Check for updates

General Article

Advances in Methods and Practices in Psychological Science
October-December 2025, Vol. 8, No. 4, pp. 1–24
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/25152459251372116
www.psychologicalscience.org/AMPPS

Meeting the Bare Minimum: Quality Assessment of Idiographic Temporal Networks Using Power Analysis and Predictive-Accuracy Analysis

Yong Zhang[1], Jordan Revol[2], Ginette Lafit[3], Anja F. Ernst[1], Josip Razum[4], Eva Ceulemans[2], and Laura F. Bringmann[1]

[1]Department Psychometrics and Statistics, University of Groningen, Groningen, the Netherlands; [2]Research Group of Quantitative Psychology and Individual Differences, KU Leuven, Leuven, Belgium; [3]Methodology of Educational Sciences Research Group, KU Leuven, Leuven, Belgium; and [4]Ivo Pilar Institute of Social Sciences, Zagreb, Croatia

Abstract

The network theory of psychopathology inspired clinicians and researchers to use idiographic networks to study how symptoms of an individual interact over time, hoping to find the target symptom(s) for intervention to most effectively break this self-sustaining network. These networks are often based on the vector-autoregressive (VAR) model and rely on intensive longitudinal data collected in patients' daily lives. Nowadays, one major challenge these networks are faced with is that they are used without sufficient quality assessments. Because VAR-based temporal networks are complex and highly parameterized, they can easily face problems of low statistical power and overfitting, especially when the time series available is short. In this study, we review existing idiographic-network studies with a focus on the number of variables and time points used in the analysis and show that the "big network, short time series" problem is prevalent. As potential solutions, we propose two simulation-based methods that aim to find the optimal number of time points to be collected: power analysis and predictive-accuracy analysis. Two applications of both methods are demonstrated: (a) "a priori"—informing the sample-size planning of future network studies and (b) "retrospective"—evaluating whether the sample size of existing network studies was large enough to avoid problems of low statistical power and overfitting. Results confirmed the observation that the sample sizes in past network studies are often insufficient, suggesting that findings of existing network studies should be critically assessed. Future idiographic-network studies are thus strongly advised to make more guided decisions on sample size using the proposed methods.

Keywords

idiographic network, vector-autoregressive modeling, overfitting, sample size, simulation, open data, open materials

Received 5/29/24; Revision accepted 6/24/25

The network theory of psychopathology considers the development of mental disorders as the result of interactions among symptoms over time (Borsboom, 2017; Borsboom & Cramer, 2013; Cramer et al., 2010). Rather than driven by a neurobiological common cause, symptoms sustain each other and eventually form a chronic disordered equilibrium (Borsboom et al., 2019). Since the introduction of the network theory, more and more researchers and clinicians have focused on studying mental-health disorders on the within-persons level through collecting intensive longitudinal data (ILD). ILD allow the exploration of the intraindividual variation of

Corresponding Author:
Yong Zhang, Department of Psychometrics and Statistics, Faculty of Behavioural and Social Sciences, University of Groningen, Groningen, the Netherlands.
Email: yong.zhang@rug.nl

Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (https://creativecommons.org/licenses/by/4.0/) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

momentary experiences (Hamaker, 2012; Molenaar, 2004) and how such experiences interact over time (Wichers, 2014).

To capture such temporal dynamics in a patient's daily life, ILD can be collected using methods such as experience-sampling methods (ESMs; Larson & Csikszentmihalyi, 2014) and ecological-momentary assessments (EMAs; Smyth & Stone, 2003). The methodological tool of idiographic (or person-specific) temporal networks has been developed for exploring the symptom interactions in ILD and has been increasingly used in clinical research and practice in recent years (e.g., Bringmann, 2021; Schemer et al., 2023; von Klipstein et al., 2020). In an idiographic temporal network, nodes represent the analyzed variables (e.g., emotions and symptoms) and are connected with directed edges (arrows). Take the hypothetical network in Bringmann (2021) as an example (Fig. 1): The arrow pointing from sad mood to sleep problems suggests that the sleep problems of this person can be predicted by the person's preceding sad mood. Such a network can be a helpful exploratory tool for clinicians to establish potential causal links among one specific patient's emotion, cognition, and physical problems. This can eventually benefit the process of case conceptualization (e.g., de Vos et al., 2017; Frumkin et al., 2021; Hall et al., 2025) and the design of personalized treatment plans (e.g., Levinson et al., 2021; Piccirillo & Rodebaugh, 2022). To summarize, the idiographic analysis aims to describe the within-persons psychological processes of one individual. This differs from the traditional nomothetic approach (i.e., multiple participants being measured once) used in psychological research that concerns only between-persons differences (Hamaker & Wichers, 2017; Molenaar, 2004).

To generate an idiographic temporal network, the lag-1 vector-autoregressive model (VAR[1]; Brandt & Williams, 2007; Lütkepohl, 2005) needs to be fitted to the ILD gathered from an individual. The VAR(1) model predicts the value of each variable at a given time point using the values of all variables in the system at the previous time point (i.e., the lag-1 value). Two types of effects are of primary interest in a VAR(1) model and become visualized in the network: the autoregressive and cross-regressive effects. The autoregressive effect refers to how the value of one variable is related to the lagged value of the same variable (e.g., the arrow pointing from sad mood to itself in Fig. 1), whereas the cross-regressive effect shows how well one variable can be predicted by the lagged value of another variable. Both types of effects represent the unique predictive value of each lagged variable.

To make meaningful interpretations of an idiographic temporal network, satisfactory quality of the VAR(1) model is necessary. However, this is often taken for granted and not examined thoroughly (Vogelsmeier et al., 2024). Therefore, our main goal in this article is to present methods that can be useful for quality assessment of the VAR(1)-based network and discuss practical methodological considerations for ensuring model quality.

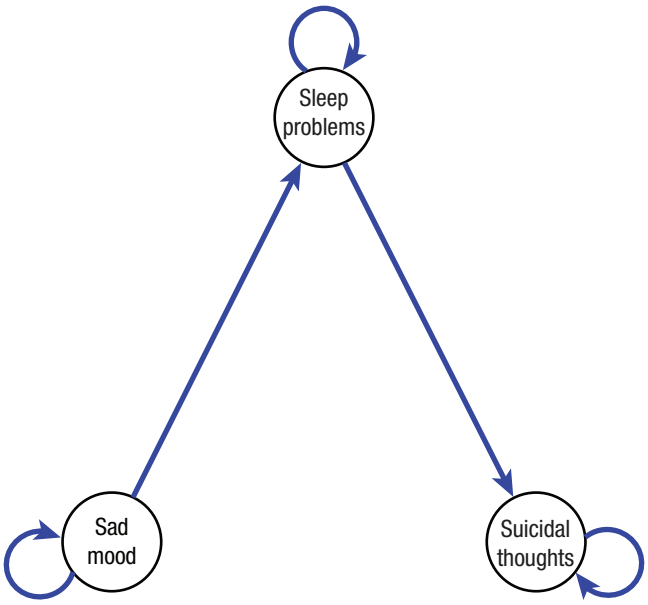

Fig. 1. A hypothetical network in Bringmann (2021).

Statistical models have the dual functionality of explanation (e.g., testing the hypothesized association among variables) and prediction (e.g., predicting the outcome variable's value based on predictors' values for an unseen observation; Shmueli, 2010). Therefore, their quality assessments should also involve both aspects. From the explanatory-modeling perspective, a key factor of model quality is the statistical power of the conducted significance tests—the probability of yielding statistically significant results when the effect of interest is true (Cohen, 1992). Significance tests with low statistical power cannot reliably detect the meaningful relationships among variables across samples and will thus hurt the model's explanatory capacity. Because power analysis is usually conducted for the significance test of one effect at a time, it is of limited value when assessing the quality of a multivariate model (Mulder, 2022; Wang & Rhemtulla, 2021). In contrast, for predictive modeling, the generalizability of the model is crucial: Can a model estimated with a time series of an individual also make accurate predictions of data observed in other similar time periods for the same individual? One common reason why this predictability is not realized is overfitting: The model mistakenly treats sample-specific noises as generalizable signals and eventually yields unreplicable estimates that provide researchers with overly an optimistic impression of the

model's performance (Kuhn & Johnson, 2013). Based on this definition, an overfitted model is unlikely to uncover and explain the true data-generating process well. This suggests that the two functionalities, explanation and prediction, should not be seen as completely disconnected from each other (Hofman et al., 2021; Rocca & Yarkoni, 2021).

Complex models with a large number of parameters (e.g., networks) are known to have a strong tendency to overfit, especially when fitted to a small sample (i.e., a short time series; Bulteel et al., 2018; Yarkoni & Westfall, 2017). Clinicians would hope that the idiographic networks they build for their patients are generalizable, implying that the network can accurately describe the patients' temporal dynamics during both the data collection and a similar time period. This way, the resulting personalized treatment plans are not guided by random noises. Careless usage of such networks without sufficient quality assessment could lead to false conclusions of how symptoms interact and can be misleading for the patients. However, overfitting is not a problem that can be instantly spotted. Overfitting will become visible only through certain validation procedures in which the estimated network gets applied to another sample and evaluated on its predictive accuracy: whether it can make accurate predictions with minimal errors for data in this other sample.

Predictive accuracy, therefore, should be considered as an important quality index of a statistical model that indicates a model's risk of overfitting. Fortunately, this quality index is gradually attracting more attention from psychological researchers (Rocca & Yarkoni, 2021; Verhagen, 2022), especially in the field of time-series analysis (Lafit et al., 2022; Loossens, Dejonckheere, et al., 2021; Loossens, Tuerlinckx, & Verdonck, 2021; Revol et al., 2024). Two studies provided first indications that the generalizability and predictive accuracy of VAR(1) networks are not satisfactory (e.g., Bulteel et al., 2018; Mansueto et al., 2023). Bulteel et al. (2018) showed that when fitted to short time series generated under a VAR(1) process, the correctly specified VAR(1) model still had a lower predictive accuracy than simpler models because of overfitting. Such findings clearly show the importance of carefully planning the number of time points to be collected from one individual to ensure sufficient predictive accuracy of the VAR(1) networks. It is thus crucial to apply methodologies that can calculate the exact number of time points required for a temporal network to be accurately retrieved with minimal risk of overfitting. Ideally, such methods are used before data collection for sample-size planning. Yet for existing network studies, the methods could also be very useful for evaluating the risk of overfitting for estimated networks and helping researchers interpret the networks with a critical perspective.

In the remainder of this article, we start by providing a detailed description of the idiographic VAR(1) model for the readers to become more familiar with the foundation of a temporal network. Then, we discuss two simulation-based sample-size planning methods for VAR(1) networks recently developed by Revol et al. (2024): (a) power analysis of individual edges and (b) predictive-accuracy analysis of the entire network. Using a set of hypothetical network parameters, we further demonstrate how to use both methods for a priori sample-size planning. Then, we show how both methods can be used retrospectively to assess the quality of idiographic networks estimated in existing studies. For this section, we begin by presenting a review of existing idiographic-network studies, focusing on the number of nodes and time points used and providing an overview of the current "big network, short time series" problem in the field. We then apply both methods to networks estimated in Bak et al. (2016) and Epskamp, van Borkulo, et al. (2018) to show whether the number of time points used in both studies was large enough to ensure sufficient power for the significance test of individual edges and network predictive accuracy. Finally, we provide a summary of the findings and discuss the future directions of improving the sample-size optimization methods and the usage of idiographic networks.

Method

The VAR(1) model and temporal network

Model specification and assumptions. A typical ESM study aiming to build an idiographic temporal network for a single patient assesses M variables, which are measured repeatedly over T time points. The value of variable y_p ($p = 1, 2, 3, \ldots, M$) at a given time point t ($t = 1, 2, 3, \ldots, T$) is denoted as $y_{p,t}$. The VAR(1) model is a multivariate model that can be seen as consisting of M separate linear-regression models. Each of the M variables at a given time point (i.e., t) is predicted by the values of all M variables at the previous time point (i.e., $t-1$). The matrix notation of a VAR(1) model is thus as follows:

$$\boldsymbol{y_t} = \boldsymbol{\Delta} + \boldsymbol{\Phi y_{t-1}} + \boldsymbol{\epsilon_t}, \tag{1}$$

where $\boldsymbol{y_t}$ and $\boldsymbol{y_{t-1}}$ are vectors containing the values of all M variables at time point t and $t-1$, respectively. $\boldsymbol{\Delta}$ is the $M \times 1$ vector with the values of all the intercepts in the M linear regressions. Φ is the $M \times M$ matrix containing all the autoregressive/cross-regressive coefficients (i.e., the network edges), representing the lagged relationships among variables. $\boldsymbol{\epsilon_t}$ is the vector for the residuals or innovations[1] in all regression models, representing the variation of all variables at the same time

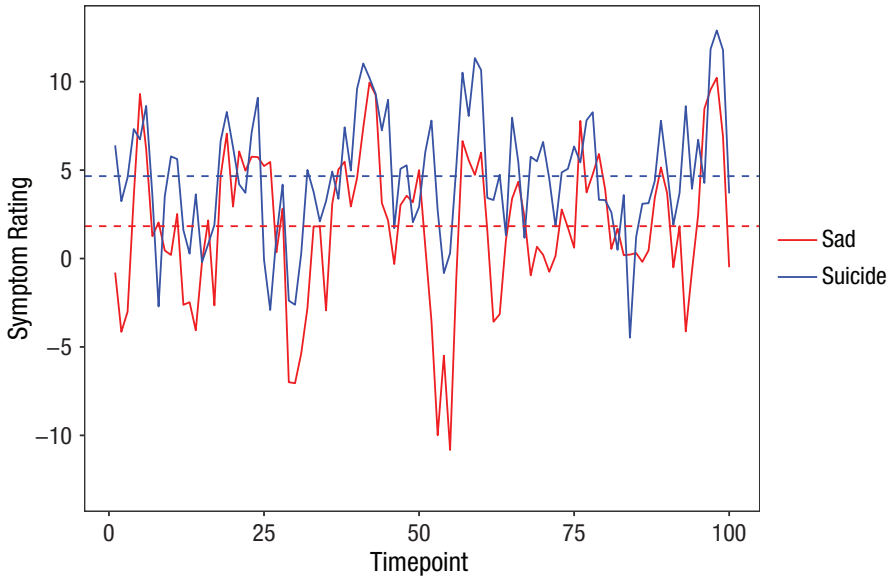

Fig. 2. A simulated bivariate time series of variables *Sad* and *Suicide*. The dashed lines denote the mean value of both variables in this time series.

point that cannot be explained by the lagged effects. The VAR(1) model assumes that the innovation terms follow a multivariate normal distribution with a mean vector of $\mathbf{0}$ and a covariance matrix Σ (Brandt & Williams, 2007). Moreover, the innovation terms are white-noise time series with zero autocorrelation or cross-correlation (Hamilton, 1994).

For example, to study the temporal relationships between sad mood and suicidal thoughts for a patient diagnosed with major depressive disorder, the following VAR(1) model can be employed:

$$\begin{aligned}\begin{bmatrix} Sad_t \\ Suicide_t \end{bmatrix} &= \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix} + \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix} \begin{bmatrix} Sad_{t-1} \\ Suicide_{t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{Sad,t} \\ \epsilon_{Suicide,t} \end{bmatrix} \\ &= \begin{bmatrix} 0 \\ 2 \end{bmatrix} + \begin{bmatrix} .5 & .3 \\ .2 & .4 \end{bmatrix} \begin{bmatrix} Sad_{t-1} \\ Suicide_{t-1} \end{bmatrix} + \begin{bmatrix} \epsilon_{Sad,t} \\ \epsilon_{Suicide,t} \end{bmatrix}, \end{aligned} \quad (2)$$

where

$$\begin{aligned}\begin{bmatrix} \epsilon_{Sad,t} \\ \epsilon_{Suicide,t} \end{bmatrix} &\sim \mathcal{N}\left(\begin{bmatrix} E\left(\epsilon_{Sad,t}\right) \\ E\left(\epsilon_{Suicide,t}\right) \end{bmatrix}, \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} \right) \\ &\sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 10 & 4 \\ 4 & 10 \end{bmatrix} \right). \end{aligned} \quad (3)$$

In this example, the autoregressive coefficient $\phi_{11} = .5$ represents the relationship between the current and lag-1 values of *Sad* after controlling for all other lag-1 variables in the model (i.e., *Suicide*). The positive value of .5 suggests that if the patient currently experiences a high level of sad mood, the sad mood will very likely remain high at the next measurement. On the other hand, the cross-regressive coefficient $\phi_{21} = .2$ represents the relationship between the current value of *Sad* and the lag-1 value of *Suicide* following the same controlling procedure. With this set of parameters, a sample bivariate time series is simulated and visualized in Figure 2.

Just like any parametric statistical model, the VAR(1) model has certain assumptions that need to be met to ensure valid inferences. Besides the common assumptions of normality, linearity, and homoscedasticity for linear-regression models (Poole & O'Farrell, 1971), another particular assumption of the VAR(1) model is covariance stationarity: All parameters in the VAR(1) model (i.e., $\mathbf{\Delta}$, $\mathbf{\Phi}$, $\mathbf{\Sigma}$) do not change over time (Lütkepohl, 2005). This assumption suggests that (a) each variable in the VAR(1) model fluctuates around its constant mean level, which we call its expected value, $E(\boldsymbol{y})$, and (b) the temporal dynamics of how variables interact do not change over time.

Visualizing the VAR(1) model. The VAR(1) model in the example can be visualized as an idiographic temporal network (see Fig. 3), which presents the temporal relationships (i.e., the autoregressive/cross-regressive coefficients $\mathbf{\Phi}$) clearly. Yet other meaningful information in the VAR(1) model, that is, the intercept vector, $\mathbf{\Delta}$, and the innovation covariance matrix, $\mathbf{\Sigma}$, are usually not visualized in such temporal networks (Epskamp, Waldorp, et al., 2018). For how such information can be visualized, see Appendix 1 in the Supplemental Material available online.

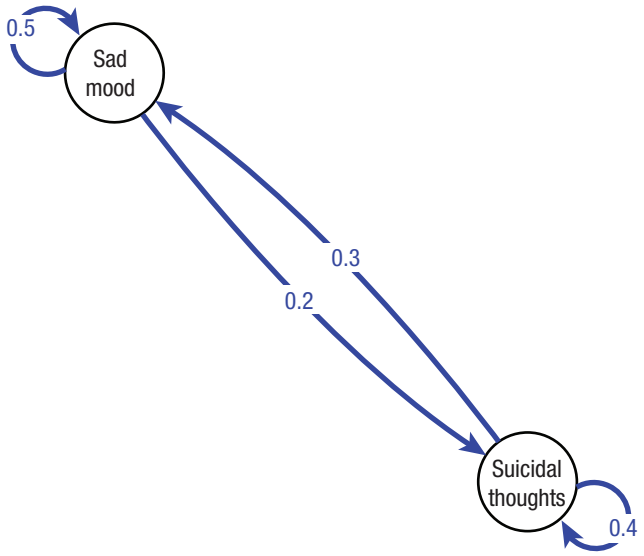

Fig. 3. A hypothetical temporal network.

Simulation-Based Power Analysis

When researchers collect a sample of ILD from a patient and estimate an idiographic network, they aim that estimated intercepts and slopes can accurately resemble the true parameters in the VAR(1) model: Ideally, nonzero effects (i.e., the autoregressive and cross-regressive effects in the previous example) can be detected, and zero effects (i.e., the intercept, $\delta_1 = 0$, in the example) yield nonsignificant estimates. An important factor to consider during sample-size planning is the statistical power of the significance test. A priori power analysis can help researchers determine the minimum sample size required to achieve adequate statistical power for a specific significance test. However, power analysis for time-series statistical models has not been implemented in commonly used software (e.g., G*Power; Faul et al., 2007) that can quickly conduct analytical power analysis for simpler models. Recently, simulation-based power analysis has been developed to assist with sample-size planning for many univariate models used in longitudinal studies (Lafit et al., 2021) and introduced for the multivariate VAR(1) model (Revol et al., 2024).[2] In this section, we show in detail how power analysis for the VAR(1) model can be conducted through simulation and used for sample-size planning when the goal is to fit an idiographic network. We describe the stepwise procedure of this approach and demonstrate it with the example specified earlier.

Stepwise procedure

Suppose that the temporal dynamics of sadness and suicidal thoughts of a patient were estimated through Equations 2 and 3. Now, researcher Jesse intends to replicate such findings in a similar patient by fitting the VAR(1) model to the ILD that will be collected from this other patient. To decide how many time points are needed to ensure sufficient power of all the significance tests to be conducted (i.e., for the individual elements in $\mathbf{\Delta}$ and $\mathbf{\Phi}$), here are the steps Jesse will take (for an overview of the procedures, see Fig. 4).

Step 1: determine model and testing parameters. First, Jesse needs to specify the hypothesized effect sizes of the VAR(1) model. This requires input on all parameters of the VAR(1) model: $\mathbf{\Delta}$, $\mathbf{\Phi}$, Σ. Given the large number of parameters to be specified, we strongly recommend that Jesse bases their decisions on the results of previous studies.

Other necessary parameters to specify for this analysis include the significance level for the test of each effect, α, the sample size (the number of time points) to test for, T, and the number of samples to simulate under each condition, n. As an example, Jesse will use the model parameters in Equations 2 and 3 and set the

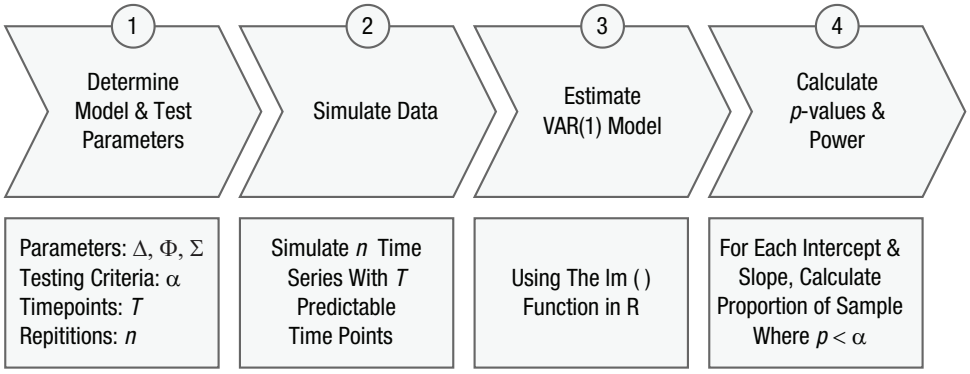

Fig. 4. The procedures of simulation-based power analysis for the lag-1 vector-autoregressive model.

Table 1. Results of Power Analysis for the Hypothetical Network

T	$\delta_1 = 0$	$\delta_2 = 2$	$\phi_{11} = .5$	$\phi_{12} = .3$	$\phi_{21} = .2$	$\phi_{22} = .4$
50	.063	**.935**	**.856**	.466	.258	.592
75	.057	**.996**	**.970**	.653	.387	**.814**
100	.055	**.999**	**.995**	.787	.521	**.922**

Note: Values in bold denote results that exceed the target performance of .8.

significance level of the tests for all parameters in Δ and Φ as $\alpha = .05$. The sample sizes to be tested include $T = 50, 75, 100$, which are commonly adopted in such studies. The number of replications for each sample size is $n = 5{,}000$.

Step 2: data simulation. With the parameter input, samples of ILD following the specified VAR(1) process can be simulated, representing the time series one can collect from the patient. For detailed procedures on data simulation, see Appendix 2 in the Supplemental Material.

Step 3: estimate VAR(1) models from simulated samples. For each of the $n = 5{,}000$ simulated time series, Jesse will fit the VAR(1) model with the ordinary least squares (OLS) estimator[3]:

$$\begin{bmatrix} \widehat{Sad_t} \\ \widehat{Suicide_t} \end{bmatrix} = \begin{bmatrix} b_{10} \\ b_{20} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} Sad_{t-1} \\ Suicide_{t-1} \end{bmatrix}. \quad (4)$$

In Equation 4, Latin letters are used to represent the estimates (e.g., b) and to be distinguished from the true model parameters specified earlier (e.g., δ, ϕ).

Step 4: extract p values and calculate power. The statistical power of each t test on the null hypothesis that a parameter is 0 can be calculated as the percentage of samples that yields estimates that are significantly different from 0 ($p < \alpha = .05$). The parameters include the intercepts, autoregressive and cross-regressive coefficients. For results of a power analysis for the example, see Table 1. If Jesse follows the convention and sets the required power to be .8 (Cohen, 1992), results suggest that the significance tests of the autoregressive coefficients for *Sad* ($\phi_{11} = .5$) and *Suicide* ($\phi_{22} = .4$) will be sufficiently powered with the sample size of 50 and 75, respectively. Although, for the two cross-regressive effects ($\phi_{12} = .3$ and $\phi_{21} = .2$), the power of the significance tests of them are still below .8 (i.e., .787 and .521, respectively) given the maximum sample size tested ($T = 100$).

As expected, in general, the power of the significance tests for larger effects is higher. As the sample size becomes larger, the power of the significance tests for all nonzero effects (e.g., all slopes and the intercept δ_2) increases as well (Cohen, 1992). With all values of T, the tests of zero effects (e.g., the intercept δ_1) still yielded significant results around 5% of the time. This reflects the Type I error rate of the test and is indeed close to the chosen significance level $\alpha = .05$.

Reflection on power analysis

Simulation-based power analysis focuses on one effect at a time and can offer recommendations on the required sample size for the significance test of each effect to be sufficiently powered. Yet as shown in the example, such recommendations usually vary for different effects in the same VAR(1) model and thus cannot always offer clear suggestions on an appropriate sample size for a whole model to be of good quality. A sample size that ensures sufficient power (i.e., higher than .8) for the significance tests of small effects (e.g., the cross-regressive effect in the previous example, $\phi_{21} = .2$) might be too large and unrealistic. Although, with a smaller sample size, relatively small effects can be too difficult to detect. Moreover, power analysis focuses only on statistical significance (i.e., p values) and does not concern the overall fit of a model. Therefore, a more holistic perspective of sample-size planning that complements power analysis is necessary as argued by Revol et al. (2024), in which the focus is placed on the quality of an entire network, especially how accurately it can predict future unseen data of the same patient.

Simulation-Based Predictive-Accuracy Analysis

To assess the predictive accuracy of a statistical model, the common workflow usually starts by collecting a sample on variables of interest. This sample is then divided into two parts: the training set and the test set. Researchers estimate ("train") the model with the training set and use the model estimates to make predictions for the outcome variables in the separate test set. Then, they can compare the predicted values for the test set and the observed values in the test set to evaluate the model's predictive accuracy as a way to assess the generalizability of the model estimates to unseen samples. For the assessment process to be unbiased, the two parts of the sample should be comparable, more specifically, generated with the same underlying processes in which the true relationships among variables are the same (Hastie et al., 2009).

In empirical research, the method of cross-validation is often applied to estimate a model's predictive accuracy such that the aforementioned division into training and test sets is carried out multiple times (e.g., Bulteel et al., 2018). Because the goal is to conduct such

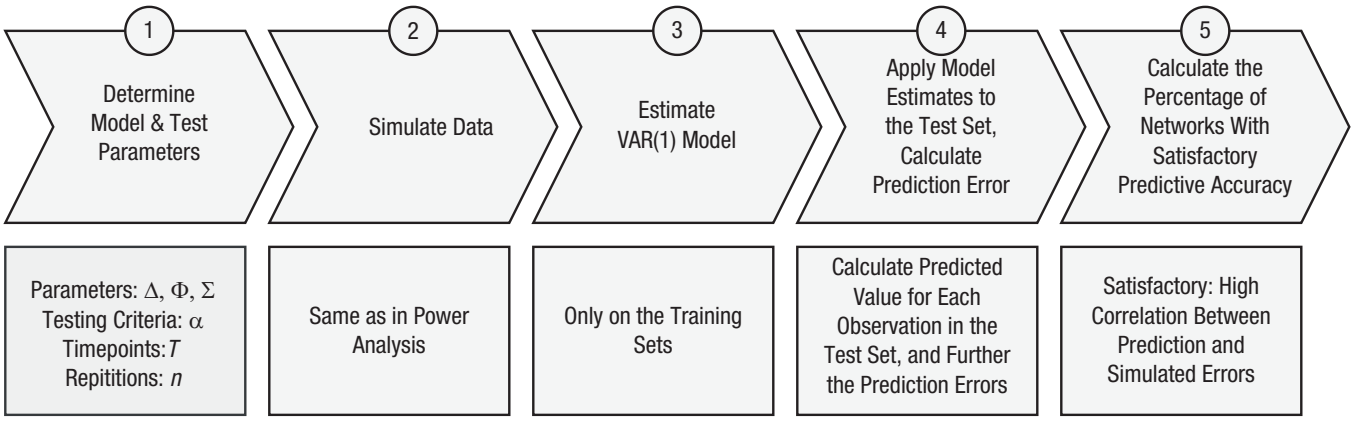

Fig. 5. The procedures of simulation-based predictive accuracy analysis for the lag-1 vector-autoregressive model.

generalizability analysis before data collection as a way to avoid overfitting, such training and test sets can be simulated (e.g., Ernst et al., 2021; Lafit et al., 2022). In the following section, we describe step by step the procedure of simulation-based predictive-accuracy analysis proposed in Revol et al. (2024) while leaving out technical details to make it easily accessible for applied researchers. For details, see Revol et al. and Appendix 2 in the Supplemental Material. We also demonstrate how this method can be applied in a relevant research setting and compare this method with power analysis using the earlier example.

Stepwise procedure

Step 1: determine model and testing parameters. Again, we provide an overview of the procedures in Figure 5. The first three steps of predictive-accuracy analysis are highly similar to those of power analysis. As a preparation for simulating the training sets (i.e., the samples on which the VAR[1] model is estimated), Jesse needs to specify (a) the hypothesized network parameters ($\boldsymbol{\Delta}$, $\boldsymbol{\Phi}$, $\boldsymbol{\Sigma}$), (b) the sample size to test for (T), and (c) the number of replications for each sample size (n). Such choices should again be based on findings from previous studies. For the test set, we simulate a very long time series using the same VAR(1) parameters as the training set to make sure it is highly representative of the underlying processes (i.e., the "population"). In the following example, 100,000 is used as the sample size of the test set.

Steps 2 and 3: data simulation; estimate VAR(1) models from training sets. The same simulated samples in power analysis can be used as training sets in predictive-accuracy analysis for a meaningful comparison between the two methods. Naturally, the estimates of the VAR(1) model acquired earlier can also be used here.

The process of simulating the test set is identical to the process described earlier. Because the test set is only for assessing the risk of overfitting for the earlier estimates, Jesse will not fit the VAR(1) model to it. This highlights a core difference between power analysis and predictive-accuracy analysis: Power analysis follows a purely in-sample approach in which no separate test set is used, whereas predictive-accuracy analysis investigates the out-of-sample performance of estimated models.

Step 4: apply model estimates to the test set and calculate prediction errors. Jesse then uses the VAR(1) estimates acquired from each training set to calculate the predicted values (e.g., $\widehat{Sad_t}$) of all variables in the test set:

$$\begin{bmatrix} \widehat{Sad_t^{Test}} \\ \widehat{Suicide_t^{Test}} \end{bmatrix} = \begin{bmatrix} b_{10} \\ b_{20} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} Sad_{t-1}^{Test} \\ Suicide_{t-1}^{Test} \end{bmatrix}. \quad (5)$$

To assess the predictive accuracy of a set of VAR(1) estimates, we examine whether the multivariate prediction errors resemble the simulated errors for each time point. Because every observation includes some random noise ($\boldsymbol{\epsilon_t}$), no model can achieve perfectly accurate predictions. Nonetheless, a high correlation between the prediction and simulated errors suggests that prediction errors stem primarily from the random noise instead of the model failing to capture meaningful patterns in the data. Thus, a high correlation between the prediction and the simulated errors indicates that the predicted values account for a large portion of the explainable variance in the observed data. This further suggests that the VAR(1) estimates successfully capture the autoregressive and cross-regressive effects in the data-generating process, demonstrating sufficient predictive accuracy.

The prediction-error vector for a given time point t in the test set, $\mathbf{e_t^{Test}}$, can be calculated as the differences between predicted and observed values:

$$\mathbf{e_t^{Test}} = \begin{bmatrix} e_{Sad,t}^{Test} \\ e_{Suicide,t}^{Test} \end{bmatrix} = \begin{bmatrix} \widehat{Sad_t^{Test}} \\ \widehat{Suicide_t^{Test}} \end{bmatrix} - \begin{bmatrix} Sad_t^{Test} \\ Suicide_t^{Test} \end{bmatrix}. \quad (6)$$

This vector of prediction errors, $\mathbf{e_t^{Test}}$, can be used to determine the aforementioned similarity between prediction errors and simulated errors. A major challenge is, however, that the prediction errors are unstandardized measures of a model's performance. For example, a prediction error of 5 is way larger for a variable measured on a 7-point Likert scale than another variable measured on a scale from 0 to 100. Moreover, the univariate prediction errors among multiple variables tend to be correlated because the simulated errors in the VAR(1) model share a nonzero covariance (as demonstrated by Equation 3). This should be corrected for when calculating the statistic that represents the predictive accuracy of a multivariate model. Both challenges can be tackled by standardizing the error vectors. By using the mean and covariance matrix of the simulated errors, we can calculate the squared Mahalanobis distance, D^2 (Mahalanobis, 2018):

$$D^2 = \left(\mathbf{y} - \bar{\mathbf{y}}\right)^T \Sigma^{-1} \left(\mathbf{y} - \bar{\mathbf{y}}\right). \quad (7)$$

The Mahalanobis distance is a standardized distance measure between one observation and the center of a multivariate normal distribution in which different weights are applied to different variables based on their (co)variances. Therefore, we can standardize the prediction-error vectors by calculating the squared Mahalanobis distance between each error vector and the center of the innovation distribution, which is considered to have the mean of $\mathbf{0}$ and the covariance matrix of $\mathbf{\Sigma}$ during data simulation. The same calculation is performed on the 100,000 simulated innovation vectors in the test set.

This calculation achieves an important dimension reduction: The vectors of both prediction errors and simulated innovation (i.e., $\mathbf{e_t^{Test}}$ and $\boldsymbol{\epsilon}_\mathbf{t}^{\mathbf{Test}}$) are reduced from $M \times 1$ to a standardized single value (denoted as $D_{pred,t}^2$ and $D_{sim,t}^2$) while accounting for the innovation (co) variances. This makes the aforementioned similarity assessment easier.

The similarity can then be quantified as the squared Pearson's correlation (R^2) between the two sets of squared Mahalanobis distances on the test set: prediction errors of a set of VAR(1) estimates and the simulated innovation vectors (i.e., $D_{pred,t}^2$ and $D_{sim,t}^2$). When the squared correlation is large, we consider that the variance of simulated innovation is accurately explained by the prediction error, suggesting that the estimated VAR(1) model does not overfit the sample. For this purpose, we need to select an appropriate threshold value for R^2. Revol et al. (2024) found that the relationship between R^2 and the required sample size is mostly linear until $R^2 = .9$ and becomes exponential after that. Therefore, we recommend using at least $R^2 = .9$ as the threshold value in this step: More than 90% of the variance of simulated innovation needs to be explained by the prediction errors. Because the growth rate of required sample size with respect to R^2 is still small given $R^2 = .9$, the recommended sample size with this threshold is considered as a lower bound—a sufficient sample size if the empirical data to be collected perfectly resemble the simulation. However, this assumption is usually unrealistic given the potential deviations from the simulation in empirical data (e.g., measurement errors and heteroscadasticity of residuals). Therefore, we recommend that researchers also use higher threshold values, specifically, $R^2 = .92$ or $R^2 = .95$, to get a more conservative sample-size recommendation—a larger and thus safer choice in view of such deviations. Such thinking certainly also applies to power analysis (e.g., using a value higher than .8 as the target performance) but will not be discussed in detail for simplicity.

Step 5: calculate the percentage of networks with satisfactory predictive accuracy. As the final step, we calculate the percentage of networks estimated from the 5,000 training sets that can be deemed not overfitting using the thresholds $R^2 = .9, .92, .95$ for each sample size. These results indicate the probability (the long-run proportion) that a network estimated from a sample will not overfit given the network parameters and the sample size. We call this proportion the "sufficient predictive accuracy probability" (SPAP). For the purpose of sample-size planning, we can set a threshold of SPAP before the analysis and search for the minimum sample size required for this goal to be reached.

The results of predictive-accuracy analysis for the example is presented in Table 2. If the threshold of SPAP is set as .8, the sample-size recommendations are 100 using the threshold of $R^2 = .9$. This means that when Jesse collects a sample of 100 time points from this patient and fits a temporal network to the data, the probability that the network does not overfit is higher than 80% (82.9%) as long as the sample presents no deviation from the model specified in Step 1. If possible, Jesse can collect more time points (125 given $R^2 = .92$ or 200 given $R^2 = .95$), which are safer choices in case of violations of the assumptions of the VAR(1) model.

Table 2. Sufficient Predictive Accuracy Probability of the Hypothetical Network

T	50	75	100	125	150	200
R^2 = .9	.393	.646	**.829**	**.904**	**.961**	**.993**
R^2 = .92	.280	.503	.713	**.823**	**.912**	**.977**
R^2 = .95	.107	.241	.423	.558	.691	**.862**

Note: Values in bold denote results that exceed the target performance of .8.

Reflection on predictive-accuracy analysis

In this section, we demonstrate the predictive-accuracy analysis using one set of VAR(1) model parameters. Such analysis can be performed rapidly: For the $T = 100$ condition in the example, conducting power and predictive-accuracy analyses takes 152.55 s in total.[4] In practice, although researchers can select the parameter values based on previous studies, uncertainty still exists regarding whether these values are indeed accurate. Such uncertainty influences the appropriateness of the sample-size suggestion made based on only one set of parameters. Therefore, we recommend that when possible, researchers can specify multiple sets of plausible parameter values as input for both power and predictive-accuracy analysis (for an example, see Lafit et al., 2025). In the next section, we show in detail how this can be achieved.

Application: Retrospective Quality Assessment of Existing Network Studies

In the previous section, we showed how a priori power analysis and predictive-accuracy analysis can be applied to support sample-size planning before the data collection of single-case network studies. Here, we demonstrate the retrospective usage of both methods. With the same methods, we can assess whether the sample size adopted in an existing network study was large enough to ensure sufficient quality of the estimated networks. This application is based on the assumption that the true network parameters are identical to the estimates. Given this assumption and the number of time points collected in an existing study, if the study is to be replicated, one can calculate through simulation (a) the achieved statistical power of the test for each edge in the network and (b) SPAP—the probability of the estimated network having sufficient predictive accuracy and not overfitting the sample.

To demonstrate this usage of both methods, we first conducted a systematic review of previous studies in the field of clinical psychology that estimated idiographic networks. For the detailed search protocol, see Appendix 2 in the Supplemental Material; for an overview of such studies, see Table 3 below. A first impression of the results is that on average, the number of variables/nodes used in these studies was large (range = 5–21, *Mdn* = 9), yet the number of time points available for analysis tended to be small.[5] This suggests a high risk of overfitting for the estimated networks (Bulteel et al., 2018). Here, we apply the two methods to two studies (i.e., Bak et al., 2016; Epskamp, van Borkulo, et al., 2018) for a closer inspection of the quality of the estimated networks. These two studies were chosen as representative examples of the field: The number of variables and predictable time points in both studies are on an average level across all studies in Table 3 below. In addition, both studies reported the research design and data-analysis procedures thoroughly and made the materials necessary for reproducing the results easily accessible.

Before we start, an important message we wish to convey is that this retrospective approach is recommended only for assessing the quality of existing studies. When planning for a new study, we strongly recommend that researchers justify their plans of sample size before data collection instead of afterward. The reason lies in the uncertainty of the effect size estimated from samples, especially from small samples (Leon et al., 2011). Without proper consideration of the uncertainty, using such estimates as parameters in retrospective power analysis can further lead to biased power estimates (Albers & Lakens, 2018) and naturally biased SPAP. To account for such uncertainty in this retrospective analysis, we incorporated the standard errors of all autoregressive/cross-regressive coefficients when deciding the value of $\mathbf{\Phi}$, which is a distinctive feature compared with the a priori use of both methods presented earlier. We discuss how to achieve this in later text.

Example 1: Bak et al.

Bak et al. (2016) collected ESM data from a patient ("Miss A") who was receiving pharmacological treatment for psychosis experiences. The patient was followed for a year in this study and asked to provide ratings for multiple symptoms 10 times per day during 4 days of each week. Such symptoms included *hearing* (i.e., hearing voices), *down*, *relaxed*, *paranoia*, and *control* (i.e., loss of control), and they were all measured on 7-point Likert scales. Throughout the year, Miss A was in a "stable state" most of the time, during which she was prescribed 350 mg of clozapine per day. Yet there were a few episodes of impending relapses and full relapses in which Miss A experienced heightened severity of her symptoms. The dosage of clozapine was increased to 400 and 450 mg/day, respectively, during the two states. Bak et al.

Table 3. An Overview of Studies Using Person-Specific Temporal Networks

Authors (year)	Diagnosis/target population	Participants, *n*	Sampling scheme	Nodes, *n*	Nonmissing time points, *n*	Time points for analysis, *n*
Nonregularized ($N = 10$)						
Bak et al. (2016)	Psychosis	1	10 beeps/day, 1 year	5	3 phases: 662, 158, 119	3 phases: 353, 86, 63
Bulteel et al. (2016)	Depression	1	1 beep/day, 6 months	11	100	78
Curtiss et al. (2023)	Depression	31	1 beep/day, 8 weeks	10	Not reported	Moving windows of 18 days
Reeves and Fisher (2020)	Posttraumatic stress disorder	20	4 beeps/day, around 30 days	20	Not reported	$M = 126.15$, $SD = 12.75$
Rowland and Wenzel (2020)	Undergraduates	125	6 beeps/day, 40 days	8	$M = 182.16$, $SD = 14.8$	Not reported
Strauss et al. (2023)	Schizophrenia and HCs	46 and 52	8 beeps/day, 6 days	2 networks: 5 and 10	$M = 33.74$; $M = 30.43$	Not reported
van Der Velden et al. (2018)	Parkinson's disease	1	10 beeps/day, 34 days	7	121	57
Voigt et al. (2018)	Bipolar disorder	1	10 beeps/day, 90 days	6	447	Not reported
Wichers et al. (2016)	Depression	1	10 beeps/day, 239 days	5	1,474	Moving windows of 30 days
Wichers et al. (2020)	Depression	6	3 beeps/day, 95–183 days	5	370	Moving windows of 30 days
Regularized ($N = 15$)						
Beck and Jackson (2020)	Undergraduates	349	4 beeps/day, 2 weeks	9	2 waves with medians of 41 and 33	Not reported
Bos et al. (2018)	Depression	50	3 beeps/day, 30 days	6	$M = 76$, $SD = 5.3$	Missing data imputed
David et al. (2018)	Depression	1	1 beep/day, 122 days	19	90	Not reported
de Vos et al. (2017)	Depression and HCs	27 and 27	3 beeps/day, 30 days	14	$M = 83.2$, $SD = 7.4$	90 (after imputation)
Epskamp, van Borkulo, et al. (2018)	Depression	1	5 beeps/day, 14 days	7	65	47
Fisher et al. (2017)	Generalized anxiety disorder, depression	40	4 beeps/day, at least 30 days	21	$M = 130.43$, $SD = 19.27$	Not reported
Frumkin et al. (2021)	Generalized anxiety disorder, depression, etc.	17	5 beeps/day, 21–24 days	12	$M = 94$ (69–117)	Not reported
Kroeze et al. (2017)	Panic disorder and depression	1	5 beeps/day, 14 days	10	66	Not reported
Lazarus et al. (2020)	Undergraduates	52	4 beeps/day, at least 15 days	9	$M = 57.8$, $SD = 5.1$	Not reported
Levinson et al. (2018)	Eating disorder	66	4 beeps/day, 7 days	10	$M = 20.72$, $SD = 6.97$	Not reported
Levinson et al. (2021)	Eating disorder	34	5 beeps/day, 15 days	Two networks: 8 and 15	$M = 54.6$	75 (missing data imputed)
McGhie and McNally (2025)	Trauma	52	5 beeps/day, 14 days	5	At least 40	Not reported
Piccirillo and Rodebaugh (2022)	Social anxiety disorder and depression	35	5 beeps/day, 30 days	12	$M = 125.43$, $SD = 19.26$	Not reported
van der Krieke et al. (2017)	General public	247	3 beeps/day, 30 days	6	At least 68 to generate a network	Not reported
van der Tuin et al. (2022)	Young adults at risk for psychosis	77	1 beep/day, 90 days	10	$M = 81.9$	Not reported

Note: HCs = healthy control subjects.

fitted a temporal network to the data in each of the three states (i.e., stable state, impending relapse, and full relapse) and was interested in exploring whether the temporal relationships among Miss A's symptoms differed among the three states.

In this section, we demonstrate the stepwise procedure of applying both retrospective power analysis and predictive-accuracy analysis to this study. Such applications will provide ways of evaluating the quality of the estimated networks and the validity of conclusions drawn by Bak et al. (2016) and from other similar network studies.

Step 1: record the actual sample size

When fitting a network to a time series, the actual number of predictable time points in the analysis is smaller than the total number of measurement prompts sent to a participant because of the exclusion of certain observations. Therefore, we cannot directly use the latter in simulation-based analyses. Here, we briefly discuss the two most common reasons for such exclusions.

First, an observation can be analyzed in the VAR(1) model only if its lagged values are not missing. Given that imputation of the missing data is not common practice in such time-series analysis yet, observations that are either missing or have missing lagged values cannot be predicted.

Second, the VAR(1) model assumes equal intervals between two consecutive observations. This is a reasonable assumption because the strength of a temporal relation depends on the intervals between the two measurements. For example, if Miss A feels down at this moment, we can be more confident in predicting that she will still feel down in 3 hr than in 3 days. This assumption of the VAR(1) model usually holds for same-day observations: Miss A received a prompt approximately every 90 min between 7:30 a.m. and 10:30 p.m. in this study. However, the overnight lag between the last observation of a day and the first observation of the next day is much longer than 90 min, yet these two observations are still considered consecutive in the time series. Therefore, the first observation of each day is not accompanied by a valid lagged value and thus is not being predicted by the previous observation. For a more nuanced discussion of this issue, see Berkhout et al. (2025).

After such data-preprocessing procedures, the number of predictable time points to be used in later data simulation is naturally smaller than the total number of prompts sent to the patient, even smaller than the number of prompts that the patient responded to. For the three states of Miss A (stable state, impending relapse, and full relapse), although the number of complete observations was 662, 158, and 119, respectively, the actual number of observations used in the network analysis was only 353, 86, and 63.

Step 2: acquire network estimates

Before starting the data simulation, we also need to specify the network parameters, for which we will use the network estimates in an existing study. In general, if researchers reported the estimates of the intercepts, slopes, and innovation (co)variances thoroughly, we can use this information as the simulation parameters. Otherwise, we should reanalyze the data by fitting the VAR(1) model to the data following the identical data-preprocessing procedures used in the original study. For the current example, we contacted one of the coauthors of Bak et al. (2016) who had access to the data of Miss A and successfully reproduced the analysis with R.

The estimated network of Miss A's stable state is shown in Figure 6a[6] with

$$\begin{bmatrix} Down_t \\ Relax_t \\ Paranoia_t \\ Hearing_t \\ Control_t \end{bmatrix} = \begin{bmatrix} 1.83 \\ 2.83 \\ 2.17 \\ 4.18 \\ .66 \end{bmatrix} + \begin{bmatrix} .29 & -.16 & .16 & -.06 & .08 \\ -.04 & .39 & -.14 & .00 & .02 \\ .27 & -.22 & .24 & -.11 & .06 \\ .08 & -.15 & .06 & .17 & -.03 \\ .19 & -.09 & .00 & .05 & .11 \end{bmatrix} \cdot \begin{bmatrix} Down_{t-1} \\ Relax_{t-1} \\ Paranoia_{t-1} \\ Hearing_{t-1} \\ Control_{t-1} \end{bmatrix} + \begin{bmatrix} e_{Down,t} \\ e_{Relax,t} \\ e_{Paranoia,t} \\ e_{Hearing,t} \\ e_{Control,t} \end{bmatrix}, \tag{8}$$

and

$$\begin{bmatrix} e_{Down,t} \\ e_{Relax,t} \\ e_{Paranoia,t} \\ e_{Hearing,t} \\ e_{Control,t} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1.87 & -.36 & .54 & .44 & .31 \\ -.36 & 1.48 & -.35 & -.33 & -.16 \\ .54 & -.35 & 2.85 & .88 & .54 \\ .44 & -.33 & .88 & 2.15 & .30 \\ .31 & -.16 & .54 & .30 & 1.84 \end{bmatrix} \right). \tag{9}$$

Likewise, the estimated networks of Miss A during the states of impending relapse and full relapse are shown in Figure 7. In the original study, nonsignificant edges were also visualized in all three networks with the corresponding point estimate (Bak et al., 2016). From visual inspections of the networks, the authors concluded that connections among symptoms became stronger during the relapse states.[7] In addition, the authors also calculated network centrality indices for each symptom and compared them among the three networks. The centrality indices of a node are calculated by aggregating the edges pointing toward it and from it. This is considered a way of quantifying the importance of a node (Johal &

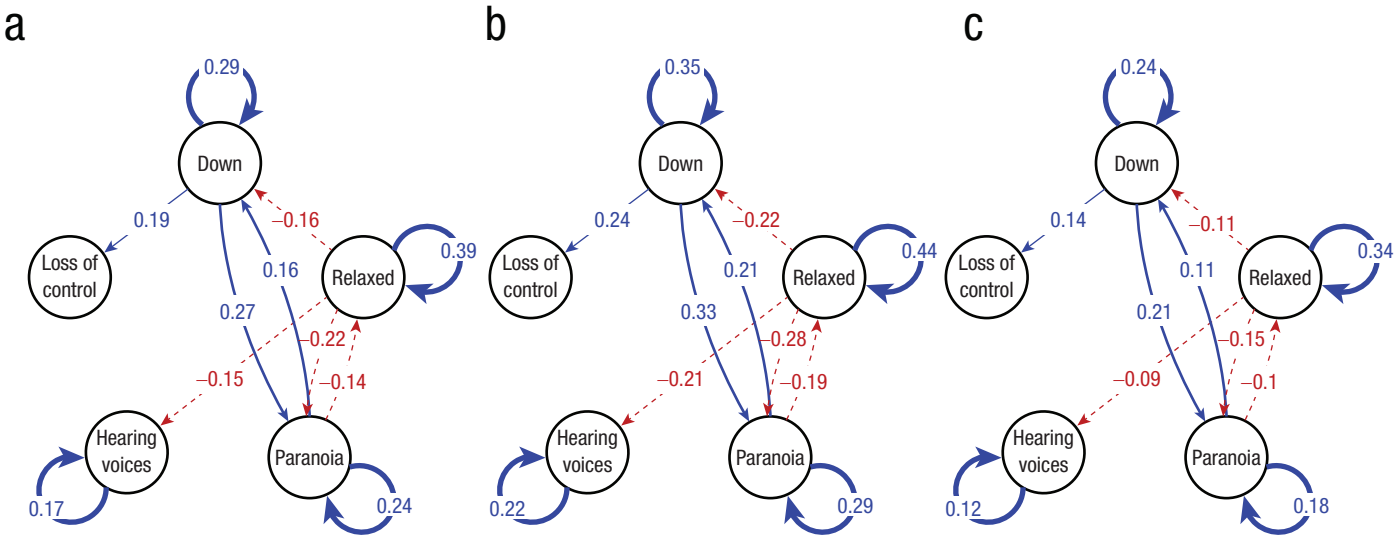

Fig. 6. The estimated networks of Miss A's stable state with uncertainty: (a) point estimates, (b) 1 *SE* larger, and (c) 1 *SE* smaller.

Rhemtulla, 2024).[8] For the symptom *Paranoia*, larger centrality indices were found during the two relapse states, which led to the conclusion that paranoia played a central role during relapse.

These two conclusions were both based on the point estimates of edges in the networks and did not consider the degree of uncertainty of these estimates—the standard error of the estimate of each edge. The uncertainty of the estimates of all lagged effects was indeed small for the stable state (all *SE*s were smaller than .08) given that the analyzed sample contained a large number of time points ($T = 353$). However, for the two relapse states, the uncertainty was large (all *SE*s were larger than .09) because of the small sample sizes (86 and 63, respectively). To facilitate an accurate interpretation of a network and related centrality indices, it is important to take the uncertainty of estimates into account and to openly communicate them (Bringmann et al., 2019). When

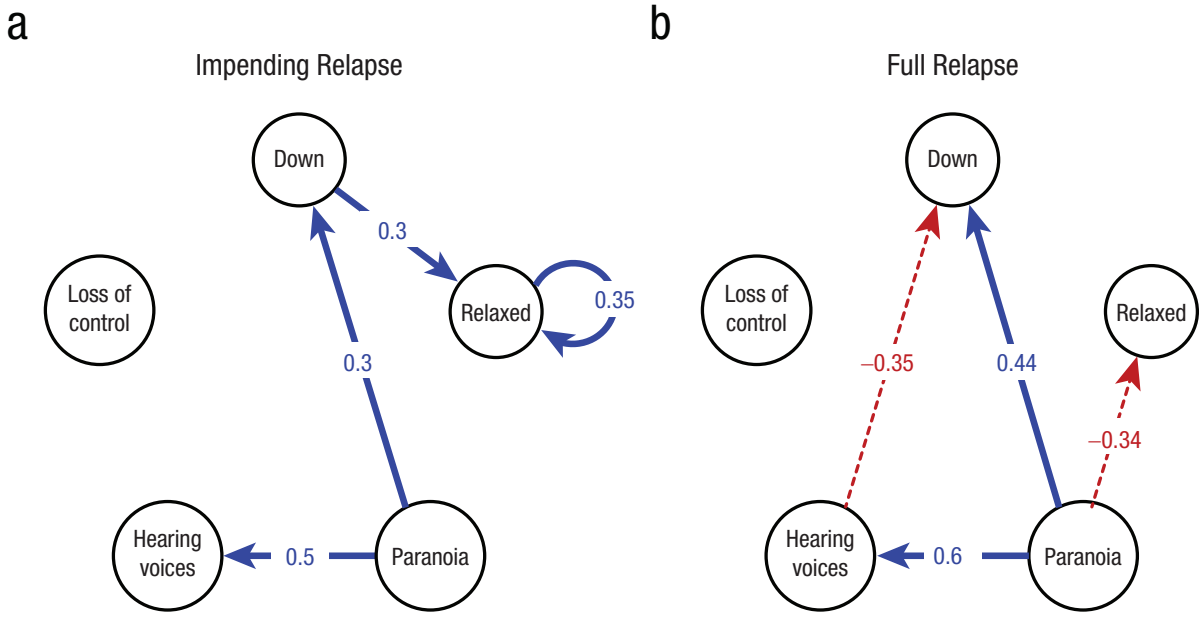

Fig. 7. The estimated networks of Miss A's impending-relapse and full-relapse states.

Table 4. Results of Power Analysis for Edges Directed to *Down* in the Network of Miss A's Stable State

T	ϕ_{11} = .29 *Down* → *Down*	ϕ_{12} = −.16 *Relaxed* → *Down*	ϕ_{13} = .16 *Paranoia* → *Down*	ϕ_{14} = −.06 *Hearing* → *Down*	ϕ_{15} = .08 *Control* → *Down*
353	**.999** (**.988,1.000**)	**.840** (.452,**.966**)	**.923** (.671,**.995**)	.179 (.053, .524)	.310 (.066, .747)
300	**.998** (**.965,1.000**)	.739 (.385,**.940**)	**.883** (.604,**.985**)	.159 (.051, .463)	.264 (.062, .686)
250	**.989** (**.937,.999**)	.663 (.321,**.885**)	**.816** (.512,**.961**)	.143 (.049, .404)	.227 (.059, .600)
200	**.967** (**.872,.994**)	.558 (.254,**.813**)	.716 (.419,**.917**)	.125 (.052, .319)	.185 (.059, .499)
175	**.939** (**.803,.990**)	.496 (.240, .756)	.651 (.385,**.871**)	.112 (.055, .286)	.165 (.059, .436)
150	**.894** (.783,**.973**)	.447 (.211, .687)	.585 (.327,**.823**)	.103 (.056, .261)	.151 (.055, .374)

Note: Values in bold denote results that exceed the target performance of .8.

runing power and predictive-accuracy analysis for networks, incorporating such uncertainty is also crucial.

Taken all together, three sets of values are used for data simulation. First, the network point estimates of each of the three states are used as the true network parameters for data simulation in later steps. To account for the uncertainty of the estimates, we used the standard error of each estimated lagged coefficient (Liu & Wang, 2019; Perugini et al., 2014) to create two additional networks for data simulation. One network has a larger effect size than the point estimates, with all lagged effects enlarged by 1 *SE*. The other network has a smaller effect size from all lagged effects being shrunk by 1 *SE*.[9] These two networks represent an upper bound and a lower bound, respectively, of the true network parameters (for the network visualizations of these two matrices for Miss A's stable state, see Figs. 6b and 6c). By running power and predictive-accuracy analyses with these two lower-bound and upper-bound matrices as the true $\mathbf{\Phi}$ for the network, we can compute a plausible range of achieved statistical power for edges in the networks and the SPAP for the estimated networks given a specific number of time points in the training sets. The choice of ±1 *SE* is roughly consistent with the 60% confidence interval used in Perugini et al. (2014). Researchers can also set the upper bound and lower bound to be farther apart (e.g., ±2 *SE*) for sample-size suggestions in a larger safety margin.

Step 3a: power analysis

With input acquired in Steps 1 and 2, we can further set the significance level of the test (e.g., $\alpha = .05$) and decide the sample size we are interested in testing for in both analyses. The primary goal of such retrospective analysis is to test whether the sample size adopted in existing studies was large enough to ensure sufficient power (e.g., power > .8) and predictive-accuracy probability (e.g., SPAP > .8). Therefore, the most relevant sample size to initialize the analysis with is naturally the actual sample sizes in the original study, which are 353, 86, and 63 for the three states of Miss A, respectively. Ideally, an adopted sample size is large enough for an estimated model to be of good quality: The significance test for effects of interest is sufficiently powered, and the SPAP of the whole model is higher than the target threshold. If this turns out to be untrue, we can gradually enlarge T and search for the appropriate sample size that can guarantee sufficient model quality. But if this is met, we can also lower T step by step in the analysis to find the minimum sample size required. Because SPAP is a single value and thus an easier to use indicator of a model's quality than statistical power, this searching process will be primarily based on the result of predictive-accuracy analysis, which is described in Step 3b.

It is important to recall that power analysis and predictive-accuracy analysis share the same training sets and can thus be conducted in parallel to each other. For Miss A's stable state, power analysis was run with a set of values for T, including 353, 300, 250, 200, 175, and 150 (see Table 4). Given the large amount of parameters to estimate in this VAR(1) model, only the statistical power of tests with *Down* being the dependent variable are reported in Table 4. In each cell of Table 4, a range is also reported, with the statistical power calculated with each coefficient being shrunk and enlarged by 1 *SE* as the lower and upper bounds. For results of power analysis for the tests of all other parameters, see Appendix 3 in the Supplemental Material. Given the number of time points available for analysis during Miss A's stable state (i.e., 353), the three significant edges with *Down* being the dependent variable (i.e., ϕ_{11}, ϕ_{12}, and ϕ_{13}) are indeed sufficiently powered. We also note that

Table 5. Results of Power Analysis for the Network of Miss A's Impending-Relapse State

T	ϕ_{13} = .30 *Paranoia → Down*	ϕ_{21} = .30 *Down → Relaxed*	ϕ_{22} = .35 *Relaxed → Relaxed*	ϕ_{43} = .50 *Paranoia → Hearing*
86	.736 (.318,**.971**)	.631 (.254,**.911**)	**.851** (.463,**.998**)	**.990** (**.897,1.000**)
100	.796 (.363,**.985**)	.692 (.292,**.944**)	**.914** (.526,**.999**)	**.996** (**.938,1.000**)
150	**.940** (.516,**.999**)	**.878** (.401,**.994**)	**.986** (.721,**1.000**)	**1.000** (**.992,1.000**)
175	**.969** (.591,**1.000**)	**.918** (.470,**.999**)	**.994** (.789,**1.000**)	**1.000** (**.997,1.000**)
200	**.984** (.647,**1.000**)	**.950** (.537,**.999**)	**.997** (**.849,1.000**)	**1.000** (**.999,1.000**)

Note: Values in bold denote results that exceed the target performance of .8.

when ϕ_{12} and ϕ_{13} are set as 1 *SE* smaller from their point estimates, statistical power of their significance tests is below .8.

For the two networks in Miss A's impending- and full-relapse states, power analysis was first conducted for each state with T being the corresponding available number of time points in the original study, 86 and 63, followed by a larger set of T for both states, including 100, 150, 175, and 200. The results of power analysis for the significant edges in both networks are shown in Tables 5 and 6. For the network of the impending-relapse state with $T = 86$, the power of the significance tests of ϕ_{13} (*Paranoia → Down*) and ϕ_{21} (*Down → Relaxed*) is below .8, whereas the testings of ϕ_{35} (*Relaxed → Relaxed*) and ϕ_{51} (*Paranoia → Hearing*) are sufficiently powered. However, when the −1 *SE* estimates are used in the analysis, the significance test of ϕ_{22} also becomes underpowered. Similar results are found for the network of the full-relapse state: When $T = 63$, the power of the testings of all relevant edges is sufficient (>.8) when their point estimates are used for the analysis, yet some become insufficient when the lower-bound $\mathbf{\Phi}$ are used as the true parameter value. With both networks, the statistical power of the significance tests of all edges increases as T gets enlarged and can mostly reach the target threshold of .8 when T is as large as 175, as displayed in Tables 5 and 6.

Step 3b: predictive-accuracy analysis

As mentioned in the previous section, the search for the optimal number of time points is primarily dependent on the results of the predictive-accuracy analysis. The search process is as follows. We set the starting value of T to be the number of time points available in each state and start the first run of predictive-accuracy analysis. From here, the search progress diverges into two directions depending on whether the threshold of .8 for SPAP is reached given $R^2 = .9$: (a) If yes, we round down T to the nearest integer multiple of 100 and start to reduce T at a step of 50 until SPAP falls below .8. When this happens, we enlarge T by 25 and end the search. (b) If

Table 6. Results of Power Analysis for the Network of Miss A's Full-Relapse State

T	ϕ_{13} = .44 *Paranoia → Down*	ϕ_{14} = −.35 *Hearing → Down*	ϕ_{23} = −.34 *Paranoia → Relaxed*	ϕ_{43} = .60 *Paranoia → Hearing*
63	**.950** (.494,**1.000**)	**.790** (.369,**.999**)	**.943** (.517,**1.000**)	**.995** (**.858,1.000**)
100	**.996** (.712,**1.000**)	**.952** (.550,**1.000**)	**.997** (.750,**1.000**)	**1.000** (**.977,1.000**)
150	**1.000** (**.875,1.000**)	**.995** (.738,**1.000**)	**1.000** (**.894,1.000**)	**1.000** (**.998,1.000**)
175	**1.000** (**.921,1.000**)	**.999** (**.810,1.000**)	**1.000** (**.936,1.000**)	**1.000** (**1.000,1.000**)
200	**1.000** (**.948,1.000**)	**1.000** (**.848,1.000**)	**1.000** (**.960,1.000**)	**1.000** (**.1000,1.000**)

Note: Values in bold denote results that exceed the target performance of .8.

Table 7. Sufficient Predictive Accuracy Probability of the Network of Miss A's Stable State

T	353	300	250	200	150	125
R^2 = .9	**1.000** (**1.000**,**1.000**)	**1.000** (**1.000**,**1.000**)	**1.000** (**.998**,**.999**)	**.982** (**.971**,**.984**)	**.807** (.757,**.804**)	.516 (.485,.581)
R^2 = .92	**1.000** (**.999**,**1.000**)	**.999** (**.997**,**.998**)	**.982** (**.976**,**.985**)	**.879** (**.850**,**.892**)	.490 (.462,.493)	.223 (.204,.258)
R^2 = .95	**.948** (**.939**,**.941**)	**.812** (.793,**.803**)	.561 (.546,.578)	.259 (.230,.258)	.046 (.041, .049)	.009 (.007, .013)

Note: The result presented in the first line of each cell represents the sufficient predictive accuracy probability when using the point estimates as the simulation parameter, Φ. The range presented in parentheses in the second line shows the sufficient predictive accuracy probability when using the upper-bound Φ (the left value) and the lower-bound Φ (the right value). Values in bold denote results that exceed the target performance of .8.

no, we round up T to the nearest integer multiple of 100 and start to increase T at a step of 50 until SPAP reaches .8. Once this is achieved, we reduce T by 25 and end the search. The last value of T that leads to a SPAP larger than .8 is the recommended number of time points.

Results of the predictive-accuracy analysis for Miss A's stable state are shown in Table 7. All networks estimated with simulated training sets with 353 time points showed sufficient predictive accuracy on the test set (SPAP = 1.000). Given the threshold of R^2 = .9, roughly only 150 time points are required for the probability of getting a set of network estimates with satisfactory predictive accuracy for Miss A's stable state to be higher than 80%. However, the results using R^2 = .92 and R^2 = .95 as the thresholds suggest that it is safer to have 200 and 300 time points, respectively, in case the collected data violate any assumptions of the VAR model.

Results based on the lower-bound and upper-bound **Φ** are, again, presented in Table 7 as a range below the SPAP when using the point estimates as **Φ**. Opposite to the results of power analysis in which larger effect size is accompanied by higher power of the significance test when the number of time points are fixed, SPAP is usually lower with larger effect size (Revol et al., 2024), potentially because the bias of the OLS estimator is larger with larger effect size (Engsted & Pedersen, 2014). Thus, the left value of the range of SPAP represents the result when using the upper-bound **Φ** for simulation, and the right value represents that when the lower-bound **Φ** is used.[10]

Tables 8 and 9 show the results of predictive-accuracy analysis for the networks during the impending- and full-relapse states. Given the actual number of analyzable time points during the two states (T = 86 and T = 63) and the threshold of R^2 = .9, the predictive accuracy of the networks estimated from simulated data is lower than the desired level (SPAP = .108 and SPAP = .009). Similar to the stable state, the number of time points in the sample needs to be at least 150 for both states so that the estimated networks are very unlikely to overfit.

Example 2: Epskamp, van Borkulo, et al.

Hoping to limit spurious edges in networks and to avoid overfitting, many network researchers have started to use a series of methods implementing regularization techniques (Bar-Kalifa & Sened, 2020; Epskamp, van Borkulo, et al., 2018). The most commonly used is the least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996). LASSO applies a downward bias to the model such that strong edges in the network get reduced

Table 8. Sufficient Predictive Accuracy Probability of the Network of Miss A's Impending-Relapse State

T	86	100	125	150	175
R^2 = .9	.108 (.111, .106)	.264 (.257, .276)	.558 (.566, .594)	**.842** (**.815**,**.834**)	**.940** (**.931**,**.952**)
R^2 = .92	.024 (.027, .024)	.082 (.077, .081)	.242 (.252, .261)	.526 (.507, .535)	.742 (.735, .764)
R^2 = .95	.001 (.000, .000)	.000 (.001, .004)	.011 (.013, .012)	.051 (.051, .052)	.131 (.137, .139)

Note: Values in bold denote results that exceed the target performance of .8.

Table 9. Sufficient Predictive Accuracy Probability of the Network of Miss A's Full-Relapse State

T	63	100	150	175	200
R^2 = .9	.009 (.016, .010)	.241 (.209, .262)	.790 (.698,**.825**)	**.918** (**.813,.956**)	**.981** (**.909,.987**)
R^2 = .92	.001 (.003, .001)	.073 (.067, .077)	.487 (.462, .509)	.694 (.608, .772)	**.863** (.777,**.893**)
R^2 = .95	.000 (.000, .000)	.002 (.002, .002)	.046 (.059, .045)	.119 (.125, .143)	.258 (.241,.257)

Note: Values in bold denote results that exceed the target performance of .8.

("shrinkage") and rather weak edges are set to 0 ("selection"), resulting in a sparser network with fewer edges. Regularized networks have shown better predictive accuracy than standard networks (Bulteel et al., 2018). However, this relative advantage over standard networks does not mean that regularized networks are guaranteed to have satisfactory quality (Zhou et al., 2024). For example, LASSO regression models suffer from a higher false-positive rate, especially when the available number of time points for the analysis is small relative to the number of predictors (Lafit et al., 2019). Moreover, LASSO regression with a strong penalty can yield an overly sparse model and show unsatisfactory predictive performance (Musoro et al., 2014).

Here, we apply the current method of predictive-accuracy analysis to a regularized network estimated by Epskamp, van Borkulo, et al. (2018) for a retrospective evaluation of the risk of having low predictive accuracy for this network. Power analysis is not conducted for this example because of the difficulty of making unbiased statistical inferences about any edge (i.e., calculating the standard-error and p values) in a regularized network without proper correction for the selection process (Taylor & Tibshirani, 2015). Statistical inference of individual edges was also not performed by Epskamp, van Borkulo, et al. Although relevant correction techniques exist (Lee et al., 2016; Waldorp & Haslbeck, 2024), they are not implemented in the commonly used software packages for network analysis yet.

Data used for the network analysis were collected from a female patient who suffered from major depressive disorder. The patient received five prompts per day at an average interval of 3 hr for 2 weeks, resulting in a total number of 70 prompts. Seven variables were used in the network analysis, including sadness, tiredness, rumination, bodily discomfort, nervousness, relaxation, and the ability to concentrate. After excluding missing data and overnight lags, 47 time points were available for the analysis.

The estimated regularized network is shown in Figure 8. Note that given the difficulty of statistical inferences with regularized models, the decisions of which

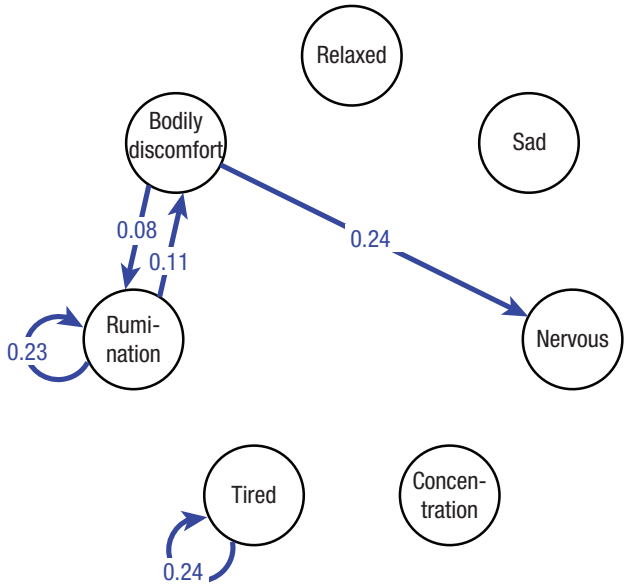

Fig. 8. The estimated regularized temporal network in Epskamp, van Borkulo, et al. (2018).

edges to visualize cannot be based on the results of the significance tests of them anymore as for the standard networks discussed previously. Thus, all nonzero edges are visualized in the network. Accordingly, only the point estimates of the edges are used as the parameter values during data simulation—no plausible range of SPAP can be calculated given each value of T.

We conducted simulation-based predictive-accuracy analysis on this network after minor changes in the network-estimation process: All the zero edges were constrained to be 0 during the estimation. This adapted estimation process has been used in related algorithms, such as relaxed LASSO (Meinshausen, 2007). All other procedures remain the same as previously described.

Results of the analysis are shown in Table 10. With samples of 47 time points, only 42% of the networks estimated with the simulated time series demonstrate sufficient predictive accuracy with the threshold of R^2 = .9, suggesting a high risk of overfitting for the

Table 10. Sufficient Predictive Accuracy Probability of the Regularized Network in Epskamp, van Borkulo, et al. (2018)

T	47	75	100
$R^2 = .9$	.420	.742	**.876**
$R^2 = .92$	.264	.567	.760
$R^2 = .95$	.068	.212	.393

Note: Values in bold denote results that exceed the target performance of .8.

network estimated in Epskamp, van Borkulo, et al. (2018). On a positive note, only 100 time points are required for the probability of estimating a nonoverfitting regularized network to be sufficiently high in this case ($\text{SPAP} > .8$), which is much less demanding than standard networks.

To summarize, results of retrospective predictive-accuracy analysis show that the estimated networks in both Bak et al. (2016) and Epskamp, van Borkulo, et al. (2018) are under considerable risk of overfitting, except for the network of Miss A's stable state. Such results further suggest that the temporal networks estimated in other studies (see Table 3) are also likely to overfit given that in most cases, more nodes and fewer time points were used in the network analysis (except for Bos et al., 2018; Wichers et al., 2016).

Discussion

As a promising tool for clinical research and practice, idiographic temporal networks have been receiving increasing interest from researchers and clinicians. However, reasonable quality concerns of such networks, for example, the replicability of each individual edge and the risk of overfitting for the whole network, have been raised but not sufficiently handled in previous literature (Bringmann, 2021). In this article, we showed how simulation-based power analysis and predictive-accuracy analysis can be used to tackle these concerns. Applying both methods before starting a new study (a priori) can help researchers determine a sufficient number of time points to collect from an individual, which helps ensure satisfactory edge replicability and network predictive accuracy. Moreover, both methods can also be applied retrospectively to quantify the risk of both quality problems for a previously estimated network.

Our key findings are as follows. First, for both power analysis and predictive-accuracy analysis, we showed that the number of time points in a sample (T) is a crucial factor that influences the quality of a temporal network: Given the same set of true network parameters, both the statistical power of the significance tests for edges and SPAP increase as T gets larger. Second, more complex networks (e.g., with a larger number of nodes) typically require a larger number of time points in the samples for estimation to reach the thresholds of sufficient statistical power and predictive accuracy. Third, previous network studies usually did not employ a sufficiently large number of time points, which leads to overfitting and low statistical power. Our findings suggest that networks estimated in previous studies should be interpreted with caution because they often suffer from such quality problems. Finally, regularization methods can effectively reduce the risk of overfitting for networks given that the seven-node regularized network in Epskamp, van Borkulo, et al. (2018) requires much fewer time points to reach the threshold of SPAP than the five-node standard networks in Bak et al. (2016). However, when the number of time points available for the network analysis is too small, even a regularized network can still overfit.

The simulation-based methods described in this article can efficiently provide sample-size suggestions in an "idealistic" setting with no assumption violation of the VAR(1) model. Thus, feasibility and practical considerations should still have a role in the sample-size-planning procedures for acquiring a valid suggestion. When power analysis and predictive-accuracy analysis are used, the suggested number of time points refers only to the number of observations that are accompanied by valid lagged values and are not missing themselves. However, the more practical question researchers and clinicians need to answer with sample-size planning is how many measurement prompts they need to send the participants. Thus, additional steps need to be taken beyond the simulation-based analysis to acquire a more practical sample-size suggestion. Imagine an example in which the power analysis and predictive-accuracy analysis suggest that a researcher should collect 100 time points so that the network to be estimated can reach the aforementioned thresholds. If the researcher decides to send five measurement prompts to the participant each day, one out of five prompts will not have valid lagged values because of overnight lags. Moreover, if the researcher expects a compliance rate of 80% from the participant, then a maximum of 40% of the observations will be either missing themselves or have missing lagged values. Given these considerations, a safe suggestion for the number of time points to collect can be calculated as:

$$T = \frac{100}{\left(1 - \frac{1}{5}\right)\left(1 - 0.4\right)} = 208, \tag{10}$$

which is much larger than the suggestion made by the idealistic simulation-based analysis. Likewise, when

conducting the retrospective analysis for a study, the number of predictable observations is needed but often not reported explicitly (see Table 3; Lafit et al., 2025). Therefore, we urge future ESM researchers to report such information explicitly.

A general impression readers get from such findings might be that the more time points in the sample, the better. This is indeed true in an ideal setting in which no model assumption is violated. However, it is unrealistic to expect that the assumption of stationarity always holds, especially when the sampling period becomes longer. For example, the change of treatment plans can potentially influence the temporal dynamics of a patient's symptoms (Wichers et al., 2016). Specifically, a more effective treatment will ideally result in a lower mean level of the symptoms and looser connections among different symptoms (Cramer et al., 2016; van Borkulo et al., 2015). On the other hand, a heightened mean level of symptoms and stronger connections among symptoms are often considered as a sign of an impending episode of mental-health crisis in the literature of early warning signals (e.g., Helmich et al., 2021; Schreuder et al., 2020; Smit et al., 2022). Thus, the number of time points in a sample and the risk of assumption violation for the VAR(1) model can become two competing factors of network quality. This delicate yet crucial balance between these two factors requires more careful consideration from researchers.

A starting point for moving forward might be to rethink the general guideline of "the more nodes, the better" and not to embrace a complexity that the statistical model is not yet ready for. Because our findings suggested that networks estimated with few nodes usually require a rather small number of time points to be of good quality (consistent with findings of Mansueto et al., 2023), we see two directions of improving the practice of temporal network analysis. First, more research should be done to assess the quality of regularized networks because using such regularization is an effective way to create sparser and, thus, smaller networks. Such research can include (a) power analysis of regularized networks with the help of proper inference methods (Dezeure et al., 2015; Waldorp & Haslbeck, 2024) and (b) extending the predictive-accuracy analysis to regularized networks. Second, researchers could consider shifting the purpose of using such networks from exploratory analysis to confirmatory analysis and from delineating a complex system with dozens of symptoms to a simple dynamic of a few symptoms. If clearer hypotheses and theories can be generated regarding the temporal dynamics in dyads or triads of symptoms (Bringmann, 2024; Bringmann et al., 2024; Eronen & Bringmann, 2021), the networks can be powerful tools to confirm or to falsify such hypotheses because they will require only a reasonable number of time points to be of good quality. With a shorter sampling period, patients will also be less burdened by the data collection, which can, in turn, appease the feasibility concerns many researchers have for long sampling duration in the first place. Besides "downsizing" the network, we also encourage researchers to think beyond the VAR(1) model and question whether the specification of this model reflects the dynamics of the symptoms accurately. In line with this, we consider it important to develop new statistical models that are more congruent with the network theory (e.g., incorporating context into analysis; Bringmann et al., 2024) and to evaluate these models by performing formal model comparisons with the VAR(1) model (Borsboom, 2022).

Reflecting on the data simulations performed in the current study, we acknowledge that these procedures tend to have rather strong assumptions. First, we assume that the parameter we specify in the a priori analysis and the parameter estimates we use in the retrospective analysis are accurate. We recommend that researchers conduct both a priori and retrospective analyses based on network estimates acquired in existing studies. Moreover, we encourage researchers to test the robustness of such network estimates via resampling approaches, such as bootstrapping, when they have access to the raw time-series data (Bringmann et al., 2013; Epskamp, Borsboom, & Fried, 2018). Second, the simulated innovation terms in the VAR(1) model are assumed to follow a multivariate normal distribution. For one of the studies we investigated in this article (Bak et al., 2016), we tested whether the empirical innovation terms followed such a multivariate normal distribution by calculating the multivariate kurtosis and skewness statistics (Cain et al., 2017). Our results suggested that even for the stable state, the state with the largest number of time points ($T = 353$), the distribution of innovation terms strongly deviates from a multivariate normal distribution. Such deviations are more severe in the other two states with smaller sample sizes and could potentially harm the validity of the simulation. To improve the validity of future simulation studies, a crucial first step should be deeper inspections of existing symptom time-series data. Such inspections can start with the features of the distributions of univariate time series, for example, whether they are unimodal (e.g., normal distribution) or multimodal (Haslbeck et al., 2023). For unimodal distributions, important questions can be whether floor or ceiling effects are present for certain symptoms (e.g., von Klipstein et al., 2023) and whether and how they might lead to bias in model estimates (e.g., Ernst & Albers, 2017; Haqiqatkhah et al., 2024; Mestdagh et al., 2018). Knowledge learned from such processes will help researchers adapt the simulation settings to be better able to embody the nonnormality and improve the ecological validity of the predictive-accuracy analysis.

With this article, we hope to emphasize the importance of careful sample-size planning when using idiographic temporal networks. For such single-case network studies, sample-size justification is equally if not more important than for cross-sectional studies, considering the complex network's tendency of overfitting. An insufficient sample size cannot result only in low statistical power—certain important lagged effects could be deemed statistically insignificant but also overfitting—the network estimates could be largely driven by noise and thus differ from the true temporal dynamics. Such inaccurate networks can be highly misleading for patients if used for personalized feedback because personalized feedback based on time-series analysis has been shown to have a considerable impact on the receiver's self-perception regardless of the feedback being accurate or not (Leertouwer et al., 2022). Therefore, we argue that sample-size justification from the perspectives of network quality should become the standard practice in studies using temporal networks. For this purpose, researchers should ideally conduct the a priori analysis during the planning phase of a study. If this is not realized, researchers should conduct the retrospective power and predictive-accuracy analysis to show whether an estimated network can reach the quality thresholds. Through defining a quality bare minimum and checking whether it can be met, such methods will advance the research practice of idiographic networks.

Transparency

Action Editor: David A. Sbarra
Editor: David A. Sbarra

Author Contributions

Yong Zhang: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Validation; Visualization; Writing – original draft; Writing – review & editing.

Jordan Revol: Conceptualization; Data curation; Formal analysis; Resources; Software; Validation; Writing – review & editing.

Ginette Lafit: Conceptualization; Supervision; Writing – review & editing.

Anja F. Ernst: Conceptualization; Supervision; Writing – review & editing.

Josip Razum: Investigation; Resources; Visualization; Writing – original draft; Writing – review & editing.

Eva Ceulemans: Conceptualization; Funding acquisition; Supervision; Writing – review & editing.

Laura F. Bringmann: Conceptualization; Funding acquisition; Methodology; Project administration; Supervision; Writing – review & editing.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

The research presented in this article was supported by research grants from the Fund for Scientific Research-Flanders (Project No. G0C9821N) and from the Research Council of KU Leuven (C14/23/062; iBOF/21/090) awarded to E. Ceulemans. The research of L. F. Bringmann was supported by the Dutch Research Council Veni Grant (NWO-Veni 191G.037). The research of Y. Zhang was supported by the University of Groningen PhD Fund of the Faculty of Behavioural and Social Sciences. Y. Zhang and L. F. Bringmann were also supported by the Stress in Action project (www.stress-in-action.nl), which is financially supported by the Dutch Research Council and the Dutch Ministry of Education, Culture and Science (NWO Gravitation Grant No. 024.005.010).

Open Practices

This article has received the badges for Open Data and Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iDs

Yong Zhang https://orcid.org/0000-0002-6313-2575
Jordan Revol https://orcid.org/0000-0001-5511-3617
Ginette Lafit https://orcid.org/0000-0002-8227-128X
Laura F. Bringmann https://orcid.org/0000-0002-8091-9935

Acknowledgments

We thank Peter de Jonge for providing valuable input to this study. We thank Marjan Drukker for kindly providing us the data necessary for the analysis and Sacha Epskamp for making the data and analysis code openly available. We thank Fridtjof Petersen for helping us design the flow diagrams used in this article. We are grateful to the editors, reviewers, Ena Vojvodic, Theodor Nowicki, and Ilse P. Peringa for providing helpful comments to the article.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/25152459251372116.

Notes

1. The residuals of a VAR(1) model represents the variance that cannot be explained by the previous time point. Thus, the residuals are considered as contemporaneous random shocks and called "innovation" in the time-series literature.

2. For an R Shiny app used for the analysis, see https://gitlab.kuleuven.be/ppw-okpiv/researchers/u0148925/shinyapp-paa_var_n1.

3. This is achieved by using the lm() function in R to estimate two regression models: Sad_t and $Suicide_t$ both being predicted by the same two predictors—Sad_{t-1} and $Suicide_{t-1}$.

4. We used a computer with an eight-core processor and 64 GB RAM for the analyses.

5. Descriptive statistics are not calculated for the number of time points available for analysis because of insufficient reporting. See further text for details.

6. Only edges that are significantly different from 0 ($p < .05$) are visualized in the network, in which each edge represents the point estimate of a lagged effect. The same applies to all networks shown in Figure 7.

7. Recently, methods have been developed for formal testing of the differences between idiographic networks, such as Hoekstra et al. (2023, 2024) and Siepe et al. (2024).

8. We remind the readers of critical viewpoints on using centrality indices for psychological networks, for example, Bringmann et al. (2019) and Siepe et al. (2025).

9. If the largest eigenvalue of either matrix exceeds 1 and thus leads to nonstationarity of the process, all coefficients for that specific matrix will be shrunk at the same ratio so that the largest eigenvalue is set as 0.99.

10. We observe that the ranges are very narrow in all cells and sometimes present an opposite comparison with what we expected: The SPAP with a larger effect size is sometimes higher than with a smaller effect size. The unexpected results are possibly due to sampling variability and show that a VAR(1) model's effect size tends to have a smaller impact on SPAP than the number of predictable time points, which is consistent with the findings of Revol et al. (2024).

References

Albers, C., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology*, *74*, 187–195. https://doi.org/10.1016/j.jesp.2017.09.004

Bak, M., Drukker, M., Hasmi, L., & Van Os, J. (2016). An n=1 clinical network analysis of symptoms and treatment in psychosis. *PLOS ONE*, *11*(9), Article e0162811. https://doi.org/10.1371/journal.pone.0162811

Bar-Kalifa, E., & Sened, H. (2020). Using network analysis for examining interpersonal emotion dynamics. *Multivariate Behavioral Research*, *55*(2), 211–230. https://doi.org/10.1080/00273171.2019.1624147

Beck, E. D., & Jackson, J. J. (2020). Consistency and change in idiographic personality: A longitudinal ESM network study. *Journal of Personality and Social Psychology*, *118*(5), 1080–1100. https://doi.org/10.1037/pspp0000249

Berkhout, S. W., Schuurman, N. K., & Hamaker, E. L. (2025). Let sleeping dogs lie? How to deal with the night gap problem in experience sampling method data. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000762

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

Borsboom, D. (2022). Possible futures for network psychometrics. *Psychometrika*, *87*(1), 253–265. https://doi.org/10.1007/s11336-022-09851-z

Borsboom, D., & Cramer, A. O. (2013). Network analysis: An integrative approach to the structure of psychopathology. *Annual Review of Clinical Psychology*, *9*(1), 91–121. https://doi.org/10.1146/annurev-clinpsy-050212-185608

Borsboom, D., Cramer, A. O. J., & Kalis, A. (2019). Brain disorders? Not really: Why network structures block reductionism in psychopathology research. *Behavioral and Brain Sciences*, *42*, Article e2. https://doi.org/10.1017/S0140525X17002266

Bos, F., Blaauw, F., Snippe, E., van der Krieke, L., de Jonge, P., & Wichers, M. (2018). Exploring the emotional dynamics of subclinically depressed individuals with and without anhedonia: An experience sampling study. *Journal of Affective Disorders*, *228*, 186–193. https://doi.org/10.1016/j.jad.2017.12.017

Brandt, P. T., & Williams, J. T. (2007). *Multiple time series models*. Sage Publications.

Bringmann, L. F. (2021). Person-specific networks in psychopathology: Past, present, and future. *Current Opinion in Psychology*, *41*, 59–64. https://doi.org/10.1016/j.copsyc.2021.03.004

Bringmann, L. F. (2024). The future of dynamic networks in research and clinical practice. *World Psychiatry*, *23*(2), 288–289. https://doi.org/10.1002/wps.21209

Bringmann, L. F., Ariens, S., Ernst, A. F., Snippe, E., & Ceulemans, E. (2024). Changing networks: Moderated idiographic psychological networks. *Advances in Psychology*, *2*, Article e658296. https://doi.org/10.56296/aip00014

Bringmann, L. F., Elmer, T., Epskamp, S., Krause, R. W., Schoch, D., Wichers, M., Wigman, J. T. W., & Snippe, E. (2019). What do centrality measures measure in psychological networks? *Journal of Abnormal Psychology*, *128*(8), 892–903. https://doi.org/10.1037/abn0000446

Bringmann, L. F., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., Borsboom, D., & Tuerlinckx, F. (2013). A network approach to psychopathology: New insights into clinical longitudinal data. *PLOS ONE*, *8*(4), Article e60188. https://doi.org/10.1371/journal.pone.0060188

Bulteel, K., Mestdagh, M., Tuerlinckx, F., & Ceulemans, E. (2018). VAR(1) based models do not always outpredict AR(1) models in typical psychological applications. *Psychological Methods*, *23*(4), 740–756. https://doi.org/10.1037/met0000178

Bulteel, K., Tuerlinckx, F., Brose, A., & Ceulemans, E. (2016). Using raw VAR regression coefficients to build networks can be misleading. *Multivariate Behavioral Research*, *51*(2), 330–344. https://doi.org/10.1080/00273171.2016.1150151

Cain, M. K., Zhang, Z., & Yuan, K.-H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior Research Methods*, *49*(5), 1716–1735. https://doi.org/10.3758/s13428-016-0814-1

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cramer, A. O. J., van Borkulo, C. D., Giltay, E. J., Van Der Maas, H. L. J., Kendler, K. S., Scheffer, M., & Borsboom, D. (2016). Major depression as a complex dynamic system. *PLOS ONE*, *11*(12), Article e0167490. https://doi.org/10.1371/journal.pone.0167490

Cramer, A. O. J., Waldorp, L. J., van der Maas, H. L. J., & Borsboom, D. (2010). Complex realities require complex

theories: Refining and extending the network approach to mental disorders. *Behavioral and Brain Sciences*, *33*(2–3), 178–193. https://doi.org/10.1017/S0140525X10000920

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

David, S. J., Marshall, A. J., Evanovich, E. K., & Mumma, G. H. (2018). Intraindividual dynamic network analysis – Implications for clinical assessment. *Journal of Psychopathology and Behavioral Assessment*, *40*(2), 235–248. https://doi.org/10.1007/s10862-017-9632-8

de Vos, S., Wardenaar, K. J., Bos, E. H., Wit, E. C., Bouwmans, M. E. J., & de Jonge, P. (2017). An investigation of emotion dynamics in major depressive disorder patients and healthy persons using sparse longitudinal networks. *PLOS ONE*, *12*(6), Article e0178586. https://doi.org/10.1371/journal.pone.0178586

Dezeure, R., Bühlmann, P., Meier, L., & Meinshausen, N. (2015). High-dimensional inference: Confidence intervals, p-values and R-software hdi. *Statistical Science*, *30*(4), 533–558. https://doi.org/10.1214/15-STS527

Engsted, T., & Pedersen, T. (2014). Bias-correction in vector autoregressive models: A simulation study. *Econometrics*, *2*(1), 45–71. https://doi.org/10.3390/econometrics2010045

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, *50*(1), 195–212. https://doi.org/10.3758/s13428-017-0862-1

Epskamp, S., van Borkulo, C. D., van der Veen, D. C., Servaas, M. N., Isvoranu, A.-M., Riese, H., & Cramer, A. O. (2018). Personalized network modeling in psychopathology: The importance of contemporaneous and temporal connections. *Clinical Psychological Science*, *6*(3), 416–427. https://doi.org/10.1177/2167702617744325

Epskamp, S., Waldorp, L. J., Mõttus, R., & Borsboom, D. (2018). The Gaussian graphical model in cross-sectional and time-series data. *Multivariate Behavioral Research*, *53*(4), 453–480. https://doi.org/10.1080/00273171.2018.1454823

Ernst, A. F., & Albers, C. J. (2017). Regression assumptions in clinical psychology research practice—A systematic review of common misconceptions. *PeerJ*, *5*, Article e3323. https://doi.org/10.7717/peerj.3323

Ernst, A. F., Timmerman, M. E., Jeronimus, B. F., & Albers, C. J. (2021). Insight into individual differences in emotion dynamics with clustering. *Assessment*, *28*(4), 1186–1206. https://doi.org/10.1177/1073191119873714

Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: How to move forward. *Perspectives on Psychological Science*, *16*(4), 779–788. https://doi.org/10.1177/1745691620970586

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fisher, A. J., Reeves, J. W., Lawyer, G., Medaglia, J. D., & Rubel, J. A. (2017). Exploring the idiographic dynamics of mood and anxiety via network analysis. *Journal of Abnormal Psychology*, *126*(8), 1044–1056. https://doi.org/10.1037/abn0000311

Frumkin, M. R., Piccirillo, M. L., Beck, E. D., Grossman, J. T., & Rodebaugh, T. L. (2021). Feasibility and utility of idiographic models in the clinic: A pilot study. *Psychotherapy Research*, *31*(4), 520–534. https://doi.org/10.1080/10503307.2020.1805133

Hall, M., Lappenbusch, L. M., Wiegmann, E., & Rubel, J. A. (2025). To use or not to use: Exploring therapists' experiences with pre-treatment EMA-based personalized feedback in the TheraNet project. *Administration and Policy in Mental Health and Mental Health Services Research*, *52*(1), 41–58. https://doi.org/10.1007/s10488-023-01333-3

Hamaker, E. L. (2012). Why researchers should think "within-person": A paradigmatic rationale. In M. R. Mehl, & T. S. Conner (Eds.), *Handbook of research methods for studying daily life* (pp. 43–61). The Guilford Press.

Hamaker, E. L., & Wichers, M. (2017). No time like the present: Discovering the hidden dynamics in intensive longitudinal data. *Current Directions in Psychological Science*, *26*(1), 10–15. https://doi.org/10.1177/0963721416666518

Hamilton, J. D. (1994). *Time series analysis*. Princeton University Press. http://www.jstor.org/stable/j.ctv14jx6sm

Haqiqatkhah, M. M., Ryan, O., & Hamaker, E. L. (2024). Skewness and staging: Does the floor effect induce bias in multilevel AR(1) models? *Multivariate Behavioral Research*, *59*(2), 289–319. https://doi.org/10.1080/00273171.2023.2254769

Haslbeck, J., Ryan, O., & Dablander, F. (2023). Multimodality and skewness in emotion time series. *Emotion*, *23*(8), 2117–2141. https://doi.org/10.1037/emo0001218

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2). Springer.

Helmich, M. A., Olthof, M., Oldehinkel, A. J., Wichers, M., Bringmann, L. F., & Smit, A. C. (2021). Early warning signals and critical transitions in psychopathology: Challenges and recommendations. *Current Opinion in Psychology*, *41*, 51–58. https://doi.org/10.1016/j.copsyc.2021.02.008

Hoekstra, R. H. A., Epskamp, S., & Borsboom, D. (2023). Heterogeneity in individual network analysis: Reality or illusion? *Multivariate Behavioral Research*, *58*(4), 762–786. https://doi.org/10.1080/00273171.2022.2128020

Hoekstra, R. H. A., Epskamp, S., Nierenberg, A. A., Borsboom, D., & McNally, R. J. (2024). Testing similarity in longitudinal networks: The Individual Network Invariance Test. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000638

Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., Margetts, H., Mullainathan, S., Salganik, M. J., Vazire, S., Vespignani, A., & Yarkoni, T. (2021). Integrating explanation and prediction in computational social science. *Nature*, *595*(7866), 181–188. https://doi.org/10.1038/s41586-021-03659-0

Johal, S. K., & Rhemtulla, M. (2024). Relating network-instantiated constructs to psychological variables through network-derived metrics: An exploratory study. *advances*.

in/psychology, *2*, Article e939409. https://doi.org/10.56296/aip00024

Kroeze, R., van der Veen, D. C., Servaas, M. N., Bastiaansen, J. A., Voshaar, R. C. O., Borsboom, D., Ruhe, H. G., Schoevers, R. A., & Riese, H. (2017). Personalized feedback on symptom dynamics of psychopathology: A proof-of-principle study. *Journal for Person-Oriented Research*, *3*(1), 1–11. https://doi.org/10.17505/jpor.2017.01

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer. https://doi.org/10.1007/978-1-4614-6849-3

Lafit, G., Adolf, J. K., Dejonckheere, E., Myin-Germeys, I., Viechtbauer, W., & Ceulemans, E. (2021). Selection of the number of participants in intensive longitudinal studies: A user-friendly shiny app and tutorial for performing power analysis in multilevel regression models that account for temporal dependencies. *Advances in Methods and Practices in Psychological Science*, *4*(1). https://doi.org/10.1177/2515245920978738

Lafit, G., Meers, K., & Ceulemans, E. (2022). A systematic study into the factors that affect the predictive accuracy of multilevel VAR(1) models. *Psychometrika*, *87*(2), 432–476. https://doi.org/10.1007/s11336-021-09803-z

Lafit, G., Revol, J., Cloos, L., Kuppens, P., & Ceulemans, E. (2025). The effect of different construct operationalizations, study duration, and preprocessing choices on power-based sample size recommendations in intensive longitudinal research. *Assessment*, *32*(2), 206–223. https://doi.org/10.1177/10731911241286868

Lafit, G., Tuerlinckx, F., Myin-Germeys, I., & Ceulemans, E. (2019). A partial correlation screening approach for controlling the false positive rate in sparse Gaussian graphical models. *Scientific Reports*, *9*(1), Article 17759. https://doi.org/10.1038/s41598-019-53795-x

Larson, R., & Csikszentmihalyi, M. (2014). The experience sampling method. In *Flow and the foundations of positive psychology* (pp. 21–34). Springer.

Lazarus, G., Sened, H., & Rafaeli, E. (2020). Subjectifying the personality state: Theoretical underpinnings and an empirical example. *European Journal of Personality*, *34*(6), 1017–1036. https://doi.org/10.1002/per.2278

Lee, J. D., Sun, D. L., Sun, Y., & Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, *44*(3), 907–927. https://doi.org/10.1214/15-AOS1371

Leertouwer, I., Vermunt, J., & Schuurman, N. K. (2022). *A pre-post design for testing insight from personalized feedback about positive affect in contexts*. PsyArXiv. https://doi.org/10.31234/osf.io/cfkrv

Leon, A. C., Davis, L. L., & Kraemer, H. C. (2011). The role and interpretation of pilot studies in clinical research. *Journal of Psychiatric Research*, *45*(5), 626–629. https://doi.org/10.1016/j.jpsychires.2010.10.008

Levinson, C. A., Hunt, R. A., Keshishian, A. C., Brown, M. L., Vanzhula, I., Christian, C., Brosof, L. C., & Williams, B. M. (2021). Using individual networks to identify treatment targets for eating disorder treatment: A proof-of-concept study and initial data. *Journal of Eating Disorders*, *9*(1), Article 147. https://doi.org/10.1186/s40337-021-00504-7

Levinson, C. A., Vanzhula, I., & Brosof, L. C. (2018). Longitudinal and personalized networks of eating disorder cognitions and behaviors: Targets for precision intervention a proof of concept study. *International Journal of Eating Disorders*, *51*(11), 1233–1243. https://doi.org/10.1002/eat.22952

Liu, X., & Wang, L. (2019). Sample size planning for detecting mediation effects: A power analysis procedure considering uncertainty in effect size estimates. *Multivariate Behavioral Research*, *54*(6), 822–839. https://doi.org/10.1080/00273171.2019.1593814

Loossens, T., Dejonckheere, E., Tuerlinckx, F., & Verdonck, S. (2021). Informing VAR(1) with qualitative dynamical features improves predictive accuracy. *Psychological Methods*, *26*(6), 635–659. https://doi.org/10.1037/met0000401

Loossens, T., Tuerlinckx, F., & Verdonck, S. (2021). A comparison of continuous and discrete time modeling of affective processes in terms of predictive accuracy. *Scientific Reports*, *11*(1), Article 6218. https://doi.org/10.1038/s41598-021-85320-4

Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer.

Mahalanobis, P. C. (2018). On the generalized distance in statistics. *Sankhyā: The Indian Journal of Statistics, Series A*, *80*, S1–S7.

Mansueto, A. C., Wiers, R. W., van Weert, J. C. M., Schouten, B. C., & Epskamp, S. (2023). Investigating the feasibility of idiographic network models. *Psychological Methods*, *28*(5), 1052–1068. https://doi.org/10.1037/met0000466

McGhie, S. F., & McNally, R. J. (2025). Posttraumatic stress disorder symptoms and positive affect: Individual and multilevel dynamic networks. *Psychological Trauma: Theory, Research, Practice, and Policy*, *17*(3), 593–602. https://doi.org/10.1037/tra0001605

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics & Data Analysis*, *52*(1), 374–393. https://doi.org/10.1016/j.csda.2006.12.019

Mestdagh, M., Pe, M., Pestman, W., Verdonck, S., Kuppens, P., & Tuerlinckx, F. (2018). Sidelining the mean: The relative variability index as a generic mean-corrected variability measure for bounded variables. *Psychological Methods*, *23*(4), 690–707. https://doi.org/10.1037/met0000153

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research & Perspective*, *2*(4), 201–218. https://doi.org/10.1207/s15366359mea0204_1

Mulder, J. D. (2022). Power analysis for the random intercept cross-lagged panel model using the powRICLPM R-package. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*(4), 645–658. https://doi.org/10.1080/10705511.2022.2122467

Musoro, J. Z., Zwinderman, A. H., Puhan, M. A., Ter Riet, G., & Geskus, R. B. (2014). Validation of prediction models based on lasso regression with multiply imputed data. *BMC Medical Research Methodology*, *14*(1), Article 116. https://doi.org/10.1186/1471-2288-14-116

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates.

Perspectives on Psychological Science, *9*(3), 319–332. https://doi.org/10.1177/1745691614528519

Piccirillo, M. L., & Rodebaugh, T. L. (2022). Personalized networks of social anxiety disorder and depression and implications for treatment. *Journal of Affective Disorders*, *298*, 262–276. https://doi.org/10.1016/j.jad.2021.10.034

Poole, M. A., & O'Farrell, P. N. (1971). The assumptions of the linear regression model. *Transactions of the Institute of British Geographers*, *52*, 145–158. https://doi.org/10.2307/621706

Reeves, J. W., & Fisher, A. J. (2020). An examination of idiographic networks of posttraumatic stress disorder symptoms. *Journal of Traumatic Stress*, *33*(1), 84–95. https://doi.org/10.1002/jts.22491

Revol, J., Lafit, G., & Ceulemans, E. (2024). A new sample-size planning approach for person-specific VAR(1) studies: Predictive accuracy analysis. *Behavior Research Methods*, *56*(7), 7152–7167. https://doi.org/10.3758/s13428-024-02413-4

Rocca, R., & Yarkoni, T. (2021). Putting psychology to the test: Rethinking model evaluation through benchmarking and prediction. *Advances in Methods and Practices in Psychological Science*, *4*(3). https://doi.org/10.1177/25152459211026864

Rowland, Z., & Wenzel, M. (2020). Mindfulness and affect-network density: Does mindfulness facilitate disengagement from affective experiences in daily life? *Mindfulness*, *11*(5), 1253–1266. https://doi.org/10.1007/s12671-020-01335-4

Schemer, L., Glombiewski, J. A., & Scholten, S. (2023). All good things come in threes: A systematic review and Delphi study on the advances and challenges of ambulatory assessments, network analyses, and single-case experimental designs. *Clinical Psychology: Science and Practice*, *30*(1), 95–107. https://doi.org/10.1037/cps0000083

Schreuder, M. J., Hartman, C. A., George, S. V., Menne-Lothmann, C., Decoster, J., Van Winkel, R., Delespaul, P., De Hert, M., Derom, C., Thiery, E., Rutten, B. P. F., Jacobs, N., Van Os, J., Wigman, J. T. W., & Wichers, M. (2020). Early warning signals in psychopathology: What do they tell? *BMC Medicine*, *18*(1), Article 269. https://doi.org/10.1186/s12916-020-01742-3

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310. https://doi.org/10.1214/10-STS330

Siepe, B. S., Kloft, M., & Heck, D. W. (2024). Bayesian estimation and comparison of idiographic network models. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000672

Siepe, B. S., Kloft, M., Zhang, Y., Petersen, F., Bringmann, L., & Heck, D. (2025). *Using features of dynamic networks to guide treatment selection and outcome prediction: The central role of uncertainty*. PsyArXiv. https://doi.org/10.31234/osf.io/2c8xf_v1

Smit, A. C., Schat, E., & Ceulemans, E. (2022). 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

Smyth, J. M., & Stone, A. A. (2003). Ecological momentary assessment research in behavioral medicine. *Journal of Happiness Studies*, *4*(1), 35–52. https://doi.org/10.1023/A:1023657221954

Strauss, G. P., Zamani-Esfahlani, F., Raugh, I. M., Luther, L., & Sayama, H. (2023). Network analysis of discrete emotional states measured via ecological momentary assessment in schizophrenia. *European Archives of Psychiatry and Clinical Neuroscience*, *273*(8), 1863–1871. https://doi.org/10.1007/s00406-023-01623-9

Taylor, J., & Tibshirani, R. J. (2015). Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, *112*(25), 7629–7634. https://doi.org/10.1073/pnas.1507583112

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Methodological*, *58*(1), 267–288. https://www.jstor.org/stable/2346178

van Borkulo, C., Boschloo, L., Borsboom, D., Penninx, B. W. J. H., Waldorp, L. J., & Schoevers, R. A. (2015). Association of symptom network structure with the course of depression. *JAMA Psychiatry*, *72*(12), 1219–1226. https://doi.org/10.1001/jamapsychiatry.2015.2079

van der Krieke, L., Blaauw, F. J., Emerencia, A. C., Schenk, H. M., Slaets, J. P., Bos, E. H., de Jonge, P., & Jeronimus, B. F. (2017). Temporal dynamics of health and well-being: A crowdsourcing approach to momentary assessments and automated generation of personalized feedback. *Psychosomatic Medicine*, *79*(2), 213–223. https://doi.org/10.1097/PSY.0000000000000378

van der Tuin, S., Balafas, S. E., Oldehinkel, A. J., Wit, E. C., Booij, S. H., & Wigman, J. T. (2022). Dynamic symptom networks across different at-risk stages for psychosis: An individual and transdiagnostic perspective. *Schizophrenia Research*, *239*, 95–102. https://doi.org/10.1016/j.schres.2021.11.018

van Der Velden, R. M., Mulders, A. E., Drukker, M., Kuijf, M. L., & Leentjens, A. F. (2018). Network analysis of symptoms in a Parkinson patient using experience sampling data: An n = 1 study: Symptom network analysis in Parkinson's disease. *Movement Disorders*, *33*(12), 1938–1944. https://doi.org/10.1002/mds.93

Verhagen, M. D. (2022). A pragmatist's guide to using prediction in the social sciences. *Socius: Sociological Research for a Dynamic World*, *8*. https://doi.org/10.1177/23780231221081702

Vogelsmeier, L. V. D. E., Jongerling, J., & Maassen, E. (2024). Assessing and accounting for measurement in intensive longitudinal studies: Current practices, considerations, and avenues for improvement. *Quality of Life Research*, *33*(8), 2107–2118. https://doi.org/10.1007/s11136-024-03678-0

Voigt, A., Kreiter, D., Jacobs, C., Revenich, E., Serafras, N., Wiersma, M., Os, J. V., Bak, M., & Drukker, M. (2018). Clinical network analysis in a bipolar patient using an experience sampling mobile health tool: An n=1 study. *Bipolar Disorder: Open Access*, *4*(1). https://doi.org/10.4172/2472-1077.1000121

von Klipstein, L., Riese, H., van der Veen, D. C., Servaas, M. N., & Schoevers, R. A. (2020). Using person-specific networks in psychotherapy: Challenges, limitations, and how we

could use them anyway. *BMC Medicine*, *18*(1), Article 345. https://doi.org/10.1186/s12916-020-01818-0

von Klipstein, L., Servaas, M. N., Lamers, F., Schoevers, R. A., Wardenaar, K. J., & Riese, H. (2023). Increased affective reactivity among depressed individuals can be explained by floor effects: An experience sampling study. *Journal of Affective Disorders*, *334*, 370–381. https://doi.org/10.1016/j.jad.2023.04.118

Waldorp, L., & Haslbeck, J. (2024). Network inference with the lasso. *Multivariate Behavioral Research*, *59*(4), 738–757. https://doi.org/10.1080/00273171.2024.2317928

Wang, Y. A., & Rhemtulla, M. (2021). Power analysis for parameter estimation in structural equation modeling: A discussion and tutorial. *Advances in Methods and Practices in Psychological Science*, *4*(1). https://doi.org/10.1177/2515245920918253

Wichers, M. (2014). The dynamic nature of depression: A new micro-level perspective of mental disorder that meets current challenges. *Psychological Medicine*, *44*(7), 1349–1360. https://doi.org/10.1017/S0033291713001979

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

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Zhou, D. J., Chahal, R., Gotlib, I. H., & Liu, S. (2024). Comparison of lasso and stepwise regression in psychological data. *Methodology*, *20*(2), 121–143. https://doi.org/10.5964/meth.11523