Predicting Symptom Trajectories of Schizophrenia using Mobile Sensing

RUI WANG, Dartmouth College
WEICHEN WANG, Dartmouth College
MIN S. H. AUNG, Cornell University
DROR BEN-ZEEV, University of Washington
RACHEL BRIAN, Dartmouth College
ANDREW T. CAMPBELL, Dartmouth College
TANZEEM CHOUDHURY, Cornell University
MARTA HAUSER, Hofstra Northwell School of Medicine
JOHN KANE, Hofstra Northwell School of Medicine
EMILY A. SCHERER, Dartmouth College
MEGAN WALSH, Northwell Health

Continuously monitoring schizophrenia patients' psychiatric symptoms is crucial for in-time intervention and treatment adjustment. The Brief Psychiatric Rating Scale (BPRS) is a survey administered by clinicians to evaluate symptom severity in schizophrenia. The *CrossCheck symptom prediction system* is capable of tracking schizophrenia symptoms based on BPRS using passive sensing from mobile phones. We present results from an ongoing randomized control trial, where passive sensing data, self-reports, and clinician administered 7-item BPRS surveys are collected from 36 outpatients with schizophrenia recently discharged from hospital over a period ranging from 2-12 months. We show that our system can predict a symptom scale score based on a 7-item BPRS within ±1.45 error on average using automatically tracked behavioral features from phones (e.g., mobility, conversation, activity, smartphone usage, the ambient acoustic environment) and user supplied self-reports. Importantly, we show our system is also capable of predicting an individual BPRS score within ±1.59 error purely based on passive sensing from phones without any self-reported information from outpatients. Finally, we discuss how well our predictive system reflects symptoms experienced by patients by reviewing a number of case studies.

CCS Concepts: •**Human-centered computing → Ubiquitous and mobile computing;** •**Applied computing → Life and medical sciences;**

Additional Key Words and Phrases: Mobile Sensing, Mental Health, Schizophrenia, BPRS

ACM Reference format:
Rui Wang, Weichen Wang, Min S. H. Aung, Dror Ben-Zeev, Rachel Brian, Andrew T. Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Emily A. Scherer, and Megan Walsh. 2017. Predicting Symptom Trajectories of Schizophrenia using Mobile Sensing. *PACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 110 (September 2017), 24 pages.
DOI: http://doi.org/10.1145/3130976

The research reported in this article is supported the National Institute of Mental Health, grant number R01MH103148.

DOI: http://doi.org/10.1145/3130976

1 INTRODUCTION

Schizophrenia is a severe chronic psychiatric disorder associated with high individual and societal costs. Psychosis is not considered as a fixed state. Rather, the majority of people with schizophrenia fluctuate between full or partial remission and episodes of symptomatic relapse [61]. Psychotic symptoms may change over months, weeks, or even days, and can be affected by both external conditions and internal states [5]. In the case of relapse, patients may find themselves suffering severe hardship if not helped, such as, homelessness, incarceration, and victimization [16, 29, 31, 38]. Patients are often hospitalized as consequence of schizophrenia relapse.

Clinicians need to track schizophrenia patients' symptom states to identify risks and adjust treatment as necessary. At the CrossCheck study [63] partner hospital, Zucker Hillside Hospital, in New York City, schizophrenia outpatients regularly schedule clinical visits with their clinicians. The time between visits varies from once a week to once a month, depending on the patients' symptom severity and risk. Clinicians use a battery of mental health tests to evaluate the patients' symptom states and adjust their treatment accordingly. In our study, clinicians administer a 7-item Brief Psychiatric Rating Scale (BPRS), a subset of the original 24-item BPRS [18, 28, 37, 62] as a part of their clinical model. The BPRS is a rating scale to measure psychiatric symptoms associated with schizophrenia, such as, depression, anxiety, hallucinations, and unusual behavior. Each symptom is rated 1-7 (1 is given if the symptom is not present to 7 extremely severe). The reliability, validity and sensitivity of the BPRS measurement has been widely examined and considered a gold standard in assessment [30]. The 7 items include grandiosity, suspiciousness, hallucinations, unusual thought content, conceptual disorganization, blunted affect, and mannerisms and posturing. The clinical research team at Zucker Hillside Hospital determines these 7 items represent the strongest predictors of deterioration in symptoms amongst all BPRS items. The total score of the 7 BPRS items measures the overall symptom severity. However, this assessment has its shortcomings. Clinicians are not aware if a patient experiences deteriorated symptoms between visits. Because of this gap of knowledge in outpatients management between visits, clinicians are more likely to miss the optimal time to intervene to treat patients who are increasingly symptomatic and experiencing increased risk of relapse. Finally, the burden of hospital visits and face to face assessments on patients and health service providers further prohibits patients from more frequent visits with their clinicians to adjust treatment or provide intervention.

In order to address these shortcomings, we develop the *CrossCheck symptom prediction system* to monitor patients' trajectory of psychiatric symptoms. The system predicts patients' weekly 7-item BPRS total scores using passive sensing and self-reported ecological momentary assessment (EMA) responses from smartphones. Other than self-reported EMAs, the 7-item BPRS is administered by a trained clinician at our study partner hospital. The scored 7-item BPRS survey serves as a clinical indicator of treatment for patients who have moderate to severe disease. The clinician is responsible for interpreting the constructs in the assessment which are technical. The CrossCheck symptom prediction system predict the total score of the 7-item BPRS every week. Weekly predictions track participants' overall psychiatric symptoms and level of risk for relapse.

In our previous research [63] from the CrossCheck project, we reported on associations between passive sensing data and self-reported EMA responses, and models that can predict the self-reported EMA scores using sensor data from phones. These weekly EMA questions developed specifically for the CrossCheck study attempt to measure several dynamic dimensions of mental health and functioning in people with schizophrenia. In this paper, we turn to predict the 7-item BPRS, which is administered by a trained clinician. We present 7-item BPRS prediction results from an ongoing randomized control trial (RCT), where passive sensor data, self-reports and clinically administered 7-item BPRS reports are collected from 36 outpatients with schizophrenia recently discharged from hospital over a period ranging from 2-12 months. We show that our system can predict 7-item BPRS using a combination of passive sensing data and self-reported EMA. *Importantly, we also show that we can predict 7-item BPRS scores purely based on passive sensing data from mobile phones.* This paper makes the following contributions:

- To the best of our knowledge, the CrossCheck symptom prediction system is the first system capable of tracking schizophrenia patients' symptom scores measured by the 7-item BPRS using passive sensing and self-report EMA from phones. The system enables clinicians to track changes in psychiatric symptoms of patients without evaluating the patient in person.
- We identify a number of passive sensing predictors of the 7-item BPRS scores. These predictors describe a wide range of behaviors and contextual environmental characteristics associated with patients. Specifically, we find features extracted from physical activity, conversation, mobility, phone usage, call logs, and ambient sound are predictive of the 7-item BPRS.
- We use leave-one-record-out and leave-one-subject-out cross validations [35] to evaluate the 7-item BPRS prediction performance for prediction using passive sensing and EMA data. With leave-one-record-out cross validation, the system predicts the 7-item BPRS scores within ±1.45 of error on average. With leave-one-subject-out cross validation, the system predicts the 7-item BPRS scores within ±1.70 of error on average.
- We predict the 7-item BPRS scores within ±1.59 of error solely based on passive sensing without self-reports. Self-report suffers from a number of problems, adherence being one. In addition, because of increasing symptoms severity patients may be incapable of providing accurate self-reports (e.g., due to cognitive impairments). A system based purely on passive sensing data alleviates this problem and resolves the adherence problem. Furthermore, it opens the way for continuous assessment of symptoms and risk as people go about their everyday lives.
- We discuss anecdotal information associated with three patients in the study. These case studies show that our system can identify patients with rising risk.

The structure of the paper is as follows. We present related work on mental health sensing in Section 2, followed by a detailed description of the CrossCheck symptom prediction system in Section 3. We detail the CrossCheck study design and dataset in Section 4. Following that, we discuss the performance results of our the symptom prediction system in Section 5. In Section 6, we discuss three case studies and show how the CrossCheck symptom prediction system reflects real-life conditions of patients. Finally, we present some concluding remarks in Section 7.

2 RELATED WORK

There is an increasing amount of research in mental health sensing using phones [2, 41, 64]. In [53] the authors demonstrate that speech and conversation occurrences extracted from audio data and physical activity infer mental and social well-being. The StudentLife study [64] investigates the relationship between passive sensing behaviors including conversation, sleep, activity and co-location with mental health outcomes, such as, depression, stress, loneliness, and flourishing for 48 college students over a 10-week term. The authors find for the first time significant correlations between passive sensing data from phones and the PHQ9 depression scale [39, 40, 59]. Researchers at Northwestern University [56] find that mobility and phone usage features extracted from phone data correlates with depressive symptom severity measured by PHQ9[39, 40, 59] for 40 participants recruited from the general community over a 2 week period. Their results show features from GPS data, including circadian movement, normalized entropy, location variance, and phone usage features, including usage duration and usage frequency, are related to depressive symptom severity. The research team reproduce the findings from their initial study [55] using the StudentLife dataset [64]. This replication of study findings using a different dataset indicates that the mobility features discussed in [56] could be translational and potentially broadly applicable in depression sensing of different communities (i.e, students on a remote college campus, people recruited in a metropolitan area). Canzian et al [11] develop an expanded set of mobility features over [55] and find that location data correlates with PHQ9 [39, 40, 59]. The authors show that the maximum distance traveled between two places strongly correlates with the PHQ score. Mobility features are used to track the trajectory of depression over

time. Recently, the same team [48] reports on the association between human-smartphone interaction features extracted from a two-week study and depression. A group from the University of Connecticut [21] also finds that behavioral data from phones (e.g., home stay duration, normalized entropy) can predict clinical depression diagnoses with the precision of 84%. In [15] the authors combine wearable and smartphone data and discuss the temporal relationship between self-reported affect and physiological variables linked to depressive symptoms.

The MONARCA project [27, 49, 50, 54] first report on finding from smartphone sensing and bipolar disorder. The authors [50] discuss correlations between the activity levels over different periods of the day and psychiatric evaluation scores associated with the mania-depression spectrum. The findings reported in [1] show the automatic inference of circadian stability as a measure to support effective bipolar management. Maxuni et al. [47] extend these insights by using speech and activity levels to successfully classify stratified levels of bipolar disorder.

There is a growing body of work on stress inference using mobile devices. For stress detection from speech, [44] infers stress with > 0.76 accuracy using acoustic features. Other studies investigate the use of location information [3], measures of social interaction derived from phone calls, SMS, and proximity data [8] to detect stress. In [32, 57, 58], the authors demonstrate using features from both smartphones and wearables to detect and track stress. A number of papers [23, 26, 36, 45] explore using physiological sensing from wearable devices to predict stress; for example, features extracted from galvanic skin response and heart rate variability are found to be good predictors of stress.

There is also a growing amount of research studying early warning signs and rising risk for people with schizophrenia. It is widely accepted that traditional clinical evaluation approaches, such as face to face interviews or periodic self-reported surveys, cannot offer continuous monitoring to detect early warnings of symptom exacerbation [7, 19, 60]. Early work in mobile mental health for schizophrenia conducted by Ben-Zeev et al. [6] first study the feasibility and acceptability of using mobile devices for behavioral sensing among individuals with schizophrenia. In [6] the authors find that participants feel comfortable using the mobile phones, accepting of passive sensing, with participants interested in receiving feedback and suggestions regarding their health. Our previous work on CrossCheck [63] finds associations between passive sensed behaviors and self-reported mental health status. Kerz et al. [34] tests feasibility and acceptability of SleepSight, a system collecting longitudinal accelerometry, heart-rate, ambient light and phone usage patterns for 15 participants diagnosed with schizophrenia living at home. To the best of our knowledge there has been no work that attempts to use sensing data passively collected from mobile phones to predict BPRS scores.

3 CROSSCHECK SYMPTOM PREDICTION SYSTEM

The CrossCheck symptom prediction system comprises the CrossCheck app running on Android phones [63] and the CrossCheck data analytics service running in the cloud. The CrossCheck app collects participants' passive sensing data and self-reported EMA data [63] and uploads it daily to the data analytics service. The CrossCheck data analytics service processes the participants' data and predicts participants' 7-item BPRS scores on a weekly basis. These weekly reports allow the research and clinical teams to reach out to patients if the system predicts rising risk. Figure 1 summarises the systems components, workflow, and outreach.

3.1 CrossCheck App and Sensing System

The CrossCheck app [63] is built based on our prior sensing work [1, 42, 64]. The app continuously infers and record participants' physical activities (e.g., stationary, in a vehicle, walking, running, cycling), sleep (duration, bed time, and rise time), and sociability (i.e., the number of independent conservations a participant is around and their duration). The app also collects audio amplitude, accelerometer readings, light sensor readings, location coordinates, application usages, and call logs. The app uses a built-in MobileEMA component [64] to administer self-reported EMAs [6]. To protect participants' privacy, the app does not collect phone numbers, content of text messages or any conversation content [64]. We remotely erase the CrossCheck data on the phone and reset it if

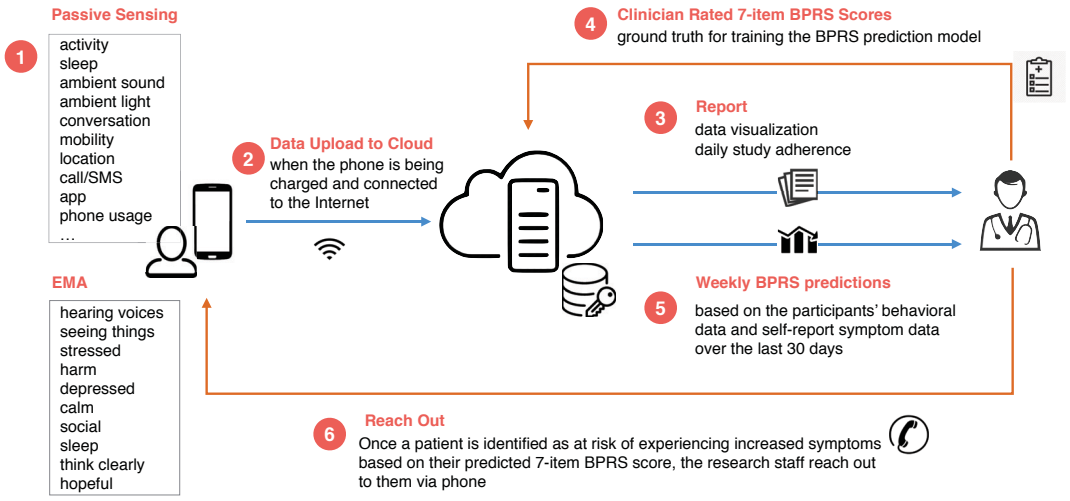


Fig. 1. System overview of the CrossCheck symptom prediction system

the phone is lost. The app uploads the data to the secured CrossCheck data analytics service in the cloud when the participant is charging their phones and under WiFi or cellular data services. The study provides participants with a Samsung Galaxy S5 Android phone and unlimited data plan for the duration of the study. The CrossCheck EMA questions are listed in Table 2. See [63] for more detailed discussion of the implementation and evaluation of the phone application and system.

3.2 CrossCheck Data Analytics System

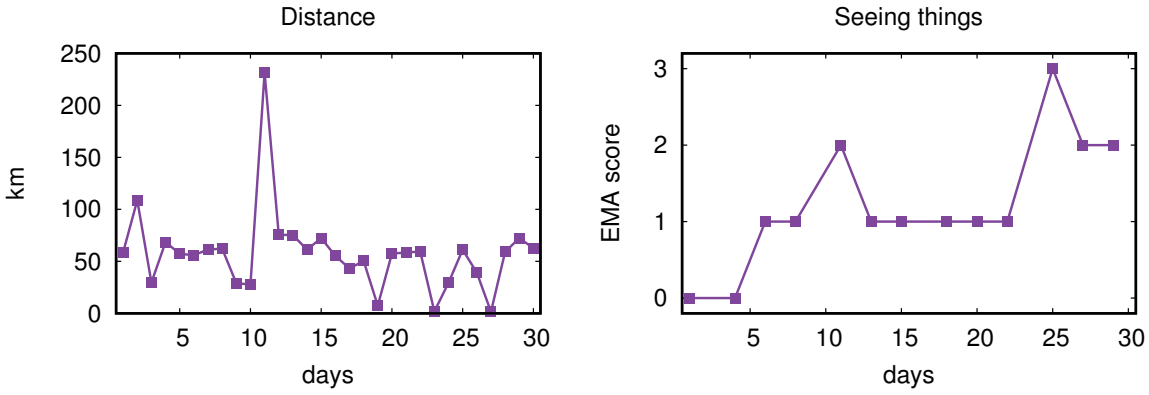
The CrossCheck data analytics system receive and process the data from the app. It generates reports and visualizations for research staff to monitor study adherence and changes in participants' behaviors. The research team periodically receive clinician rated BPRS scores (i.e., ranging from weekly to monthly depending on patients' condition severity) and the system predicts every participant's 7-item BPRS score every week.

Smartphone data processing. The analytics system receives the passive sensing and EMA data from the CrossCheck app and stores the data to a distributed mongoDB database. The analytics system automatically generates behavioral features from raw passive sensing and EMA data. The behavioral features are the basis for visualizing behavior changes, monitoring study adherence, and predicting BPRS scores.

Clinician rated BPRS scores. Research staff input participants' clinician rated 7-item BPRS scores to the system. The clinician rated 7-item BPRS scores are used as the ground truth for training the 7-item BPRS prediction model.

Passive Sensing and EMA Data visualization. The system generates plots to show how participants' behaviors and self-report symptoms change over time. For example, Figure2 shows the distance traveled by a participant, and

self-reported *visual hallucinations* (i.e., the seeing things EMA item) symptom over 30 days. These visualizations help research staff evaluate participants' symptoms in addition to 7-item BPRS predictions.



(a) distance traveled

(b) self-reported visual hallucinations symptom

Fig. 2. Example data visualization used for assessment showing the changes in distance traveled and self-reported visual hallucinations symptom over a 30 day period. Our research staff uses these plots to better understand behavioral trends associated with 7-item BPRS predictions.

Weekly 7-item BPRS predictions. The system predicts every participant's 7-item BPRS scores at the beginning of each week and emails prediction reports to research staff to review. The weekly interval gives research staff enough time to respond (i.e., reach out to patients and their clinical team) if needed. The prediction is based on the participants' passive sensing data and their self-reported symptom data over the last 30 days. The reports contain the predicted 7-item BPRS score for the last three weeks and the changes in the predicted 7-item BPRS scores. Research staff use the 7-item BPRS prediction reports to identify participants at increased risk. The criteria to identify a participant at risk is discussed in Section 4. Once a participant is considered at rising risk, the research staff reach out to the patient to check if they are indeed experiencing increased symptoms. The research staff reach out to clinicians to notify them of any potential risk allowing clinicians to take actions to help patients (e.g., arrange for a caregiver to contact them, schedule an immediate clinical visit). The 7-item BPRS prediction model is described in detail in Section 5.

Daily study adherence reports. To monitor patients' study adherence and detect if any participants are experiencing technical issues that prevent the app from uploading the passive sensing data, the system sends a daily report on how many hours of different sensor data are collected for the last few days. These daily reports label participants who have not uploaded any data. Researchers rely on these daily reports to identify participants who are having problems with the system so they would call non-compliant participants to give assistance and get them back on track – we deem this a technical outreach and not a clinical outreach associated with BPRS prediction. Examples of non-adherence because of technical problems include not using the phone because of problems with the device, no data coverage, can't recharge, or lost or stolen.

4 CROSSCHECK STUDY AND THE DATASET

The CrossCheck symptom prediction system is deployed in an on-going randomized controlled trial [12] conducted in collaboration with a large psychiatric hospital, Zucker Hillside Hospital, in New York City [63]. In what follows, we discuss the study design and dataset.

4.1 CrossCheck Study

The study aims to recruit 150 participants for 12 months using rolling enrollment. The participants are randomized into one of two study arms: smartphone (n=75) or treatment-as-usual (n=75) [63]. This study is approved by the Committee for Protection of Human Subjects at Dartmouth College and Institutional Review Board at Northwell Health System. The CrossCheck symptom prediction system development and deployment is supported by NIH's EUREKA (Exceptional Unconventional Research Enabling Knowledge Acceleration) program, grant number R01MH103148.

Recruiting participants. We use the study hospital's electronic medical record to identify potential study candidates. A candidate is a patient who is 18 or older, met DSM-IV or DSM-V criteria for schizophrenia, schizoaffective disorder, psychosis not otherwise specified, and had psychiatric hospitalization, daytime psychiatric hospitalization, outpatient crisis management, or short-term psychiatric hospital emergency room visits within 12 months before study entry. The candidate should be able to use smartphones and have at least 6th grade reading determined by the Wide Range Achievement Test 4 [65]. Individuals with a legal guardian are excluded. Staff members contact potential study candidates who meet the study criteria to gauge their interest in the study and if interested individuals sign the consent form to join the study. In order to make sure the participants understand the consent form to provide informed consent, the study candidates need to pass a competency screener. After consent is given, enrolled participants are administered a baseline assessment, then are randomly assigned to either the smartphone arm or treatment-as-usual arm. Note, patients assigned to the treatment-as-usual arm do not get a phone and therefore no remote sensing is done. Participants in the smartphone arm are given a Samsung Galaxy S5 Android phone equipped with the CrossCheck app and receive a tutorial on how to use the phone. To ensure the acquired sensing data has broad coverage of behaviors, participants personal phone numbers are migrated over to the new phone and they are provided with an unlimited data plan for data uploading. In this manner, we attempt to increase the likelihood that they will use the study phone as their primary phone. Participants are asked to keep the phone turned on and to carry it with them as they go about their day and charge it close to where they sleep at night.

Monthly 7-item BPRS evaluations. Participants schedule monthly visits with their clinicians. During their visits, clinicians administer the 7-item BPRS. The 7-item BPRS score ranges from 7 to 49. Higher scores are associated with more severe symptoms. The construct of the 7-item BPRS is discussed in Section 4.2.1. Our study staff input the clinician rated 7-item BPRS scores to the CrossCheck data analytics system.

Weekly 7-item BPRS predictions and rising risk criteria. The prediction system sends out a 7-item BPRS prediction report every week to research staff by email. The prediction report shows the participant's predicted 7-item BPRS scores for the last 3 weeks. Our research staff use the predicted 7-item BPRS scores over the preceding 2-week period to identify any patients who may be potentially at risk. A patient at risk is one whose predicted 7-item BPRS score is above 12 or experiences an increase of 10% or more since their last predicted 7-item BPRS score. The research and clinical teams determined the rising risk threshold criteria (i.e., the score cut off and percent change) by studying the historical BPRS scores from patients who experienced relapse; that is, we analyzed scores in time periods prior to relapse to determine the cut-off and, in addition, because some patients' data prior to relapse showed a lower cut off but large increasing percent changes we also determined the additional criteria of the 10% change or greater between two predictions as a red flag.

4.2 CrossCheck Dataset

The dataset comprises the participants' monthly 7-item BPRS scores rated by their clinicians, behavioral features extracted from passive sensing, and symptom features extracted from self-report EMAs. We use 30 days of sensing and self-report EMA data to predict a 7-item BPRS score. The 30-day time frame is called the 7-item BPRS prediction time frame. The 30-day time frame matches the interval of clinician rated 7-item BPRS, which is 30 days on average. The passive sensing features summarize the level of behaviors (e.g., the average conversation duration per day in the 30-day time frame) and behavior changes (e.g., increase or decrease in conversation duration and the dynamics – for example direction and steepness – of change) in the 7-item BPRS prediction time frame. To compute a feature for the prediction time frame, we first compute the daily feature time series from the raw sensing data. We then compute the 30-day features from the daily feature time series. In what follows, we discuss the construction of the dataset in detail.

4.2.1 The 7-item Brief Psychiatric Rating Scale. The BPRS [18, 28, 37, 51, 62] survey is a 24-item rating scale that is is a validated tool administered by clinicians to evaluate symptom severity in schizophrenia. The reliability, validity, and sensitivity of the BPRS measurement has been widely examined [30]. BPRS is rated by a trained rater, usually a clinician. The rater scores each BPRS item based on the patient' responses to questions, observed behavior, and speech.

Table 1 lists all the BPRS items [51]. Each item is rated from 1-7 (1 is given if the symptom is not present to 7 extremely severe). The clinical team at our partner hospital administers 7-item BPRS, which is a subset of the 24-item BPRS. Specifically, the 7 items are grandiosity, suspiciousness, hallucinations, unusual thought content, conceptual disorganisation, blunted affect, and mannerisms and posturing. The clinical research team at Zucker Hillside Hospital chooses these 7 items as a part of their clinical model because they represent the strongest predictors of deterioration in symptoms. A 7-item BPRS total score is computed by summing up the scores from the 7 items. The total score ranges from 7 to 49, where higher score indicates deteriorating symptoms. The 7-item BPRS total score is the outcome of our predictions.

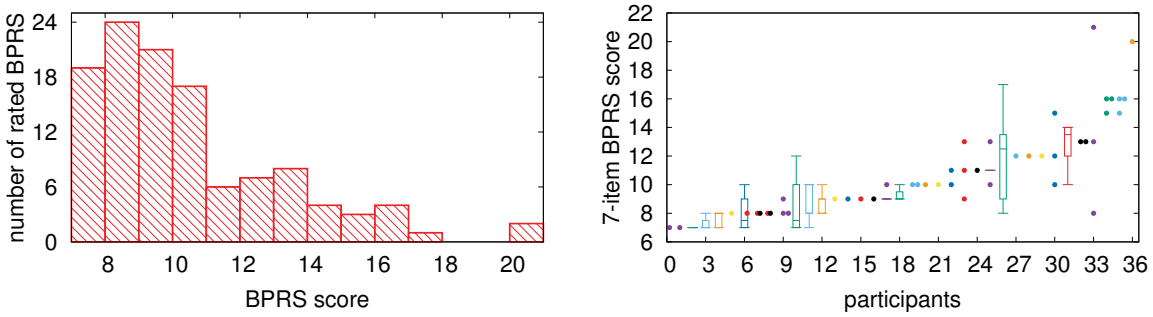
Table 1. Brief Psychiatric Rating Scale Items

Somatic concern, anxiety, depression, suicidality, guilt, hostility
Elevated mood, **grandiosity**, **suspiciousness**, **hallucinations**, **unusual thought content**
Bizarre behavior, self-neglect, disorientation, **conceptual disorganization**
Blunted affect, emotional withdrawal, motor retardation, tension, uncooperativeness
Excitement, distractibility, motor hyperactivity, **mannerisms and posturing**
Note, only items in bold are evaluated during monthly clinic visits.

As mentioned only 7 out of 24 items (marked in bold in Table 1) are evaluated during clinic visits. The 7-item BPRS total score is the sum score of items in bold, which ranges from 7 to 49. The CrossCheck study clinicians enter a score for each of 7 term that best describes the patient's condition at the time of the face to face visit. In what follows we briefly explain each item. For more detail on the BRPS form see [9] and the theory behind the BPRS see [30, 51]. The "grandiosity" item assesses exaggerated self-opinion, arrogance, conviction of unusual power or abilities. The "suspiciousness" item captures mistrust, belief others harbor malicious or discriminatory intent. The "hallucinations" survey item, measures the patient's perceptions without normal external stimulus correspondence. The item that relates to "unusual thought content" gauges unusual, odd, strange, bizarre thought content that the patient has experienced or exhibits. "Conceptual disorganization" relates to how patients' thought processes might be confused, disconnected, disorganized, disrupted. The item associated with "blunted affect" captures reduced emotional tone, reduction in formal intensity of feelings, flatness the patient exhibits

during assessment. The final item, "mannerisms and posturing" rates any peculiar, bizarre, unnatural motor behavior (not including tic) displayed by the patient.

Figure 3(a) shows the distribution of the 7-item BPRS total scores in the dataset. The 7-item BPRS data is from 36 participants over a period of 2-12 month period. There are a total of 116 administered BPRS reports. The BPRS scores range from 7 to 21. The mean score is 10.0 and the standard deviation is 2.86. The score cutoff for symptom deterioration is 12, which is determined by looking at the clinician rated 7-item BPRS scores closest in time to symptomatic relapse for participants who previously relapsed. Figure 3(b) shows the within-individual BPRS score variation. We list participants according to their average BPRS scores. We give greater participant IDs to participants rated with higher average BPRS scores. Some participants record the same BPRS score (e.g., participant 1, 2, and 5) whereas other participants record larger ranges of BPRS scores (e.g., participant 26).



(a) The distribution of all participants' 7-item BPRS total scores

(b) The distribution of every participant's 7-item BPRS scores

Fig. 3. The distribution of 7-item BPRS total scores from 36 participants administered over a of 2-12 month period. In total 116 surveys were administered during this period. The 7-item BPRS scores from the participants ranges from 7 to 21. The mean BPRS score is 10.0, and the standard deviation is 2.86. (a) shows that participants are rated with low 7-item BPRS scores most of the time. However, some cases show higher 7-item BPRS scores, meaning the participants experienced deteriorated symptoms during the span of the study. (b) shows the the within-individual BPRS score variation. Some participants record the same BPRS score (e.g., participant 1, 2, and 5) whereas other participants record larger range of the BPRS scores (e.g., participant 26). The order of participants in (b) is based on their average BPRS score (i.e., participants with greater participant id are rated higher BPRS score on average).

.

4.2.2 Daily Passive Sensing Features. The CrossCheck app collects a wide range of behavioral passive sensing data from the phone. These data capture physical activity, sociability (based on speech and conversational data), mobility, sleep, phone usage, and characteristics of the ambient environment in which participants dwell. We extract features from the multimodal behavioral data to capture the characteristics of participants behaviors that are predictive of 7-item BPRS scores. Features are computed on a daily basis and for specific periods of the day; that is, in addition to daily features, we also compute features over four epoch periods [63]: morning (6 am to 12 pm), afternoon (12 pm to 6 pm), evening (6 pm to 12 am), and night (12 am to 6 am). This allows us to get important behavioral insights during specific periods of a person's day (e.g., conversation during the evening, phone usage at night, etc.).

Activity. The app infers physical activities using the Android activity recognition API [24, 63]. Activity recognition infers whether a user is on foot, stationary, in vehicle, on bicycle, tilting, or doing an unknown activity. Our

evaluation [63] of the Android activity recognition API inferences of walking, in vehicle, and stationary is 95% accurate [63]. In addition to recording activity classes, we compute the durations of walking states, stationary state, and stationary plus in vehicle state. People in a vehicle are typically not physically active in terms off movement. In order to capture the time when the participant is not actively moving, we sum up the stationary duration and the in vehicle duration to estimate the true non-physically active duration. We use the activity recognition classes from Android and our aggregate features for analysis.

Speech and conversational interaction. The CrossCheck app infers the amount of speech and conversation a patient is around [63]. On a daily basis we compute the number of independent conversations and their duration as a proxy for social interaction. We also compute the ratio of detected human voice label observed (e.g., amongst all inferred audio frames during a day, for example, 10% human voice). In [63, 64] we discuss the detailed in design of the conversational classifier that continuously runs on the phone in an energy efficient manner (i.e., duty cycled). In brief, it represents a two level classifier. At the lowest level we detect speech segments and at the higher level we determine if the set of segments represent a unique conversation. Note, we do not record any speech on the phone or upload to the cloud and all audio signal processing and feature extraction is based on privacy preserving algorithms [66–68] destructively processed on the phone. Because we do not do speaker identification we do not know that the patient is actively involved in the conservations or not (e.g., they could be sitting at a table or next to a table in a cafe where others are speaking). We consider the CrossCheck conversational features as a proxy for social interaction and engagement [64]. Finally, the audio classifier is designed, implemented and tuned not to be confused by other audio sources, such as music, conservation on the TV or radio [64].

Location and mobility. We compute distance traveled, the number of places visited, and location entropy from the location data [11, 56, 64]. In order to remove the noise caused by GPS drifting, we use DBSCAN [46] to cluster GPS coordinates collected during the day. DBSCAN groups GPS coordinates that are close together as a significant place where the participant visits. Some coordinates are not grouped to any significant places. These coordinates are collected when the participant is mobile. We compute the number of places visited as a feature. We also compute the distance traveled and location entropy using the centroid coordinates of visited places and coordinates collected when the participant is mobile.

Sleep. We use a novel method to infer sleep data from the phone without the user doing anything outside their normal routine or using any form of makers from self-report. We introduced this "best effort" sleep technique and validated in [14]. Since then it has been used in a number of studies [1, 63, 64]. The classifier infers sleep duration, sleep onset time, and wake time each 24 hour period day[14, 63, 64]. The classifier does not infer naps. It simply computes the longest period of inferred sleep. The sleep classifier approximates sleep duration with in +/- 30 minutes. The sleep inference is based on four sensors on the phone: ambient light, audio amplitude, activity, and screen on/off [14]. In addition, we can observe disrupted sleep by measuring phone activity during the sleep period (e.g., lock/unlock data). It is well-known sleep and changes in sleep patterns are strongly correlated with mental health; a result confirmed using mobile passive sensing in the StudentLife study [64] for depression. We report in [4] that atypical activity at night (indicating fractured sleep) was linked to one of the relapse events in the CrossCheck study.

Phone usage, calls, and texting. We compute the number of phone lock/unlock events and the duration that the phone is unlocked. We also compute the number and duration of incoming and outgoing phone calls and the number of incoming and outgoing SMS messaging.

Ambient environment. CrossCheck activity infers behavioral characteristics of not only the user but also attempts to quantify where the user resides (e.g., significant places such as home, work, etc) and the contextual information associated with a location. We compute features to measure the ambient sound and light conditions of the user's surrounding environment. We compute the mean audio amplitude to determine the acoustic conditions

ranging from quiet to loud environments. We also compute the standard deviation of the audio amplitude to gauge if the audio environment is ambient (e.g., loud air conditioning) or active (e.g., people talking). Similarly, we compute mean and standard deviation of light amplitude to capture the characteristics of the light environment.

4.2.3 Self-reported Ecological Momentary Assessment. Patients periodically respond to a set of short questions related to their symptoms and functioning using their phones. The CrossCheck app administers a 10-item EMA [63] every Monday, Wednesday and Friday. The EMA items are listed in Table 2. Each item scores from 0-not at all to 3-extremely. Amongst the 10 items, 5 items are positive items (i.e., feeling calm, social, sleeping well, think clearly, and hopeful about the future) and the remaining 5 items are negative items associated with symptoms (i.e., bothered by voice, visual hallucinations, feeling stressed, worry about been harmed by other people, and depressed). We use each item's score as a self-report feature to capture symptoms and affect. We also calculate the EMA negative score, positive score, and sum score from the responses [63]. The positive and negative scores represent the sum of all positive and negative items, respectively. The sum score is the positive score minus the negative score.

Table 2. The CrossCheck self-reported EMA questions [63]

Prompt: "Just checking in to see how you've been doing over the last few days".
Have you been feeling CALM?
Have you been SOCIAL?
Have you been bothered by VOICES?
Have you been SEEING THINGS other people can't see?
Have you been feeling STRESSED?
Have you been worried about people trying to HARM you?
Have you been SLEEPING well?
Have you been able to THINK clearly?
Have you been DEPRESSED?
Have you been HOPEFUL about the future?
Options: 0- Not at all; 1- A little; 2- Moderately; 3- Extremely.

4.2.4 Feature Set Constructions. The dataset has two sets of features: passive sensing features and self-report EMA features. In order to construct the feature sets, we first compute the daily feature time series. For example, for a participant evaluated on day d, we compute the daily sensing features and EMA features from day $d-31$ to day $d-1$. Take conversation duration as an example, we compute the total conversation durations over the 24-hour day and four 6-hour epochs everyday from day $d-31$ to day $d-1$. The result is four conversation duration time series. We then compute four features from each of the time series: a mean to capture the level of behavior and three slopes to capture behavior changes.

Time series means. We compute the mean of the daily feature time series. The mean describes the average behavior or self-reported symptoms during the 30-day period. For example, the mean conversation duration over the 30-day period is the average conversation duration the participant is around everyday.

Time series slopes. We compute the slopes of the daily feature time series to describe how behavior or self-reported symptoms change overtime. We fit the feature time series with a linear regression model and use the regression coefficient as the slope. The slope describes the direction and the steepness of the change. For example, a positive slope in conversation duration indicates the participant is around more and more conversations, whereas a negative slope indicates the surrounding conversations decrease over time. The absolute value of the

slope shows how fast the conversation duration changes. We compute three slopes for each time series: over the prediction time frame (slope), over the first 15 days of the prediction time frame (slope 1), and over the last 15 days of the prediction time frame (slope 2). In summary, we extract 486 features from the passive sensing and EMA data. The passive sensing feature set has 434 features and the EMA feature set has 52 features. We use the same feature extraction method to compute features for weekly 7-item BPRS prediction.

4.2.5 Passive Sensing Inclusion Criteria. We define a "good day" as a day with more than 19 hours of the sensing data. In order to avoid missing data skewing the time series features in the prediction time frame, we need to control the data completeness in the 30-day time frame. We include time frames with more than 20 *good days* of the sensing data. As a result, we take a conservative approach to collection of data to increase the fidelity of the data signal. In addition, we use 116 7-item BPRS records and corresponding features from 36 participants for evaluating the 7-item BPRS prediction performance. For the 36 participants included in the analysis, 17 participants are females and 19 are males. 14 patients are African American, 1 Asian, 9 Caucasian, 3 Pacific Islander, 8 Multiracial, and 1 did not disclose. The average age of the 36 participants is 35 years. 8 participants reported they previously owned basic cell phones, 9 did not own any type of cell phone, 19 previously owned a smartphone.

5 PREDICTION MODEL AND RESULTS

In this section, we present the CrossCheck 7-item BPRS prediction model and its prediction performance. We compare the prediction accuracy between three different feature setups: (i) using both the passive sensing feature set and the EMA feature set to predict 7-item BPRS; (ii) using just the passive sensing feature set to predict 7-item BPRS; and (iii) using just the EMA feature set to predict 7-item BPRS. We report the prediction accuracy obtained by two cross validation methods: leave-one-record-out cross validation and leave-one-subject-out cross validation. We then discuss the most significant features selected by the prediction models. We use regression analysis to explore the linear relations between the selected features and the 7-item BPRS score. Finally, we present an example of predicting a participant's weekly 7-item BPRS scores.

5.1 Predicting BPRS Scores

We use Gradient Boosted Regression Trees (GBRT) [22, 52] to predict the 7-item BPRS scores. GBRT is an ensemble method that trains and combines several weak regression trees to make accurate predictions. It builds base estimators (i.e., regression trees) sequentially. Each estimator tries to reduce the bias of the previously combined estimators. GBRT has many advantaged inherited from the tree based classification and regression models: that is, it is less sensitive to outliers [33] and robust to overfitting [20]. It computes feature importance measures, which can be used for feature selection.

In order to understand the prediction accuracy of the three different feature setups, we train three models with (i) using both the passive sensing features and the EMA features; (ii) using just the passive sensing features; and (iii) using just the EMA features. We evaluate the prediction accuracy with leave-one-record-out cross validation and leave-one-subject-out cross validation. The leave-one-record-out cross validation leaves one 7-item BPRS example out from the dataset as the testing example and use the rest of the examples for training the model. The results from the leave-one-record-out cross validation show the prediction accuracy of predicting an existing participant's 7-item BPRS score. The participant's previous clinician rated 7-item BPRS scores are available to the system to improve the prediction accuracy by incorporating the data to the training examples. The leave-one-subject-out cross validation trains the model with data from subjects other than the testing subject and tests on the testing subject's data. The results from the leave-one-subject-out cross validation shows the prediction accuracy of predicting a new participant who just joined the study when their clinician rated 7-item BPRS scores are not available to the system.

Feature selection. Considering the high dimensionality of the feature space (i.e., 486 features) and the relatively small number of training examples that exist (i.e., 116 BPRS surveys), we need to significantly reduce the feature space dimensionality so that the prediction model can be properly trained. We select features based on GBRT feature importance. GBRT computes feature importance by averaging the number of times a particular feature is used for splitting a branch across the ensemble trees, higher values are deemed as more important. The feature importance value ranges from 0 to 1, where higher values indicate more important features. We train the GBRT model on all the 7-item BPRS data. We select features with a feature importance greater than the average importance value of all features. We repeat this process until no more than 20 features are left. The heuristic 20 feature rule is based on our experiments in which we find we get higher training error with a lower or higher threshold. We repeat this process for the three feature set setups as discussed above: that is, passive sensing and EMA, passive sensing only, and EMA only.

Prediction performance. We use mean absolute error (MAE), the Pearson's r, and generalized estimating equations (GEE) [10, 17, 43, 69] to evaluate the prediction performance. MAE describes the bias of the predictions. The Pearson correlation treats the predicted BPRS scores as independent variables. The Pearson's r describes how well the predictions capture the outcome's variance. GEE focuses on estimating the average response over the population [43]. It is a more robust method to evaluate correlations between repeated measures. The GEE coefficient shows the direction of the correlation and the p-value indicates the statistical significance of the coefficient.

Table 3. Prediction performance

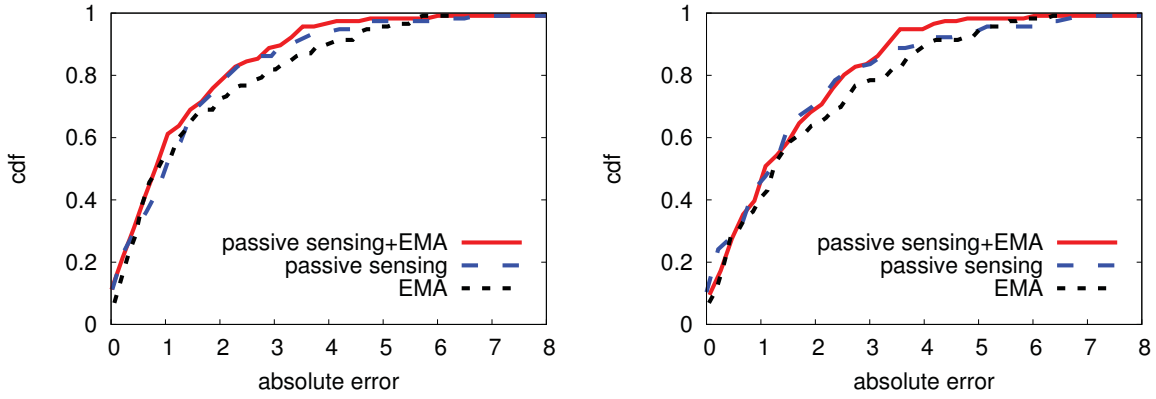
		passive sensing + EMA	passive sensing	EMA
leave-one-record-out	MAE	1.45	1.59	1.62
	Pearson's r	0.70*	0.63*	0.62*
	GEE coeff	1.05*	1.11*	0.81*
leave-one-subject-out	MAE	1.70	1.80	1.90
	Pearson's r	0.61*	0.48*	0.50*
	GEE coeff	0.99*	0.93*	0.81*

* $p < 0.0001$

Table 3 shows the mean absolute error, the Pearson's r, and GEE coefficient for all models predicting the BPRS score. The leave-one-record-out cross validation with both passive sensing and EMA features achieves the best result with MAE = 1.45, meaning we can predict the 7-item BPRS score with on average ±1.45 error (3.5% of the scale). The predicted 7-item BPRS scores strongly correlate with the 7-item BPRS ground truth (i.e., clinician scored BPRS surveys) with $r = 0.70, p < 0.0001$. The result shows our existing system can accurately predict patients' 7-item BPRS scores. The result gives us confidence to track symptoms every week. The prediction performance for leave-one-record-out cross validation using only passive sensing or EMA features is MAE = $1.59, r = 0.63, p < 0.0001$ (3.8% of the scale) and MAE = $1.62, r = 0.62, p < 0.0001$ (3.9% of the scale), respectively. The leave-one-subject-out cross validation offers the best prediction performance using both passive sensing and EMA features with MAE = 1.70 (4.0% of the scale). The predicted 7-item BPRS scores strongly correlate with the 7-item BPRS ground truth with $r = 0.61, p < 0.0001$.

The prediction performance for leave-one-subject-out cross validation using only passive sensing or EMA features is MAE = $1.80, r = 0.48, p < 0.0001$ (4.3% of the scale) and MAE = $1.90, r = 0.50, p < 0.0001$ (4.5% of the scale), respectively. When comparing using both passive sensing and EMA features, results in a 0.1 and 0.2 increase in absolute errors, respectively. Again, passive sensing features outperforms EMA features in term of MAE. Figure 4 shows the cumulative distribution function (CDF) of the absolute error of the 7-item BPRS

predictions in greater detail. In both cross validations, we see combining the passive sensing and EMA features performs better than just using passive sensing features, which in turn outperforms EMA features.



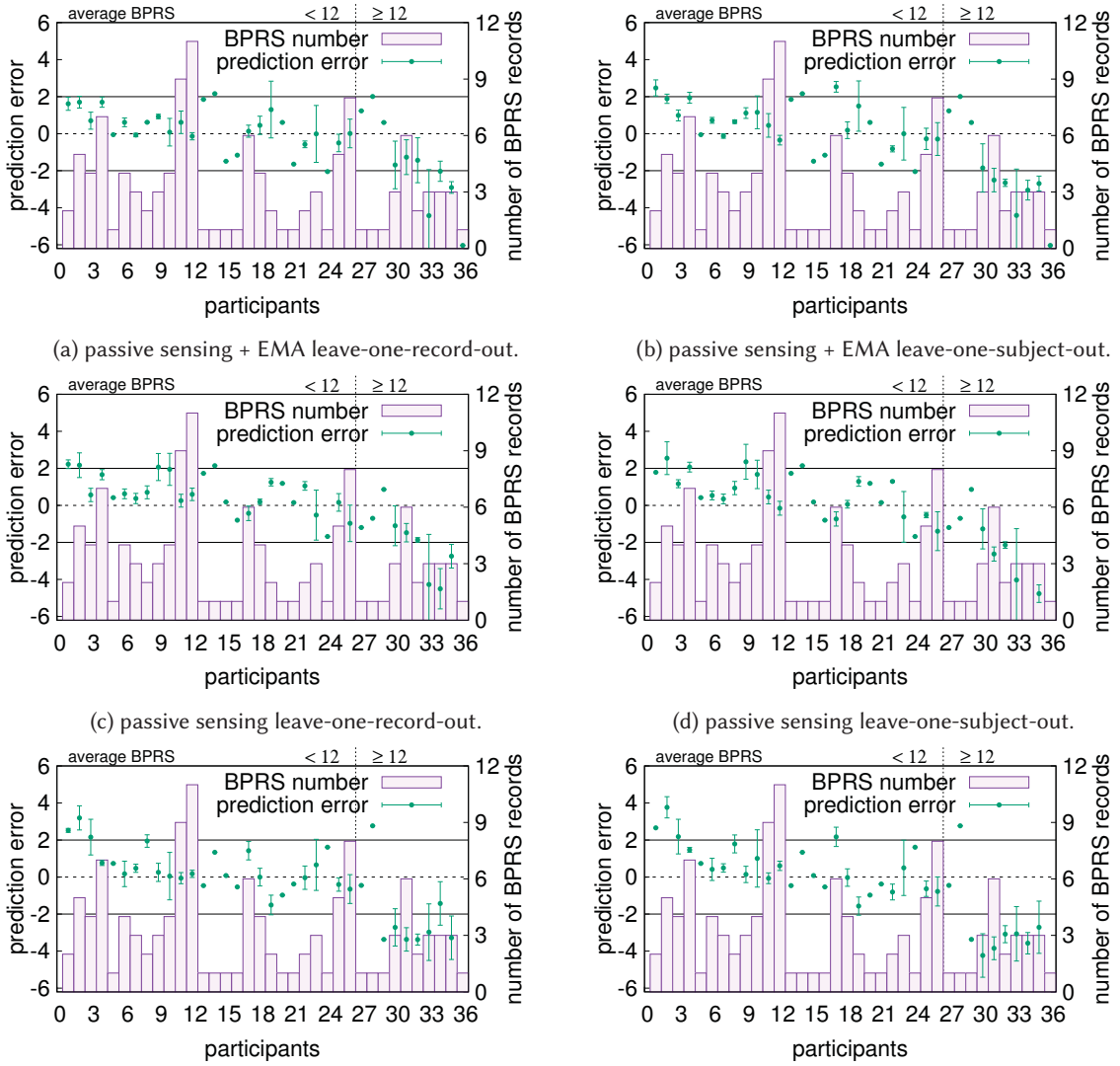
(a) leave-one-record-out. (b) leave-one-subject-out.

Fig. 4. The cumulative distribution function (CDF) of the absolute errors for leave-one-record-out cross validation and leave-one-subject-out cross validation. Using both passive sensing and EMA features results in the best prediction performance, followed closely by passive sensing alone, whereas using only EMA features presents the worst prediction performance.

Within-individual prediction errors. Figure 5 shows the average within-individual prediction error of the six models with the different feature setups and cross-validation methods. The order of the participants shown in the plots is determined by their average clinician rated BPRS scores. We observe that all six models archive lower prediction errors for participants with lower clinician rated BPRS scores but higher errors for participants with higher clinician rated BPRS scores. Judging from Figure 3(a) most of the clinician rated BPRS scores are between 7 and 12. Therefore, the dataset is unbalanced and skews to lower BPRS scores (< 12). The GBRT models are undertrained for higher BPRS scores (≥ 12). As a result, the models underestimate high-BPRS-score participants' scores (i.e., participants with average BPRS ≥ 12). The prediction models need more high-BPRS-score participants' data to improve the prediction performance. The impact of prediction errors on clinical practice is discussed in Section 5.3.

5.2 Interpreting Selected Features

We use bivariate regression analysis to understand the linear relationship between the features selected by GBRT feature importance measures and the 7-item BPRS scores. Considering the longitudinal nature of our dataset; that is, because data from the same subject is likely correlated we apply generalized estimating equations (GEE) [10, 17, 43, 69] to determine associations between each of the selected features and the 7-item BPRS scores. In order to better understand the regression results, we normalize each feature to zero mean and one standard deviation so that the coefficients are within a reasonable range. Table 4 shows the 13 selected features from the model using both passive sensing and EMA features based on the feature selection criteria described in Section 5.1. Out of the 13 features, two are self-reported EMA features associated with hearing voices and being social and 11 are passive sensing features. The passive sensing features cover a broad range of behaviors: unlocking phones, conservation, being stationary, visiting different locations, and being in different ambient acoustic environments. Interesting enough, the model selects the duration that a patient is stationary or in vehicle (we consider the patient as stationary while in a vehicle) as a predictor rather than simply the duration of being stationary. It



(a) passive sensing + EMA leave-one-record-out.

(b) passive sensing + EMA leave-one-subject-out.

(c) passive sensing leave-one-record-out.

(d) passive sensing leave-one-subject-out.

(e) EMA leave-one-record-out.

(f) EMA leave-one-subject-out.

Fig. 5. The average within-individual prediction error of the six models. The patients are ordered by their average rated BPRS scores. The vertical dashed line separate patients with average BPRS score ≤ 12 and patients with BPRS score > 12. The horizontal lines labels the region with prediction error more than -2 and less than 2. Patients with higher rated BPRS scores get worse predictions. This is because the dataset is skewed to patients with lower BPRS scores.

indicates that the combination of the stationary label and the in vehicle label gives a stronger 7-item BPRS predictor. GEE finds four significant associations between the selected features and the 7-item BPRS scores. For example, patients who report hearing voices, tend to use their phone more often during the evening period, are typically around more voices (i.e., more audio frames are labeled as human voice) in the morning, spend an

increasing amount of time in more active ambient sound environments (i.e., there is more variation in audio volume over time) in the morning during the first 15 days of the 7-item BPRS prediction time frame; these patients are more likely to have higher BPRS scores. The model selects seven slope features and six mean features for prediction, which shows that behavior changes are good 7-item BPRS predictors.

Table 4. Selected features for the passive sensing and EMA model, features with $p < 0.05$ are in bold.

feature	GEE coefficient	p value
unlock number slope 1	0.248	0.505
ambient sound volume afternoon slope	0.236	0.484
voices EMA	0.994	0.006
unlock duration evening	0.914	0.025
ambient sound volume evening slope 2	0.491	0.100
call out number evening	0.791	0.086
social EMA slope 2	1.308	0.173
ambient sound volume standard deviation morning slope 1	0.600	0.008
stationary and in vehicle duration night slope	0.346	0.272
conversation duration slope	-0.015	0.975
ratio of voice frames morning	0.647	0.041
number of visited locations	-0.196	0.556
ratio of voice frames	0.310	0.356

Table 5 shows the 18 features selected from the prediction model for passive sensing features only. The features cover a wide range of behaviors. In comparison with the features selected using a combined model for passive sensing and EMA, the pure passive sensing model selects four additional features related to phone calls. GEE finds six significant associations between the selected features and the 7-item BPRS scores. Specifically, participants who decrease phone usage during the night, have more phone usage during the afternoon, stay in louder acoustic environments with more human voice, and show increasing visits to more places (i.e., locations) in the morning during the second 15 days of the 7-item BPRS prediction time frame; these patients are more likely to have higher 7-item BPRS scores. Out of the 18 features, 12 are associated with behavior changes. Again, we observe that behavior changes are strongly predictive of 7-item BPRS scores. Specifically, conversation duration is not selected as a predictor whereas the change in conversation duration is considered a predictor of 7-item BPRS.

Table 6 shows the selected EMA features by the EMA model. GEE finds 4 significant associations between the selected features and the 7-item BPRS scores. Specifically, patients who report decreasing sociability, feeling less calm, and increases in hearing voices and the feeling of being harmed are more likely to have higher BPRS scores.

In summary, a wide range of behaviors captured by phones are predictors of the 7-item BPRS score. We find that changes in behavior are more predictive than the absolute level of behaviors. The bivariate regression analysis, however, does not confirm that every selected feature is linearly associated with the 7-item BPRS scores. This is because the regression analysis finds only linear association whereas the GBRT model is non-linear – capturing non-linear relations between features and outcomes. Furthermore, prior work [25] has found features with little power at predicting outcomes when combined with other features can provide significant performance improvements.

5.3 Application of Predicting Weekly 7-item BPRS Scores

We use our prediction model to predict the 7-item BPRS scores each week. The model predicts weekly BPRS scores using both passive sensing and EMA features. The model is updated weekly using any new BPRS evaluations

Table 5. Selected features for the passive sensing model, features with $p < 0.05$ are in bold.

feature	GEE coefficient	p value
unlock number night slope	-0.896	< 0.001
ratio of voice frames slope 2	0.422	0.114
stationary and in vehicle duration night slope	0.346	0.272
unlock duration afternoon	0.857	0.039
conversation duration afternoon slope	-0.145	0.624
call out duration slope	0.112	0.597
ambient sound volume night	0.158	0.603
ambient sound volume morning	0.906	0.006
call in duration morning	0.083	0.714
call out number afternoon slope	0.179	0.643
ratio of voice frames morning	0.647	0.041
ambient sound volume	0.769	0.008
ambient sound volume evening slope 2	0.491	0.100
number of visited locations morning s2	0.485	0.022
call out duration slope 2	-0.266	0.164
number of visited locations slope 2	0.065	0.835
conversation duration morning slope	-0.368	0.081
distance traveled evening slope	0.141	0.178

from clinicians. Figure 6 shows an example of a patient's predicted BPRS scores over a 10-week period. The scores are computed weekly and reflected in weekly prediction reports automatically sent out every Sunday evening to all researcher staff in the study. In this case, a clinician evaluated the patient during week 4 and week 8 scoring a BRPS value of 7 (i.e., no symptoms) for each clinic visit. Please note, the differences between clinician's scores and the predicted scores may due to the different days the patient was rated and the predictions were made. The predictions, however, show that the participants' BPRS scores are slightly higher before the first evaluation, between two evaluations, and after the second evaluation. The result shows that the prediction may capture nuanced changes in the patient's state that could not be observed with out the predictive reports and mobile sensing. This example highlights the strength of our approach and vision. It may provide opportunities for clinicians and care givers to reach out to outpatients.

Our research staff use the weekly 7-item BPRS predictions to determine if a patient is at risk and requires reachout from the clinical team. A patient is at risk if the predicted 7-item BPRS score is above 12 or experiences an increase of 10% or more since their last predicted 7-item BPRS score. The prediction errors would affect how research staff determine whether a patient is at risk. Suppose a patient's true BPRS score is 11, research staff would correctly identify this patient as not at risk if the predicted score is below 12 (e.g., a negative error). However, if the predicted score is more than 12 (i.e., a positive error greater than 1) the research staff would incorrectly identify the patient as at risk (i.e., a false positive). Conversely, if a patient's true BPRS score is 15, which is above the cutoff, the research staff would correctly identify that this patient is at risk if the prediction has a positive error. A negative error of more than 3 (i.e., predicted BPRS score is less than 12) would make the research staff incorrectly identify the patient as not at risk. Figure 5(a) shows that the average absolute errors of the predictions are below 2 for participants whose average BPRS scores are below 12, which indicates that they are not likely to be incorrectly identified as at risk. For participants whose average BPRS scores are above 12, the predicted scores are likely to be lower than the true score. However, the errors are not big enough to make the

Table 6. Selected features for EMA model, features with $p < 0.05$ are in bold.

feature	GEE coefficient	p value
social	-0.027	0.949
social slope 2	1.308	0.173
sum score	-0.414	0.354
sleeping slope	-0.211	0.360
calm slope	-0.715	0.038
voices	0.994	0.006
harm	0.806	0.030
negative score	0.699	0.089
depressed	0.228	0.661
calm	0.013	0.976
think	-0.116	0.796
stressed slope	0.217	0.440
sum score slope	-0.340	0.266
stressed	0.222	0.616
social slope	-0.369	0.264
negative score slope	0.239	0.499
think slope	0.101	0.829
positive score slope 2	-0.055	0.895
hopeful slope	0.242	0.152
voices slope	0.257	0.344

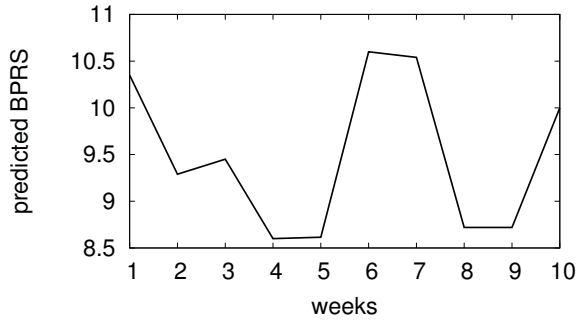
predicted scores below the cutoff. For example, participants 34,35's lowest true BPRS scores are 15 whereas the errors are below -3 thus the predicted scores are still above the cutoff 12.

In addition to the cutoff, the research staff use the changes in the predicted BPRS scores to identify patients at risk. The predicted BPRS scores highly correlate with clinician rated BPRS scores. Therefore, how the BPRS scores change over time is a symptom deterioration indicator. We combine both the score cutoff and changes in two consecutive scores as symptom deterioration indicators for reachout. In the next section, we present a number of case studies that show that the predictive system reflects what is going on in patients lives.

The study is ongoing and we are evolving the BPRS prediction based patient-at-risk criteria described above to reduce false positives and false negatives.

6 PATIENT CASE STUDIES

The CrossCheck prediction model is re-trained each week if new clinician rated 7-BPRS scores are available. Each week our research staff review the weekly prediction scores of all patients in the smartphone arm of the randomized control trial to determine if patients are at risk – based on the criteria discussed previously. Once we identify that a patient might be at risk, our research staff outreach and contact the participant and the clinical team at the hospital. A natural question when building predictive infrastructure is how well the system predicts time varying symptoms of patients, and how well the system reflects the current symptoms and risk experienced by outpatients living out in the community. In what follows, we provide insights into the lives of three patients at the time our system indicates increasing symptoms. We show through anecdotal information from research staff and the clinical team when reaching out to patients when the prediction system indicates rising risk.



111

Fig. 6. A participant's predicted 7-item BPRS score over 10 weeks. The clinician gave a score 7 twice for this participant in week 4 and week 8. The BPRS predictions, however, shows the scores changes during the two evaluations.

The first patient is a 55 year-old African American male diagnosed with schizophrenia, paranoid type. He was clinically flagged on August 22, 2016 based on an elevated predicted 7-item BPRS score of 12.86. When our research staff at the hospital contacted the patient on August 24, 2016, he endorsed symptom increases over the past three months with increasing intensity over the past three weeks. He discussed negative thoughts he'd had about his deceased mother who had passed away five years earlier. He said that he sees images of his mother in his mind when she was in her 30s. However, she was in her 70s when she passed away. The patient also said he believed these thoughts were present to make him feel "emotionally sick." The on-site researcher (who is located at the hospital) and patient discussed coping mechanisms, such as, relating these thoughts in therapy to support persons, positive self-talk, and writing his thoughts down on paper as a reality check. Once the researcher determined that the patient was not in any imminent danger, the researcher encouraged him to share all these symptoms with his treatment team and then brought the call to an end.

As the study protocol dictates, once a researcher reaches out to a patient, the researcher then contacts the clinical team to inform them of the CrossCheck prediction assessment and the symptoms reported by the patient. In this case, the patient's psychiatrist reviewed the new information and told the researcher that the patient had been experiencing difficulty scheduling his next outpatient medication management appointment. Because of this new information provided by the CrossCheck team the psychiatrist immediately reached out to the patient's case manager to coordinate an in-person visit, which occurred less than a week after the initial research outreach. The psychiatrist determined during the clinical visit that the patient was below his baseline level of functioning and adjusted his medication accordingly. This case shows the predictive system, outreach and clinical assessment all concur strongly.

The next patient is a 63 year-old caucasian male diagnosed with schizophrenia. He was clinically flagged on October, 17 2016 based on an increase in predicted BPRS score. Although the predicted 7-item BPRS score was 10.63, which fell short of the 12 point cut-off marker, the BPRS score represented an increased of 16% over the previous week. In addition, self-reported EMAs indicated deteriorating symptoms and the passive sensing data signaled limited sleep. On October 18, 2016 the patient was contacted by phone by a member of the research staff. During the conversation, the patient reported he'd told his therapist the day before, October, 17 2016 that he planned to kill himself on December 1, 2016. He endorsed thoughts of killing himself several times a day, but he was able to hold off on these thoughts. He said he felt hopeless and negative about the future. In addition, he said he had disturbed sleep due to bad dreams and flashbacks, which he experienced several times during the night, most nights. The researcher and patient discussed ways to manage his thoughts, including mindfulness, therapy, and attending his day program. He endorsed medication adherence with no disturbances in appetite.

The patient's treatment team, consisting of his psychiatrist and therapist, were notified immediately after the call of the patient's mental status and symptom exacerbation. The patient informed his therapist the day prior to his plan to commit suicide on a specified date, and together they were able to complete a safety assessment, which included working on coping skills in treatment. Through the care coordination efforts by the research team and clinical team, the patient was placed on a high-risk list and monitored more closely by his treatment providers.

The researcher assessed the patient for safety and determined he was not in imminent danger to self or others. The researcher and patient discussed ways to manage his thoughts, including mindfulness, therapy, and attending his day program. He endorsed medication compliance with no disturbances in appetite. The patient's treatment team, consisting of his psychiatrist and therapist, were notified immediately after the call of the patient's mental status and symptom exacerbation. The patient informed his therapist the day prior to his plan to commit suicide on a specified date, and together they were able to complete a safety assessment, which included working on coping skills in treatment. Through the care coordination efforts by the research team and clinical team, the patient was placed on a high risk list and monitored more closely by his treatment providers. This case represents an example where the predicted score was just below the risk threshold but the weekly percentage change was significant to flag the patient as potentially at risk.

The final patient is a 19 year-old Asian male diagnosed with psychosis not otherwise specified. On November 1, 2016, our research staff noticed his predicted BPRS score was 16.71, which was a 12.7% increase from the last predicted score. The researcher was able to reach the participant for a verbal check-in on November 2, 2016. During the call, the patient denied all symptoms to the researcher, including any sleep disturbances, changes in eating behaviors, or auditory/visual hallucinations. He said that he was socializing well with his friends and family and endorsed medication adherence, denying any thoughts of self-harm or harm to others. The patient reported that he felt tired that morning and was unable to attend school that day. The researcher thanked the patient for his time and encouraged him to share any symptoms with his treatment provider. After several weeks of this patient rescheduling his weekly session with his therapist, on November 29, 2016, the therapist confirmed symptom increases. The therapist told the researcher, that the patient was symptomatic, experiencing psychosis in the form of auditory hallucinations, poor concentration, and distractibility. This case shows our system predicting increased symptoms, the patient not concurring with this assessment, but clinical experts confirming the prediction.

7 CONCLUDING REMARKS

The CrossCheck system discussed in this paper shows promise in using mobile phones and passive sensing to predict symptoms of schizophrenia for people living out in the community. The system and models show good performance using passive sensing and self-reports as well as just using passive sensing. A system based purely on passive sensing opens the way for continuous assessment of symptoms and risk as people go about their everyday lives.

The CrossCheck study is on-going. We plan to continue using the CrossCheck symptom prediction system to track and better understand patients' symptoms during the remainder of the randomized control trial. Our system allows research staff and clinical teams to use the periodic reports to reach out and assess better how the predictions reflect real life experiences of our study participants; that is, we will further evaluate the system's effectiveness at identifying participants experiencing deteriorated symptoms. We will use information from future case studies as feedback to refine our models and system. Ultimately we aim to develop a system capable of not just indicating rising risk, but rather, we aim to develop a model and associated systems to accurately predict relapse and through intervention and treatment adjustment keep patients healthy and out of hospital. As things stand today we do not have sufficient cases of relapse and hospitalizations in our cohort to build statistically significant models.

We also recognize limitations of our work. We only had 116 BPRS clinician scored surveys to train our model. Typically, in the lifetime of the study clinicians administer BRPS once per month on average for each patient – 12 per year for each patient. In our on-going study, outpatients do not experience severe symptoms often and thus mostly report lower 7-item BPRS scores. Therefore, the current dataset is unbalanced and skews toward lower BPRS scores, as discussed earlier in the paper. The unbalanced dataset causes our prediction models to underestimate the BPRS scores of patients with higher clinician rated BPRS scores. However, we show that clinicians can adjust the score cutoff for symptom deterioration and leverage the changes in predicted BPRS scores to reduce the false negatives. To further improve BPRS prediction, we need to collect more data, especially from patients with more severe symptoms. We would also need to apply re-sampling techniques, such as SMOTE [13], to balance the dataset. While our initial results are promising we plan to address these limitations at the end of the CrossCheck randomized controlled trial.

Another possible limitation is that all patients live in a large dense city and the models may not generalize to other locations, such as, patients living in rural communities. Study adherence is also an issue. Patients break, loose, lend, and neglect to use or charge their phones. In some cases they experience persistent cellular or WiFi coverage issues for our system to successfully upload their data in a timely manner (i.e., once per day when they are charging their phones and under cellular or WiFi). We continually try to think of innovative ideas to deal with these issues and currently rely on technical outreach (as distinct from outreaches associated with increased symptoms) and removing incomplete data if we do not have sufficient per day (i.e., at least 19 hours per day as discussed in Section 4 for reasons of model performance. Finally, we discussed three case studies that showed the predicted system correctly reflected increasing risk; that is, the CrossCheck symptom prediction system accurately captured the changing conditions of these patients as reported by the research and clinical teams that reached out to them or interacted with them during subsequent clinical visits, respectively. These results look very promising.

ACKNOWLEDGMENTS

The research reported in this article is supported the National Institute of Mental Health, grant number R01MH103148.

REFERENCES

[1] Saeed Abdullah, Mark Matthews, Ellen Frank, Gavin Doherty, Geri Gay, and Tanzeem Choudhury. 2016. Automatic detection of social rhythms in bipolar disorder. *Journal of the American Medical Informatics Association* 23, 3 (2016), 538–543.

[2] Min Aung, Faisal Alquaddoomi, Cheng-Kang Hsieh, Mashfiqui Rabbi, Longqi Yang, JP Pollak, Deborah Estrin, and Tanzeem Choudhury. 2016. Leveraging Multi-Modal Sensing for Mobile Health: a Case Review in Chronic Pain. *IEEE Journal of Selected Topics in Signal Processing* 10, 5 (2016), 1–13.

[3] Gerald Bauer and Paul Lukowicz. 2012. Can smartphones detect stress-related changes in the behaviour of individuals?. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2012 IEEE International Conference on*. IEEE, 423–426.

[4] D Ben-Zeev, R Brian, R Wang, W Wang, AT Campbell, MS Aung, M Merrill, VW Tseng, T Choudhury, M Hauser, and others. 2017. CrossCheck: Integrating Self-Report, Behavioral Sensing, and Smartphone Use to Identify Digital Indicators of Psychotic Relapse. *Psychiatric rehabilitation journal* (2017).

[5] Dror Ben-Zeev, Gregory J McHugo, Haiyi Xie, Katy Dobbins, and Michael A Young. 2012. Comparing retrospective reports to real-time/real-place mobile assessments in individuals with schizophrenia and a nonclinical comparison group. *Schizophrenia bulletin* 38, 3 (2012), 396–404.

[6] Dror Ben-Zeev, Rui Wang, Saeed Abdullah, Rachel Brian, Emily A Scherer, Lisa A Mistler, Marta Hauser, John M Kane, Andrew Campbell, and Tanzeem Choudhury. 2015. Mobile behavioral sensing for outpatients and inpatients with schizophrenia. *Psychiatric services* 67, 5 (2015), 558–561.

[7] Max Birchwood, Jo Smith, Fiona Macmillan, Bridget Hogg, Rekha Prasad, Cathy Harvey, and Sandy Bering. 1989. Predicting relapse in schizophrenia: the development and implementation of an early signs monitoring system using patients and families as observers, a preliminary investigation. *Psychological Medicine* 19, 03 (1989), 649–656.

[8] Andrey Bogomolov, Bruno Lepri, Michela Ferron, Fabio Pianesi, and Alex Sandy Pentland. 2014. Daily stress recognition from mobile phone data, weather conditions and individual traits. In *Proceedings of the ACM international conference on multimedia*. ACM, 477–486.

[9] BPRS 2017. Brief Psychiatric Rating Scale (BPRS) Expanded Version (4.0). (2017). http://www.public-health.uiowa.edu/icmha/outreach/documents/bprs_expanded.pdf

[10] P Burton, L Gurrin, and P Sly. 1998. Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. *Statistics in medicine* 17, 11 (jun 1998), 1261–91. http://www.ncbi.nlm.nih.gov/pubmed/9670414

[11] Luca Canzian and Mirco Musolesi. 2015. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 1293–1304.

[12] Thomas C Chalmers, Harry Smith, Bradley Blackburn, Bernard Silverman, Biruta Schroeder, Dinah Reitman, and Alexander Ambroz. 1981. A method for assessing the quality of a randomized control trial. *Controlled clinical trials* 2, 1 (1981), 31–49.

[13] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[14] Zhenyu Chen, Mu Lin, Fanglin Chen, Nicholas D Lane, Giuseppe Cardone, Rui Wang, Tianxing Li, Yiqiang Chen, Tonmoy Choudhury, and Andrew T Campbell. 2013. Unobtrusive sleep monitoring using smartphones. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2013 7th International Conference on*. IEEE, 145–152.

[15] Philip Chow, Wesley Bonelli, Yu Huang, Karl Fua, Bethany A Teachman, and Laura E Barnes. 2016. DEMONS: an integrated framework for examining associations between physiology and self-reported affect tied to depressive symptoms. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. ACM, 1139–1143.

[16] Jeremy W Coid, Simone Ullrich, Paul Bebbington, Seena Fazel, and Robert Keers. 2016. Paranoid ideation and violence: meta-analysis of individual subject data of 7 population surveys. *Schizophrenia bulletin* 42, 4 (2016), 907–915.

[17] Peter Diggle, Patrick Heagerty, Kung-Yee Liang, and Scott Zeger. 2013. *Analysis of longitudinal data*. OUP Oxford.

[18] Lisa Dixon, Gretchen Haas, Peter J Weiden, John Sweeney, and Allen J Frances. 1991. Drug abuse in schizophrenic patients: clinical correlates and reasons for use. *Am J Psychiatry* 148, 2 (1991), 224–230.

[19] Enrique Echeburúa, Montserrat Gómez, and Montserrat Freixa. 2017. Prediction of Relapse After Cognitive-Behavioral Treatment of Gambling Disorder in Individuals With Chronic Schizophrenia: A Survival Analysis. *Behavior Therapy* 48, 1 (2017), 69–75.

[20] Jane Elith, John R Leathwick, and Trevor Hastie. 2008. A working guide to boosted regression trees. *Journal of Animal Ecology* 77, 4 (2008), 802–813.

[21] Asma Ahmad Farhan, Chaoqun Yue, Reynaldo Morillo, Shweta Ware, Jin Lu, Jinbo Bi, Jayesh Kamath, Alexander Russell, Athanasios Bamis, and Bing Wang. 2016. Behavior vs. Introspection: Refining prediction of clinical depression via smartphone sensing data. In *7th Conference on Wireless Health, WH*.

[22] Jerome H. Friedman. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* 29, 5 (2001), 1189–1232. http://www.jstor.org/stable/2699986

[23] Martin Gjoreski, Hristijan Gjoreski, Mitja Luštrek, and Matjaž Gams. 2016. Continuous Stress Detection Using a Wrist Device: In Laboratory and Real Life. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct (UbiComp '16)*. ACM, New York, NY, USA, 1185–1193. DOI: http://dx.doi.org/10.1145/2968219.2968306

[24] Google Activity Recognition Api. 2017. Google Activity Recognition Api. https://developers.google.com/android/reference/com/google/android/gms/location/ActivityRecognitionApi. (2017).

[25] Isabelle Guyon and André Elisseeff. 2003. An introduction to variable and feature selection. *Journal of machine learning research* 3, Mar (2003), 1157–1182.

[26] Katrin Hänsel, Akram Alomainy, and Hamed Haddadi. 2016. Large Scale Mood and Stress Self-assessments on a Smartwatch. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct (UbiComp '16)*. ACM, New York, NY, USA, 1180–1184. DOI: http://dx.doi.org/10.1145/2968219.2968305

[27] C Haring, R Banzer, A Gruenerbl, S Oehler, G Bahle, P Lukowicz, and O Mayora. 2015. Utilizing Smartphones as an Effective Way to Support Patients with Bipolar Disorder: Results of the Monarca Study. *European Psychiatry* 30 (2015), 558.

[28] Steven A Harvey, Elliot Nelson, John W Haller, and Terrence S Early. 1993. Lateralized attentional abnormality in schizophrenia is correlated with severity of symptoms. *Biological Psychiatry* 33, 2 (1993), 93–99.

[29] William B Hawthorne, David P Folsom, David H Sommerfeld, Nicole M Lanouette, Marshall Lewis, Gregory A Aarons, Richard M Conklin, Ellen Solorzano, Laurie A Lindamer, and Dilip V Jeste. 2012. Incarceration among adults who are in the public mental health system: Rates, risk factors, and short-term outcomes. *Psychiatric Services* 63, 1 (2012), 26–32.

[30] James Hedlund. 1980. *The brief psychiatric rating scale (BPRS): A comprehensive review*.

[31] Kahyee Hor and Mark Taylor. 2010. Review: Suicide and schizophrenia: a systematic review of rates and risk factors. *Journal of psychopharmacology* 24, 4_suppl (2010), 81–90.

[32] Karen Hovsepian, Mustafa al'Absi, Emre Ertin, Thomas Kamarck, Motohiro Nakajima, and Santosh Kumar. 2015. cStress: Towards a Gold Standard for Continuous Stress Assessment in the Mobile Environment. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. ACM, New York, NY, USA, 493–504. DOI: http://dx.doi.org/10.1145/2750858.2807526

[33] Peter J Huber and others. 1964. Robust estimation of a location parameter. *The Annals of Mathematical Statistics* 35, 1 (1964), 73–101.

[34] Maximilian Kerz, Amos Folarin, Nicholas Meyer, Mark Begale, James MacCabe, and Richard J Dobson. 2016. SleepSight: a wearables-based relapse prevention system for schizophrenia. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. ACM, 113–116.

[35] Ron Kohavi and others. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, Vol. 14. Stanford, CA, 1137–1145.

[36] Helen Koo, Ivan Hebrio, Megan Johnston, Nicholas Hosein, and Kris Fallon. 2016. Stresssense: Skin Conductivity Monitoring Garment with a Mobile App. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct (UbiComp '16)*. ACM, New York, NY, USA, 487–490. DOI: http://dx.doi.org/10.1145/2968219.2971348

[37] Alex Kopelowicz, Joseph Ventura, Robert Paul Liberman, and Jim Mintz. 2007. Consistency of Brief Psychiatric Rating Scale factor structure across a broad spectrum of schizophrenia patients. *Psychopathology* 41, 2 (2007), 77–84.

[38] Ai Koyanagi, Andrew Stickley, and Josep Maria Haro. 2015. Psychotic-like experiences and nonsuidical self-injury in England: results from a national survey. *PloS one* 10, 12 (2015), e0145533.

[39] Kurt Kroenke and Robert L Spitzer. 2002. The PHQ-9: a new depression diagnostic and severity measure. *Psychiatric Annals* 32, 9 (2002), 509–515.

[40] Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The Phq-9. *Journal of general internal medicine* 16, 9 (2001), 606–613.

[41] Nicholas D Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew T Campbell. 2010. A survey of mobile phone sensing. *Communications Magazine, IEEE* 48, 9 (2010), 140–150.

[42] Nicholas D Lane, Mashfiqui Mohammod, Mu Lin, Xiaochao Yang, Hong Lu, Shahid Ali, Afsaneh Doryab, Ethan Berke, Tanzeem Choudhury, and Andrew Campbell. 2011. Bewell: A smartphone application to monitor, model and promote wellbeing. In *5th international ICST conference on pervasive computing technologies for healthcare*. 23–26.

[43] Kung-Yee Liang and Scott L Zeger. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* 73, 1 (1986), 13–22. DOI: http://dx.doi.org/10.1093/biomet/73.1.13

[44] Hong Lu, Denise Frauendorfer, Mashfiqui Rabbi, Marianne Schmid Mast, Gokul T Chittaranjan, Andrew T Campbell, Daniel Gatica-Perez, and Tanzeem Choudhury. 2012. Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*. ACM, 351–360.

[45] Naoki Maeda, Yuko Hirabe, Yutaka Arakawa, and Keiichi Yasumoto. 2016. COSMS: Unconscious Stress Monitoring System for Office Worker. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct (UbiComp '16)*. ACM, New York, NY, USA, 329–332. DOI: http://dx.doi.org/10.1145/2968219.2971397

[46] Jörg Sander Martin Ester, Hans-Peter Kriegel and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD '96*. AAAI Press, 226–231.

[47] Alban Maxhuni, Angélica Muñoz-Meléndez, Venet Osmani, Humberto Perez, Oscar Mayora, and Eduardo F Morales. 2016. Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients. *Pervasive and Mobile Computing* (2016).

[48] Abhinav Mehrotra, Robert Hendley, and Mirco Musolesi. 2016. Towards multi-modal anticipatory monitoring of depressive states through the analysis of human-smartphone interaction. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. ACM, 1132–1138.

[49] Venet Osmani. 2015. Smartphones in mental health: detecting depressive and manic episodes. *IEEE Pervasive Computing* 14, 3 (2015), 10–13.

[50] Venet Osmani, Alban Maxhuni, Agnes Grünerbl, Paul Lukowicz, Christian Haring, and Oscar Mayora. 2013. Monitoring activity of patients with bipolar disorder using smart phones. In *Proceedings of International Conference on Advances in Mobile Computing & Multimedia*. ACM, 85.

[51] John E Overall and Donald R Gorham. 1962. The brief psychiatric rating scale. *Psychological reports* 10, 3 (1962), 799–812.

[52] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[53] Mashfiqui Rabbi, Shahid Ali, Tanzeem Choudhury, and Ethan Berke. 2011. Passive and in-situ assessment of mental and physical well-being using mobile sensors. In *Proceedings of the 13th international conference on Ubiquitous computing*. ACM, 385–394.

[54] Matthia Sabatelli, Venet Osmani, Oscar Mayora, Agnes Gruenerbl, and Paul Lukowicz. 2014. Correlation of significant places with self-reported state of bipolar disorder patients. In *Wireless Mobile Communication and Healthcare (Mobihealth), 2014 EAI 4th International Conference on*. IEEE, 116–119.

[55] Sohrab Saeb, Emily G Lattie, Stephen M Schueller, Konrad P Kording, and David C Mohr. 2016. The relationship between mobile phone location sensor data and depressive symptom severity. *PeerJ* 4 (2016), e2537.

[56] Sohrab Saeb, Mi Zhang, Christopher J Karr, Stephen M Schueller, Marya E Corden, Konrad P Kording, and David C Mohr. 2015. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *Journal of medical Internet research* 17, 7 (2015).

[57] Akihide Sano and Rosalind W Picard. 2013. Stress recognition using wearable sensors and mobile phones. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 671–676.

[58] Moushumi Sharmin, Andrew Raij, David Epstien, Inbal Nahum-Shani, J. Gayle Beck, Sudip Vhaduri, Kenzie Preston, and Santosh Kumar. 2015. Visualization of Time-series Sensor Data to Inform the Design of Just-in-time Adaptive Stress Interventions. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. ACM, New York, NY, USA, 505–516. DOI: http://dx.doi.org/10.1145/2750858.2807537

[59] Robert L Spitzer, Kurt Kroenke, Janet BW Williams, Patient Health Questionnaire Primary Care Study Group, and others. 1999. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. *Jama* 282, 18 (1999), 1737–1744.

[60] Marie Stentebjerg-Olesen, Stephen J Ganocy, Robert L Findling, Kiki Chang, Melissa P DelBello, John M Kane, Mauricio Tohen, Pia Jeppesen, and Christoph U Correll. 2015. Early response or nonresponse at week 2 and week 3 predict ultimate response or nonresponse in adolescents with schizophrenia treated with olanzapine: results from a 6-week randomized, placebo-controlled trial. *European child & adolescent psychiatry* 24, 12 (2015), 1485–1496.

[61] Gregory P Strauss, Martin Harrow, Linda S Grossman, and Cherise Rosen. 2010. Periods of recovery in deficit syndrome Schizophrenia: a 20-year multi–follow-up longitudinal study. *Schizophrenia bulletin* 36, 4 (2010), 788–799.

[62] Cheryl D Swofford, John W Kasckow, Geri Scheller-Gilkey, and Lawrence B Inderbitzin. 1996. Substance use: a powerful predictor of relapse in schizophrenia. *Schizophrenia research* 20, 1 (1996), 145–151.

[63] Rui Wang, Min S. H. Aung, Saeed Abdullah, Rachel Brian, Andrew T. Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Michael Merrill, Emily A. Scherer, Vincent W. S. Tseng, and Dror Ben-Zeev. 2016. CrossCheck: Toward Passive Sensing and Detection of Mental Health Changes in People with Schizophrenia. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*. ACM, New York, NY, USA, 886–897. DOI: http://dx.doi.org/10.1145/2971648.2971740

[64] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell. 2014. StudentLife: Assessing Mental Health, Academic Performance and Behavioral Trends of College Students Using Smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '14)*. ACM, New York, NY, USA, 3–14. DOI: http://dx.doi.org/10.1145/2632048.2632054

[65] Gary S Wilkinson and GJ Robertson. 2006. Wide range achievement test (WRAT4). *Psychological Assessment Resources, Lutz* (2006).

[66] Danny Wyatt, Tanzeem Choudhury, and Jeff A Bilmes. 2007. Conversation detection and speaker segmentation in privacy-sensitive situated speech data.. In *INTERSPEECH*. 586–589.

[67] Danny Wyatt, Tanzeem Choudhury, Jeff A Bilmes, and Henry A Kautz. 2007. A Privacy-Sensitive Approach to Modeling Multi-Person Conversations.. In *IJCAI*, Vol. 7. 1769–1775.

[68] Danny Wyatt, Tanzeem Choudhury, and Henry Kautz. 2007. Capturing spontaneous conversation and social dynamics: A privacy-sensitive data collection effort. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, Vol. 4. IEEE, IV–213.

[69] S L Zeger and K Y Liang. 1992. An overview of methods for the analysis of longitudinal data. *Statistics in medicine* 11, 14-15 (Jan 1992), 1825–39. http://www.ncbi.nlm.nih.gov/pubmed/1480876

Received February 2017; revised May 2017; accepted July 2017