Received: 27 July 2020 | Revised: 31 January 2021 | Accepted: 5 February 2021

DOI: 10.1002/brb3.2077

ORIGINAL RESEARCH

Brain and Behavior Open Access WILEY

Estimating longitudinal depressive symptoms from smartphone data in a transdiagnostic cohort

Amelia M. Pellegrini[1] | Emily J. Huang[2] | Patrick C. Staples[4] | Kamber L. Hart[1] | Jeanette M. Lorme[4] | Hannah E. Brown[3] | Roy H. Perlis[1] | Jukka-Pekka J. Onnela[4]

[1]Center for Quantitative Health, Massachusetts General Hospital, Boston, MA, USA

[2]Department of Mathematics and Statistics, Wake Forest University, Winston-Salem, NC, USA

[3]Department of Psychiatry, Boston Medical Center, Boston, MA, USA

[4]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

Correspondence

Roy H. Perlis, Center for Quantitative Health, Massachusetts General Hospital, 185 Cambridge Street, Simches Research Building, 6th Floor, Boston, MA 02114, USA.
Email: rperlis@partners.org

and

Jukka-Pekka Onnela, Department of Biostatistics, Harvard T.H. Chan School of Public Health, 655 Huntington Avenue, Building 2, Room 423, Boston, MA 02115, USA.
Email: onnela@hsph.harvard.edu

Funding information

RHP was supported by NIMH & NHGRI grant no. 1P50MH106933; EJH, PCS, JML, and JPO were supported by NIH grant no. 1DP2MH103909. The sponsors had no role in study design, writing of the report, or data collection, analysis, or interpretation.

Abstract

Background: Passive measures collected using smartphones have been suggested to represent efficient proxies for depression severity, but the performance of such measures across diagnoses has not been studied.

Methods: We enrolled a cohort of 45 individuals (11 with major depressive disorder, 11 with bipolar disorder, 11 with schizophrenia or schizoaffective disorder, and 12 individuals with no axis I psychiatric disorder). During the 8-week study period, participants were evaluated with a rater-administered Montgomery–Åsberg Depression Rating Scale (MADRS) biweekly, completed self-report PHQ-8 measures weekly on their smartphone, and consented to collection of smartphone-based GPS and accelerometer data in order to learn about their behaviors. We utilized linear mixed models to predict depression severity on the basis of phone-based PHQ-8 and passive measures.

Results: Among the 45 individuals, 38 (84%) completed the 8-week study. The average root-mean-squared error (RMSE) in predicting the MADRS score (scale 0–60) was 4.72 using passive data alone, 4.27 using self-report measures alone, and 4.30 using both.

Conclusions: While passive measures did not improve MADRS score prediction in our cross-disorder study, they may capture behavioral phenotypes that cannot be measured objectively, granularly, or over long-term via self-report.

KEYWORDS

bipolar disorder, ecological momentary assessment, major depressive disorder, mobile applications, schizophrenia, self report, smartphone

1 | INTRODUCTION

In modern psychopharmacology, the gold standard for measurement of depressive symptoms, as with most psychiatric outcomes, has been clinician-rated scales. However, reliance on such measures introduces substantial limitations: the need for trained clinician raters increases the cost of assessment (despite enthusiasm for measurement-based care and recognition of the importance

Amelia M. Pellegrini and Emily J. Huang are joint first authors.

Roy H. Perlis and Jukka-Pekka J. Onnela are joint senior authors.

of quantifying treatment outcomes); the time required for clinical evaluation has precluded widespread use in practice; and clinician ratings contain sources of variance that are unrelated to underlying clinical affect. These scales themselves have been criticized for measuring a narrow range of symptoms that may be overly weighted toward specific illness features, neglecting the multidimensional nature of psychopathology (Insel et al., 2010).

The ubiquity of smartphones presents an opportunity to measure different social, cognitive, and behavioral markers in naturalistic settings. As of February 2018, 95% of Americans own a cellphone of some kind, with 77% owning a smartphone, up from just 35% in 2011 (Mobile Fact Sheet, 2018). With 6.3 billion smartphone subscriptions expected globally by 2021 (Cerwall, 2016), this technology offers an unprecedented opportunity to objectively measure human behavior in naturalistic settings outside of research laboratories and clinics.

We have previously defined the concept of digital phenotyping as the “moment-by-moment quantification of the individual-level human phenotype in situ using data from personal digital devices, in particular smartphones” (Onnela & Rauch, 2016; Torous et al., 2015). Others have defined similar concepts (Glenn & Monteith, 2014; Jain et al., 2015; Monteith et al., 2015), and there is a small but growing number of studies in mental health using smartphone data (Alvarez-Lozano et al., 2014; Benson et al., 2011; Faurholt-Jepsen et al., 2015; Gruenerbl et al., 2014; Miskelly, 2005; Saeb et al., 2015; Torous et al., 2015; Wang et al., 2016) and other electronic devices (De Choudhury et al., 2013; Dickerson et al., 2011; Gulbahce et al., 2012; Jashinsky et al., 2014; Kane et al., 2013; Kappeler-Setz et al., 2013; Katikalapudi et al., 2012; Matic et al., 2012; McIntyre et al., 2009; Minassian et al., 2010; Roh et al., 2012). Smartphones are well-suited as an instrument for digital phenotyping given their widespread adoption, the extent to which users engage with the devices, and the richness of their data. Being able to accomplish this without the expense and burden associated with additional specialized equipment makes the approach attractive to researchers (Onnela & Rauch, 2016). Smartphone-based digital phenotyping encompasses the collection of a range of different social and behavioral data, including but not limited to spatial trajectories (via GPS), physical mobility patterns (via accelerometer), social networks and communication dynamics (via call and text logs), and voice samples (via microphone) (Onnela & Rauch, 2016; Torous et al., 2015).

The performance of digital phenotyping has rarely been directly compared to clinical rating scales in trial-like settings, nor has it been examined in a transdiagnostic cohort. To address these gaps, we conducted an 8-week study among psychiatric outpatients with mood and psychotic disorders, as well as healthy controls. Our aim was to assess whether digital phenotyping may be used as a complement for in-person psychiatric assessments of depressive symptoms, using the MADRS, in a clinical population. We sought to assess to what extent it might be possible to predict a future clinician-rated score on the Montgomery–Åsberg Depression Rating Scale (MADRS) from baseline assessments of MADRS, surveys administered on the phone (here, Patient Health Questionnaire), passively collected smartphone data (here, GPS and accelerometer), or a combination of these measures. More generally, we sought to quantify data completeness, a critical but commonly overlooked question in digital phenotyping, examining GPS, accelerometer, and phone survey data over the course of the 8-week study.

2 | MATERIALS AND METHODS

2.1 | Study design and cohort description

This study used a prospective cohort design and aimed to recruit equal-sized groups of outpatients with major depressive disorder (n = 11), bipolar I or II disorder (n = 11), schizophrenia or schizoaffective disorder (n = 11), and screened healthy controls with no axis I psychiatric disorder (n = 12). Each participant's primary diagnosis was confirmed by the Structured Clinical Interview for DSM-IV (SCID) Modules A-D (*Diagnostic & statistical manual of mental disorders: DSM-IV*, 2000). Demographic features of the study cohort are shown in Table 1.

Participants were recruited from outpatient clinics of the Massachusetts General Hospital (Boston, MA) and via advertisements seeking healthy control participants between 2015 and 2018. All participants signed written informed consent prior to participation. The study protocol was reviewed and approved by the Partners HealthCare Institutional Review Board (protocol #: 2015P000666). Participants were compensated $50 after the initial baseline visit and an additional $100 upon completion of the study. If a participant withdrew from the study before completing the full 8 weeks, they were compensated $25 in addition to the initial $50. Participants received reimbursement for reasonable parking and travel expenses for each in-person study visit.

All participants were 18 years or older and owned a smartphone running an iOS or Android operating system and were judged likely able to comply with study procedures by the site investigator's estimation. Participants installed the Beiwe application at the baseline visit and provided demographic information. Participants then returned for four follow-up visits over the course of the 8 weeks (for a total of five in-person visits, scheduled approximately every two weeks).

2.2 | Longitudinal assessments

At baseline and each follow-up visit, trained raters (AMP and KLH) certified and supervised by psychiatric clinical trialists (HEB and RHP) administered the MADRS. The overall MADRS score ranged from 0 to 60, and the following cutoff points were usually applied: 0 to 6 (not depressed), 7 to 19 (mild depression), 20 to 34 (moderate depression), and above 34 (severe depression). Participants also responded daily to a 4-question in-app Likert scale survey on overall

TABLE 1 Description of cohort (*n* = 41)

Baseline Covariate	
Sex	
Male	37% (15/41)
Female	63% (26/41)
Age (years)	
Mean (*SD*)	43 (12)
Min, Q1, Q2, Q3, Max	21, 33, 45, 52, 68
Diagnosis	
Healthy control	27% (11/41)
Major depressive disorder	24% (10/41)
Bipolar disorder	24% (10/41)
Schizophrenia/schizoaffective	24% (10/41)
Race	
White	71% (29/41)
African-American	20% (8/41)
Asian	7% (3/41)
Other	2% (1/41)
Baseline MADRS Score (mean (*SD*))	
Healthy control	0.7 (1.2)
Major depressive disorder	20.0 (12.7)
Bipolar disorder	9.7 (10.6)
Schizophrenia/schizoaffective	6.2 (5.4)

Abbreviations: MADRS, Montgomery–Åsberg Depression Rating Scale; Q1, Q2, and Q3 represent the first, second, and third quartiles; sd, standard deviation.

mood, social interest, sleep quality, and activity level (Table S1). They were also prompted once a week on Saturdays to take an in-app Patient Health Questionnaire (PHQ-8) survey (Kroenke et al., 2009). The question assessing suicidality included in PHQ-9 was omitted because the Partners HealthCare IRB has previously determined that its inclusion would require real-time evaluation of patient data, deemed by the investigators to be infeasible in the present design. To remain enrolled in the study, participants were required to respond to the surveys at least five times a week.

2.3 | Beiwe research platform

In this study, we used the Beiwe application for data collection, which is the front-end component of the Beiwe research platform. We have previously described an earlier version of the Beiwe research platform for high-throughput smartphone-based digital phenotyping in biomedical research use (Torous et al., 2016). The front end of Beiwe consists of smartphone apps for iOS (by Apple) and Android (by Google) devices. The back-end system, which enables data collection and data analysis and supports study management, makes use of Amazon Web Services (AWS)-based cloud computing infrastructure. While data collection is arguably becoming easier with developing technology, analysis of the collected data is increasingly identified as the main bottleneck in research settings (Iniesta et al., 2016; Kubota et al., 2016; Kuehn, 2016). For this reason, Beiwe consists of a growing suite of data analysis and modeling tools triggered by the Beiwe data analysis pipeline.

Reproducibility remains a challenge in the biomedical sciences, as fewer than 10% of studies have been found fully reproducible (Prinz et al., 2011). To enhance reproducibility, all Beiwe data collection settings for both active (smartphone surveys and audio samples) and passive (smartphone sensors and logs) data are captured in a single JSON-formatted configuration file, which can be imported to future studies to enable them to use identical data collection. The configuration for this present study is also available.

2.4 | Data collection, storage, and security

Each study participant was assigned a randomly generated 8-character Beiwe User ID and a temporary password, and study staff assisted participants with app installation and activation at the time of enrollment. Data collected by the Beiwe application were immediately encrypted and stored on the smartphone until the phone was connected to Wi-Fi, at which point the data were uploaded to the study server and expunged from the phone. The reason for configuring Beiwe to use Wi-Fi rather than cellular data in this study was to avoid charges associated with uploading large volumes of data, roughly 1GB per subject-month, to the cloud. Any potentially identifying data were hashed on the mobile device, and all data were encrypted while stored on the phone awaiting upload, while in transit, and while on the server.

2.5 | Processing of passive data: Phone gps and accelerometer data

During the time period between the baseline visit and the last follow-up visit, accelerometer and GPS data from participants' smartphones were collected using Beiwe. The GPS measured the phone's latitude/longitude coordinates, while the accelerometer measured its acceleration along three orthogonal axes. To preserve the battery life of the phone (mainly due to GPS) and to reduce data volume (mainly due to accelerometer), each sensor alternated between an on-cycle and off-cycle according to a predefined schedule (10 s on, 10 s off for the accelerometer; 2 min on, 10 min off for GPS). We selected a longer on-period for the GPS as it required time to locate the satellites required for positioning its location, and we correspondingly selected a longer off-period to reduce battery drain. In Supplemental Material, we describe our procedure for generating covariates for MADRS prediction from raw GPS and accelerometer data. Roughly speaking, the data were first summarized at a daily level and then the daily summaries were aggregated by type of day (weekend versus weekday). For Android users, in addition to accelerometer and GPS data, we collected anonymized communication logs, which we used to derive two summary statistics: the number of outgoing calls and the number of unique phone numbers dialed. Plots and summary statistics of the Android communication log data are presented in Results section.

2.6 | Statistical analysis

We used linear mixed models for MADRS prediction. Linear mixed models are an extension of standard linear regression to clustered data, where the clusters here are multiple MADRS assessments over time for each subject. Importantly, linear mixed models can handle clusters of varying size due to missing data. We considered four main model specifications. Each of them included the baseline MADRS score and the demographic variables as predictors (Table 2). We included the baseline MADRS score as a predictor based on the following rationale. One would ideally like to predict MADRS scores from passive data only, but this would require a large sample size and may not even be possible. The next best approach is to predict future MADRS scores from passive data and some baseline MADRS data. We assumed this latter approach because this approach, if successful, could reduce the number of times the MADRS score needs to be evaluated, which would help economize healthcare resources. The models differed by which smartphone-based covariates were included as additional predictors: Model A used phone-based PHQ-8 surveys, Model B used weekly summaries of passive smartphone data, Model C used both PHQ-8 surveys and weekly summaries of passive smartphone data, and Model D used neither. In Models A and C, when including the phone-based PHQ-8 survey score as a predictor, we used the survey that was closest in time preceding the MADRS assessment in question. We chose to include the PHQ-8 survey score as a predictor because of the ease of completion on a mobile phone by patients, and because of its widespread use as a screen in primary care settings. For Models B and C, we sought to predict MADRS score based on passive smartphone data collected in the seven days preceding the MADRS assessment. We computed summary statistics using raw GPS and accelerometer data. Our previous work has shown that one needs to impute missing GPS data when constructing summary statistics from GPS data. To generate summary statistics from GPS data, we first imputed missing GPS trajectories using a resampling method that has previously been demonstrated to result in a 10-fold reduction in the error averaged across all mobility features compared to simple linear interpolation of data by Barnett and Onnela (2018). After imputing missing data, we then computed several GPS summaries proposed by Canzian and Musolesi (2015), Saeb et al. (2015), and Barnett and Onnela (2020). There were 32 candidate summary statistics computed from smartphone passive data (GPS and accelerometer) (see Table 2 and Table S2). As many of these statistics were correlated, rather than including all 32 statistics as predictors in the models that used passive data, we performed a principal component analysis (PCA) on the 32 summary statistics and used the first principal component as a predictor. For each model, we performed leave-one-subject-out cross-validation to evaluate its prediction accuracy. This entailed holding out the data from each participant in turn, fitting the model with the data from the other participants, and using the fixed effects portion of the fitted model to predict the MADRS scores of the held-out participant. At the model-fitting step, we excluded data points with missing values for one or more of the predictors. As our accuracy metric, we computed the root-mean-squared error (RMSE) for each participant and then took the average across all participants. To compute the RMSE for each participant, we took the squared error between the predicted

TABLE 2 Predictors used in the study

Baseline predictors	Predictors based on phone surveys	Predictors based on passive smartphone data[a]
• Age • Sex • Diagnostic category • MADRS score	• Score on the PHQ-8 survey closest in time preceding the MADRS assessment	GPS-based[b] • Number of significant locations visited • Time spent at home • Distance traveled • Maximum diameter • Maximum home distance • Radius of gyration • Average flight length • Standard deviation of flight length • Average flight duration • Standard deviation of flight duration • Probability of pause • Significant location entropy • Circadian routine • Weekend–weekday routine • Number of minutes with missing data Accelerometer-based[b] • Activity level • Number of minutes with missing data

Abbreviations: MADRS, Montgomery–Åsberg Depression Rating Scale; PHQ-8, Patient Health Questionnaire-8.

[a]These predictors are defined in Methods S1.

[b]Except for number of minutes with missing data, all other GPS-based or accelerometer-based predictors were computed separately for weekdays versus weekends.

and actual MADRS score for each visit, averaged the squared errors across all visits for the given participant, and finally took the square root of this quantity. A lower RMSE indicates more accurate predictions. Of note, as a preliminary investigation, we elect to present model fit rather than statistical comparisons of models.

3 | RESULTS

3.1 | Participant baseline covariates and MADRS scores

Of the 45 consented participants, we excluded four participants who elected to cease study participation at or before the first follow-up visit (Figure 1). All other participants (n = 41) were included in the analysis, of whom three participants dropped out after the first follow-up visit and 38 fully completed the 8-week study. Table 1 shows the baseline features of these 41 study participants, including age, sex, diagnostic category, race, and baseline MADRS score. There were no missing data for these features. For the participants who completed the study, MADRS scores were available at baseline and at each of the four follow-up visits. Among the three participants who dropped out after the first follow-up visit, MADRS scores were assessed for two participants at the first follow-up visit. For descriptive purposes, Figure S1a, b shows the participants' MADRS trajectories over time, and a scatterplot of the average of the MADRS scores versus the standard deviation of the MADRS scores for each subject.

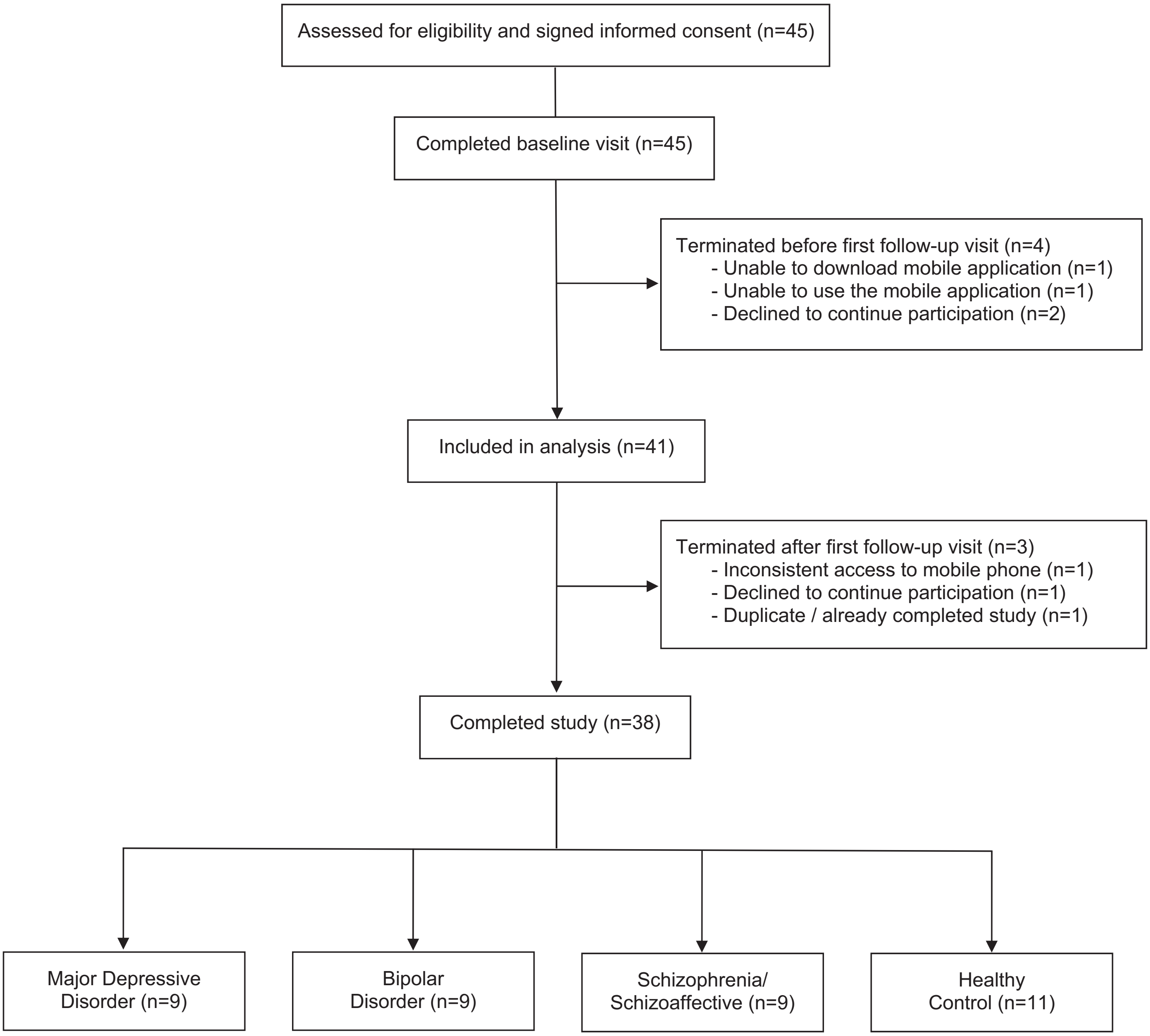
Assessed for eligibility and signed informed consent (n=45)
Completed baseline visit (n=45)
Terminated before first follow-up visit (n=4)
- Unable to download mobile application (n=1)
- Unable to use the mobile application (n=1)
- Declined to continue participation (n=2)
Included in analysis (n=41)
Terminated after first follow-up visit (n=3)
- Inconsistent access to mobile phone (n=1)
- Declined to continue participation (n=1)
- Duplicate / already completed study (n=1)
Completed study (n=38)
Major Depressive Disorder (n=9)
Bipolar Disorder (n=9)
Schizophrenia/ Schizoaffective (n=9)
Healthy Control (n=11)

FIGURE 1 CONSORT flow diagram

3.2 | Assessing completeness of phone data

The completeness of the accelerometer and GPS data was assessed at the participant level. For the accelerometer, we divided the number of minutes of data actually collected by the number of minutes of data expected to be collected. We examined the time period ranging from the day after the baseline visit to the day before the last follow-up visit, i.e., the time period including all full days in the study. Since accelerometer data were scheduled to be collected every minute, the expected number of minutes with data was the number of minutes in the time period. The completeness of GPS data was assessed analogously, except the expected number of minutes of data was 1/6 of the number of minutes in the time period (the 2-min on-cycle is 1/6 of the total cycle). The proportions for accelerometer and GPS are shown in Figure S2a. The proportions are variable, ranging from 0 to 0.99 for accelerometer and from 0 to 0.87 for GPS. For the accelerometer, 23 out of 41 (56%) participants had proportions of 0.5 or higher. For GPS, 16 out of 41 (39%) participants had proportions of 0.5 or higher. The proportions tended to be greater for accelerometer data than for GPS data. Despite the missingness, a large amount of data was captured over the course of the study, including 674,969,086 accelerometer measurements and 14,733,731 GPS measurements. The quantity of collected data for iOS phones tended to be greater on average than for Android phones.

Figure S2b shows the completion rate for each PHQ-8 survey, indicated by the solid black line. Given a specific survey, its completion rate was defined as the proportion of participants who completed the survey. If a participant completed Survey t after Survey $t + 1$ had been sent, they were counted as not having completed Survey t but were counted as having completed Survey $t + 1$. The completion rate was 95% for the first survey and 80% for the last survey, which took place approximately two months after the baseline visit. Figure S3 shows a histogram of the number of weeks that the participant completed one or more PHQ-8 surveys. If a participant completed more than one survey during some week (i.e., the participant was late on the previous week's survey), the multiple surveys only contributed 1 to the participant's tally. Overall, 78% of participants completed PHQ-8 surveys on 8 or more weeks.

As an example of passive data, Figure 2a-d plots the average activity level hour-by-hour (from 12:00 a.m. to 11:59 p.m.) for four randomly chosen participants in the schizophrenia/schizoaffective group on weekdays and weekends. For each participant, the curves were computed using accelerometer data collected throughout their follow-up as described in detail in Supplemental Material. For any given 1-hr window, the average activity level estimates the proportion of time that the participant was active (e.g., walking, using stairs) compared to stationary (e.g., sitting, standing, lying down) during this hour of the day. On weekdays, the participant in Panel A had low activity levels overnight, which began rising around 7 a.m., and hit their highest levels between 9 a.m. and 1 p.m., followed by a decline over the course of the evening. On weekends, their activity level was lower in the morning than on weekdays and was highest at 1 p.m. In interpreting these plots, a caveat is that the participant's activity was missed if the phone was not carried (e.g., it was left on a table). Thus, differences between the participants could be due to differences in their activity patterns, as well as differences in their phone use habits (e.g., how often each participant carried their phone).

Data completeness for each passive modality, and for self-report, is summarized in Supplemental Results. In addition, we collected smartphone communication logs from Android devices (no iOS devices were included in this part of the analysis). Figure 3 shows the cumulative distribution functions for the number of outgoing phone calls and the number of unique phone numbers dialed over Weeks 2–7, stratified by status (healthy control versus schizophrenia/schizoaffective, bipolar, or major depressive disorder). All individuals included here had communication log data collected throughout Weeks 2–7. Among this subset of participants (n = 19), the median age was 33 years (IQR: 29 – 41) for healthy controls (n = 7) and 52 years (IQR: 43–55) for others (n = 12). The proportions of female participants were 43% and 83%, respectively. The median number of outgoing calls was 56 (IQR: 24–79) for the healthy controls compared to 121 (IQR: 42–195) for those with a psychiatric diagnosis. The median number of unique phone numbers was also lower for the healthy controls at 18 (IQR: 12–24) versus 28 (IQR: 21–41) for those with a psychiatric diagnosis.

3.3 | MADRS prediction

Figure 4a–d shows the predicted MADRS scores compared to the clinician-rated MADRS scores. Panels A–D correspond to Models A–D: Panel/Model A (baseline MADRS & demographics & PHQ-8); Panel/Model B (baseline MADRS & demographics & passive data); Panel/Model C (baseline MADRS & demographics & PHQ-8 & passive data); and Panel/Model D (baseline MADRS & demographics). For the models that included passive data, we performed a principal component analysis and used the first principal component as a predictor. When principal component analysis was applied without excluding any subjects, the first principal component explained 46% of the variance in the data and the highest weights came from GPS-based features. In Table S3, we provide the weighting for each sensor-based feature in the first principal component. The predicted MADRS scores were computed using leave-one-subject-out cross-validation, as described above. The average RMSE was 4.27 for Model A, 4.72 for Model B, 4.30 for Model C, and 4.66 for Model D. That is, incorporation of passive variables in Model B did not meaningfully improve the average RMSE compared to using only the baseline MADRS score and demographics in Model D.

Models A–D each included both baseline MADRS and demographics as predictors. Although basic demographic variables can be easily collected and incorporated in the model, baseline MADRS scores might not be commonly available. To assess prediction accuracy in this modified setting, we next omitted the baseline MADRS from each model and otherwise proceeded as above. The average RMSE was 5.46 for Model A' (demographics & PHQ-8), 6.99 for Model B' (demographics & passive data), 5.46 for Model C' (demographics & PHQ-8 & passive data), and 6.91 for Model D' (demographics). The inclusion of baseline

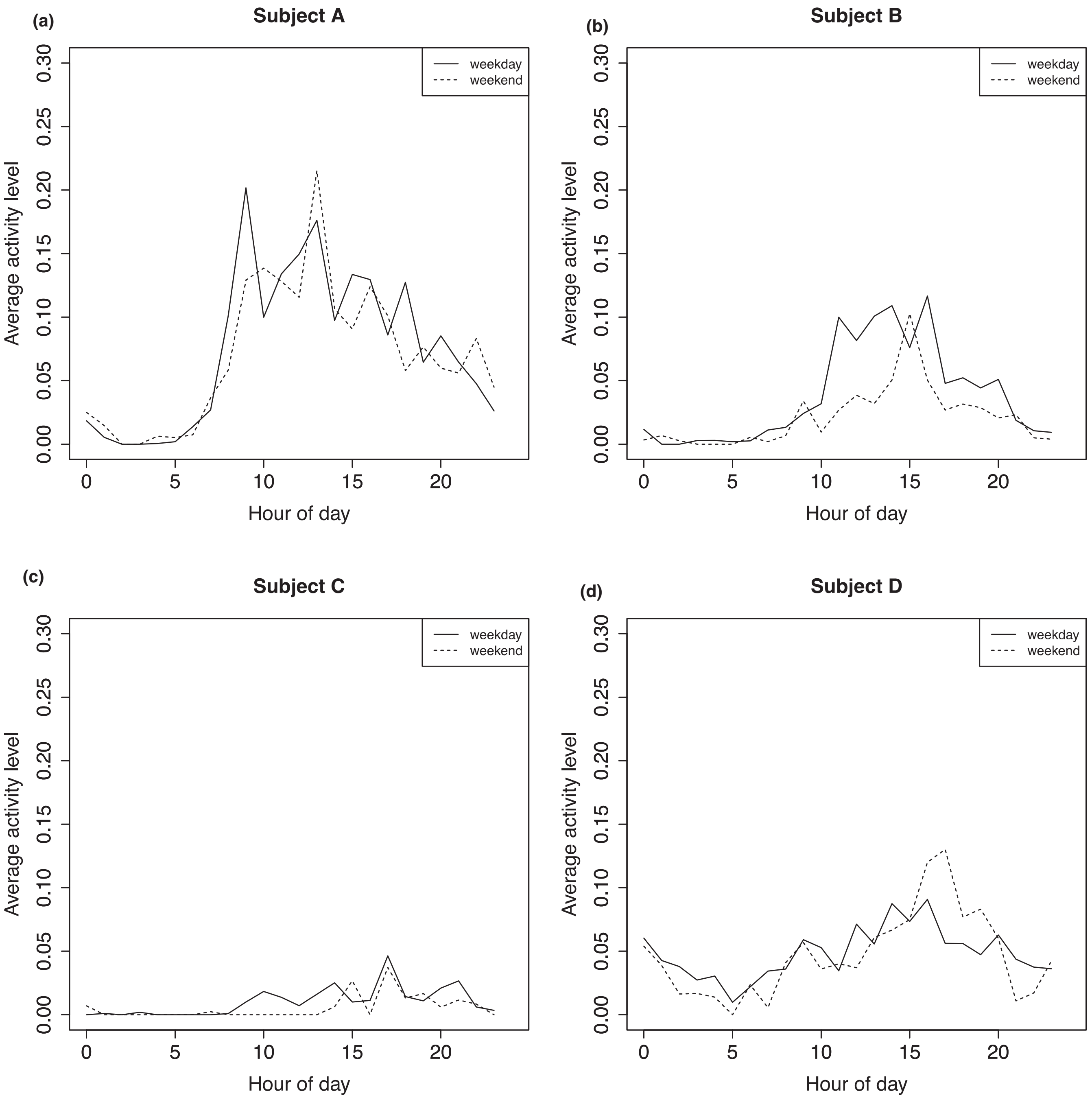

FIGURE 2 (a-d). Average activity level from 12:00 a.m. to 11:59 p.m. on weekdays and weekends for four randomly selected participants in the schizophrenia/schizoaffective diagnostic group. The solid line corresponds to the weekday, and the dotted line to the weekend. The x-axis origin of hour = 0 corresponds to 12:00 a.m. See Methods S1 for details on how these curves were computed

MADRS scores improves the average RMSE by approximately 1 point if PHQ-8 is included and 2 points if PHQ-8 is not included. Finally, as a sensitivity analysis, if demographics are also omitted, we obtain the following RMSE values: 5.69 for Model A'' (PHQ-8), 7.94 for Model B'' (passive data), 5.72 for Model C'' (PHQ-8 & passive data), 7.95 for Model D'' (no predictors, only an intercept). Results for different models are summarized in Table 3.

As an exploratory analysis, we evaluated the effect of including the second principal component (PC) as a predictor, which we call Model E. The results are shown in Table 3 in the row entitled Model E. Comparing the average RMSE's after adding the second PC (Model E) relative to having the first PC only (Model B), the average RMSE slightly improves when there are no other variables in the model or when the other variables are demographics, but slightly worsens when baseline MADRS and demographics are included. We conducted a separate exploratory analysis in which we identified the variables that had the highest loadings in the first PC: distance traveled maximum diameter, maximum home distance, and radius of gyration on the weekend. Since it is the most interpretable of the four, we used distance traveled on the weekend as the single passive predictor in a new model called Model F,

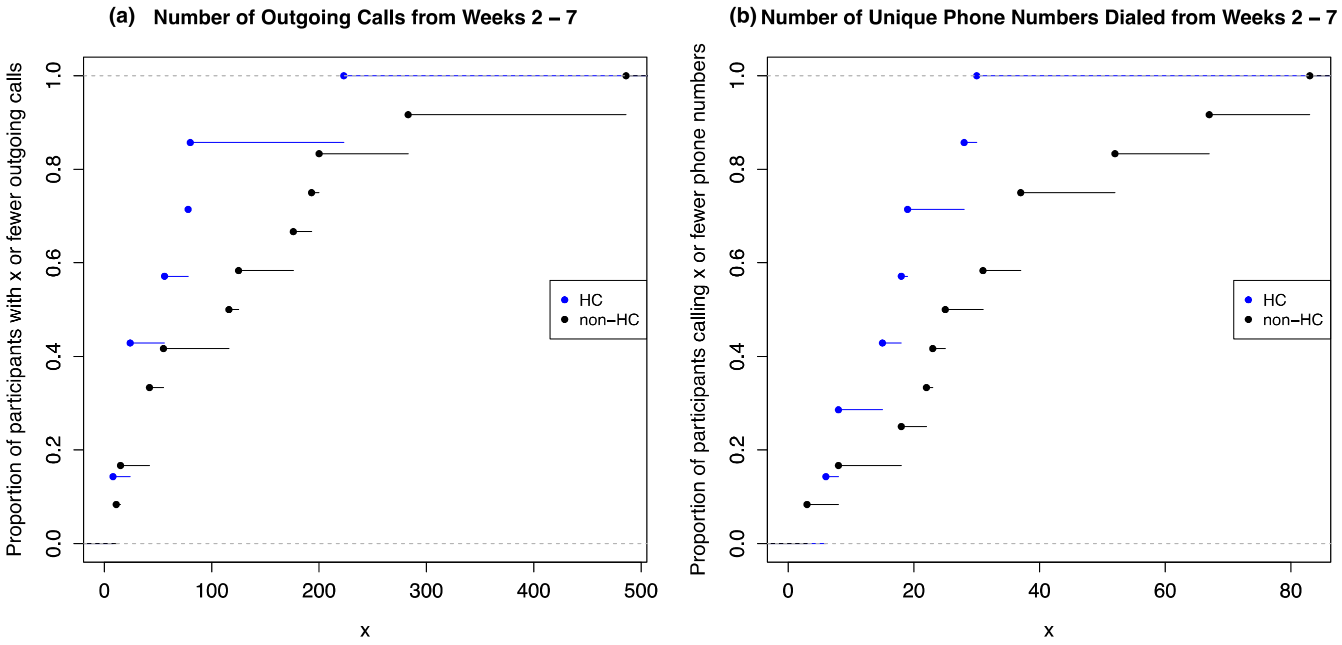

FIGURE 3 Cumulative distribution functions for the number of outgoing calls (a) and the number of dialed phone numbers (b) among Android participants, stratified by healthy control status. HC, healthy control; non-HC, individuals with a psychiatric diagnosis

which had no PC's included. This led to similar average RMSE's as using the first PC (Model B) and using the first and second PC (Model E).

4 | DISCUSSION

In this cross-disorder investigation, we found that including passive data as a predictor did not improve the prediction of clinician-rated MADRS scores. While the participant payment employed in this study precludes strong conclusions about acceptability, the high retention rate suggests that, with compensation, participants are willing to adopt this technology as part of a standard clinical assessment model. A similar approach has successfully been used in other settings, such as to study patients with schizophrenia, where the subjects were not paid for app use, not given additional support for app use, and not provided with check-in calls or study staff reminders to use the app (Barnett et al., 2018).

Both academic researchers and pharmaceutical leaders have suggested that passive measures may replace clinical evaluation in clinical trials as a means of improving signal detection (Harvey et al., 2018). Setting aside the need for clinician involvement to ensure participant safety, our results suggest that more work will be required to replace clinical raters for assessment of MADRS.

Although passive data did not perform as well as phone-based PHQ-8 in terms of average RMSE, it is important to stress that the passive approach requires only a one-time installation of the application which, even if less precise, may be valuable in settings where individuals are unlikely to adhere to a survey protocol, especially for extended time periods and in the absence of financial or other incentives.

One possible explanation for why incorporating passive variables in Model B did not improve the average RMSE compared to using only the baseline MADRS score and demographics in Model D is the varying data quality among participants. For example, Figure S4 shows the availability of accelerometer data for three participants. For each hour over the course of the follow-up, we plot the proportion of minutes with accelerometer data collected. A shading of white corresponds to 0 (no data collected during that hour), black to 1 (data collected at every minute), and different shadings of gray to in-between values. The *x*-axis shows the week of the follow-up, and the y-axis shows the day of the week with the tick marks occurring at 12:00 a.m. The participant in the top panel had high data quality throughout their follow-up. The participant in the middle panel had high data quality during most of the study with some long gaps with no data. The participant in the bottom panel had some medium data quality periods interspersed with periods with no data. Using incomplete passive data to predict the MADRS score can be challenging since the timing of the missing gaps may not be random (Figure S4). When deriving our predictors from passive data, we avoided the naïve approach of taking averages across the available data, which would overweight time intervals during which data tended to be collected. Instead, we utilized a more robust method for handling missingness, which is described in Methods S1. However, the predictors may be inaccurate when the proportions of data collected are low (Figure S2a).

In a meta-analysis of seven smartphone-based digital phenotyping studies, there was no significant difference found in levels of missing data by sex, age, educational background, and phone operating system for either accelerometer or GPS data (Kiang et al., 2019). Another study found that levels of missing GPS and accelerometer data were predictive of future clinical survey scores in a cohort of

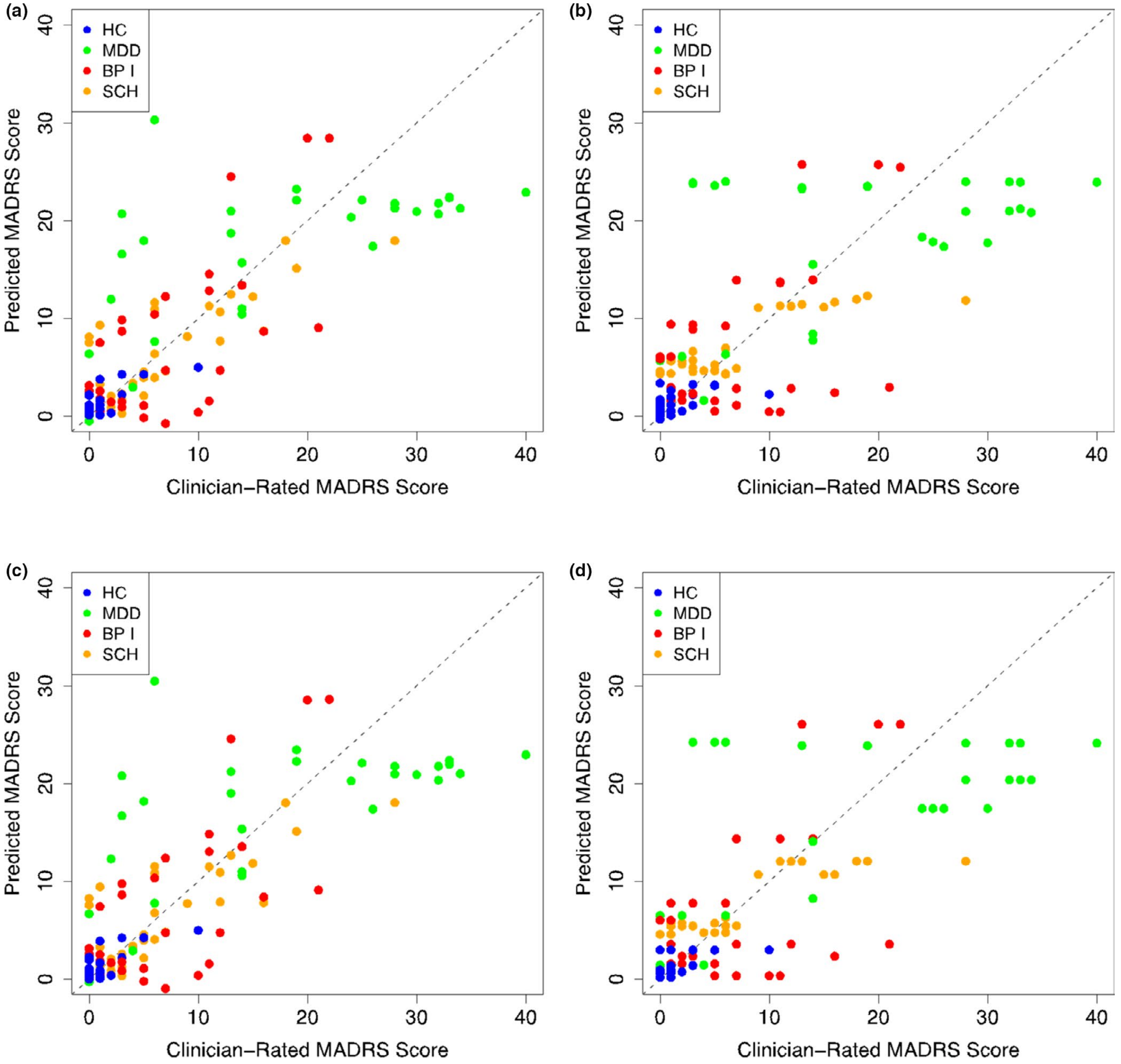

FIGURE 4 Montgomery–Åsberg Depression Rating Scale (MADRS) score predictions using four models. All models included age, sex, diagnostic category, and baseline MADRS score as predictors. They differed by which phone-based variables were used as predictors: Patient Health Questionnaire-8 survey scores only (a), passive data only (b), both (c), and neither (d). BP I, bipolar disorder; HC, healthy control; MDD, major depressive disorder; SCH, schizophrenia/schizoaffective disorder

patients with schizophrenia (Torous et al., 2018), which presents a potential future extension of the analyses presented here.

We note multiple important limitations in considering our results. First, the study design precludes conclusions about application of smartphone apps in longer-term studies or those using 'lighter touch' designs without in-person visits. Second, we cannot exclude the possibility that additional passive measures, or alternate means of analyzing such measures, will yield better prediction of clinician ratings. Indeed, our work should encourage other investigators to apply our open-source platform and further develop our analytic methodologies. Our analyses mix between-person and within-person variation in MADRS scores. Since these are distinct types of variation, a potential area of future research is to separately assess within-person changes from between-person differences. Third, because of the IRB-mandated omission of the PHQ suicide item, we likely underestimate the ability of this measure to capture more severe depression. Fourth, as a pilot study, sample size is modest and thus the result that passive measures do not significantly contribute to predicting MADRS must be viewed as preliminary. In future studies, strategies to reduce missing data (for example, by monitoring data missingness for each participant during the course of the study and intervening where required) merit consideration. Higher data quality may help improve the utility of passive measures.

TABLE 3 The average root-mean-squared error (RMSE) in predicting the MADRS score (scale 0–60) using three different variants of Models A–D

		+ demographics	+ baseline MADRS + demographics
Model A (PHQ–8)	5.69	5.46	4.27
Model B (passive data)	7.94	6.99	4.72
Model C (PHQ–8 & passive)	5.72	5.46	4.30
Model D (no phone-based predictors)	7.95 (intercept only)	6.91	4.66
Model E (first and second principal component)	7.86	6.97	4.75
Model F (distance traveled on weekend)	7.96	6.97	4.72

Abbreviations: MADRS, Montgomery–Åsberg Depression Rating Scale; PHQ-8, Patient Health Questionnaire-8.

We also emphasize strengths of using passively collected smartphone data in psychiatric settings. Passive data likely capture depressive features that are not well-measured by clinical raters, such as physical activity levels, spatial isolation (as measured via GPS-based home time), and social isolation (as measured via communication logs). Investigation of this hypothesis represents an important priority for clinical investigators seeking to develop a next generation of pragmatic trials. In other words, rather than simply replacing clinical raters, passive measures may themselves represent useful biomarkers, but only if they can be validated for this role.

We elected to conduct a cross-disorder study to recognize that categorical diagnosis fails to capture the dimensional nature of psychopathology, consistent with the NIMH's Research Domain Criteria framework (Insel et al., 2010). That is, it may be useful to capture negative valence symptoms such as depression across a range of disorders, not just in major depressive disorder. While such symptoms may be attributed to different underlying processes (e.g., negative symptoms in schizophrenia), our results suggest the ability of a single platform to measure across disorders.

5 | CONCLUSION

While passively collected smartphone data did not improve the prediction of MADRS scores in our cross-disorder study, we demonstrate its application to capture features of patients' daily functioning—such as physical activity, social isolation, and spatial isolation—that are otherwise difficult to capture with surveys. These various behavioral phenotypes, which are listed in Table 2 and defined in the Supplement, can describe participants' physical activity (e.g., from the accelerometer data), spatial isolation (e.g., time spent at home, computed from GPS data), and social isolation (e.g., number of outgoing calls from Android call log data).

HUMAN SUBJECTS ETHICS STATEMENT

All participants signed written informed consent prior to their inclusion in the study. The study protocol was reviewed and approved by the Partners HealthCare Institutional Review Board (protocol #: 2015P000666).

CONFLICT OF INTEREST

AMP, EJH, KLH, and JML have no disclosures to report. PCS is employed by Mindstrong Health, Inc. HEB receives research funding from Janssen Pharmaceutica, Acadia Pharmaceuticals, and the Stanley Medical Research Institute. RHP receives research funding from NIMH, NHLBI, NHGRI, and Telefonica Alfa. RHP holds equity in Psy Therapeutics and Outermost Therapeutics; serves on the scientific advisory board of Genomind, Psy Therapeutics, and Outermost Therapeutics; consults to RID Ventures and Takeda; and receives salary support from *JAMA Network-Open* for service as Associate Editor. JPO receives research funding from NIH, Harvard University, Otsuka Pharmaceutical, and Apple; he has also received an unrestricted gift from Mindstrong Health, Inc.

AUTHOR CONTRIBUTIONS

All authors have contributed meaningfully to this work and gave final approval to submit for publication.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/brb3.2077.

DATA AVAILABILITY STATEMENT

IRB approval does not permit public data release in light of concerns about reidentifiability (McCoy & Hughes, 2018). Investigators seeking access to data are encouraged to contact the authors.

ORCID

Amelia M. Pellegrini https://orcid.org/0000-0001-9312-7673
Roy H. Perlis https://orcid.org/0000-0002-5862-6757

REFERENCES

Alvarez-Lozano, J., Osmani, V., Mayora, O., Frost, M., Bardram, J., Faurholt-Jepsen, M., & Kessing, L. V. (2014). *Tell me your apps and I will tell you your mood: Correlation of apps usage with bipolar disorder state*. Proceedings of the 7th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '14), Article No. 19. Rhodes, Greece.

Barnett, I., & Onnela, J.-P. (2020). Inferring mobility measures from GPS traces with missing data. *Biostatistics*, *21*(2), e98–e112. https://doi.org/10.1093/biostatistics/kxy059

Barnett, I., Torous, J., Staples, P., Sandoval, L., Keshavan, M., & Onnela, J.-P. (2018). Relapse prediction in schizophrenia through digital phenotyping: A pilot study. *Neuropsychopharmacology*, *43*(8), 1660–1666. https://doi.org/10.1038/s41386-018-0030-z

Benson, E., Haghighi, A., & Barzilay, R. (2011). *Event discovery in social media feeds*. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, 389–398. Association for Computational Linguistics.

Canzian, L., & Musolesi, M. (2015). *Trajectories of depression: Unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis*. Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15), 1293–1304. https://doi.org/10.1145/2750858.2805845

Cerwall, P. (2016). *Ericsson mobility report, mobile world congress edition*. Ericsson.

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). *Predicting depression via social media*. Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, 128–137. Association for the Advancement of Artificial Intelligence.

Diagnostic and statistical manual of mental disorders: DSM-IV-TR. , 4th. ed. (2000) American Psychiatric Association.

Dickerson, R. F., Gorlin, E. I., & Stankovic, J. A. (2011, October 10). *Empath: A continuous remote emotional health monitoring system for depressive illness*. Presented at the Proceedings of the 2nd Conference on Wireless Health, San Diego, California.

Faurholt-Jepsen, M., Frost, M., Ritz, C., Christensen, E. M., Jacoby, A. S., Mikkelsen, R. L., Knorr, U., Bardram, J. E., Vinberg, M., & Kessing, L. V. (2015). Daily electronic self-monitoring in bipolar disorder using smartphones – the MONARCA I trial: A randomized, placebo-controlled, single-blind, parallel group trial. *Psychological Medicine*, *45*(13), 2691–2704. https://doi.org/10.1017/S0033291715000410

Glenn, T., & Monteith, S. (2014). New measures of mental state and behavior based on data collected from sensors, smartphones, and the Internet. *Current Psychiatry Reports*, *16*(12), 523. https://doi.org/10.1007/s11920-014-0523-3

Gruenerbl, A., Osmani, V., Bahle, G., Carrasco, J. C., Oehler, S., & Mayora, O., & Lukowicz, P. (2014). *Using smart phone mobility traces for the diagnosis of depressive and manic episodes in bipolar patients*. Proceedings of the 5th Augmented Human International Conference (AH '14), 1–8. https://doi.org/10.1145/2582051.2582089

Gulbahce, N., Yan, H., Dricot, A., Padi, M., Byrdsong, D., Franchi, R., Lee, D.-S., Rozenblatt-Rosen, O., Mar, J. C., Calderwood, M. A., Baldwin, A., Zhao, B. O., Santhanam, B., Braun, P., Simonis, N., Huh, K.-W., Hellner, K., Grace, M., Chen, A., … Barabási, A.-L. (2012). Viral perturbations of host networks reflect disease etiology. *PLoS Computational Biology*, *8*(6), e1002531. https://doi.org/10.1371/journal.pcbi.1002531

Harvey, P., Farchione, T., Keefe, R., & Davis, M. (2018). 2018 Autumn Abstract: Innovative Uses of Technology for Measuring Outcomes in Clinical Trials - (Parts 1 and 2). Retrieved from The International Society for CNS Clinical Trials and Methodology website: https://isctm.org/abstract-innovative-uses-of-technology-for-measuring-outcome-in-clinical-trials-parts-1-and-2/

Iniesta, R., Stahl, D., & McGuffin, P. (2016). Machine learning, statistical learning and the future of biological research in psychiatry. *Psychological Medicine*, *46*(12), 2455–2465. https://doi.org/10.1017/S0033291716001367

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., Sanislow, C., & Wang, P. (2010). Research domain criteria (RDoC): Toward a new classification framework for research on mental disorders. *American Journal of Psychiatry*, *167*(7), 748–751. https://doi.org/10.1176/appi.ajp.2010.09091379

Jain, S. H., Powers, B. W., Hawkins, J. B., & Brownstein, J. S. (2015). The digital phenotype. *Nature Biotechnology*, *33*(5), 462–463. https://doi.org/10.1038/nbt.3223

Jashinsky, J., Burton, S. H., Hanson, C. L., West, J., Giraud-Carrier, C., Barnes, M. D., & Argyle, T. (2014). Tracking suicide risk factors through Twitter in the US. *Crisis*, *35*(1), 51–59. https://doi.org/10.1027/0227-5910/a000234

Kane, J. M., Perlis, R. H., DiCarlo, L. A., Au-Yeung, K., Duong, J., & Petrides, G. (2013). First experience with a wireless system incorporating physiologic assessments and direct confirmation of digital tablet ingestions in ambulatory patients with schizophrenia or bipolar disorder. *The Journal of Clinical Psychiatry*, *74*(6), e533–540. https://doi.org/10.4088/JCP.12m08222

Kappeler-Setz, C., Gravenhorst, F., Schumm, J., Arnrich, B., & Tröster, G. (2013). Towards long term monitoring of electrodermal activity in daily life. *Personal and Ubiquitous Computing*, *17*(2), 261–271. https://doi.org/10.1007/s00779-011-0463-4

Katikalapudi, R., Chellappan, S., Montgomery, F., Wunsch, D., & Lutzen, K. (2012). Associating Internet usage with depressive behavior among college students. *IEEE Technology and Society Magazine*, *31*(4), 73–80. https://doi.org/10.1109/MTS.2012.2225462

Kiang, M. V., Chen, J. T., Krieger, N., Buckee, C. O., & Onnela, J.-P. (2019). Human Factors, Demographics, and Missing Data in Digital Phenotyping: Boston, Massachusetts 2015-2018.

Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B. W., Berry, J. T., & Mokdad, A. H. (2009). The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, *114*(1–3), 163–173. https://doi.org/10.1016/j.jad.2008.06.026

Kubota, K. J., Chen, J. A., & Little, M. A. (2016). Machine learning for large-scale wearable sensor data in Parkinson's disease: Concepts, promises, pitfalls, and futures. *Movement Disorders*, *31*(9), 1314–1326. https://doi.org/10.1002/mds.26693

Kuehn, B. M. (2016). FDA's foray into big data still maturing. *JAMA*, *315*(18), 1934–1936. https://doi.org/10.1001/jama.2016.2752

Matic, A., Mehta, P., Rehg, J. M., Osmani, V., & Mayora, O. (2012). Monitoring dressing activity failures through RFID and video. *Methods of Information in Medicine*, *51*(1), 45–54. https://doi.org/10.3414/ME10-02-0026

McCoy, T. H., & Hughes, M. C. (2018). Preserving patient confidentiality as data grow: Implications of the ability to reidentify physical activity data. *JAMA Network Open*, *1*(8), e186029. https://doi.org/10.1001/jamanetworkopen.2018.6029

McIntyre, G., Gocke, R., Hyett, M., Green, M., & Breakspear, M. (2009). *An approach for automatically measuring facial activity in depressed subjects*. 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, 1–8. https://doi.org/10.1109/ACII.2009.5349593

Minassian, A., Henry, B. L., Geyer, M. A., Paulus, M. P., Young, J. W., & Perry, W. (2010). The quantitative assessment of motor activity in mania and schizophrenia. *Journal of Affective Disorders*, *120*(1–3), 200–206. https://doi.org/10.1016/j.jad.2009.04.018

Miskelly, F. (2005). Electronic tracking of patients with dementia and wandering using mobile phone technology. *Age and Ageing*, *34*(5), 497–499. https://doi.org/10.1093/ageing/afi145

Mobile Fact Sheet. (2018, February 5). Retrieved from Pew Research Center website: http://www.pewinternet.org/fact-sheet/mobile/

Monteith, S., Glenn, T., Geddes, J., & Bauer, M. (2015). Big data are coming to psychiatry: A general introduction. *International Journal of Bipolar Disorders*, *3*(1), 21. https://doi.org/10.1186/s40345-015-0038-9

Onnela, J.-P., & Rauch, S. L. (2016). Harnessing Smartphone-Based Digital Phenotyping to Enhance Behavioral and Mental Health. *Neuropsychopharmacology*, *41*(7), 1691–1696. https://doi.org/10.1038/npp.2016.7

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, *10*(9), 712. https://doi.org/10.1038/nrd3439-c1

Roh, T., Bong, K., Hong, S., Cho, H., & Yoo, H.-J. (2012). *Wearable mental-health monitoring platform with independent component analysis and nonlinear chaotic analysis*. 2012 34th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 4541–4544. https://doi.org/10.1109/EMBC.2012.6346977

Saeb, S., Zhang, M., Karr, C. J., Schueller, S. M., Corden, M. E., Kording, K. P., & Mohr, D. C. (2015). Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study. *Journal of Medical Internet Research*, *17*(7), e175. https://doi.org/10.2196/jmir.4273

Torous, J., Kiang, M. V., Lorme, J., & Onnela, J.-P. (2016). New Tools for New Research in Psychiatry: A Scalable and Customizable Platform to Empower Data Driven Smartphone Research. *Journal of Medical Internet Research: Mental Health*, *3*(2), e16. https://doi.org/10.2196/mental.5165

Torous, J., Staples, P., Barnett, I., Sandoval, L. R., Keshavan, M., & Onnela, J.-P. (2018). Characterizing the clinical relevance of digital phenotyping data quality with applications to a cohort with schizophrenia. *Npj Digital Medicine*, *1*(1), 15. https://doi.org/10.1038/s41746-018-0022-8

Torous, J., Staples, P., Shanahan, M., Lin, C., Peck, P., Keshavan, M., & Onnela, J.-P. (2015). Utilizing a Personal Smartphone Custom App to Assess the Patient Health Questionnaire-9 (PHQ-9) Depressive Symptoms in Patients With Major Depressive Disorder. *Journal of Medical Internet Research: Mental Health*, *2*(1), e8. https://doi.org/10.2196/mental.3889

Wang, R., Scherer, E. A., Tseng, V. W. S., Ben-Zeev, D., Aung, M. S. H., Abdullah, S., & Merrill, M. (2016). *CrossCheck: Toward passive sensing and detection of mental health changes in people with schizophrenia*. Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16), 886–897. https://doi.org/10.1145/2971648.2971740

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Pellegrini AM, Huang EJ, Staples PC, et al. Estimating longitudinal depressive symptoms from smartphone data in a transdiagnostic cohort. *Brain Behav.* 2022;12:e2077. https://doi.org/10.1002/brb3.2077