CrossCheck: Toward passive sensing and detection of mental health changes in people with schizophrenia

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

Dartmouth College, Cornell University[†], Hofstra Northwell School of Medicine[‡],
{ruiwang, campbell}@cs.dartmouth.edu, {mhauser, jkane2}@northwell.edu
{msa242, sma249, tanzeem.choudhury, mam546, wt262}@cornell.edu
{dror.ben-zeev, rachel.m.brian, emily.a.scherer}@dartmouth.edu

ABSTRACT

Early detection of mental health changes in individuals with serious mental illness is critical for effective intervention. CrossCheck is the first step towards the passive monitoring of mental health indicators in patients with schizophrenia and paves the way towards relapse prediction and early intervention. In this paper, we present initial results from an ongoing randomized control trial, where passive smartphone sensor data is collected from 21 outpatients with schizophrenia recently discharged from hospital over a period ranging from 2-8.5 months. Our results indicate that there are statistically significant associations between automatically tracked behavioral features related to sleep, mobility, conversations, smartphone usage and self-reported indicators of mental health in schizophrenia. Using these features we build inference models capable of accurately predicting aggregated scores of mental health indicators in schizophrenia with a mean error of 7.6% of the score range. Finally, we discuss results on the level of personalization that is needed to account for the known variations within people. We show that by leveraging knowledge from a population with schizophrenia, it is possible to train accurate personalized models that require fewer individual-specific data to quickly adapt to new users.

Author Keywords

Mobile Sensing; Mental Health

ACM Classification Keywords

H.1.2 Models and Principles: User/Machine Systems; J.4 Computer Applications: Social and Behavioral Sciences

INTRODUCTION

Schizophrenia is a severe and complex psychiatric disorder that develops in approximately 1% of the world's population [49]. Although it is a chronic condition, its symptom presentation and associated impairments are not static. Most people with schizophrenia vacillate between periods of relative remission and episodes of symptom exacerbation and relapse. Such changes are often undetected and subsequent interventions are administered at late stages and in some cases after the occurrence of serious negative consequences. It is well understood that observable behavioral precursors can manifest prior to a transition into relapse [2]. However, these precursors can manifest in many different ways. Studies have shown these to include periods of social isolation, depression, stressed interactions, hearing voices, hallucinations, incoherent speech, changes in psychomotor and physical activity and irregularities in sleep [13, 26]. Evidence also suggests that clinical intervention at an early enough stage is effective in the prevention of transitions into a full relapse state. This directly reduces the need for hospitalization and can also lead to faster returns to remission [40].

Existing clinical practices are inefficient in detecting early precursors. Standard methods are based on face to face interactions and assessments with clinicians, conducted at set times and locations. This has major limitations due to a high dependency on patient attendance as well as the resources of clinical centers in terms of time and expertise. Moreover, such assessments have limited ecological validity with a heavy reliance on accurate patient recall of their symptoms and experiences. As such, the data from standard assessments can only be considered as single snapshots rather than a true record of dynamic behavior. This static data does little to inform the robust detection of early warning signs as they emerge longitudinally, especially if there is low adherence to follow-up visits.

To this end, research has begun in the use of mobile devices to achieve more dynamic assessments in schizophrenia [31], though the use of smartphones for this use is still in its infancy. This, in part, is due to the associated risks which necessitated studies to demonstrate feasibility, acceptability and usability within this population. Ben-Zeev et al. developed

[*] contributed equally to this work.

UbiComp '16, September 12-16, 2016, Heidelberg, Germany
© 2016 ACM. ISBN 978-1-4503-4461-6/16/09…$15.00
DOI: http://dx.doi.org/10.1145/2971648.2971740

the FOCUS self management app [6] that provides illness self-management suggestions and interventions in response to participants' rating of their clinical status and functioning. This system received high acceptance rates among users and is shown to be usable by this population [7]. A pilot study in the efficacy of tracking patients [9] over two weeks shows that sensing using smartphones is acceptable to both inpatients and outpatients. These results paves the way for new sensing and inference systems to passively monitor and detect mental health changes using commercially available smartphones.

In this paper, we analyze preliminary data from a randomized control trial of CrossCheck, a smartphone sensing system currently deployed to outpatients with schizophrenia. CrossCheck is the first system to use continuous passive sensing and periodic self-reports to monitor and assess mental health changes in schizophrenia. The ultimate goal of the project is to develop sensing, inference and analysis techniques capable of dynamically assessing mental health changes and predicting the risk of relapse without the need for retrospective recall or self-reports. Another future aim of CrossCheck is to implement new invention techniques to automatically alert clinicians in time to prevent or reduce the severity of relapse. In this paper, we are not directly addressing relapse or intervention, but take a first step towards these goals by investigating: (i) the relationships between passively tracked behavior and self-reported measures, and (ii) how much personalization of the system is required given the observed variability between individual patients.

Specifically, the contributions of this exploratory study are:

- CrossCheck, the first system to use passive sensing data to monitor and predict indicators of mental health for 21 outpatients diagnosed with schizophrenia recently discharged from hospital; CrossCheck monitors these outpatients for periods between 64 and 254 days.
- Meaningful associations between passively tracked data and indicators or dimensions of mental health in people with schizophrenia (e.g., stressed, depressed, calm, hopeful, sleeping well, seeing things, hearing voices, worrying about being harmed) to better understand the behavioral manifestation of these measures and eventually develop a real-time monitoring and relapse prevention system.
- Models that can predict participants' aggregated ecological momentary assessment (EMA) scores that measure several dynamic dimensions of mental health and functioning in people with schizophrenia.
- Level of personalization that is needed to account for the known variations within people. We show that by leveraging knowledge from a population with schizophrenia, it is possible to train personalized models that require fewer individual-specific data to quickly adapt to a new user.

RELATED WORK

There is growing interest in using smartphones to monitor and assess wellbeing and mental health [23]. Smartphones are a natural platform to monitor and assess behavioral patterns that manifest over long periods. Such longitudinal tracking is essential for addressing mental health states that have low frequency changes taking days, weeks or even months [4, 32, 50].

There has been no prior work in the prediction of changes in mental health using passive sensing data from smartphones in schizophrenia. Previous work conducted in populations with depression and bipolar disorder informs our schizophrenia-focused efforts. For depression, early work by [17] uses location, social interaction, activity and mood inferred from a range of sensors to assess depression. Saeb et al. [45] explore the relationships between a wide range of features derived from sensing and show that variation in location as well as phone usage significantly correlates with depressive symptoms. Canzian and Musolei [19] show significant correlations between various measures of mobility derived from location traces with depressive mood. In modeling bipolar disorder, the findings reported in [1] show the automatic inference of circadian stability as a measure to support effective bipolar management. The MONARCA project [41] demonstrate correlations between accelerometer based activity levels over different periods of the day and psychiatric evaluation scores for the mania-depression spectrum. Maxuni et al. [38] add to this by utilizing speech along with activity levels to successfully classify stratified levels of bipolar disorder. For stress detection, [35] detects stress with >0.76 accuracy using acoustic features. Other studies investigate the use of location information [5], measures of social interaction derived from phone-call, SMS, and proximity data [14] to detect stress. In [29, 46, 47], the authors demonstrate using features from both smartphones and wearables to detect and track stress.

The use of smartphone data has also been used to model broader measures of well being over long periods. In [43] the authors demonstrate that speech and conversation occurrences extracted from audio data and physical activity infer mental and social well being. The Studentlife [50] study investigates correlations between conversation, sleep, activity and co-location with a range of wellness scores relating to stress, loneliness, flourishing and depression within the context of a university campus over a single term. This led on from BeWell [33], which inferred sleep, social interaction and activity from smartphones, as a means of promoting wellness.

CROSSCHECK STUDY DESIGN

The CrossCheck study is a randomized control trial (RCT) [20] conducted in collaboration with a large psychiatric hospital in Long Island, NY. The study aims to recruit 150 participants for 12 months using rolling enrollment. The participants are randomized to one of two arms: CrossCheck (n=75) or treatment-as-usual (n=75). The participants from the CrossCheck smartphone arm enrolled to date are the focus of this paper. We report on inferring indicators of mental health and not relapse prediction as there is only a small number of relapses cases (7) observed at present. Given previous data on this type of study population, we expect that at the end of the year long RCT there will be a larger cohort of patients that have experienced relapse to make robust relapse prediction viable.The study has been approved by the Committees for the Protection of Human Subjects at Dartmouth College and Human Services and the Institutional Review Board at Zucker Hillside Hospital. In what follows, we discuss participant recruitment, the sensing system, and the detailed study procedure.

Identifying Participants

The study hospital's Electronic Medical Record is used to identify potential study candidates who are then approached by a staff member to gauge their interest in the study. If interested, a research interview is scheduled. Research flyers are also posted at the study site with the research coordinator's phone number. A candidate is a patient who is 18 or older, met DSM-IV or DSM-V criteria for schizophrenia, schizoaffective disorder or psychosis, 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 [51]. Individuals with a legal guardian are excluded.

Recruiting Participants

The staff at the recruitment hospital first screened candidates based on criteria described in Identifying Participants. Then the staff contacted candidates in person at the study site or by phone to provide a complete description of the study. Interested individuals review the consent form with study staff and are administered a competency screener to verify that they understand what is being asked of them and are able to provide informed consent. After consent, enrolled participants are administered the baseline assessment, then are randomly assigned to CrossCheck or the treatment-as-usual arm where no sensing is done. Participants in the smartphone arm are loaned 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 data has a broad coverage of behaviors, participants personal phone numbers are migrated to the new phone and they are provided with an unlimited data plan for data uploading. 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. As of February 2, 2016, 48 participants are randomized to the CrossCheck arm, with 14 who dropped out. The primary reason for dropping out is due to leaving treatment at the study site. A few participants dropped out due to not being interested in participating anymore. In the 34 remaining, 17 participants are females and 17 are males (11 African American, 2 Asian, 19 Caucasian, 1 Multiracial and 1 did not disclose).

CrossCheck System

The CrossCheck sensing system is built based on our prior sensing work [1, 33, 50] that uses smartphone sensing and self-report tools. Compared with the StudentLife sensing system [50], the CrossCheck app uses the Android activity recognition API instead of the self developed classifier to infer activities. The CrossCheck app collects sensor data continuously and does not require the participant's interaction. The CrossCheck app automatically infers activity (stationary, walking, running, driving, cycling), sleep duration, and sociability (i.e., the number of independent conservations and their durations). The app also collects audio amplitude, accelerometer readings, light sensor readings, location coordinates, and application usages. CrossCheck uses a built in MobileEMA module [50] to administer EMAs [9]. During the collection phase, participants are asked to respond to EMA questions every Monday, Wednesday, and Friday (see CrossCheck Dataset). This paper focuses on the EMA data as symptom measures. CrossCheck is published in Google Play Store's beta testing channel to control access. Google Play Store is used to remotely update the sensing system when necessary. The inferences, the sensor data, and the EMA responses are temporarily stored on the phone and are efficiently uploaded to a secured server when users recharge their phones. Figure 2 gives an overview of the data collection and analysis workflow.

Data collection monitoring. CrossCheck includes management scripts that automatically produce statistics on compliance. It sends a daily report on how many hours of sensor data had been collected for the last few days. The daily report labels participants who have not uploaded any data. CrossCheck also sends out weekly reports with visualizations of participants' sensing data (e.g., distance traveled, sleep and conversation duration) and EMA responses for the most recent week. Daily reports and weekly reports help researchers to identify participants who are collecting data or are having problems with the system. Research staff would call noncompliant participants to give assistance and get them back on track.

Privacy considerations. In order to protect participants' personal information, each participant is given a random study ID. Any identifiable information is stored securely in locked cabinets and secured servers. The participant's personal information, such as phone number and email address, is not collected by the sensing app. Participants' data is uploaded to a secured server using encrypted SSL connections. If a participant's phone is lost we remotely erase the data on the phone and reset it.

CROSSCHECK DATASET

The dataset includes behavioral features and inferences from raw sensor data, EMA responses, and combined indicator scores calculated from EMA responses. We select behavioral features based on participants' behaviors (e.g., physical activity, sociability, sleep , mobility) that are associated with dimensions of mental health state [1,19,33,38,41,43,45,50]. We use self-reported EMA data as mental health state indicators of schizophrenia patients.

Timescale and Epochs

Behavioral features are computed on a daily basis. For example, the daily conversation frequency is the number of conversations a participant is around over a 24-hour period. In addition, a day is partitioned evenly into four epochs: 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), we also compute behavioral features for these four epochs to explore behavioral patterns within different phases in a day.

Behavioral Sensing Features

A wide range of behavioral sensing features from the raw sensor data and behavioral inferences are collected by the CrossCheck app. These features describe patterns of participants' physical activity, sociability, mobility, phone usage, sleep, and the characteristics of the ambient environment in which the participant dwells. Below, we discuss these features and the rationale behind using them for our analysis.

Activity. We use the Android activity recognition API that includes: on foot, still, in vehicle, on bicycle, tilting, and unknown. CrossCheck gives an activity update every 10 seconds when the user is moving, or every 30 minutes when the user is stationary. We compute the durations of stationary state and walking states per day and within each of the four epochs as physical activity features. Our scale evaluation shows that the Android activity recognition API infers walking and stationary with 95% accuracy.

Speech and conversation. Previous studies [33, 43, 50] have shown that the detection of conservations and human voice is related to wellness and mental health. We compute the number and duration of detected conversational episodes per day and over each of the four epochs. We also compute the number of occurrences of human voice and non human voice along with their respective durations per day.

Calls and SMS. To further inform the level of social interaction and communication we consider phone calls and SMS activities. We compute the number and duration of incoming and outgoing calls over a day and the number of incoming and outgoing SMS.

Sleep. Changes in sleep pattern or the onset of unusual sleep behavior may indicate changes in mental health [13]. Sleep related features that are derived from the sleep inferences are: overall duration of sleep, going to sleep time, and wake time for each day [21, 50].

Location. Prior studies have shown that a user's mobility patterns from geo-location traces are associated with mental health and wellness [19, 45, 50]. In schizophrenia, for example, it is not uncommon for people to be isolated and stay at home with little external contact especially when individuals are experiencing distressing psychotic symptoms. We calculate the following set of location features on a daily basis: total distance traveled, maximum distance travelled between two tracked points, maximum displacement from the home, standard deviation of distances, location entropy, duration of time spent at primary location, duration of time spent at secondary location. Finally, we compute a locational routine index over seven days to quantify the degree of repetition in terms of places visited with respect to the time of day over a specific period of time. These features stem from the works on depression in [19, 45]. Further we propose the number of new places visited in a day by using the number of new locations in a day that have not been seen previously. Sampled location readings/coordinates are clustered in to primary, secondary or other location using the DBSCAN clustering method [37] with a minimum of ten points per cluster and a minimum cluster radius of ten meters over the entirety of a single user's data. The first and second largest clusters are labeled as the primary and secondary locations, respectively.

Phone and app usage. User interaction with the phone is potentially indicative of general daily function. For a coarse measure, we compute the number of times the phone is unlocked per day, as well as the duration in which the phone is unlocked per day and within each of the four epochs. We also create more nuanced measures by leveraging information about the types of apps that are running. Given the wide variety of apps, we classify each app into one of the three broad categories: social, engagement, and entertainment. These categories were chosen as they are indicative of sociability and daily function which in turn may potentially be indicative of mental health changes. We use the meta-information from Google Play's categorizations and bin all active apps into one of the three categories. The social category is a combination of social and communication apps, examples include Facebook and Twitter. The engagement category consists of health & fitness, medical, productivity, transportation and finance apps, examples include Calendar and Runkeeper. The entertainment category consists of news & magazines, media & video, music & audio, and entertainment apps. Examples of apps in this category are YouTube and NetFlix. We compute the total number of apps that belong to each of these three categories every 15 minutes from the process stack. We then calculate the increases in the number of apps that belong to each category which is indicative of how often the participant launches an app in one of the categories.

Ambient environment. We compute features to measure the ambient sound and light environment. The mean levels of ambient volume per day and within four epochs reflect the ambient context of the participant's acoustic environment, for example quiet isolated places versus noisy busy places. Similarly, we consider the ambient light levels to get more information about the environmental context of the participant, for example dark environment versus well illuminated environment. We acknowledge that the phone cannot detect the ambient light when in the pocket. However, we found that the phone can opportunistically sense the ambient light environment that can be used to help infer sleep [21]. We use the mean illumination over a day and within the four epochs.

Ecological Momentary Assessments

There are several dynamic dimensions of mental health and functioning in people with schizophrenia that are of interest. These include items such as visual and auditory hallucinations, incoherent speech delusion, social dysfunction or withdrawal, disorganized behavior, and inappropriate affect [3]. Other possible indicators of changes in mental health include variations in sleep, depressive mood and stress. EMA has shown to be a valid approach to capture mental health states amongst people with schizophrenia [27]. The set of EMA questions we use in CrossCheck are based on self-reported dimensions defined in previous schizophrenia research [8]. The EMA has 10 questions, which can be grouped into two categories: positive item questions and negative item questions. Higher score in positive questions indicates better outcomes whereas higher scores in negative item questions indicates worse outcomes. Positive questions ask a participant if they have been feeling calm, been social, been sleeping well, been able to think clearly, and been hopeful about the future. Negative questions ask a participant if they have been depressed, been feeling stressed, been bothered by voices, been seeing things other people can't see, and been worried about being harmed by other people. The questions are framed as simple one sentence questions with a 0-3 multiple choice answers (for specific phrasing see Table 1). The MobileEMA user interface is designed to be simple and easy to use. It shows the questions one by one. The participant responds to the question by touching a big button associated with their response.

We calculate the EMA negative score, positive score, and sum score from the responses. The EMA positive score is the sum of all positive questions' score, the negative score is the sum of all negative questions' score, and the sum score is the positive score minus the negative score. The positive and negative score range from 0 to 15 and the sum score ranges from -15 and 15.

Table 1: EMA questions related indicators of mental health

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.

ANALYSIS AND RESULTS

We identify a number of important associations between phone-based behavioral features described in CrossCheck Dataset and dynamic dimensions of mental health and functioning in terms of EMA scores (e.g., feeling depressed, hearing voices or thinking clearly). Also in this section, we present results on the use of predictive models on aggregated EMA scores. We test the level of personalization needed for accurate modeling and for predicting longer term underlying trends in the scores.

Methods overview

We first run bivariate regression analysis to understand associations between the measures of interest in schizophrenia from the EMA scores and passively tracked behavioral features. The regression results are presented in Bivariate Regression Analysis. We then run prediction analysis using Gradient Boosted Regression Trees (GBRT) [25, 42] to evaluate the feasibility of predicting EMA sum scores, which is discussed in Prediction Analysis. Finally, we generate person specific models using Random Forest (RF) [15] to gain insight into predicting smoothed EMA sum scores that characterize underlying trends.

Data cleaning. Given that our analysis is based on data that are aggregated over a day (e.g., distance traveled during a day), missing data during a day would skew derived values and may misrepresent behavior. Therefore, the proportion of three forms of continuously sampled data (activity, location, and audio) are used to determine how many hours of data is sensed in a day. Days with fewer than 19 hours of sensing data are discarded. Since recruitment of outpatients and data collection is an ongoing process, participants join the study at different times leading to varying amounts of data. We include participants who have been in the study for longer periods and are compliant when answering EMAs. Specifically, we select participants who have more than 60 days of sensor data as of February 2nd 2016 and completed at least 50% of the EMAs. 21 out of 34 participants in the CrossCheck arm of the RCT satisfy this criteria. As a result we analyze 2809 days of sensing data and 1778 EMA responses for 21 participants. All participants are in the study for a minimum of 64 days. The total number of days ranges from 64 to 254 days. On average, each participant in the study provides 133.76 days (19 weeks) of sensing data and 84.7 EMA responses.

Data preparation. Given that the EMA module launches a set of questions every 2-3 days, we aggregate the sensed data from the days within this interval by taking the mean. Figure 1 shows the daily data aggregation strategy used to predict EMA scores. For example, if a participant gave EMA responses on day 3, 6, and 9, we compute the mean of each feature data (e.g., the mean sleep duration and the mean distanced traveled) from day 1 to 3 to predict the EMA score at day 3, the mean from day 4 to 6 to predict the EMA score at day 6, and the mean from day 7 to 9 to predict the EMA on day 9.

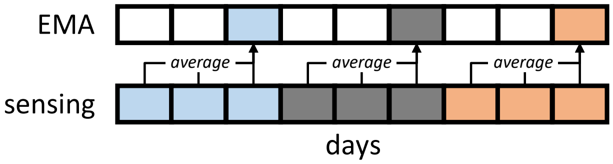


Figure 1: Feature/EMA preparation

Feature Space Visualization

To gain an insight into the feature space, the data from all participants is mapped using the t-Distributed Stochastic Neighbor Embedding (t-SNE) [36] method. The t-SNE [36] is an emerging technique for dimensionality reduction that is particularly well suited to visualize high-dimensional datasets. It projects each high-dimensional data point to a two-dimensional point such that similar data points in the high-dimensional space are projected to nearby points in the two-dimensional space and dissimilar data points are projected to distant points. The feature visualization is shown in Figure 3.

Figure 3(a) shows the mapped features on a two-dimensional space. Each data point represents a subject's behavioral features used to predict EMA responses. We observe data points are grouped into different clusters. By color-coding each point per participant, it can be clearly seen that each cluster is predominantly participant specific. This important finding

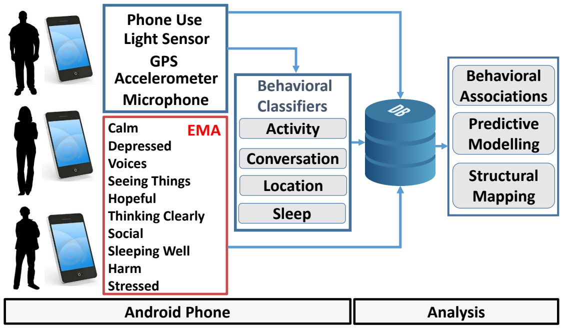


Figure 2: CrossCheck sensing and analysis system.

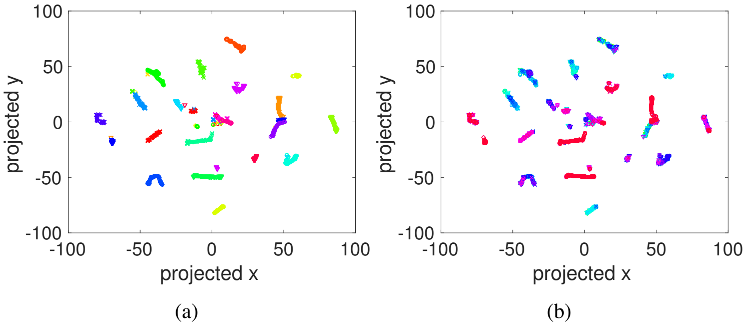


Figure 3: Feature visualization using t-SNE. (a) Data is color coded by user ID. Individual subject's data clusters together. (b) Data is color coded by EMA sum scores. Data with same score tend to cluster within subject.

is interesting because it shows that our features captures behavioral difference between different individuals and that the data is highly person dependent. Figure 3(b) shows a further color coding of the data; this time by EMA sum scores. In this case, the colors are intermixed. However, we observe that data points associated with the same score are also clustered together, though the purity of such clusters are not as high as shown in Figure 3(a). This observation gives us confidence in predicting participants' EMA sum scores using personalized models. These insights govern the analysis discussed in the remainder of this section.

Bivariate Regression Analysis

Standard statistical analysis methods such as correlation analysis and ordinary regression analysis assume independence between observations. However, our longitudinal dataset violates this independence assumption: data from the same subject are likely to be correlated. Models that do not account for intra-subject correlations can lead to misleading results. To addrress this, we apply generalized estimating equations (GEE) [18, 22, 34, 52] – a model specifically designed to analyze longitudinal datasets – to determine associations between each of the features and their EMA responses.

The GEE method is a marginal model, in which the regression and within-subject correlation are modeled separately. The marginal expectation of subject i's response Y_{it} at time t is $E(Y_{it}) = \mu_{it}$. This is related to the features x_{it} by function $g(\mu_{it}) = \beta_0 + \beta x_{it}$, where g is a link function. From initial inspection we assume the EMA responses have Poisson distributions leading to the use of $\log$ as the link function. The β coefficients corresponding to feature vector x_{it}, which indicates the association between the features and the outcome Y_{it}, where β_0 is the intercept. The p-value associated with each β indicates the probability of the feature coefficient β being zero (i.e., the feature does not associate with the outcome). In addition, GEE does not rely on strict assumptions about distribution and is robust to deviation from assumed distribution. The GEE analysis describes differences in the mean of the response variable Y across the population, which is informative from the population perspective.

The resultant β values indicate the direction and strength of the association between a behavioral feature and an EMA score. A unit increase in the feature value is associated with e^{β} increase in the associated EMA value. To allow for interperson comparability, each feature is normalized per participant to a zero mean with one standard deviation. Therefore, the resultant features values are indicative of feature deviation from the mean. A positive β indicates that a greater feature value is associated with a greater EMA score, whereas a negative β indicates that a greater feature value is associated with a smaller EMA score. The most significant β values are selected using the corresponding p-value from each feature-EMA combination.

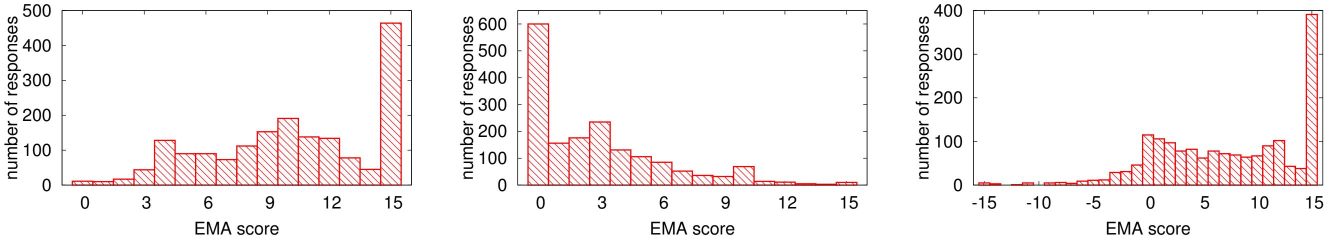
We apply a bivariate regression using GEE to all 610 combinations of the 61 features and 10 EMA questions. We apply the Benjamini-Hochberg procedure (BH) proposed in [10, 11] to inform the false discovery rate (FDR) in our exploratory regression analysis. The BH procedure finds a threshold for the p value given the target false discovery rate by exploring the distribution of the p-values. We find 88 regressions with $p < 0.05$, which corresponds to FDR $< 32.8\%$, meaning associations with $p < 0.05$ has at most 32.8% chance of being false discoveries. We find 12 regressions with $p < 0.0016$, FDR $<$ 0.1, and 7 regressions with $p < 0.00025$, FDR < 0.05.

Table 2: Positive questions regression results

EMA item	associated behavior
calm	sleep end time (-), conversation number (-), conversation number afternoon (-), conversation number night (-), call in (-), call out (-), increase in entertainment app use (-), ambient light afternoon (-), ambient sound volume night (-)
hopeful	**call out (-)**, **call out duration (-)**, sms in (-), sms out (-)
sleeping	conversation duration evening (-), conversation number evening (-), ambient sound volume morning (-)
social	walk duration evening (-), sleep duration (-), sleep end time (-), ambient light evening (-)
think	conversation duration night (-), call in (-), call in duration (-), call out (-), sms in (-), increase in entertainment app use (-), durations of non-voice sounds (-), number of non-voice sounds (-), number of voice sounds (-)

(-):negative association, (+):positive association
all associations with $p < 0.05$.
FDR < 0.1 in **bold** and FDR < 0.05 in ***bold italic***.

Positive Questions. Table 2 shows features that are associated with the five positively worded questions (*viz. calm*, *social*, *thinking clearly*, *sleeping well* and *hopeful*). A higher



(a) Positive score distribution.

(b) Negative score distribution.

(c) Sum scores distribution.

Figure 4: EMA aggregated score distributions

score indicates a more positive mental health state. The reported associations' feature β values are within $-0.04 < \beta < -0.02$ with $p < 0.05$. We find in general, higher scores in positive questions are associated with waking up earlier, having fewer conversations, fewer phone calls, and fewer SMS. Specifically, higher *calm* scores are associated with fewer number of conversations, fewer phone calls, and staying in quieter environment at night and darker environment in the afternoon. Higher *hopeful* scores are associated with making fewer phone calls, and sending and receiving fewer SMS. Higher *sleeping well* scores are associated with fewer conversations, and staying in quieter environment in the morning. Higher *social* scores are associated with walking less in the evening, sleeping less, waking up earlier, and staying in darker environment in the evening. Finally, higher ability to *think clearly* is associated with fewer conversations at night, having fewer calls and SMS, and using fewer entertainment apps.

Negative Questions. Table 3 shows features that are associated with the five negatively worded questions (*viz. hearing voices*, *seeing things*, *stress*, *harm* and *depressed*). A higher score indicates a more negative mental health state. The reported associations' feature β values are within $-0.22 < \beta < 0.2$ with $p < 0.05$. We find in general, higher scores in negative questions are associated with staying stationary more in the morning but less in the evening, visiting fewer new places, being around fewer conversations but making more phone calls and SMS, and using the phone less. In addition, we find higher *depressed* scores are associated with using the phone less in the morning; higher *harmed* scores are associated with using fewer engagement apps; higher hearing *voices* scores are associated with staying in quieter environments, especially in the morning period.

Prediction Analysis

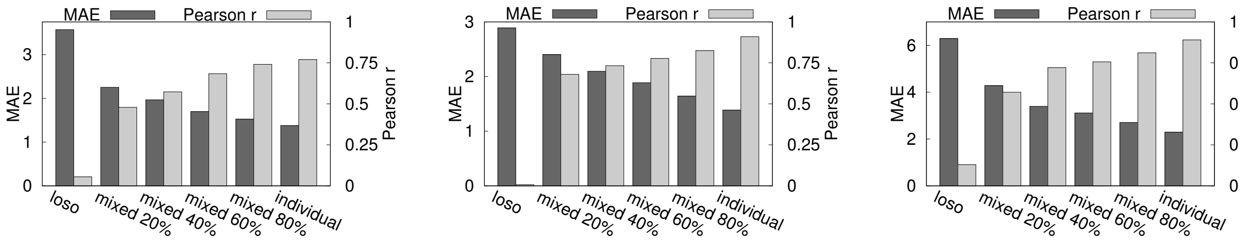
In this section, we discuss two supervised learning schemes for predicting aggregated EMA scores. The first scheme explores the level of personal data needed for accurate prediction. We use different training sets with various proportions taken from one participant of interest along with instances taken from the general population, we then test the model on the scores of the said participant. The second scheme is a further analysis on a set of wholly personalized models to test the difference in predicting smoothed versus raw aggregated EMA and the effect on accuracy by varying temporal proximity between training and testing data. The distribution of EMA positive scores, negative scores, and sum scores are shown in Figure 4.

Personalized EMA Predictions

Predicting the aggregate EMA scores is a regression task. We use Gradient Boosted Regression Trees (GBRT) [25,42] to predict EMA scores. GBRT is an ensemble method which 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. More formally, GBRT is an additive model with the following form [42]: $F(x) = \sum_{m=1}^{M} \gamma_m h_m(x)$, where $h_m(x)$ are the basis functions and γ_m are the step length for gradient decent. Building the additive model can be viewed as gradient descent by adding $h_m(x)$. This addition is based on a forward stagewise fashion where the model at stage m is $F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$. The additional term $\gamma_m h_m(x)$ is determined by solving $F_m(x) = F_{m-1}(x) + \underset{h}{\operatorname{argmin}} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) - h(x))$, where L is the Huber loss [25,30] also GBRT is less sensitive to outliers [30].

Ideally, an EMA score prediction system should be able to predict a new user's scores accurately. However, the visualization of participants' data (Figure 3) shows that there are clear separations between different subjects' behavioral data. Therefore, a certain level of model personalization is needed. We personalize a predictive model by training the model with the subject's data. In order to understand the effectiveness of the model personalization, we train three models with different training data setups to predict each of the three aggregate EMA scores: leave-one-subject-out models, mixed models, and individual models.

A *leave-one-subject-out model* (LOSO) is trained to predict a particular subject's EMA scores. The model is trained on the data from other study participants with the subject's data left out. This model emulates a new unseen user starting to use the system that has learned on data from other people. A *mixed model* personalizes the training data by introducing a small amount of the subject's data to a larger population data. The idea is to leverage knowledge from the population to help training so fewer examples of the subject's data are needed. Specifically, we train a model for a particular subject with data from the population plus some data from the subject. We want to understand how much data from a subject is needed to train an accurate model. We test models with dif-



(a) Predicting positive scores.

(b) Predicting negative scores.

(c) Predicting sum scores.

Figure 5: EMA aggregated scores prediction MAE and Pearson r. loso: leave-one-subject-out model, mixed: mixed model, individual: individual model. The results show that the model without personalization does not work. The prediction performance improves as more data from the subject is included in the training set.

Table 3: Negative questions regression results

EMA item	associated behavior
depressed	still duration morning (+), walk duration (-), walk duration morning (-), sleep start time (+), new places visited (-), call in duration (+), **call out (+)**, call out duration (+), sms in (+), sms out (+), unlock duration morning (-)
harm	still duration morning (+), walk duration (-), walk duration night (-), walk duration morning (-), walk duration evening (+), sleep start time (+), new places visited (-), conversation duration morning (-), call in (+), call in duration (+), call out (+), number of non-voice sounds (+), number of voice sounds (+), unlock duration (-), unlock duration morning (-), unlock duration afternoon (-), increase in engagement app use (-)
seeing things	still duration evening (-), walk duration evening (+), walk duration morning (-), ***sleep start time (+)***, conversation duration morning (-), call in duration (+), call out (+), ***number of non-voice sounds (+)***, ***number of voice sounds (+)***, unlock duration (-), unlock duration afternoon (-), unlock duration evening (-)
stressed	***still duration morning (+)***, ***walk duration morning (-)***, sleep start time (+), conversation duration afternoon (-), conversation duration morning (-), call in duration (+), call out duration (+), unlock duration morning (-)
voices	still duration morning (+), walk duration night (-), sleep start time (+), **new places visited (-)**, conversation duration morning (-), call in (+), ***call in duration (+)***, ***unlock duration afternoon (-)***, **unlock duration morning (-)**, ambient sound volume (-), ambient sound volume morning (-)

(-):negative association, (+):positive association
all associations with $p < 0.05$.
FDR < 0.1 in **bold** and FDR < 0.05 in ***bold italic***.

ferent amount of data from a subject while keeping the population data fixed. Specifically, we use 20%, 40%, 60%, and 80% of a subject's data plus the population data to train and evaluate four models. This model emulates a system making predictions for a new user by leveraging knowledge from the population plus a small amount of the subject's behavioral patterns. Please note, the leave-one-subject-out model is a special case of the mixed model, where we use 0% of a subject's data for training. An *individual model* is a fully personalized model, which uses data from only the subject to train the model.

We use a 10-fold blocked cross validation method [12, 16, 28, 48] to evaluate the prediction performance of the individual models and mixed models. We define a block as a temporally continuous segment of the data. This ensures that test data stems from a different block of time to those in the training data. Moreover, for additional rigour, we also omit boundary instances in the training set that are temporally close to the test set based on the h-block cross-validation as proposed in [16], which was designed to evaluate time dependent observations. Training instances that are less than or equal to h time points from the test block are not used in training. This ensures that temporally the test instances are always at least h time points from instances used in the training set.

To evaluate the individual model, we use $n - 1$ blocks as the training set and the remaining block as the test set. As stated, we remove h observations in the training set preceding and following the observation in the test. In order to make use of all the data, we iteratively select each block for testing, as suggested in [12]. As the data collection is ongoing, there are different amounts of data from each subject leading to different sized test sets for different subjects. The number of observations in the testing set ranges from 5 to 13 with median of 9. We choose $h = 6$ for our cross-validation (i.e., 2 weeks of data because we administer 3 EMAs a week). The value of 6 for h is used as it is ~50% of the block size of the subject with the most data.

For the mixed models, we use the same h-block cross validation method. The mixed-model's training data has two parts: the population data and the subject's data. The population data does not contain any data from the subject and is the same for all folds. The training data from the target subject follows the similar h-block cross validation principle as in the individual model. Again, we test using 20%, 40%, 60%, and 80% of the data from the subject (i.e., 2 blocks, 4 blocks, 6 blocks, and 8 blocks) plus the population data for training. We test on the rest of the subject's data. Similar to the individual model, the training and test data are from time-continuous blocks and $h = 6$ observations are removed from the subject's training data that are at either side of the test data. For every fold, we shift the training data from the subject 1 block forward, and test on the rest. For example, if we run cross-validation with 20% from the subject, we first train the model with block 1 and 2 plus the population data, and test on blocks 3 to 10. In the second fold, we train the model with block 2 and 3 plus the population data, and test on block 1 and blocks 4 to 10.

Prediction performances. Figure 5 shows the mean absolute error (MAE), and the Pearson's r for all models predicting EMA positive, negative, and sum scores. For the positive scores, we get the best prediction performance from the *individual model*, where MAE $= 1.378$. The prediction strongly correlates with the outcome with $r = 0.77$ and $p < 0.001$. We get the worst prediction performance from the *leave-one-subject-out model*, where MAE $= 3.573$ and the predicted scores do not correlate with the ground-truth. This supports our observation from Figure 3 for the need for personalization in building the model. In *mixed models*, we see consistent prediction performance improvement as we include more data from the subject in the training set. With 20% of the subject's data as the training data plus the population data, the MAE of the mixed model is reduced to 2.254 comparing with the LOSO model. The predicted scores correlate with the ground-truth with $r = 0.479$ and $p < 0.001$. The MAE further reduces and the predicted scores are more correlated with ground-truth as we use more data from the subject for training. With 80% of the data from the subject as training data, the MAE drops to 1.525.

This same trend occurs with the negative scores and the sum score (Figure 5b), the LOSO models are not predictive. However, the negative score mixed models trained with 20% of an individual's data starts to be able to make predictions with MAE $= 2.401$, $r = 0.680$, and $p < 0.001$. The prediction performance steadily improves as we use more data from the subject for training. The individual model achieves the best prediction performance with MAE $= 1.383$, $r = 0.856$, and $p < 0.001$.

Please note that the EMA sum score has a larger scale than the positive score and the negative score, where the sum score ranges from -15 to 15 and the positive and negative scores range from 0 to 15. By taking the different score scales into consideration, we find that the individual model predicts the sum score (MAE $\times\ 0.5 = 1.15$) more accurately than the positive score (MAE $= 1.378$) and negative scores (MAE $= 1.383$). We suspect that the sum score better captures individuals' mental health state in general. Again, the results from mixed models show that including 20% of the subject's data in the training set bolsters performance and the prediction performance steadily improves as more data from the subject is used.

Our results show that model personalization is required to build EMA score prediction systems. With small amount of training data from the subject (20%) plus the population's data we can make relevant EMA predictions that are correlated with the ground truth. Therefore, we can quickly build an EMA prediction model for a new user when we do not have much data from them. The predictions would be more accurate as more data from the subject becomes available. *These results provide confidence that our ultimate goal of building a schizophrenia relapse prediction systems is likely feasible.*

Relative feature importance. We examine which features are relatively more important in predicting EMA positive, negative, and sum scores. In GBRT models, this is calculated 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. We average the feature importance across all individual models to find the top-10 most important features for predicting the EMA positive, negative, and sum scores, as shown in Table 4.

Compared with the regression analysis results, we find that four of the top-10 features (i.e., durations of non-voice sounds, walk duration evening, call in duration, and ambient sound volume night) to predict the positive score are associated with positive EMA items. To predict the negative score, six of the top-10 features (i.e., sleep start time, walk duration morning, conversation duration morning, call out duration, call in, and call in duration) that are associated with negative EMA items. For the sum score, two of the top-10 important features (i.e., ambient sound volume afternoon and ambient light night) are not associated with any EMA items. We also observe that epoch behavioral features are more important than corresponding daily features. For example, the predictive models find conversational features during the morning is more predictive than daily conversational features. This supports our initial decision to divide the day into 4 equal epochs to explore the data. We suspect that epoch features better capture behavioral changes when an individual experiences changes in mental health state.

Table 4: Feature importance

	top-10 important features
positive score	durations of non-voice sounds, ambient light night, unlock duration night, walk duration evening, sleep start time, call in duration, ambient sound volume night, walk duration, location entropy, duration at primary location
negative score	sleep start time, call out duration, max dist travelled btwn 2 location points, ambient light morning, unlock number, call in, call in duration, walk duration morning, stdev of distances travelled, conversation number morning
sum score	call out duration, ambient sound volume afternoon, walk duration, conversation number morning, unlock duration evening, sleep start time, durations of non-voice sounds, call in, ambient light night, call in duration

Predicting Underlying EMA Trends

In this section, we investigate the prediction of underlying trends in the EMA score specific to each participant. Figure 6 shows lower frequency trends in the aggregated EMA score which are especially apparent for outpatients who are in the study for longer durations. To extract these underlying trends we apply a Savitzky-Golay filter (with polynomial order of 2) to the sum EMA score only. Smoothing is not applied to the feature values. Compared with other adjacent averaging techniques, this method better preserves the signal's characteristics (e.g., relative maxima, minima and width). For prediction, we train a set of Random Forest regression (RF) [15] models. Training is done using person specific data to generate a set of individual models. We consider data points that are temporally closer would be more similar to each other than data points taken further in time. We also consider that such temporal dependencies to be personalised, hence the use of individual models only in this experiment. For example, the amount of staying at home in cold months may be high and may decrease as months get warmer, however the rate of this change will be dependent on each person's circumstances. Similar to the evaluation in the previous section, we evaluate the models using a time blocked cross validation approach. We set the block size to be a variable interval length in terms of multiples of training instances m, this can be interpreted in real terms since a unit m spans 2-3 days.

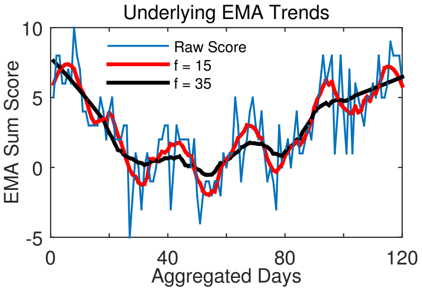


Figure 6: Examples of smoothing on EMA sum score from one participant where f is the frame size of the Savitzky-Golay filter.

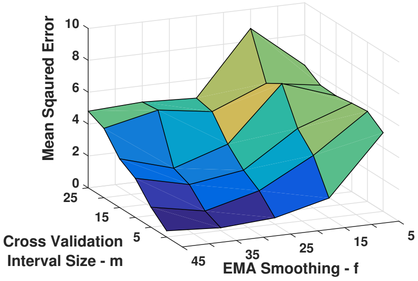


Figure 7: Mean Squared Error from Leave-One-Interval-Out validation for interval sizes versus smoothing level.

We implement a grid search between different levels of smoothing (i.e., the Savitzky-Golay frame size parameter) and different time interval sizes which we will call the leave-one-interval-out validation. We choose the Savitzky-Golay frame size parameter f as one of $\{5, 15, 25, .., 45\}$ and the time interval sizes m as one of $\{1, 5, 10, .., 25\}$. We train models with different f and m combinations, and evaluate their prediction performance using Mean Squared Error (MSE). Figure 7 shows an example of the MSE of a model trained on one participant's data. The MSE is taken from the leave-one-interval-out validation. It can be seen that where m is smaller the MSE is better, demonstrating that smaller intervals which contain data that is closer in time between the training and test sets leads better to MSE scores, but as m increases the MSE score gets worse. However, the grid search also reveals that smoothing the target score has the effect of countering this limitation. This is due to the model predicting a more stable underlying trend which is more predictable. For example in Figure 7 a smoothed outcome with $f = 45$ and $m = 25$ has a similar MSE to a model at $f = 5$ and $m = 1$. This can be interpreted as: if the interval is 3 days long (time between EMA scores), a model for a smoothed score ($f = 45$) trained on data up to 75 days ago (25 x 3) is as good as a model for an non-smoothed score ($f = 5$) trained on data up to 3 days ago. Within the personal models we find that additive increases in the smoothing parameter f by 10 increases the time span within which the tracked data is relevant and predictive by 10-15 days.

DISCUSSION AND CONCLUDING REMARKS

CrossCheck is the first system to use passive sensing data from smartphones to find significant associations with mental health indicators and to accurately predict mental health functioning in people with schizophrenia. We find lower levels of physical activity are associated with negative mental health, which is consistent with previous work [24]. In terms of sociability, our results show that patients around fewer conversations during the morning and afternoon periods are more likely to exhibit negative feelings. However, we also find participants who make more phone calls and send more SMS messages also have significant associations with negative dimensions of mental health. This may suggest that the participants prefer to use the phone instead of face-to-face communication when exhibiting a negative mental state. In terms of locations, our findings show that outpatients are likely to visit fewer new places when in a negative state. Our "new places visited" measure adds to the emerging knowledge in the use of location data for mental well being [19, 45]. For sleep, getting up earlier is associated with positive mental health, whereas going to bed later is associated with negative feelings; this also relates to a promising new direction in considering a person's chronotype and changes in sleep rhythm [44] for mental health assessment. However, we would like to note that we do not yet understand the cause and effect of these associations.

The predicted mental health indicators (i.e., aggregated EMA scores) strongly correlates with ground-truth, with $r = 0.89, p < 0.001$ and MAE $= 2.29$. We also find that by leveraging data from a population with schizophrenia it is possible to train personalized models that require fewer individual-specific data thereby adapting quickly to new users. The predictive power of participants' data decreases when temporally more distant data are included in the training of the models. However, this can be countered by predicting underlying lower frequency trends instead.

CrossCheck shows significant promise in using smartphones to predict changes in the mental health of outpatients with schizophrenia. We believe that CrossCheck paves the way toward real-time passive monitoring, assessment and intervention systems. This would include models capable of predicting the mental health outcomes discussed in this paper but also the detection of impending relapse. Finally, although the participants in the CrossCheck study are drawn from a population with schizophrenia, we firmly believe that our app, methods, and findings are relevant to the emerging field of *mHealth for mental health* [39].

ACKNOWLEDGEMENTS

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

REFERENCES

1. Abdullah, S., Matthews, M., Frank, E., Doherty, G., Gay, G., and Choudhury, T. Automatic detection of social rhythms in bipolar disorder. *Journal of the American Medical Informatics Association* (2016), 538–543.
2. Ascher-Svanum, H., Zhu, B., Faries, D. E., Salkever, D., Slade, E. P., Peng, X., and Conley, R. R. The cost of relapse and the predictors of relapse in the treatment of schizophrenia. *BMC psychiatry 10*, 1 (2010), 2.
3. Association, A. P., et al. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.
4. Aung, M., Alquaddoomi, F., Hsieh, C.-K., Rabbi, M., Yang, L., Pollak, J., Estrin, D., and Choudhury, T. 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.
5. Bauer, G., and Lukowicz, P. Can smartphones detect stress-related changes in the behaviour of individuals? In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2012 IEEE International Conference on* (2012), IEEE, pp. 423–426.
6. Ben-Zeev, D., Brenner, C. J., Begale, M., Duffecy, J., Mohr, D. C., and Mueser, K. T. Feasibility, acceptability, and preliminary efficacy of a smartphone intervention for schizophrenia. *Schizophrenia bulletin* (2014), sbu033.
7. Ben-Zeev, D., Kaiser, S. M., Brenner, C. J., Begale, M., Duffecy, J., and Mohr, D. C. Development and usability testing of focus: A smartphone system for self-management of schizophrenia. *Psychiatric rehabilitation journal 36*, 4 (2013), 289.
8. Ben-Zeev, D., McHugo, G. J., Xie, H., Dobbins, K., and Young, M. A. 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.
9. Ben-Zeev, D., Wang, R., Abdullah, S., Brian, R., Scherer, E. A., Mistler, L. A., Hauser, M., Kane, J. M., Campbell, A., and Choudhury, T. Mobile behavioral sensing for outpatients and inpatients with schizophrenia. *Psychiatric Services* (2015).
10. Benjamini, Y., and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)* (1995), 289–300.
11. Benjamini, Y., and Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics* (2001), 1165–1188.
12. Bergmeir, C., and Benítez, J. M. On the use of cross-validation for time series predictor evaluation. *Information Sciences 191* (2012), 192–213.
13. Birchwood, M., Spencer, E., and McGovern, D. Schizophrenia: early warning signs. *Advances in Psychiatric Treatment 6*, 2 (2000), 93–101.
14. Bogomolov, A., Lepri, B., Ferron, M., Pianesi, F., and Pentland, A. S. Daily stress recognition from mobile phone data, weather conditions and individual traits. In *Proceedings of the ACM international conference on multimedia* (2014), ACM, pp. 477–486.
15. Breiman, L. Random forests. *Machine learning 45*, 1 (2001), 5–32.
16. BURMAN, P., CHOW, E., and NOLAN, D. A cross-validatory method for dependent data. *Biometrika 81*, 2 (jun 1994), 351–358.
17. Burns, M. N., Begale, M., Duffecy, J., Gergle, D., Karr, C. J., Giangrande, E., and Mohr, D. C. Harnessing context sensing to develop a mobile intervention for depression. *Journal of medical Internet research 13*, 3 (2011), e55.
18. Burton, P., Gurrin, L., and Sly, P. 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.
19. Canzian, L., and Musolesi, M. 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* (2015), ACM, pp. 1293–1304.
20. Chalmers, T. C., Smith, H., Blackburn, B., Silverman, B., Schroeder, B., Reitman, D., and Ambroz, A. A method for assessing the quality of a randomized control trial. *Controlled clinical trials 2*, 1 (1981), 31–49.
21. Chen, Z., Lin, M., Chen, F., Lane, N. D., Cardone, G., Wang, R., Li, T., Chen, Y., Choudhury, T., and Campbell, A. T. Unobtrusive sleep monitoring using smartphones. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2013 7th International Conference on* (2013), IEEE, pp. 145–152.
22. Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S. *Analysis of longitudinal data*. OUP Oxford, 2013.
23. Donker, T., Petrie, K., Proudfoot, J., Clarke, J., Birch, M.-R., and Christensen, H. Smartphones for smarter delivery of mental health programs: a systematic review. *Journal of medical Internet research 15*, 11 (2013), e247.
24. Fox, K. R. The influence of physical activity on mental well-being. *Public health nutrition 2*, 3a (1999), 411–418.
25. Friedman, J. H. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics 29*, 5 (2001), 1189–1232.
26. Gleeson, J. F., Rawlings, D., Jackson, H. J., and McGorry, P. D. Early warning signs of relapse following a first episode of psychosis. *Schizophrenia research 80*, 1 (2005), 107–111.
27. Granholm, E., Loh, C., and Swendsen, J. Feasibility and validity of computerized ecological momentary assessment in schizophrenia. *Schizophrenia bulletin 34*, 3 (2008), 507–514.
28. Hammerla, N. Y., and Plötz, T. Let's (not) stick together: Pairwise similarity biases cross-validation in activity recognition. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (New York, NY, USA, 2015), UbiComp '15, ACM, pp. 1041–1051.
29. Hovsepian, K., al'Absi, M., Ertin, E., Kamarck, T., Nakajima, M., and Kumar, S. 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* (New York, NY, USA, 2015), UbiComp '15, ACM, pp. 493–504.
30. Huber, P. J., et al. Robust estimation of a location parameter. *The Annals of Mathematical Statistics 35*, 1 (1964), 73–101.
31. Kimhy, D., Myin-Germeys, I., Palmier-Claus, J., and Swendsen, J. Mobile assessment guide for research in schizophrenia and severe mental disorders. *Schizophrenia bulletin* (2012), sbr186.
32. Lane, N. D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., and Campbell, A. T. A survey of mobile phone sensing. *Communications Magazine, IEEE 48*, 9 (2010), 140–150.

33. LANE, N. D., MOHAMMOD, M., LIN, M., YANG, X., LU, H., ALI, S., DORYAB, A., BERKE, E., CHOUDHURY, T., AND CAMPBELL, A. Bewell: A smartphone application to monitor, model and promote wellbeing. In *5th international ICST conference on pervasive computing technologies for healthcare* (2011), pp. 23–26.

34. LIANG, K.-Y., AND ZEGER, S. L. Longitudinal data analysis using generalized linear models. *Biometrika 73*, 1 (1986), 13–22.

35. LU, H., FRAUENDORFER, D., RABBI, M., MAST, M. S., CHITTARANJAN, G. T., CAMPBELL, A. T., GATICA-PEREZ, D., AND CHOUDHURY, T. Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing* (2012), ACM, pp. 351–360.

36. MAATEN, L. V. D., AND HINTON, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research 9* (2008), 2579–2605.

37. MARTIN ESTER, HANS-PETER KRIEGEL, J. S., AND XU, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD '96* (1996), AAAI Press, pp. 226–231.

38. MAXHUNI, A., MUÑOZ-MELÉNDEZ, A., OSMANI, V., PEREZ, H., MAYORA, O., AND MORALES, E. F. Classification of bipolar disorder episodes based on analysis of voice and motor activity of patients. *Pervasive and Mobile Computing* (2016).

39. mhealth for mental health program. `http://www.mh4mh.org/`.

40. MORRISS, R., VINJAMURI, I., FAIZAL, M. A., BOLTON, C. A., AND MCCARTHY, J. P. Training to recognize the early signs of recurrence in schizophrenia. *Schizophrenia bulletin 39*, 2 (2013), 255–256.

41. OSMANI, V., MAXHUNI, A., GRÜNERBL, A., LUKOWICZ, P., HARING, C., AND MAYORA, O. Monitoring activity of patients with bipolar disorder using smart phones. In *Proceedings of International Conference on Advances in Mobile Computing & Multimedia* (2013), ACM, p. 85.

42. PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHEL, V., THIRION, B., GRISEL, O., BLONDEL, M., PRETTENHOFER, P., WEISS, R., DUBOURG, V., VANDERPLAS, J., PASSOS, A., COURNAPEAU, D., BRUCHER, M., PERROT, M., AND DUCHESNAY, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12* (2011), 2825–2830.

43. RABBI, M., ALI, S., CHOUDHURY, T., AND BERKE, E. Passive and in-situ assessment of mental and physical well-being using mobile sensors. In *Proceedings of the 13th international conference on Ubiquitous computing* (2011), ACM, pp. 385–394.

44. ROENNEBERG, T. Chronobiology: the human sleep project. *Nature 498*, 7455 (2013), 427–428.

45. SAEB, S., ZHANG, M., KARR, C. J., SCHUELLER, S. M., CORDEN, M. E., KORDING, K. P., AND MOHR, D. C. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *Journal of medical Internet research 17*, 7 (2015).

46. SANO, A., AND PICARD, R. W. Stress recognition using wearable sensors and mobile phones. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on* (2013), IEEE, pp. 671–676.

47. SHARMIN, M., RAIJ, A., EPSTIEN, D., NAHUM-SHANI, I., BECK, J. G., VHADURI, S., PRESTON, K., AND KUMAR, S. 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* (New York, NY, USA, 2015), UbiComp '15, ACM, pp. 505–516.

48. SNIJDERS, T. A. *On cross-validation for predictor evaluation in time series*. Springer, 1988, pp. 56–69.

49. VOS, T., BARBER, R. M., BELL, B., BERTOZZI-VILLA, A., BIRYUKOV, S., BOLLIGER, I., CHARLSON, F., DAVIS, A., DEGENHARDT, L., DICKER, D., ET AL. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the global burden of disease study 2013. *The Lancet 386*, 9995 (2015), 743–800.

50. WANG, R., CHEN, F., CHEN, Z., LI, T., HARARI, G., TIGNOR, S., ZHOU, X., BEN-ZEEV, D., AND CAMPBELL, A. T. 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* (New York, NY, USA, 2014), UbiComp '14, ACM, pp. 3–14.

51. WILKINSON, G. S., AND ROBERTSON, G. Wide range achievement test (wrat4). *Psychological Assessment Resources, Lutz* (2006).

52. ZEGER, S. L., AND LIANG, K. Y. An overview of methods for the analysis of longitudinal data. *Statistics in medicine 11*, 14-15 (Jan 1992), 1825–39.