


Evaluation of a temporal causal model for predicting the mood of clients in an online therapy

Dennis Becker ¹, Vincent Bremer,¹ Burkhardt Funk,¹ Mark Hoogendoorn,² Artur Rocha,³ Heleen Riper⁴

¹Institute of Information Systems, Leuphana University of Lüneburg, Lüneburg, Germany
²Department of Computer Science, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

³Centre for Information Systems and Computer Graphics, INESC TEC, Porto, Portugal

⁴Department of Clinical, Neuro- & Developmental Psychology, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Correspondence to

Mr Dennis Becker, Institute of Information Systems, Leuphana University of Lüneburg, Lüneburg 21335, Germany; dbecker@leuphana.de

Received 29 November 2019

Revised 10 January 2020

Accepted 10 January 2020

ABSTRACT

Background Self-reported client assessments during online treatments enable the development of statistical models for the prediction of client improvement and symptom development. Evaluation of these models is mandatory to ensure their validity.

Methods For this purpose, we suggest besides a model evaluation based on study data the use of a simulation analysis. The simulation analysis provides insight into the model performance and enables to analyse reasons for a low predictive accuracy. In this study, we evaluate a temporal causal model (TCM) and show that it does not provide reliable predictions of clients' future mood levels.

Results Based on the simulation analysis we investigate the potential reasons for the low predictive performance, for example, noisy measurements and sampling frequency. We conclude that the analysed TCM in its current form is not sufficient to describe the underlying psychological processes.

Conclusions The results demonstrate the importance of model evaluation and the benefit of a simulation analysis. The current manuscript provides practical guidance for conducting model evaluation including simulation analysis.

INTRODUCTION

Mobile devices provide new possibilities to deliver internet-based cognitive behavioural therapy (CBT)^{1 2} and to measure clients' mental health, behaviour and activities.^{3 4} Ecological momentary assessment (EMA) is the term used to describe the assessment of clients' mood and behaviour throughout the day in their natural environment.^{5–7} EMA can encompass a diversity of data such as diaries, open-text and questions regarding the clients' symptoms and experiences using Likert-scaled responses.⁸ These collected time series data are a gateway to model symptom interaction and understand the psychological dynamics that occur in individuals over time.^{9 10} The collected EMA data provides patterns, which can be used to model relationships between symptoms and predict clients' future well-being.

With the increasing development of predictive models for both diagnostic and prognostic predictions, there is an intensified interest in the methodology on model evaluation.^{11 12} Besides statistical model evaluation using study data, clinical evaluation is required.^{13 14} However, clinical model evaluation

requires substantial effort and money, therefore only a fraction of available models can be evaluated in practice.^{15 16} Guidelines for model development and statistical evaluation provide methods to improve their validity and identify invalid models early. These guidelines encompass the definitions of prediction targets, predictors, statistical model evaluation and reporting.¹⁷ Suboptimal adherence to evaluation guidelines can limit the reliability and applicability of predictive models.¹⁸ For the statistical model evaluation, such guidelines suggest the use of cross-validation and bootstrapping.^{14 17} Cross-validation indicates the expected model performance of unobserved samples from the same study and bootstrapping allows to infer the significance of parameters and variance of predictions. We, however, argue that these methods are not sufficient for statistical model evaluation. Therefore, we suggest the inclusion of a simulation analysis. The simulation analysis is used to estimate the expected model performance on study data and the model's sensitivity regarding changes in the data. Accordingly, the simulation can be utilised to investigate specific reasons for poor model prediction and provide insights into the models'

Summary box

What is already known about this subject?

- ▶ Statistical model evaluation on study data is suggested, but results obtained on a single data set can be misleading.
- ▶ Simulation analysis has shown to be beneficial to analyse model behaviour for otherwise difficult to realise scenarios.
- ▶ There is a need for evaluation guidelines.

What are the new findings?

- ▶ Multiobjective models require more consideration in the literature of statistical model evaluation.
- ▶ A simulation analysis allows to evaluate the results obtained on study data and should be utilised for statistical model evaluation.
- ▶ A thorough model evaluation is required to increase trust into these models for use in online interventions.

How might it impact on clinical practice in the foreseeable future?

- ▶ New model evaluation guidelines that consider simulation analysis and account for models that predict multiple objectives.



© Author(s) (or their employer(s)) 2020. No commercial re-use. See rights and permissions. Published by BMJ.

To cite: Becker D, Bremer V, Funk B, et al. *Evid Based Ment Health* 2020;**23**:27–33.

prediction performance under varying study conditions. Furthermore, treatment decisions can consider multiple objectives, where an improvement in one dimension might lead to a reduction in another. Likewise, predictive models become increasingly complex and allow to predict multiple objectives simultaneously. Therefore, methods for comparing these models are required as well.

In this paper, we demonstrate a thorough model evaluation combining performance estimation on study data and simulation analysis. The examined predictive model is the so-called social integration model (SIM),¹⁹ which in a preliminary model evaluation was suggested to provide reliable predictions for clients' future mood levels. It describes the relationship between social interactions of study participants and their mental well-being. The SIM is a temporal-causal model (TCM)^{20 21} that allows to predict the course of multiple EMA factors. In general, TCMs are continuous dynamic network models that describe a graph of connected states using differential equations. They allow to universally model any dynamic system, simulate its change over time and have been shown to be applicable for a wide range of domains.^{22 23}

For the estimation of the model performance on study data, we use the complete EMA data collected in a Europe-wide depression study.²⁴ A framework for TCM comparisons²⁵ is used for performance estimation and comparison to a reference model. The framework further allows to compare the performance with respect to all EMA measures and utilises client individual model parameters, which have been shown to provide more accurate results than using the same model parameters for each client.^{26–28} The simulation analysis allows to estimate the theoretical model performance on the study dataset and to investigate reasons for differences in the performance. We investigated the literature regarding potential downsides of EMA measures and identified the influence of measurement noise^{29 30} and too few factor assessments^{31–33} as possible reasons that can lead to low model performances. Therefore, we utilise the simulation analysis to investigate these reasons and use the study data to inform the simulation analyses. By systematically altering the noise and missing values on the simulated data, the model's sensitivity to these influences can be assessed. Specifically, for the analysis of noise, data with the same assessment frequency as the study data are generated. To analyse the influence of fewer EMA assessments, the EMA measures' SD from the study data is used for data generation.

This study demonstrates a thorough model evaluation. We show that besides results obtained on study data with using client individual model parameters and a comparison to a reference model, a simulation analysis to assess the model's robustness for varying conditions is required. The simulation analysis is designed to reflect the study conditions and allows to infer the theoretical model performance that would be expected on the study data. If the results obtained in both analyses contradict each other, then there is reason to believe that the model does not represent the underlying dynamics accurately. The simulation further allows to investigate and eliminate possible reasons for the obtained differences in both analyses. By employing these methods, models that do not provide reliable predictions can be identified early.

METHOD

Social integration modelling

The SIM¹⁹ is a TCM that describes the relationship between clients' social contact and mood and allows to simulate their future behaviour. Social integration exhibits a relation to well-being and social isolation can foster mental health issues.³⁴ People that are socially well-integrated and have more social contact are usually happier than individuals with limited social

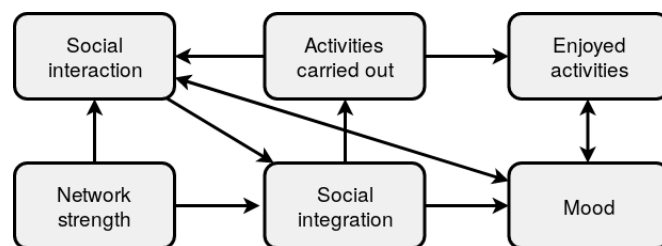


Figure 1 Visualisation of the social integration model.

contacts.³⁵ Contrary, people with depression tend to report feelings of loneliness, are less likely to engage in social activities, and have fewer social contacts.³⁶ These relationships provide the basis for the model and an overview of the interaction is shown in figure 1.

In the SIM, the mood level is influenced by the daily social interactions and how well the person is integrated into their social network. Specifically, the number of social activities and the perception of how enjoyable these activities have been. People with depression often perceive social interactions as less rewarding while interactions with close friends are more valued.^{37 38} Likewise, the daily mood level can influence the motivation to conduct activities and the perception of social interactions. A low mood level sensitises peoples' perception regarding social interactions and they are more likely to emphasise negative social interactions.³⁹ This can create a vicious cycle, where a lower mood can lead to fewer conducted activities and less social contacts, which eventually can result in low model levels and depression.^{40 41}

The strength of the influence among the different factors, which are the model's parameters, can significantly vary between healthy and unhealthy individuals and among them in general.^{42 43} Therefore, the model parameters are estimated for each client to provide client individual predictions. The model predictions are derived for each EMA factor on a daily basis. This results in a multiobjective optimisation problem for the parameter estimation and prediction error estimation for each measured EMA factor.

Evaluation of temporal predictive models

For model evaluation, we utilised a framework proposed by van Breda *et al.*²⁵ This framework assesses temporal models in the domain of mental health and allows to infer performance measures. These performance measures consider the models' fit and prediction of all EMA measures, and provide a comprehensive summary by considering each objective. It, therefore, provides methods to compare multiobjective models.

To infer these performance measures, client-specific model parameters are estimated by minimising the root mean square error (RMSE) of each objective. This results in a multiobjective optimisation problem. For solving this optimisation problem, the framework utilises the NSGA2 (non-dominated sorting genetic algorithm II)⁴⁴ optimisation algorithm. NSGA2 is a genetic algorithm for multiobjective optimisation that optimises with respect to each model objective. This process results in a set of valid solutions. Each solution provides an error for each optimised objective which can be considered as a point in Euclidean space. These points represent the Pareto optimal front, where an improvement in one dimension leads to a reduction in another dimension. Therefore, optimising one dimension leads to a trade-off among the conflicting optimisation objectives. To summarise how well the optimisation objective has been met, the dominated

hyper-volume measures the multidimensional volume covered by the Pareto optimal front and a reference point. Specifically, the dominated hyper-volume provides a measure of how well the data can be approximated by the model.

This dominated hyper-volume is utilised by the framework for the estimation of a models' *descriptive performance*. The *descriptive performance* is calculated by normalising the mean over all dominated hyper-volumes resulting from the estimated parameters for each client by its SD. In other words, this performance measure indicates the models' fit to the data with respect to all objectives.

Similarly, a *predictive performance* is estimated by combining the prediction error of unobserved data and their correlation to the models' fit of the data. To represent both performances, the *predictive performance* consists of the mean of the absolute and relative predictive performance. The absolute predictive performance describes the fit of unobserved future data using the estimated parameters. For deriving this measure, the prediction error over all optimisation objectives and estimated models is summarised by the mean and variance of the prediction error. It provides an estimate of how well the model represents future data. The relative predictive performance is calculated using the correlation between the fitting error and the prediction error. This measure ensures that models with a low fitting and prediction error receive a high relative predictive performance. In summary, the *descriptive performance* describes the fit of the model to the data and the *predictive performance* the models' capability of predicting future data. The performance measures are bound between 0 and 1, where a higher value indicates better performance. The performance measures provided by the framework are point estimates, therefore, we applied bootstrapping for CI estimation.

Regarding the framework parameters, we chose a population size of 80 and 100 generations for NSGA2 for individual model parameter estimation. The remaining parameters for the NSGA2 algorithm were set in accordance with the suggestion of the framework (crossover probability of 0.7 and a mutation probability of 0.2). We inspected the improvement of the dominated hyper-volume over the consecutive generations to confirm that the algorithm converges to a stable solution, which ensures a robust performance measure estimation. Further, we chose to execute five independent runs for each client, which reduces variability in the estimated performances. For more information regarding this framework, see van Breda *et al.*²⁵

Experimental setting and data

The utilised EMA data for model evaluation in this study originates from the European Union funded project E-COMPARED (European COMPARative Effectiveness research on online Depression),²⁴ which compared blended CBT (experiment group) and face-to-face treatment (control group) for depression. For the participants in the experimental group, EMA data were gathered and CBT was provided using a mobile phone between February 2015 and February 2018. An overview of the analysed data is provided in table 1.

The EMA measures have been assessed with varying granularity over the treatment period. The factor mood was inquired once a day at a random time between 10:00 and 22:00. The remaining factors were prompted once a week on a random day. However, during the first and last week of the treatment, the assessment was intensified and all factors were inquired daily. Since the study was conducted in eight different European countries, the suggested application use period differed among the

Table 1 Utilised ecological momentary assessment data

| Concept | Assessment question | Assessment frequency |
|------------------------|---|---|
| Mood | How is your mood right now? | Daily |
| Activities carried out | To what extent have you carried out enjoyable activities today? | Daily during first and last week; random day in other weeks |
| Enjoyed activities | How much did you enjoy activities today? | First and last week; random day in other weeks |
| Social interaction | How much were you involved in social interactions today? | Daily during first and last week; random day in other weeks |

participating countries and ranged between 6 and 20 weeks. Furthermore, clients were allowed to contribute mood measures at any time which allows for additional measures besides the one defined in the assessment protocol.

Study data analysis

To assess the model's performance on the study data, we chose a period of 6 weeks for model training and aimed to predict the data of the seventh week of treatment. Considering the prediction of the whole upcoming week might be a suitable time frame for therapists to inform short-term treatment decisions such as identifying drop-out risk or maximising short-term outcomes. The dataset consists of 324 clients. However, by considering the first 7 weeks of treatment for all clients, 112 clients provide data for model training and prediction error estimation. In the case of multiple mood measures a day, the corresponding mean value for that day was used. Afterward, all EMA data were normalised between 0 and 1.

For deriving performance measures on the study data, we utilised the framework for model comparison to compute the descriptive and predictive performance of the SIM for each client. These performances summarise the models' capabilities with respect to each EMA measure and allowed a comparison of the models' descriptive and predictive power to a reference model. This reference model, called *mean model* here, was defined as the mean value of the training data for each clients' psychological factor, which was then used as prediction. Comparison to a reference model allows an objective evaluation of the results. Furthermore, the RMSE for each factor was additionally utilised to compare the performance on each EMA measure individually.

Simulation analysis

This simulation allows to estimate the expected performance on the study data and helps to analyse reasons for a low model performance. From the literature on EMA, we identified that self-reported measures are affected by a (1) *high noise level*, which has an impact on the model performance. Potential causes for measurement errors include a lack of question comprehension by questionnaire participants, the influence of question order or the number of response alternatives.²⁹ To provide a meaningful answer to the EMA questions, clients' typically utilise contextual information to infer what the researcher might be interested in.³⁰ Their interpretation of the question, therefore, goes further than their literal meaning. Additionally, clients could consider different periods to answer the question ranging from right now to the last few hours. This variability in the interpretation of the question suggests that clients could include relatively minor events into their answers which translates to a variance in ratings. A related issue is the clients' scale usage. They anchor the endpoints of the scale with a low and high event that they experienced. This leads to a relative ranking of the current event

Table 2 Average SD for each ecological momentary assessment factor

| Concept name | Mood level | Enjoyed activities | Social interaction | Activities carried out |
|------------------------|---------------|--------------------|--------------------|------------------------|
| Average deviation (SD) | 0.137 (0.047) | 0.173 (0.060) | 0.200 (0.066) | 0.181 (0.066) |

regarding the most extremes. Therefore, events can receive a lower rating the more intense the previous events have been that serve as a high anchor.^{45 46} Similarly, concurrent ratings will be ranked according to previously still highly memorable ratings and clients tend to adapt their rating according to their currently poor health.⁴⁷

Another reason for a low predictive performance, can result from (2) *too few data points* which can lead to biased parameter estimates and poorer model performance.⁴⁸ EMA data are affected to a varying degree by missing data when clients struggle to respond to the inquiries.^{31 32} The reported compliance rates among studies considerably vary with an expected compliance rate of 75%.⁴⁹ Types of missing values are typically categorised as: missing completely at random, missing at random and missing not at random.⁵⁰ For values missing completely at random, there are no systematic differences between observed and unobserved values. Measures that are missing at random imply a likelihood that can be derived from the data, such as people with higher age might be more forgetful about reporting their measures. Missing not at random describes data where the likelihood of not observing the value depends on itself. Consequently, a rating of depression could be missing because of severe depressive symptoms, which resulted in a missed inquisition.³³ Further, this could also be the case if the assessment protocol did not define an assessment on that particular day.

The simulation analysis allowed us to control and inspect high noise levels and fewer data points separately. For both simulation analyses, we used the SIM to simulate 100 time series with parameters randomly chosen from the parameters estimated in the study data analysis. This further links the simulation analysis to the study data analysis by using parameters that are likely to be encountered from clients in a study. For the analysis of noise, we utilise the same number of measures per week as defined in the study protocol. Specifically, the simulated data includes five measures of mood and one measure of every other factor per week. To the simulated data, we add stepwise increasing Gaussian noise with a SD ranging from 0 to 0.5 with a step size of 0.01. To analyse the influence of a varying number of weekly assessments, we simulated data ranging from 1 to 7 factor measurements per week. For this analysis, the amplitude of the additional noise was estimated for each EMA factor from the study data. Where we assume that the EMA factors in the study data are constant over the considered period of 7 weeks. Following, we estimate the SD for each factor and utilise these to generate the additional noise in the simulated data. It is unlikely that these concepts are indeed constant over the considered period, thus, we consider the estimated noise levels as a worst-case estimate. The estimates are shown in table 2. These simulations enabled to estimate the models' sensitivity to these influences and to compare the predictive capabilities on the study data and the results of the simulation. If, however, neither a (1) *a high noise level* or (2) *too few data points* can explain a low model performance on the study data, it might be plausible that (3) *the SIM does not represents the dynamics of mood development sufficiently*.

Table 3 Performance measures on the study data

| Model | Descriptive performance (95% CI) | Predictive performance (95% CI) |
|--------------------------|----------------------------------|---------------------------------|
| Social integration model | 0.853 (0.837 to 0.869) | 0.492 (0.483 to 0.502) |
| Mean model | 0.839 (0.823 to 0.856) | 0.502 (0.447 to 0.561) |

RESULTS AND DISCUSSION

Study data analysis

In the following, we estimated the predictive and descriptive performance of the social integration and *mean model* on the study data, which are shown in table 3.

The performance scores indicate that both models provide similar predictive and descriptive performance. However, the descriptive performance of the SIM is slightly higher whereas the mean model has a higher predictive performance. If we consider the 95% CIs of both scores, we cannot suggest a significant difference among both models.

For comparing the average fitting and prediction error of each factor, we estimated the RMSE on each EMA measure on the training and test data. The estimated errors are illustrated in table 4.

Inspection of the training errors suggests that the SIM provides a slightly better fit to the training data compared with the *mean model* as suggested by the descriptive performance. A t-test on the average fit of both models does not provide evidence that there is a significant difference (p value=0.0698) between both models' fit to the data. There is, further, no evidence that the prediction error of the SIM is lower than the prediction error of the mean model, which was similarly indicated by the predictive performance measure.

Therefore, the performance measures, as well as the RMSE, indicate the same trend: the SIM might provide a closer fit to the data but does not provide a better prediction performance compared with the *mean model*. Although the SIM has higher complexity in terms of free parameters than the mean model, it does not provide more accurate future predictions. The two possible reasons for this finding, which we previously defined ((1) high measurement noise and (2) too few data points) are examined more closely in the simulation analysis.

Simulation analysis of measurement noise

Figure 2 illustrates the influence of noise on the model performance and prediction error in the simulation analysis. With an increase in noise, the descriptive and predictive performance of the SIM and *mean model* decreased. However, the predictive performance of the SIM was higher than the predictive performance of the mean model even though they were approaching each other with an increase in noise. The same finding applied to the RMSE of both models, where up to a noise of 0.24 we estimated a significant difference using a one-tailed t-test (p value=0.0002), with a higher noise level both models' prediction RMSE are indistinguishable.

When comparing the average prediction RMSE on the simulated data with the averaged error on the study data (0.187), we noticed that at a noise level with a SD of 0.4, the prediction RMSE of the SIM approached the prediction RMSE from the study data. The simulation showed that for higher noise levels, the SIM does not provide an average prediction RMSE below the error of the mean model. The same observation can be made for the predictive performance, which also considers the relationship between the model's fit to the data and prediction error. At a noise with a SD of 0.35 the predictive performance approaches

Table 4 Average prediction root mean square error for each factor

| Concept name | Mood level (SD) | Enjoyed activities (SD) | Social interaction (SD) | Activities carried out (SD) | Average (SD) |
|-----------------------------------|-----------------|-------------------------|-------------------------|-----------------------------|---------------|
| Training root mean square error | | | | | |
| Social integration model | 0.129 (0.014) | 0.152 (0.018) | 0.175 (0.023) | 0.167 (0.022) | 0.157 (0.020) |
| Mean model | 0.134 (0.014) | 0.163 (0.020) | 0.192 (0.024) | 0.172 (0.025) | 0.167 (0.024) |
| Prediction root mean square error | | | | | |
| Social integration model | 0.142 (0.021) | 0.189 (0.045) | 0.208 (0.061) | 0.203 (0.075) | 0.187 (0.030) |
| Mean model | 0.146 (0.022) | 0.179 (0.040) | 0.186 (0.043) | 0.183 (0.051) | 0.174 (0.019) |

the results on the study data. At this noise level, the predictive performance of the SIM is above the predictive performance of the mean model in the simulation.

Additionally, a noise level with a SD of 0.4 or 0.35 appears to be too high. For example, if we assume the rating to be normally distributed with a mean value of 0.5 and a SD of 0.4, then there is approximately 21% of the probability mass outside the valid range of the ratings [0,1]. This would result in many maximal and minimal ratings. However, in the study data, we observed that only 2.1% of the analysed measures are such extreme points (ratings of 0 or 1). According to our assumption that the ratings are normally distributed with a mean of 0.5, this would rather suggest a SD of 0.22. At this noise level, the simulation results in a lower prediction RMSE for the SIM than the mean model.

The simulation showed that with data comparable to the study data, in terms of sparsity and additional noise, the SIM should provide a lower prediction error than the mean model. Since this was not the case in our analysis, we conclude that high noise level was not responsible for the low prediction performance of the SIM.

Simulation analysis of weekly assessed measures

For analysing the effect of missing values per week, we utilised the estimated SD from the study data and simulated data with an increasing number of missing values per week. The estimated model performance on the simulated data is illustrated in figure 3.

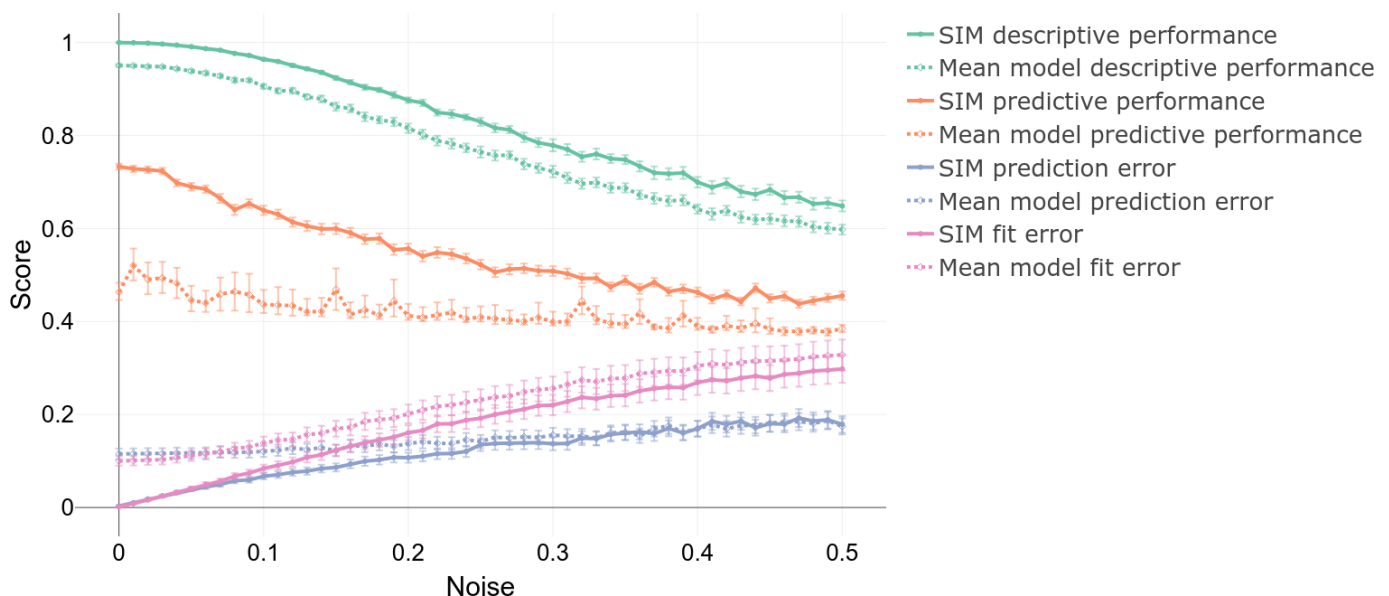
A reduction in simulated measures led to a slight increase in the descriptive performance for the SIM because the model can fit the reduced number of data points more accurately. The descriptive performance of the mean model appears mainly unaffected.

The predictive performance dropped with an increase in missing values for the SIM because the model parameters that provide a close fit to the data do not allow accurate future predictions. The predictive performance of the mean model was slightly reduced with fewer measures. With fewer measures, the noise has a stronger impact on the estimated mean value. The change in prediction performance is also reflected in the prediction RMSE of both models. The RMSE of the fit to the data of the SIM is smaller than for the mean model. Further, the predictive performance of the SIM in the simulation is higher than on the study data. In summary, by comparing the prediction RMSE and predictive performance in the simulation for one through six missing values per week to the results estimated on the study data, one notices that the prediction errors in the simulation are lower than on the study data.

This simulation shows that a reduction in measures cannot explain the lower performance of the SIM on the study data. The simulation analyses suggested that for both analysed cases the SIM should provide higher predictive performance than the mean model, which is not the case on the study data. We, therefore, conclude by the process of elimination that only the third reason for the low predictive performance of the SIM remains as a valid option. That is, the SIM in its current form does not represent the real-world relationships of psychological concepts to a degree that supports predictions.

CONCLUSION

In this analysis, we demonstrated a detailed evaluation of the predictive capabilities of a temporal-causal model. We employed

**Figure 2** Influence of increasing noise on the performance measures. SIM, social integration model.

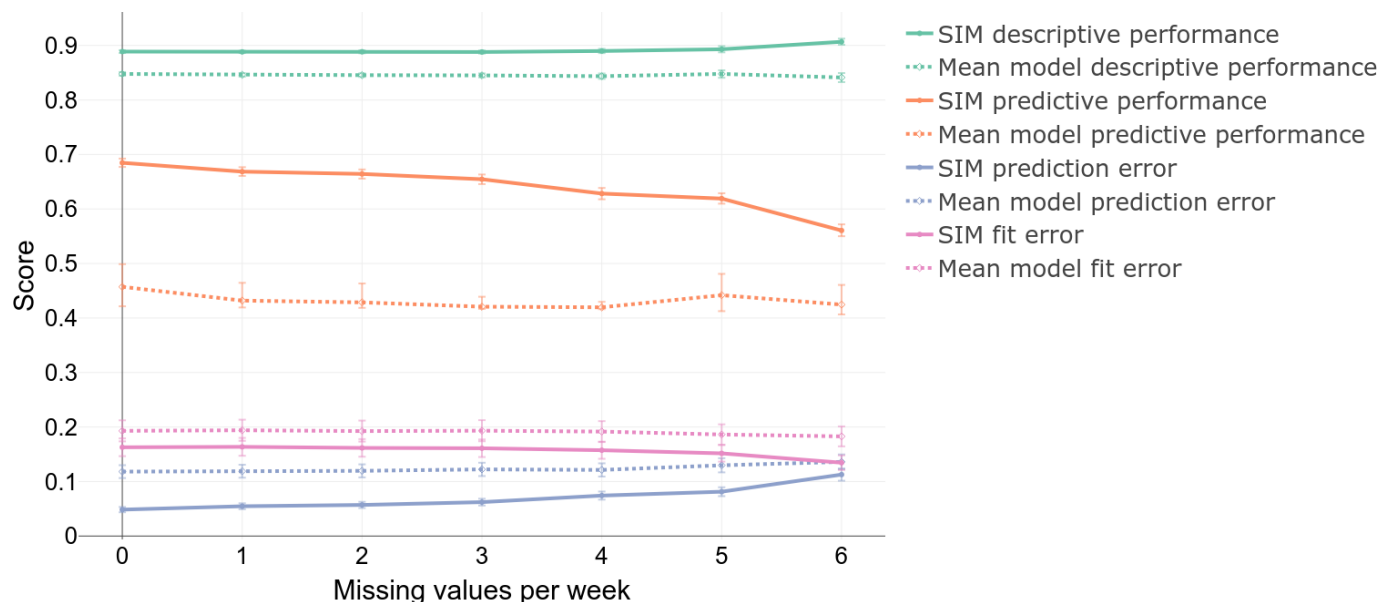


Figure 3 Influence of a reduction of Weekly measures on the performance measures. SIM, social integration model.

an EMA dataset for predicting clients' upcoming EMA ratings for a week in advance using the SIM.¹⁹ For the evaluation, we demonstrated the use of a simulation analysis and compared the SIM to a reference model (ie, the mean of EMA ratings observed so far). We evaluated the descriptive and predictive performance for both models on the study data and in a simulation analysis. Subsequently, the predictive capability of the SIM was not superior to the mean model prediction on the study data. We argued that this finding could be explained as follows: (1) measurement errors (noise) in the EMA data, (2) sparsity of measurements (eg, due to missing values), (3) or that the SIM does not fully represent the psychological dynamics. To investigate these reasons, the performance of both models was analysed for an increasing measurement error (1) and fewer weekly assessed measures (2) in the simulation analysis. For both simulations, the study data were used to inform the analysis. In the case of measurement noise, we simulated an increasing noise level until we matched the error on the study data and approximated the expected noise level based on clients' extreme ratings (rating of 0 or 1). In the case of fewer measures (2) we used the average variance of each factor among clients as an estimate for the measurement noise. For the simulated data we showed that regarding the measurement noise and reduced weekly measurements, the SIM should provide a better descriptive as well as predictive performance than the reference model. Since this was not the case on the study data, we excluded (1) measurement errors and (2) the sparsity of observations as potential reasons for the experimental results that is the SIM is not superior to the reference model. Thus, we conclude that the SIM in its current form is limited when the goal is predicting future mental states of clients.

We proposed and applied a systematic model evaluation that combines study data and simulation analysis. For both evaluations, the use of client individual parameters and comparison to a reference model is required. This is accomplished by using a model evaluation framework that provides methods for multi-objective model performance estimation and model comparison. The simulation analysis is used to analyse reasons for low model performances described in the literature. Further, the simulation is designed to reflect the study conditions to enable a comparison of the results. This provides insight into

the theoretical model performance which should be in an agreement with the results obtained on study data to provide evidence for the model to be accurate. These analyses also provide insights into the model robustness under varying study conditions. These insights can be used to state requirements on the study data assessment necessary for the model performance. We demonstrated the benefits of a simulation analysis and suggest that a simulation can complement guidelines and requirements for statistical model evaluation. The presented setup enables researchers to examine the impact of measurement errors and missing values on the predictive capabilities of their model. We thus hope to provide a solid setup for model evaluation in subsequent studies.

Acknowledgements We thank the participating associations and organisations for their contribution.

Contributors Implementation of the algorithms and analysis: DB. Writing on the publication: DB, VB, BF. Contributions to the psychological aspect: HR. Initial idea: BF. Data made available: HR. Support in the methodical aspects: MH, AR. Reviewing during the writing process and additional ideas and remarks: MH, HR, AR.

Funding The European Comparative Effectiveness Research on Internet-based Depression Treatment (E-COMPARED) is a project with funding from the European Union Seventh Framework Programme (grant agreement No: 603098).

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request. The utilized data can be made available for researchers upon request and via the webpage: <https://psydata.nl/>

ORCID iD

Dennis Becker <http://orcid.org/0000-0002-1437-6127>

REFERENCES

- de Graaf LE, Gerhards SAH, Arntz A, et al. Clinical effectiveness of online computerised cognitive-behavioural therapy without support for depression in primary care: randomised trial. *Br J Psychiatry* 2009;195:73–80.
- Lambert MJ. The outcome Questionnaire-45. *Integrating Science and Practice* 2012;2:24–7.
- Aung MH, Matthews M, Choudhury T. Sensing behavioral symptoms of mental health and delivering personalized interventions using mobile technologies. *Depress Anxiety* 2017;34:603–9.

- 4 Mohr DC, Zhang M, Schueller SM. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annu Rev Clin Psychol* 2017;13:23–47.
- 5 Robinson MD, Clore GL. Belief and feeling: evidence for an accessibility model of emotional self-report. *Psychol Bull* 2002;128:934–60.
- 6 Stone AA, Shiffman S. Ecological Momentary assessment (EMA) in behavioral medicine. *Ann Behav Med* 1994;16:199–202.
- 7 aan het Rot M, Hogenelst K, Schoevers RA. Mood disorders in everyday life: a systematic review of experience sampling and ecological momentary assessment studies. *Clin Psychol Rev* 2012;32:510–23.
- 8 Gibbons CJ. Turning the page on pen-and-paper questionnaires: combining ecological momentary assessment and computer adaptive testing to transform psychological assessment in the 21st century. *Front Psychol* 2017;7:1–4.
- 9 Fisher AJ, Boswell JF. Enhancing the Personalization of psychotherapy with dynamic assessment and modeling. *Assessment* 2016;23:496–506.
- 10 Bak M, Drukker M, Hasmi L, et al. An n=1 clinical network analysis of symptoms and treatment in psychosis. *PLoS ONE* 2016;11:1–15.
- 11 Moons KGM, Royston P, Vergouwe Y, et al. Prognosis and prognostic research: what, why, and how? *BMJ* 2009;338:b375.
- 12 Ivanescu AE, Li P, George B, et al. The importance of prediction model validation and assessment in obesity and nutrition research. *Int J Obes* 2016;40:887–94.
- 13 Kappen TH, van Klei WA, van Wolfswinkel L, et al. Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. *Diagnostic and Prognostic Research* 2018;2:1–11.
- 14 Altman DG, Vergouwe Y, Royston P, et al. Prognosis and prognostic research: validating a prognostic model. *BMJ* 2009;338:b605.
- 15 Moons KGM, Altman DG, Vergouwe Y, et al. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ* 2009;338:b606.
- 16 Steyerberg EW, Moons KGM, van der Windt DA, et al. Prognosis research strategy (progress) 3: prognostic model research. *PLoS Med* 2013;10:e1001381.
- 17 Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J* 2014;35:1925–31.
- 18 Bouwmeester W, Zuihthoff NPA, Mallett S, et al. Reporting and methods in clinical prediction research: a systematic review. *PLoS Med* 2012;9:e1001221.
- 19 Altaf Hussain Abro MK. validation of a computational model for mood and social integration. *Lecture Notes in Computer Science* 2016;2016:361–75.
- 20 Treur J. Dynamic modeling based on a temporal-causal network modeling approach. *Biologically Inspired Cognitive Architectures* 2016;16:131–68.
- 21 EfdM A, Treur J. Analysis and refinement of a temporal-causal network model for absorption of emotions. In: *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*. , 2016: vol. 9875, 27–39.
- 22 Treur J, Ziabari SSM. *Computational collective intelligence*. vol. 11056. Springer International Publishing, 2018.
- 23 Franke A, Hosain RW. A Temporal-Causal model for spread of messages in disasters 2017;10449.
- 24 Kleiboer A, Smit J, Bosmans J, et al. European comparative effectiveness research on blended depression treatment versus treatment-as-usual (E-COMPARED): study protocol for a randomized controlled, non-inferiority trial in eight European countries. *Trials* 2016;17:387.
- 25 van Breda W, Hoogendoorn M, Eiben AE, et al. Assessment of temporal predictive models for health care using a formal method. *Computers in Biology and Medicine* 2016;2017:347–57.
- 26 Mikus A, Hoogendoorn M, Rocha A, et al. Predicting short term mood developments among depressed patients using adherence and ecological momentary assessment data. *Internet Interventions* 2018;12:105–10.
- 27 Jaques N, Taylor S, Sano A, et al. Predicting Tomorrow's Mood, Health, and Stress Level using Personalized Multitask Learning and Domain Adaptation 2017.
- 28 Constantinides M, Busk J, Matic A, et al. Personalized versus Generic Mood Prediction Models in Bipolar Disorder. In: *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers - UbiComp '18*. New York, New York, USA: ACM Press, 2018: 1700–7.
- 29 Sudman S, Bradburn NM, Schwarz N. *Thinking about answers: the application of cognitive processes to survey methodology*. San Francisco, CA, US: Jossey-Bass, 1996.
- 30 Clark HH, Schober MF. *Asking questions and influencing answers. in: questions about questions: inquiries into the cognitive bases of surveys*. New York, NY, US: Russell Sage Foundation, 1992: 15–48.
- 31 Courvoisier DS, Eid M, Lischetzke T. Compliance to a cell phone-based ecological momentary assessment study: the effect of time and personality characteristics. *Psychol Assess* 2012;24:713–20.
- 32 Sokolovsky AW, Mermelstein RJ, Hedeker D. Factors predicting compliance to ecological momentary assessment among adolescent smokers. *Nicotine & Tobacco Research* 2014;16:351–8.
- 33 Papageorgiou G, Grant SW, Takkenberg JIM, et al. Statistical primer: how to deal with missing data in scientific research?†. *Interact Cardiovasc Thorac Surg* 2018;27:153–8.
- 34 Nicholson Jr. NR. Social isolation in older adults: an evolutionary concept analysis. *J Adv Nurs* 2009;65:1342–52.
- 35 Gariépy G, Honkaniemi H, Quesnel-Vallée A. Social support and protection from depression: systematic review of current findings in Western countries. *Br J Psychiatry* 2016;209:284–93.
- 36 Cacioppo JT, Hawkley LC, Crawford LE, et al. Loneliness and health: potential mechanisms. *Psychosom Med* 2002;64:407–17.
- 37 Nezlek JB, Hampton CP, Shean GD. Clinical depression and day-to-day social interaction in a community sample. *J Abnorm Psychol* 2000;109:11–19.
- 38 Nezlek JB, Imbrie M, Shean GD. Depression and everyday social interaction. *J Pers Soc Psychol* 1994;67:1101–11.
- 39 Steger MF, Kashdan TB. Depression and everyday social activity. *J Couns Psychol* 2010;56:289–300.
- 40 Cacioppo JT, Hawkley LC, Thisted RA. Perceived social isolation makes me sad: 5-year cross-lagged analyses of loneliness and depressive symptomatology in the Chicago health, aging, and social relations study. *Psychol Aging* 2010;25:453–63.
- 41 Taylor HO, Taylor RJ, Nguyen AW, et al. Social isolation, depression, and psychological distress among older adults. *J Aging Health* 2018;30:229–46.
- 42 Spanakis G, Weiss G, Boh B, et al. Network Analysis of Ecological Momentary Assessment Data for Monitoring and Understanding Eating Behavior. In: *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*. , 2016: 9545, 43–54.
- 43 Bremer V, Funk B, Riper H. Heterogeneity matters: predicting self-esteem in online interventions based on ecological Momentary assessment data. *Depress Res Treat* 2019;2019:1–9.
- 44 Deb K, Pratap A, Agarwal S, et al. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* 2002;6:182–97.
- 45 Parducci A. *Category judgment: a range-frequency model*. US: American Psychological Association, 1965.
- 46 Daamen DDL, de Bie SE. Serial context effects in survey interviews. in: *context effects in social and psychological research*. New York, NY: Springer New York 1992:97–113.
- 47 Riis J, Loewenstein G, Baron J, et al. Ignorance of hedonic adaptation to hemodialysis: a study using ecological momentary assessment. *J Exp Psychol* 2005;134:3–9.
- 48 Davey A. Issues in evaluating model fit with missing data. *Struct Equ Modeling: A Multidisciplin J* 2005;12:578–97.
- 49 Jones A, Remmerswaal D, Verveer I, et al. Compliance with ecological momentary assessment protocols in substance users: a meta-analysis. *Addiction* 2019;114:609–19.
- 50 Sterne JAC, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009;338:b2393