ORIGINAL ARTICLE

A digital self-report survey of mood for bipolar disorder

Tijana Sagorac Gruichich[1] | Juan Camilo David Gomez[1] | Gabriel Zayas-Cabán[1] |
Melvin G. McInnis[2] | Amy L. Cochran[1]

[1]University of Wisconsin, Madison, WI, USA

[2]University of Michigan, Ann Arbor, MI, USA

Correspondence
Amy L. Cochran, University of Wisconsin, Madison, WI, USA.
Email: cochran4@wisc.edu

Funding information
Richard Tam Foundation; National Institute of Mental Health, Grant/Award Number: K01MH112876; Prechter Bipolar Research Program

Abstract

Objectives: Bipolar disorder (BP) is commonly researched in digital settings. As a result, standardized digital tools are needed to measure mood. We sought to validate a new survey that is brief, validated in digital form, and able to separately measure manic and depressive severity.

Methods: We introduce a 6-item digital survey, called digiBP, for measuring mood in BP. It has three depressive items (depressed mood, fidgeting, fatigue), two manic items (increased energy, rapid speech), and one mixed item (irritability); and recovers two scores (m and d) to measure manic and depressive severity. In a secondary analysis of individuals with BP who monitored their symptoms over 6 weeks ($n = 43$), we perform a series of analyses to validate the digiBP survey internally, externally, and as a longitudinal measure.

Results: We first verify a conceptual model for the survey in which items load onto two factors ("manic" and "depressive"). We then show weekly averages of m and d scores from digiBP can explain significant variation in weekly scores from the Young Mania Rating Scale ($R^2 = 0.47$) and SIGH-D ($R^2 = 0.58$). Lastly, we examine the utility of the survey as a longitudinal measure by predicting an individual's future m and d scores from their past m and d scores.

Conclusions: While further validation is warranted in larger, diverse populations, these validation analyses should encourage researchers to consider digiBP for their next digital study of BP.

KEYWORDS
bipolar disorder, depression, digital research, Mania, mood, surveys and questionnaires

1 | INTRODUCTION

Research conducted digitally is a growing part of how we study bipolar disorder (BP). This fact is underscored by the present risk of in-person interviews due to Covid-19. Digital research offers many opportunities beyond providing an alternative to in-person data collection. The most obvious are mobile apps to manage mood and deliver treatment for BP.[1–6] Other opportunities include the ability to build BP cohorts not limited by geography and to phenotype BP at finer time scales[7–11] with greater, in-the-moment context on an individual's life including their physiology.[12–17] To improve the successes of digital research, the community is looking for standardized reporting tools for BP symptoms in a digital setting.[18–20] However, there are reasons, which we discuss below, why current tools being used in digital settings may not always be satisfactory. Thus, the present paper seeks to establish a new digital self-report survey for measuring mood in BP, which we call digiBP.

Decades of BP research is founded upon measures of mania and depression severity such as the Young Mania Rating Scale (YMRS),[21] the Hamilton Rating Scale for Depression,[22] and its structured version: the Structured Interview Guide for the Hamilton Rating Scale for Depression (SIGH-D).[23] These surveys, in which a trained interviewer evaluates symptoms on Likert scales, have been extensively validated for measuring mania and depression severity.[21,23–26] Furthermore, their ubiquity allows for the comparison of findings between different studies. Self-report surveys are also available, including the Patient Health Questionnaire-9[27] and Altman Mania Rating Scale.[28] While perhaps not as precise as interview-based surveys, self-report surveys have undergone extensive validation, making them a mainstay of BP research in situations when interviews are not feasible. Digital surveys of mood would ideally attain the same level of validation and standardization as interview-based and self-report measures of mood.

Although prior measures could be used *verbatim*, designing a new digital survey could have several advantages. The first is brevity. This is desirable because a digital setting is well suited for measuring mood frequently (e.g., daily), and certain measures may demand too much of an individual's time to complete regularly. For this reason, shortened versions of surveys are designed with only a few fundamental items. For example, the Patient Health Questionnaire-2 contains two[29] of the most informative items of the longer 9-item Patient Health Questionnaire-9. Similar brief surveys would be useful in a digital setting.

A second advantage would be user engagement, that is, an ability to get users to complete the survey frequently when delivered in digital form. Although user engagement is a common problem in all forms of digital self-monitoring,[30,31] many current digital methods have maintained user contact. For example, ChronoRecord has demonstrated an ability to measure manic and depressive severity and to get users to complete the survey.[32,33] Mood 24/7 is a text-message system which has been validated to measure depressive severity and to get users to complete the survey.[34(p.7)] The NIMH-Life-Chart Method and its electronic version is another short survey validated for BP.[35,36] The MONORCA system collects self-report measures of mood in addition to more objective measures.[2] In sum, various digital surveys for BP have previously been validated for measuring mood.

Yet, these digital surveys deviate from more traditional surveys in one important aspect: they use a single scale for manic and depressive symptoms. ChronoRecord measures mania and depression severity on a single visual analog scale; Mood 24/7 tracks mood on a single 1–10 scale; and the NIMH-Life-Chart Method and MONARCA measures mood on a single scale from −4 to 4. A potential drawback of a one-dimensional scale is the inability to capture mixed states, i.e. situations when individuals have symptoms of mania and depression. Additionally, a one-dimensional conceptualization of mood in BP may not be empirically supported, since manic and depressive symptoms commonly arise together in many individuals.[8] Indeed, many research studies use one measure for mania (e.g., YMRS) and a separate measure for depression (e.g., SIGH-D). Thus, another

advantage of a new digital survey could be an ability to separately measure the severity of manic and depressive symptoms.

With these possible advantages in mind, we developed a digital, self-report survey of mood in BP called digiBP. The survey contains six items selected from the SIGH-D and YMRS. Three items measure depressive symptoms (depressed mood, fidgeting, and fatigue), two items measure manic symptoms (increased energy, rapid speech), and one item measures both (irritability). The survey recovers two scores: a *d* score for measuring severity of depressive symptoms and an *m* score for measuring severity of manic symptoms. We delivered this survey twice daily via mobile app over the course of a 6-week study of how individuals with BP engage with digital self-monitoring of symptoms. The frequency and duration of survey delivery was designed to capture meaningful times scales for monitoring mood in BP without overburdening the participant. In particular, twice daily delivery was used in an attempt to capture rapid mood shifts[37] and diurnal patterns,[38] whereas a 6-week study period was used, since mood episodes[39] and medication waiting periods[40] require at least a week of monitoring. In our primary analysis, which we previously published, we found that individuals with BP on average recorded at least six symptoms (out of a max of 12) in the app on 81.8% of days in the study.[41] Thus, this survey has the ability to engage individuals with BP in daily monitoring of their symptoms.

The present study seeks to validate measurement properties of digiBP using the previously analyzed sample of individuals with BP (*n* = 43). Our goals are to provide internal and external validation of its measurement properties and to further demonstrate its utility as a longitudinal measure. Confirmatory and exploratory factor analyses are performed to test and refine a hypothesized conceptual factor model of item responses from digiBP. We then examine whether weekly averages of *m* and *d* scores are significantly associated with weekly YMRS and SIGH-D scores. Lastly, autoregression is performed to evaluate whether person-specific models can be developed for predicting future *m* and *d* scores based on past *m* and *d* scores.

2 | MATERIALS AND METHODS

Data was collected from a 6-week study on engagement in self-monitoring of symptoms by individuals with BP. The protocol has been published.[42] As explained in the protocol paper, primary analyses focused on engagement in self-monitoring,[41] whereas an exploratory analysis was planned to validate a mobile survey. Validating this survey is the goal of the present paper.

2.1 | Participants

Participants were recruited from the Prechter Longitudinal Study of Bipolar Disorder.[43] Participants were included if they had a smartphone, were an adult (18+ years), and had a previous diagnosis of BP of either Type I, Type II, or not otherwise specified (NOS). Diagnoses

in the Prechter Study were made using the Diagnostic Interview for Genetic Studies,[44] with clinicians reviewing diagnoses as a response to change in symptoms or discrepancy between interviewers.[43] Individuals provided informed written consent to participate in this study. The study was approved by Institutional Review Boards at the University of Michigan (HUM126732) and University of Wisconsin (2017-1322).

2.2 | Data collection

Over 6 weeks, participants logged their symptoms twice a day, once in the morning and once in the evening, through a mobile app Lorevimo (Log, Review, and Visualize your Mood). Like its name suggests, Lorevimo allows participants to log and then later review and visualize symptoms in various formats. At each time point, participants could log symptoms related to mania and/or depression, which comprise the 6-item survey digiBP. Data was collected twice daily for weeks in an effort to capture important BP characteristics such as diurnal patterns,[38] rapid mood shifts,[37] manic and depressive episodes,[39] and medication waiting periods.[40] Morning and evening were defined respectively, as the 6-h window spanning 2 h before to 4 h after their typical wake time and the 6-h window spanning 4 h before to 2 h after their typical bed time. Typical bedtimes or wake times are entered in-app by the participant and can be changed anytime. To remind participants, push notifications were sent by the app at 2-h intervals for individuals who have yet to log symptoms, are within the appropriate window, and not too close to their typical bedtime or wake time.

Phone interviews were also conducted by a trained research technician at study start and at the end of each study week for a total of seven interviews. At each interview, manic symptoms were assessed with the YMRS[21] and depressive symptoms with the SIGH-D.[23] While not examined in the present paper, general health

was assessed at study start and end with the 36-Item Short Form Health Survey,[45] and engagement in self-monitoring was evaluated at study end with a 17-item survey that we designed. Participants were also provided a Fitbit Alta HR to collect data on physical activity, heart rate, and sleep, and research technicians also reviewed daily self-reported data with a random sample of participants at the end of study weeks 1 through 5, which was used in primary analyses to determine whether reviewing would increase engagement.[41]

2.3 | digiBP

The proposed survey (digiBP) comprises six items. Figure 1A shows how items were presented in the mobile app. Three items (depressed mood, fatigue, fidgeting) measured common depressive symptoms, two items (increased energy, rapid speech) measured common manic symptoms, and one item (irritability) measured a common symptom of both mania and depression. Examples of each symptom were provided to the user. These symptoms account for varying manifestations of mood in BP, such as *anxious* depression (depressed mood + fidgeting); *anhedonic* depression (depressed mood + fatigue); *euphoric* mania (increased energy + rapid speech); *irritable* mania (irritability + increased energy); *mixed* state (depressed mood + increased energy); and *diurnal* patterns (changes in mood from morning to evening).

2.3.1 | Item scale

Participants rate symptoms on an ordinal scale: 0 = absent/normal, 1 = mild, 2 = moderate, 3 = severe. The same scale was used for each item to make it easier for users to respond. While mobile apps can use a visual analog scale, ordinal scales were used to be consistent with YMRS and SIGH-D. Additionally, an ordinal scale allows for
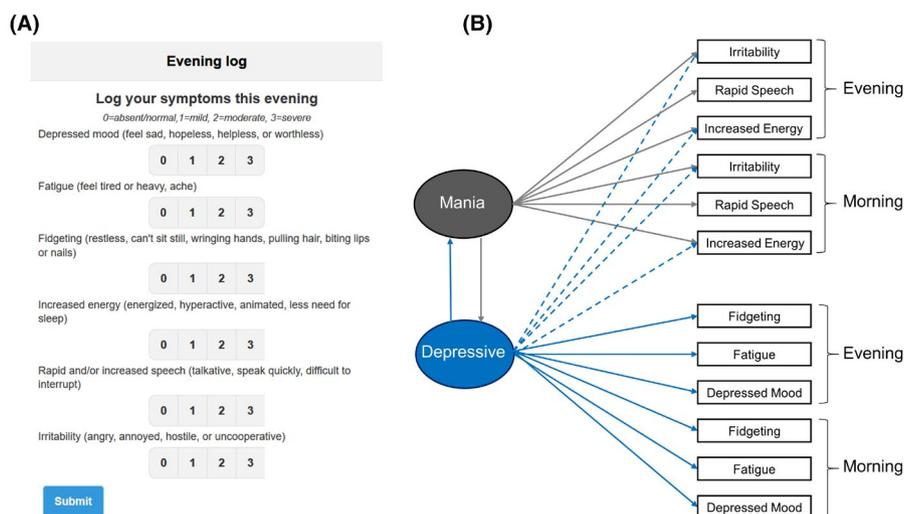

FIGURE 1 Survey items. (A) Screen shot of the digital survey of mood for bipolar disorder (digiBP). (B) Hypothesized factor model of survey items collected in the morning and evening

easier discrimination between choices and for a clearer definition of absence of symptoms.

2.3.2 | Item selection

How we selected items is detailed in the published protocol[42] and briefly reviewed here. To have a brief survey, we decided to select exactly six items. More items were thought to be too burdensome, whereas fewer were thought to poorly capture both manic and depressive symptoms. We only considered items from YMRS and 17-item SIGH-D for a total of 28 items in an effort to predict YMRS/SIGH-D scores. Using unpublished data collected from 27 individuals with either BP I or BP II from the Prechter Study, we regressed total YMRS scores and total SIGH-D scores onto responses from exactly 6 items. Linear regression models were constructed from every possible set of six items from the collection of 28 items (over 300,000 possible sets). Since individuals contributed multiple YMRS and/or SIGH-D scores, samples were weighted inversely by the number of samples per person in the regression. The resulting six items were selected, because they had yielded the smallest sum of mean-squared errors from the regression models for YMRS and SIGH-D scores.

2.3.3 | Scoring

We propose items can be aggregated into two scores to reflect depressive symptom severity and manic symptom severity. Before we present these scores, we point out that scoring (e.g., which items contribute to depressive vs. manic severity and how much they are weighted) was determined only after we had performed internal validation analyses. Thus, we leave our justification for these scores to the "Internal Validation/Results" section. A d score is defined according to the following formula:

$$d = 2 \times \lfloor \text{depressed mood} \rfloor + 2 \times \lfloor \text{fidgeting} \rfloor + 2 \times \lfloor \text{fatigue} \rfloor + \lfloor \text{irritability} \rfloor.$$

Similarly, an m score is defined according to:

$$m = 2 \times \lfloor \text{increased energy} \rfloor + 2 \times \lfloor \text{rapid speech} \rfloor + \lfloor \text{irritability} \rfloor.$$

Higher d and m scores reflect more severe depressive and manic symptoms, respectively. Since item responses range from 0 to 3, the d score ranges from 0 (no symptoms) to 21 (all depressive symptoms are severe), whereas the m score ranges from 0 (no symptoms) to 15 (all manic symptoms are severe).

2.4 | Internal validation

Our first analyses investigated whether item responses exhibit a pattern that is consistent with our conceptual model.

2.4.1 | Hypothesized factor model

Figure 1B illustrates the hypothesized factor model of morning and evening responses to digiBP. This structure embeds several hypotheses. Considering that items were selected to measure mania and depression, we hypothesized that responses would exhibit a two-dimensional factor structure: one factor representing "mania" and another representing "depression." A corollary is that we did not expect diurnal patterns to be sufficiently prominent to justify separate factors for morning and evening symptoms. Based on the regression models used to select items (see Section 2.3.2 above), we hypothesized that depressed mood, fatigue, and fidgeting load positively onto the depressive factor; and to a lesser extent, increased energy loads negatively onto the depressive factor and irritability loads positively onto the depressive factor. Increased energy, rapid speech, and irritability were hypothesized to load positively onto a manic factor. Notice these hypotheses include cross-loadings for increased energy and irritability.

2.4.2 | Factor dimension

To validate the factor dimension, we examined the eigenvalues of the empirical correlation matrix. Factor dimension was compared to the number of eigenvalues larger than 1, a long-standing simple heuristic for selecting factor dimension.[46]

2.4.3 | Factor structure

Confirmatory factor analysis tested the hypothesized factor model using Mplus version 8 software.[47] Parameters were estimated using mean and variance adjusted weighted least squares (WLSMV). Since participants completed surveys multiple times, standard errors were adjusted for non-independence using participant as a clustering variable with the COMPLEX option in MPlus. Missing data was handled using pairwise deletion. Model fit was assessed using root mean square error of approximation (RMSEA), for which values less than 0.05 were considered a good fit,[48] and comparative fit index (CFI), for which values greater than 0.95 were considered a good fit.[49] Additionally, modification indices tested modeling assumptions whereby large modification indices (>3.84) identified possible loadings that should differ from zero. As a final assessment, we compared model fit of the hypothesized factor structure with model fit of unrestricted factor model (i.e., exploratory factor analysis) in which two factors could load on all items. This exploratory factor analysis was similarly performed in MPlus with WLSMV and adjusted for non-independence. Output code from MPlus is available at https://github.com/cochran4/digiBP. Significance was considered an alpha level of 0.05.

2.5 | External Validation

We then aimed to confirm that m and d scores agree with YMRS and SIGH-D scores. To account for time scale differences

(digiBP was administered twice daily, YMRS/SIGH-D weekly), we averaged m and d scores, first over each day, and then over each week. Missing m or d scores were removed before either averaging step. A linear mixed effects model was constructed regressing the corresponding YMRS and SIGH-D score at the end of each week onto these weekly averaged m and d scores. Participant was included as a random effect to account for non-independence due to repeated measures. Since many studies might want to ask individuals to log symptoms only once a day, we repeated the analysis averaging only morning symptoms and then only evening symptoms. Models were compared using Akaike Information Criteria (AIC) and R-squared values. Lastly, since studies might also want to know how well m and d scores predict YMRS and SIGH-D scores that meet clinically meaningful thresholds, we also repeated the same analysis as before except that linear mixed effects regression was replaced with mixed effects logistic regression and dependent variables were replaced with binary variables indicating whether the SIGH-D score is greater than or equal to 8, which has been used to signify mild to severe depression,[50] and whether the YMRS score is greater than or equal to 9, which has been used to signify mania that has not remitted.[51] Sensitivity, specificity, positive predictive value, and negative predictive value, were then calculated after thresholding predicted probabilities of these binary variables by 0.50.

2.6 | Longitudinal Analysis

Our final analysis explores whether m and d scores can accurately forecast future m and d scores. We averaged m and d scores over each day, ignoring any missing scores. For each individual, these daily m and d scores were regressed onto their corresponding score on preceding days. Note, we did not center m or d scores to the mean score for that individual. For each outcome variable (m or d), we repeated regression analysis making three types of changes. First, we varied the number of preceding days from which to recover predictors between 1, 2, and 3 days. Second, we added the other mood score (i.e. the m score if d is being regressed and the d score if the m is score is being regressed). Third, we varied how many data points were required to construct the model between 25, 30, and 35 data points. Model forecasting was assessed using cross-validation whereby for each daily score, only prior measurements from the individual were used to build a linear regression model and mean square prediction error was measured using the built model. Predictions were restricted to a feasible range (i.e., 0–21 for d scores; 0–15 for m scores). To reduce the number of models considered, we only considered models with the same number of lags (either 1, 2, or 3 days) for each predictor. We report the average difference in model error and associated standard errors between each model and a model with an intercept but no predictors. Significance was assessed with a one-sample t test.

TABLE 1 Characteristics of sample population (n = 43)

Variable	
Age, years (mean ± SD)	41.58 ± 10.47
Female (n %)	22 (51)
Race (n %)	
White	38 (88)
Black or African American	3 (7)
Asian	1 (2)
American Indian or Alaskan Native	0
More than one	1 (2)
Hispanic (n, %)	5 (12)
Diagnosis (n, %)	
Bipolar I	32 (70)
Bipolar II	8 (23)
Bipolar NOS	3 (7)

3 | RESULTS

3.1 | Sample Characteristics

Out of 50 individuals recruited with BP, seven people were excluded from analysis. One experienced discomfort while wearing the Fitbit and did not continue to participate beyond the first week; another passed away between the consent date and study start; and another did not continue to follow-up with the research team after the consent date. Four additional participants were excluded because they did not self-report any mood data on the app. Table 1 summarizes the characteristics of the remaining participants. They had an average (SD) age of 41.58 (10.47) years and were about 51% female. The majority of the population were white (88%), non-Hispanic (88%), and 74% were diagnosed with Bipolar I.

3.2 | Internal Validation

The largest five eigenvalues of the empirical correlation matrix were 5.23, 2.74, 0.99, 0.82, and 0.61. Thus, the two largest eigenvalues of the empirical correlation matrix were markedly larger than the other eigenvalues and were the only eigenvalues larger than one. This suggests a two-dimensional factor model is appropriate for describing morning and evening responses to digiBP.

Confirmatory analysis found the data was partly coherent with the hypothesized factor structure. RMSEA suggested a good model fit (i.e., >0.05) but CFI, which was 0.92, suggested the model fit was less than good (i.e., <0.95) (Table 2). Morning/evening increased energy did not significantly load onto the depressive factor, as was hypothesized. Modification indices (MI) suggested that the morning response to fidgeting, the evening response to fidgeting, and the evening response to depressed mood should load onto the manic factor (MI = 7.86, 6.33, and 6.94, respectively). The remaining loadings, however, were consistent with the hypothesized structure.

TABLE 2 Loadings and model fit of hypothesized two-factor model (confirmatory) and unrestricted two-factor model (exploratory)

| | | Loadings: | | | |
| | | Confirmatory | | Exploratory | |
Time	Variable	Depressive	Manic	Depressive	Manic
Morning	Depressed mood	0.83[*]		0.84[*]	−0.01
	Fatigue	0.77[*]		0.77[*]	0.01c
	Fidgeting	0.75[*]		0.64[*]	0.32[*]
Evening	Depressed mood	0.83[*]		0.87[*]	−0.12
	Fatigue	0.77[*]		0.79[*]	−0.01
	Fidgeting	0.75[*]		0.65[*]	0.28[*]
Morning	Increased energy	−0.10	0.84[*]	−0.09	0.83[*]
	Rapid speech		0.85[*]	0.01	0.86[*]
	Irritability	0.48[*]	0.54[*]	0.51[*]	0.56[*]
Evening	Increased energy	−0.14	0.91[*]	−0.11	0.89[*]
	Rapid speech		0.84[*]	0.04	0.83[*]
	Irritability	0.51[*]	0.43[*]	0.54[*]	0.45[*]

| | | Model fit: | | |
		Confirmatory	Exploratory	
	RMSEA (90% CI)	0.033 (0.026, 0.040)	0.035 (0.028, 0.042)	
	CFI	0.92	0.92	

Abbreviations: CFI, Comparative Fit Index; RMSEA, root mean square error of approximation.
*$p < 0.05$.

TABLE 3 Model fit and coefficients of regression models for SIGH-D and YMRS scores as a function of weekly averaged d and m scores

Model	Variable	Estimate	95% CI	t	df	p	AIC	R^2
		SIGH-D:						
Morning and evening	Intercept	2.13	0.75, 3.51	3.04	246	0.003	1400.4	0.58
	d score	1.05	0.77, 1.33	7.35	246	<0.001		
	m score	−0.65	−1.22, −0.08	−2.24	246	0.026		
Only morning	Intercept	2.53	1.13, 3.93	3.56	246	<0.001	1406.2	0.58
	d score	1.05	0.75, 1.35	6.86	246	<0.001		
	m score	−0.76	−1.34, −0.18	−2.59	246	0.010		
Only evening	Intercept	1.98	0.63, 3.34	2.88	246	0.004	1401.3	0.57
	d score	0.91	0.66, 1.16	7.24	246	<0.001		
	m score	−0.34	−0.86, 0.17	−1.32	246	0.19		
		YMRS:						
Morning and evening	Intercept	1.20	0.16, 2.24	2.26	245	0.024	1265.4	0.47
	d score	−0.37	−0.59, −0.16	−3.45	245	<0.001		
	m score	1.51	1.08, 1.94	6.93	245	<0.001		
Only morning	Intercept	1.44	0.46, 2.42	2.88	245	0.004	1276.1	0.43
	d score	−0.30	−0.52, −0.09	−2.76	245	0.006		
	m score	1.27	0.85, 1.69	6.00	245	<0.001		
Only evening	Intercept	1.03	−0.05, 2.10	1.89	245	0.060	1264.2	0.47
	d score	−0.35	−0.54, −0.15	−3.50	245	<0.001		
	m score	1.43	1.03, 1.84	7.01	245	<0.001		

Abbreviations: AIC, Akaike information criteria; SIGH-D, Structured Interview Guide for the Hamilton Rating Scale for Depression; YMRS, Young Mania Rating Scale.

TABLE 4 Average difference in cross-validated mean square error between various models and a corresponding model without any predictors

	d score				*m* score			
Lags	Predictors	25 days	30 days	35 days	Predictors	25 days	30 days	35 days
1	*d*	−1.1 (0.7)	−1.1 (0.5)*	−1.3 (0.5)*	*m*	−0.4 (0.2)	−0.5 (0.3)	−0.5 (0.4)
1	*d, m*	−0.9 (0.8)	−1.0 (0.5)	−1.1 (0.5)*	*m, d*	−0.2 (0.3)	−0.3 (0.3)	−0.4 (0.4)
2	*d*	−0.3 (1.3)	−0.9 (0.6)	−1.8 (0.5)*	*m*	−0.2 (0.2)	−0.4 (0.3)	−0.5 (0.4)
2	*d, m*	−0.0 (1.2)	−0.3 (0.7)	−1.5 (0.5)*	*m, d*	−0.0 (0.2)	−0.2 (0.3)	−0.3 (0.4)
3	*d*	−0.5 (1.0)	−0.2 (1.0)	−1.8 (0.6)*	*m*	−0.0 (0.2)	−0.1 (0.3)	−0.3 (0.4)
3	*d, m*	1.0 (1.8)	1.7 (2.2)	−0.9 (0.5)	*m, d*	0.4 (0.4)	0.3 (0.4)	−0.0 (0.5)

*$p < 0.05$.

Most loadings were around 0.80 except for loadings associated with morning/evening irritability which was closer to 0.50. Since mania and depression are sometimes conceptualized as being negatively correlated, it is important to note the depressive factor was positively correlated with the manic factor ($r = 0.24$, $p = 0.053$).

Compared to the hypothesized factor model, an unrestricted two-factor model did not provide better model fits in terms of RMSEA or CFI (Table 2). The morning and evening responses to fidgeting did load significantly onto the manic factor, as was indicated by the modification indices for the hypothesized factor model, but the evening response to depressed mood did not load significantly onto the manic factor. These loadings on fidgeting were each about 0.3, which while significant, is a common threshold for deciding whether to include in a model. Other loadings were similar between hypothesized and unrestricted factor models, including the fact that the morning or evening responses of increased energy did not significantly load onto the depressive factor.

Erring on the side of a simple model, we revised our hypothesized model to remove the loadings of morning/evening increased energy onto the depressive factor but did not include the loadings of morning/evening fidgeting onto the manic factor. This model motivated our choice of scoring in which a *d* score is recovered from depressed mood, fatigue, fidgeting, and irritability, and an *m* score is recovered from increased energy, rapid speech, and irritability. With irritability loading onto each factor to a lesser degree than other items, we down-weighted irritability by one-half compared to the other items.

3.3 | External Validation

Using the proposed *m* and *d* scores, we found that weekly averages of the *m* and *d* scores were significant predictors of the interview-based SIGH-D and YMRS scores (Table 3). Regression models explained variability in the data slightly better for SIGH-*D* scores ($R^2 = 0.58$) over YMRS ($R^2 = 0.47$). Based on regression coefficients, one point increase in the averaged *d* score led to an estimated average increase in the SIGH-*D* score of 1.05 points (95% CI: [0.77, 1.33]), and one point increase in the averaged *m* score translated into an estimated average decrease in the SIGH-D score of 0.65 points

(95% CI: [0.08, 1.22]). Similarly, one point increase in the averaged *d* score translated into an estimated average decrease in the YMRS score of 0.37 points (95% CI: [0.16, 0.59]) and one point increase in the averaged *m* score translated into an estimated average increase in the YMRS score of 1.51 points (95% CI: [1.08, 1.94]).

Using only morning scores or only evening scores did not dramatically reduce model fits (at worst, $R^2 = 0.43$). However, if one had to choose, using evening scores yielded better model fits than model fits using morning scores and were within 0.01 of the R^2 values and within 1 of the AIC values associated when using both morning/evening scores.

Upon trying to predict SIGH-D and YMRS scores after thresholding (SIGH-D ≥8; YMRS ≥9) and thresholding predicted probabilities by 0.5, we achieved 66.7% sensitivity, 94.2% specificity, 84.8% positive predictive value, and 85.2% negative predictive value for the SIGH-D; and 54.2% sensitivity, 99.6% specificity, 92.9% positive predictive value, and 95.4% negative predictive value for the YMRS.

3.4 | Longitudinal Analysis

A personalized model with the previous *d* score was best able to predict current *d* score when requiring 25 data points for model building. This model, however, did not yield significantly better model fits than a model without any predictors in terms of average crossvalidated mean square error (Table 4). Given that more data points may be needed to build accurate models, we also examined model fit requiring 30 and 35 data points for modeling building. In these cases, a model with the previous *d* score did yield significantly better models fits than a model without any predictors. Moreover, a model with *d* scores from the past 2 days yielded even better fits. Adding prior *m* scores did not help with prediction.

The *m* score was more difficult to predict. A personalized model with an individual's prior *m* score was best able to predict their current *m* score regardless of how data points were required for model building. This model, however, did not yield significantly better model fits than a model without any predictors in terms of average crossvalidated mean square error (Table 4). Adding prior *d* scores did not help with prediction.

4 | DISCUSSION

Standardized digital measures are needed to support digital research of BP. In this paper, we establish the validity of a digital self-report survey to measure severity of manic and depressive symptoms in BP (digiBP). This brief 6-item survey consists of three depressive items, two manic items, and one mixed item from which two scores, m and d, can be calculated to assess manic and depressive severity, respectively. Using survey data collected twice daily over 6 weeks ($n = 43$), we confirmed the validity of a two-dimensional factor model for survey responses. We demonstrated weekly averages of m and d scores were significantly related to weekly interview-based YMRS and SIGH-D scores. Lastly, we showed that digiBP can be useful in longitudinal settings, allowing past m and d scores to be predictive of future m and d scores.

Given the importance of both mania and depression in BP, mood is studied using surveys to measure each domain of mood. The YMRS and SIGH-D are notable examples.[21,23] Current digital surveys of mood such as the ChronoRecord, NIMH Life-Chart-Method, or Mood 24/7 place mood on a one-dimensional scale.[32,34,35] A one-dimensional scale, however, cannot capture mixed states, that is, times when both depressive and manic symptoms manifest.

By contrast, digiBP is designed to capture mixed states by including items related to depression and mania. Irritability, for example, was hypothesized to measure both mania and depression. The present study helps show that mood, as measured by our survey, exhibits a two-dimensional factor structure with one factor representing manic symptoms, another factor representing depressive symptoms, and with irritability crossloading onto both factors. This factor analysis is a critical step in justifying the aggregation of survey responses into the proposed two scores (m and d). Although our survey yielded a conceptual model with manic and depressive factors, we point out the possibility that other surveys may yield other two-dimensional conceptual models of mood in BP such as negative-positive affect and pleasure-activation, among others.[52]

One result from our factor analyses was the absence of factors to capture intraday patterns. That is, even though the factor model described the 12 survey responses collected in the morning and evening, our data did not support adding factors beyond a manic and depressive factor. Moreover in our regression analyses, we found that evening scores, and to a lesser extent morning scores, could explain YMRS and SIGH-D scores nearly as well as combined evening and morning scores. These findings are interesting given that diurnal patterns are disrupted in BP.[38,53] Hence, our analyses were insensitive to diurnal patterns. The practical implication is that digiBP may only need to be filled out once a day, preferably in the evening.

Given the interest in using digital surveys to measure mood repeatedly, we felt it was important to demonstrate that digiBP could recover useful longitudinal information. Over time, repeated measures can help in two ways: to understand symptom dynamics in BP and to forecast future moods.[8–11,54,55] Focusing on the latter, we examined whether personalized models could be built from an individual's past m and d scores in order to predict their future scores.

Our results suggest that d scores, and to a lesser extent m scores, can be predicted from past scores, provided enough data is available for building personalized models. An ability to forecast future mood could be useful when trying to respond quickly to an individual's mood before they end up in the hospital. We remark that surveys are commonly validated in cross-sectional settings, which aim to differentiate between individuals based on mood severity. Less frequently are they validated in longitudinal settings, which aim to differentiate between times within an individual's life based on mood severity. Thus, the utility in longitudinal settings is an added benefit of digiBP.

There are several possible reasons why m scores were more difficult to forecast than d scores. It could be an issue with measurement validity either specifically for digiBP, in that m scores do not measure mania as well as d scores measure depression, or generally for mania, in that mania surveys do not measure mania as well as depression surveys measure depression. For example, upon shortening the YMRS, digiBP misses important features of mania such as impulsivity, flight of ideas, delusions, and grandiosity. Alternatively, it could be a general issue with predicting mania longitudinally. This latter explanation is plausible, because mania is both rarer than depression and shorter in duration.[39,56] Hence, manic symptoms are more of a "moving target" than depressive symptoms, which could mean they are more difficult to predict.

While the present paper evaluated measurement properties of digiBP, we had previously evaluated the ability of this self-report survey to engage individuals with BP in self-monitoring.[41] Other digital surveys have similarly demonstrated user engagement.[2,32,34,35] To summarize the main findings from our prior work, we found that participants with BP adhered to recording at least half of their symptoms (6 out of 12) within the app for an average (95% CI) of 81.8% (73.1%-90.4%) days in the study. Over 6 weeks, we estimated that these adherence rates declined by about 6.1% over the study. Notably, these adherence rates for digiBP were higher and declined less than adherence rates for a Fitbit in terms of wearing a Fitbit at least 12 h a day. We note that we sent push notifications to participants to remind them to log symptoms and forgetfulness was endorsed as the most likely barrier to self-monitoring. So while digiBP can be delivered repeatedly, we recommend that strategies, such as push notifications, are employed to remind individuals to complete the survey.

When regarding the conclusions of the present paper, it is important to keep several limitations in mind. Our sample was small ($n = 43$) and predominately white. It is well-documented that responses to psychiatric surveys may depend on an individual's racial, cultural, and ethnic background. Thus, the findings from this study, in particular the factor model, may not generalize to a more diverse population. The small sample might also limit the ability to uncover more subtle patterns in item responses (e.g., diurnal patterns). More to this point, our inability to detect diurnal patterns might reflect a limitation of who was in our sample or what and when items were assessed. To this latter point, it is important to recognize that participants did not have to log symptoms immediately after they woke up and before they went to bed, a logging schedule likely to be optimized to detect

diurnal patterns. In the morning, participants had 4 h after their wake time to log symptoms. In the evening, they had 4 h before their bed time to log symptoms. So, participants on a 8 AM wake/10 PM bed schedule could have logged their symptoms right before noon and right after 6 PM, leaving little time between logging time points to observe diurnal patterns.

Phone interviews may also be considered a limitation. Although phone interviews are cost effective and typically shorter than in-person interviews, they require the participant to answer the phone when called and remain on the line until the interview is over. In addition, the interviewer is unable to receive visual cues that would otherwise be considered as additional information to consider for diagnosing during an in-person interview. Another limitation is possible harm caused by push notifications. Push notifications may remind participants to log symptoms, but if they are notified too often, they might be considered intrusive or might cause distress. Another limitation was that digiBP was not validated for screening for either BP, moderate to severe mania, or moderate to severe depression. One reason for this limitation was that manic and depressive symptoms were rarely moderate to severe in our sample as measured on the YMRS and SIGH-D. Another reason was that we had focused on validating digiBP as a way to measure manic and depressive symptoms dimensionally. Hence, digiBP would benefit from further validation as a diagnostic or screening tool. Further validation would also be useful for analyzing additional properties such as internal consistency or an ability to engage users over time frames longer than 6 weeks.

In conclusion, this study validated a digital self-report survey, digiBP, to measure mood in BP. Overall, we performed internal and external validation and demonstrated its usefulness as a longitudinal measure. As a digital option, digiBP is shorter than traditional interview-based surveys, but translates to similar YMRS and SIGH-D scores. It can be feasibly used on a regular basis, for example, once daily like we suggest. Unlike other digital methods, digiBP exhibits a two-dimensional factor structure allowing one to measure depressive, manic, and mixed states. This work complements efforts to provide standardized reporting tools for studying BP that can be delivered digitally. In the future, we hope to validate digiBP to larger populations.

AUTHOR CONTRIBUTIONS
TSG and ALC conceptualized, designed, and initially drafted the paper. MGM and ALC contributed to securing funding and the acquisition of data. JCDG, ALC, and GZC conducted data analyses. All authors contributed to the interpretation of the data and the results from data analysis. All authors have participated in the drafting and the critical revision of the paper and given final approval of the version being published.

ORCID
Melvin G. McInnis 🔟 https://orcid.org/0000-0002-0375-6247
Amy L. Cochran 🔟 https://orcid.org/0000-0001-6024-796X

REFERENCES
1. Daus H, Kislicyn N, Heuer S, Backenstrass M. Disease management apps and technical assistance systems for bipolar disorder: investigating the patients´ point of view. *J Affect Disord.* 2018;229:351–357. https://doi.org/10.1016/j.jad.2017.12.059
2. Faurholt-Jepsen M, Vinberg M, Christensen EM, Frost M, Bardram J, Kessing LV. Daily electronic self-monitoring of subjective and objective symptoms in bipolar disorder—the MONARCA trial protocol (MONitoring, treAtment and pRediCtion of bipolAr disorder episodes): a randomised controlled single-blind trial. *BMJ Open.* 2013;3(7):e003353. https://doi.org/10.1136/bmjopen-2013-003353
3. Faurholt-Jepsen M, Busk J, Þórarinsdóttir H, et al. Objective smartphone data as a potential diagnostic marker of bipolar disorder. *Aust N Z J Psychiatry.* 2019;53(2):119–128. https://doi.org/10.1177/0004867418808900
4. Faurholt-Jepsen M, Frost M, Christensen EM, Bardram JE, Vinberg M, Kessing LV. The effect of smartphone-based monitoring on illness activity in bipolar disorder: the MONARCA II randomized controlled single-blinded trial. *Psychol Med.* 2020;50(5):838–848. https://doi.org/10.1017/S0033291719000710
5. Linardon J, Cuijpers P, Carlbring P, Messer M, Fuller-Tyszkiewicz M. The efficacy of app-supported smartphone interventions for mental health problems: a meta-analysis of randomized controlled trials. *World Psychiatry.* 2019;18(3):325–336. https://doi.org/10.1002/wps.20673
6. Miklowitz DJ, Price J, Holmes EA, et al. Facilitated integrated mood management for adults with bipolar disorder. *Bipolar Disord.* 2012;14(2):185–197. https://doi.org/10.1111/j.1399-5618.2012.00998.x
7. Chang S-S, Chou T. A dynamical bifurcation model of bipolar disorder based on learned expectation and asymmetry in mood sensitivity. *Comput Psychiat.* 2018;2:205–222. https://doi.org/10.1162/cpsy_a_00021
8. Cochran AL, Schultz A, McInnis MG, Forger DB. Testing frameworks for personalizing bipolar disorder. *Transl Psychiatry.* 2018;8(1):36. https://doi.org/10.1038/s41398-017-0084-4
9. Ortiz A, Bradler K, Garnham J, Slaney C, Alda M. Nonlinear dynamics of mood regulation in bipolar disorder. *Bipolar Disord.* 2015;17(2):139–149. https://doi.org/10.1111/bdi.12246
10. Sperry SH, Walsh MA, Kwapil TR. Emotion dynamics concurrently and prospectively predict mood psychopathology. *J Affect Disord.* 2020;261:67–75. https://doi.org/10.1016/j.jad.2019.09.076
11. Vazquez-Montes MDLA, Stevens R, Perera R, Saunders K, Geddes JR. Control charts for monitoring mood stability as a predictor of severe episodes in patients with bipolar disorder. *Int J Bipolar Disord.* 2018;6(1):7. https://doi.org/10.1186/s40345-017-0116-2
12. Cho C-H, Lee T, Kim M-G, In HP, Kim L, Lee H-J. Mood prediction of patients with mood disorders by machine learning using passive digital phenotypes based on the circadian rhythm: prospective observational cohort study. *J Med Internet Res.* 2019;21(4):e11029. https://doi.org/10.2196/11029
13. Matton K, McInnis MG, Provost EM. Into the wild: transitioning from recognizing mood in clinical interactions to personal conversations for individuals with bipolar disorder. Interspeech.

2019;ISCA:2019:1438–1442. https://doi.org/10.21437/Inter speech.2019-2698

14. Perna G, Riva A, Defillo A, Sangiorgio E, Nobile M, Caldirola D. Heart rate variability: can it serve as a marker of mental health resilience? *J Affect Disord*. 2020;263:754–761. https://doi.org/10.1016/j.jad.2019.10.017

15. Pratap A, Atkins DC, Renn BN, et al. The accuracy of passive phone sensors in predicting daily mood. *Depress Anxiety*. 2019;36(1):72–81. https://doi.org/10.1002/da.22822

16. Torous J, Gershon A, Hays R, Onnela J-P, Baker JT. Digital phenotyping for the busy psychiatrist: clinical implications and relevance. *Psychiatr Ann*. 2019;49(5):196–201. https://doi.org/10.3928/00485713-20190417-01

17. Zulueta J, Piscitello A, Rasic M, et al. Predicting mood disturbance severity with mobile phone keystroke metadata: a biaffect digital phenotyping study. *J Med Internet Res*. 2018;20(7):e241. https://doi.org/10.2196/jmir.9775

18. Faurholt-Jepsen M, Geddes JR, Goodwin GM, et al. Reporting guidelines on remotely collected electronic mood data in mood disorder (eMOOD)—recommendations. *Transl Psychiatry*. 2019;9(1):162. https://doi.org/10.1038/s41398-019-0484-8

19. Malhi GS, Hamilton A, Morris G, Mannie Z, Das P, Outhred T. The promise of digital mood tracking technologies: are we heading on the right track? *Evid Based Ment Health*. 2017;20(4):102–107. https://doi.org/10.1136/eb-2017-102757

20. Torous J, Firth J, Huckvale K, et al. The emerging imperative for a consensus approach toward the rating and clinical recommendation of mental health apps. *J Nerv Ment Dis*. 2018;206(8):662–666. https://doi.org/10.1097/NMD.0000000000000864

21. Young RC, Biggs JT, Ziegler VE, Meyer DA. A rating scale for Mania: reliability, validity and sensitivity. *Br J Psychiatry*. 1978;133(5):429–435. https://doi.org/10.1192/bjp.133.5.429

22. Hamilton M. A rating scale for depression. *J Neurol Neurosurg Psychiatry*. 1960;23:56–62. https://doi.org/10.1136/jnnp.23.1.56

23. Williams JBW. A structured interview guide for the Hamilton Depression Rating Scale. *Arch Gen Psychiatry*. 1988;45(8):742. https://doi.org/10.1001/archpsyc.1988.01800320058007

24. Double DB. The factor structure of manic rating scales. *J Affect Disord*. 1990;18(2):113–119. https://doi.org/10.1016/0165-0327(90)90067-I.

25. Gibbons RD, Clark DC, Kupfer DJ. Exactly what does the Hamilton depression rating scale measure? *J Psychiatr Res*. 1993;27(3):259–273. https://doi.org/10.1016/0022-3956(93)90037-3

26. Hamilton M. Development of a rating scale for primary depressive illness. *Br J Soc Clin Psychol*. 1967;6(4):278–296. https://doi.org/10.1111/j.2044-8260.1967.tb00530.x

27. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: validity of a brief depression severity measure. *J Gen Intern Med*. 2001;16(9):606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

28. Altman EG, Hedeker D, Peterson JL, Davis JM. The Altman Self-Rating Mania Scale. *Biol Psychiatry*. 1997;42(10):948–955. https://doi.org/10.1016/S0006-3223(96)00548-3

29. Kroenke K, Spitzer RL, Williams JBW. The patient health questionnaire-2: Validity of a two-item depression screener. *Med Care*. 2003;41(11):1284–1292. https://doi.org/10.1097/01.MLR.0000093487.78664.3C

30. Melville KM, Casey LM, Kavanagh DJ. Dropout from Internet-based treatment for psychological disorders. *Br J Clin Psychol*. 2010;49(Pt 4):455–471. https://doi.org/10.1348/014466509X472138

31. Yeager CM, Benight CC. If we build it, will they come? Issues of engagement with digital health interventions for trauma recovery. *Mhealth*. 2018;4:37. https://doi.org/10.21037/mhealth.2018.08.04

32. Bauer M, Grof P, Gyulai L, Rasgon N, Glenn T, Whybrow PC. Using technology to improve longitudinal studies: self-reporting with ChronoRecord in bipolar disorder. *Bipolar Disord*. 2004;6(1):67–74. https://doi.org/10.1046/j.1399-5618.2003.00085.x

33. Bauer M, Wilson T, Neuhaus K, et al. Self-reporting software for bipolar disorder: validation of ChronoRecord by patients with mania. *Psychiatry Res*. 2008;159(3):359–366. https://doi.org/10.1016/j.psychres.2007.04.013

34. Kumar A, Wang M, Riehm A, Yu E, Smith T, Kaplin A. An automated mobile mood tracking technology (Mood 24/7): validation study. *JMIR Ment Health*. 2020;7(5):e16237. https://doi.org/10.2196/16237

35. Denicoff KD, Leverich GS, Nolen WA, et al. Validation of the prospective NIMH-Life-Chart Method (NIMH-LCM-p) for longitudinal assessment of bipolar illness. *Psychol Med*. 2000;30(6):1391–1397. https://doi.org/10.1017/s0033291799002810

36. Schärer LO, Hartweg V, Valerius G, et al. Life charts on a palm-top computer: first results of a feasibility study with an electronic diary for bipolar patients. *Bipolar Disord*. 2002;4(Suppl 1):107–108. https://doi.org/10.1034/j.1399-5618.4.s1.51.x

37. Kramlinger KG, Post RM. Ultra-rapid and ultradian cycling in bipolar affective illness. *Br J Psychiatry*. 1996;168(3):314–323. https://doi.org/10.1192/bjp.168.3.314

38. Wirz-Justice A. Diurnal variation of depressive symptoms. *Dialogues Clin Neurosci*. 2008;10(3):337–343.

39. Solomon DA, Leon AC, Coryell WH, et al. Longitudinal course of bipolar I disorder: duration of mood episodes. *Arch Gen Psychiatry*. 2010;67(4):339–347. https://doi.org/10.1001/archgenpsychiatry.2010.15

40. Goodwin FK, Jamison KR. *Manic-depressive illness: bipolar disorders and recurrent depression*. Oxford, UK: Oxford University Press; 2007;2:699–720.

41. Van Til K, McInnis MG, Cochran A. A comparative study of engagement in mobile and wearable health monitoring for bipolar disorder. *Bipolar Disord*. 2020;22(2):182–190.Published online October 14, 2019. doi:10.1111/bdi.12849

42. Cochran A, Belman-Wells L, McInnis M. Engagement strategies for self-monitoring symptoms of bipolar disorder with mobile and wearable technology: protocol for a randomized controlled trial. *JMIR Res Protoc*. 2018;7(5):e130. https://doi.org/10.2196/resprot.9899

43. McInnis MG, Assari S, Kamali M, et al. Cohort profile: the Heinz C. Prechter longitudinal study of bipolar disorder. *Int J Epidemiol*. 2018;47(1):28–28n. https://doi.org/10.1093/ije/dyx229

44. Nurnberger JI. Diagnostic interview for genetic studies: rationale, unique features, and training. *Arch Gen Psychiatry*. 1994;51(11):849. https://doi.org/10.1001/archpsyc.1994.03950110009002

45. Ware JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care*. 1992;30(6):473–483.

46. Kaiser HF. The application of electronic computers to factor analysis. *Educ Psychol Measur*. 1960;20(1):141–151. https://doi.org/10.1177/001316446002000116

47. Muthén L, Muthén B. Mplus user's guide. Eighth Edition. Muthén & Muthén; 1998. Accessed May 23, 2019. https://www.statmodel.com/download/usersguide/MplusUserGuideVer_8.pdf

48. Browne MW, Cudeck R. Alternative ways of assessing model fit. *Sociol Methods Res*. 1992;21(2):230–258. https://doi.org/10.1177/0049124192021002005

49. Rigdon EE. CFI versus RMSEA: a comparison of two fit indexes for structural equation modeling. *Struct Eq Model Multi J*. 1996;3(4):369–379. https://doi.org/10.1080/10705519609540052

50. Zimmerman M, Martinez JH, Young D, Chelminski I, Dalrymple K. Severity classification on the Hamilton Depression Rating Scale. *J Affect Disord*. 2013;150(2):384–388. https://doi.org/10.1016/j.jad.2013.04.028

51. Earley W, Durgam S, Lu K, et al. Clinically relevant response and remission outcomes in cariprazine-treated patients with bipolar I disorder. *J Affect Disord*. 2018;226:239–244. https://doi.org/10.1016/j.jad.2017.09.040

52. Barrett LF, Russell JA. The structure of current affect: controversies and emerging consensus. *Curr Dir Psychol Sci*. 1999;8(1):10–14. https://doi.org/10.1111/1467-8721.00003

53. Carr O, Saunders KEA, Bilderbeck AC, et al. Desynchronization of diurnal rhythms in bipolar disorder and borderline personality disorder. *Transl Psychiatry*. 2018;8(1):79. https://doi.org/10.1038/s41398-018-0125-7

54. Bonsall MB, Geddes JR, Goodwin GM, Holmes EA. Bipolar disorder dynamics: affective instabilities, relaxation oscillations and noise. *J R Soc Interface*. 2015;12(112):20150670. https://doi.org/10.1098/rsif.2015.0670

55. Moore PJ, Little MA, McSharry PE, Goodwin GM, Geddes JR. Mood dynamics in bipolar disorder. *Int J Bipolar Disord*. 2014;2(1):11. https://doi.org/10.1186/s40345-014-0011-z

56. Cochran AL, McInnis MG, Forger DB. Data-driven classification of bipolar I disorder from longitudinal course of mood. *Transl Psychiatry*. 2016;6(10):e912. https://doi.org/10.1038/tp.2016.166