *Psychotherapy and Psychosomatics*, *85*(2), 114–116. https://doi.org/10.1159/000441458

Wichers, M., Smit, A. C., & Snippe, E. (2020). Early warning signals based on momentary affect dynamics can expose nearby transitions in depression: A confirmatory single-subject time-series study. *Journal for Person-Oriented Research*, *6*(1), 1–15. https://doi.org/10.17505/jpor.2020.22042

Wright, A. G. C., & Woods, W. C. (2020). Personalized models of psychopathology. *Annual Review of Clinical Psychology*, *16*(1), 49–74. https://doi.org/10.1146/annurev-clinpsy-102419-125032

Zetsche, U., Bürkner, P. C., Bohländer, J., Renneberg, B., Röpke, S., & Schulze, L. (2021). *Daily affect regulation in borderline personality disorder and major depression*. *PsyArXiv*. https://journal.um-surabaya.ac.id/index.php/JKM/article/view/2203

Received June 17, 2023
Revision received December 20, 2023
Accepted February 8, 2024 ■

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.